

 Access

4

Cited by

25 - Artificial Intelligence

from Part V - Intelligence and Information Processing

Published online by Cambridge University Press: **13 December 2019**

By Ashok K. Goel and Jim Davies

Edited by Robert J. Sternberg 

Book contents

Buy print copy

Summary

Artificial intelligence (AI) is a scientific discipline that seeks to understand intelligence through the design and construction of intelligent machines. AI and cognitive science have a strong two-way relationship: Cognitive psychology often has inspired AI theories, and AI research has

led to new theories of cognition that have been tested through psychological experimentation. While AI theories of cognition often are under-constrained, cognitive theories of AI tend to be over-constrained. Nevertheless, AI is useful for cognitive psychologists both as a source of new ideas and insights, and an experimental testbed. In this chapter, we describe some of the basic concepts and methods of AI by taking robot navigation in a city as an illustrative example. We also briefly discuss the history of AI, methods for assessing progress in AI, and some of AI's potential impacts on society.

Keywords

artificial intelligence AI and cognitive science psychological AI engineering AI learning

robotics AI and society history of AI

Type
Chapter
Information
The Cambridge Handbook of Intelligence , pp. 602 - 625
DOI: https://doi.org/10.1017/9781108770422.026
Publisher: Cambridge University Press
Print publication year: 2020

Artificial intelligence (AI) is the field of research that strives to understand, design, and build intelligent computational artifacts. From computer programs that can beat top international grand masters at chess to robots that can help detect improvised explosive devices in war, from intelligent agents that can answer questions in customer service to computing systems that can automatically detect credit card fraud, AI has had many well-known successes. In fact, modern societies are based on AI systems; without AI agents, advanced industrialized economies may quickly grind to a halt (Kurzweil, 2005).

As the above examples illustrate, the field of AI has a very broad scope. However, two features unify all of AI as a discipline. First, AI is united in the core belief that intelligence is a kind of computation. Thus, although in principle the design of any intelligent artifact might be classified as an AI, in practice AI agents are almost always computers or computer programs,

and AI laboratories typically are found in departments of computer science. Second, its main methodology is the exploration of the principles of intelligence by building computational artifacts.

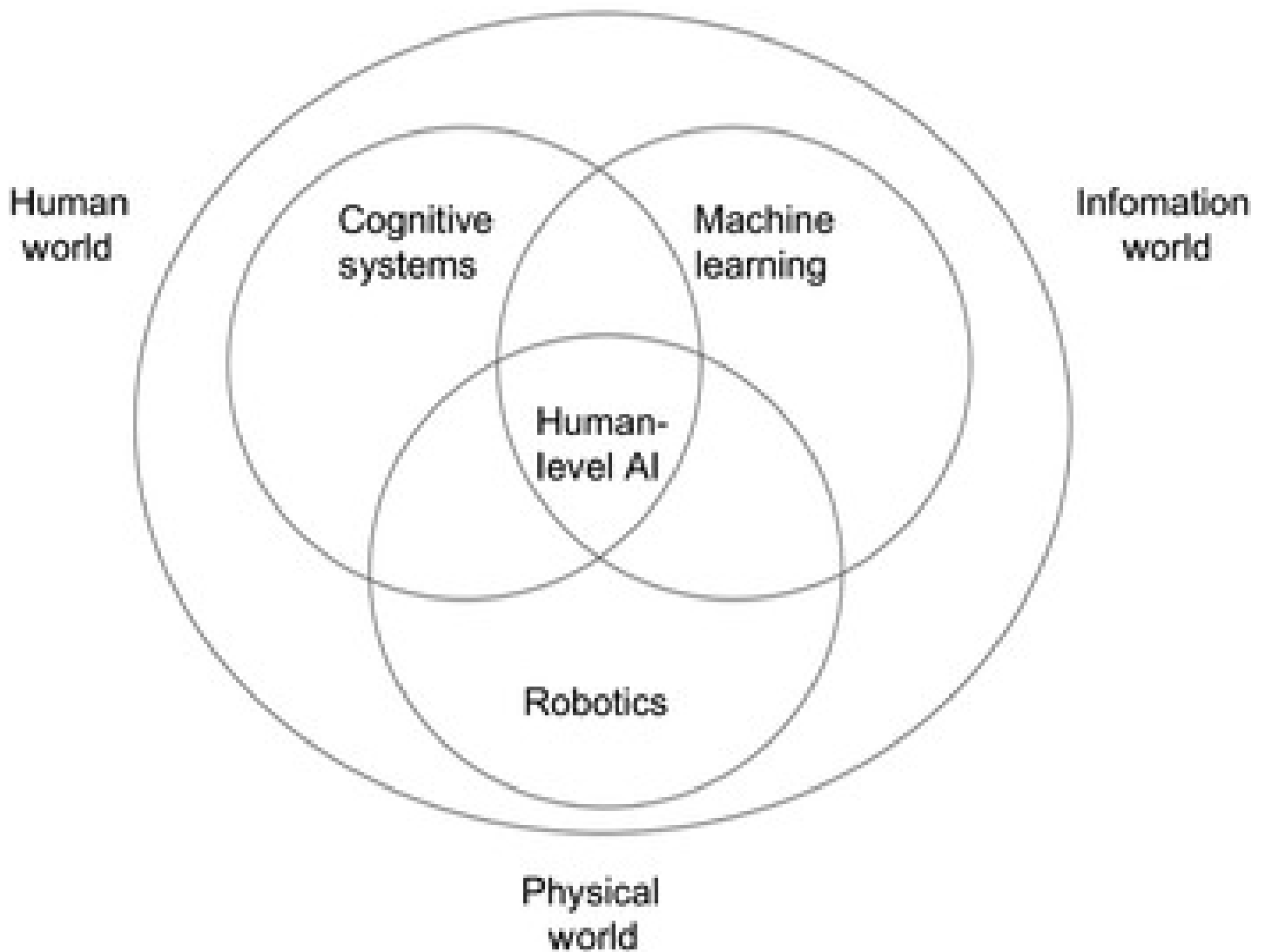
Broadly speaking, AI includes three large subfields: *robotics*, *machine learning*, and *cognitive systems*. Figure 25.1 depicts the relationship between AI and the three subfields. Robotics deals with embodied AI agents that interact with the physical world. Thus, action and perception are important parts of robotic AI agents: A robot that can detect improvised explosive devices, for example, must be able to sense and perceive the physical world as well as act on it. Machine learning typically pertains to computer programs that can detect and exploit patterns in data. To detect a fraudulent credit card transaction, for example, a computer program may first need to be trained on a set of credit card payments including both regular and irregular purchases so that the program can learn patterns indicative of a fraud. Cognitive systems, such as game-playing programs (for example, DeepBlue and AlphaGo) and conversational agents (for example, Siri and Watson) pertain to higher-level cognition, and interact with human and social worlds. Research on cognitive systems sometimes is also called cognitive computing.

Figure 25.1 The field of artificial intelligence includes the subfields of robotics, machine learning, and cognitive systems.

Of course, the three subfields of AI – robotics, machine learning, and cognitive systems – have considerable overlap. For example, the emerging area of human-robot interaction combines cognitive systems and robotics. Thus, a robot for detecting improvised explosive devices may have an embedded cognitive system for conversing with humans about its findings, receiving instructions, and asking questions. At the intersection of cognitive systems and machine learning, an interactive game-playing AI agent might have been trained on data from a large number of previously played games. Similarly, at the intersection of robotics and machine learning, a robot may learn a policy for taking action in the world through many trials. The center of Figure 25.1 indicates the intersection of all of these subfields, which describes work on general intelligence, including the construction of human-level AI dreamt by the founders of the field (McCarthy et al., 1955/2006). General AI likely will require major advances in all three subfields: robotics, machine learning, and cognitive systems.

It is also helpful to distinguish AI research into two main paradigms. *Engineering AI* is concerned with how to design the smartest intelligent artifacts possible, regardless of whether

Artificial intelligence



the processes implemented reflect those found in people (or other animals). The vast majority of AI research on robotics and machine learning falls into this category. *Psychological AI*, in contrast, endeavors to design artifacts that think the way people do (or sometimes groups of people do). Much, but not all, research on cognitive systems belongs to this paradigm, though it is possible to design cognitive systems, such as Siri and Watson, that interact with humans but do not necessarily reason as people do. In this chapter, we will focus mostly on the paradigm of psychological AI because the original dream of AI was to develop human-level intelligence, this handbook is intended for an audience of cognitive scientists, and we ourselves work in this paradigm.

AI and Cognitive Science

Cognitive science engages both AI and psychology. Cognitive psychology is mostly concerned with understanding of intelligence found naturally in humans and other animals, whereas, in addition, AI is interested in the understanding of intelligence in agents it designs. From the AI perspective, the concept of intelligence is not limited to the abilities of humans or even animals in general, but covers potentially any kind of intelligent system, be it human, computer, animal, or alien. Albus (1991, p. 474) put it eloquently: "A useful definition of intelligence ... should include both biological and machine embodiments, and these should span an intellectual range from that of an insect to that of an Einstein, from that of a thermostat to that of the most sophisticated computer system that could ever be built."

AI and cognitive psychology have a rich two-way relationship. In one direction, cognitive psychology has often inspired AI theories, and AI systems have acted as testbeds for experimenting with and evaluating the theories. Examples include theories of schema (Piaget, 1952), mental models (Craik, 1943), and learning by imitation (Tomasello, 1999), whose origins lie in psychology but have deeply influenced AI research. In the other direction, AI research has resulted in theories of cognition to be tested through psychological experimentation. Examples include semantic networks (Quillian, 1968), scripts (Schank & Abelson, 1977), and Bayesian networks (Pearl, 1988). AI models of cognition differ from other models in psychology in that AI models always implement *information-processing* theories. That is, the theory describes intelligence in terms of the content, representation, access, use, and acquisition of information, as opposed to, say, a statistical model of the influences of age or nutrition on IQ in a population.

Critics of AI from psychology sometimes view many AI programs as being psychologically implausible. Indeed, psychological claims of AI theories typically are under-constrained by empirical human data, and thus, for the most part, these criticisms of AI from psychology are not inaccurate. Most AI is engineering AI, and even psychological AI must go out on limbs simply because there are just not enough data to constrain all the choices AI scientists need to make. However, AI contributes to the understanding of intelligence in several ways (Davies & Francis, 2013).

First, although they can be under constrained, *AI programs demonstrate what kinds of data need to be collected*. Because AI programs work at a very precise level of detail, they bring to light

theoretical ambiguities that psychology might not immediately or explicitly acknowledge. For example, it is one thing to say that a person can only comprehend one speaking voice heard at a time; it is quite another to create a computer implementation of this attentional effect – to do so requires making decisions about the interaction and influences of volume, which one voice you are listening to first, what factors affect attentional switching, among many other issues. The level of detail that makes AI programs under-constrained is the very quality that brings to light previously unrecognized factors.

Humans obviously have only limited information and information-processing resources, and, thus, their rationality is intrinsically bounded (Simon, 1996). However, it is also true that many cognitive problems people routinely solve are computationally intensive. For example, deciding how to design a poster for a concert offers more possibilities than can possibly be considered. *AI approaches to solving these kinds of problems shed light on what ways will not work.* If AI shows that a means for solving a problem will take too long to be practical, then AI has shown that people cannot be doing it that way, at least not routinely.

On the other hand, *AI can show that certain methods are possible.* Though showing that something is possible is far from proving that it *is*, many current theories in psychology do not have such proof. AI serves a valuable function as creating proofs-of-concept.

Another thing AI is particularly good at is *exploring the benefits and limitations of various ways to represent and organize knowledge in memory.* Many of these benefits are clear only when dealing with a strict information-processing level of detail. Are beliefs represented as words, pictures, or something else? Given all of the cognitive tasks memories are supposed to contribute to, AI is in a good position to shed light on such issues. As we will describe in more detail later, this subfield of AI is known as “knowledge representation.”

Finally, once there is an AI program that resembles some part of human thinking to a researcher’s satisfaction, *it is possible to run experiments on the program that are either unethical or too expensive (in terms of time or money) to run on living beings.* In simulation you can run thousands of experiments in a day, with exquisite control over all variables.

If AI theories of cognition are under-constrained, theories of AI with roots in cognitive psychology can be over-constrained (Langley, 2012). In general, there is no particular reason why an AI cognitive system must imitate each and every microstructure of cognition to

manifest intelligent behavior. As an example, while the notion of a working memory originating in cognitive psychology is important in AI for, among other things, focusing attention, the size of the working memory in AI cognitive systems need not be limited to the famous seven plus or minus two typical of adult humans.

In both cognitive psychology and AI cognitive systems, researchers over the years have tried to build theories of intelligence of two different kinds at two different levels of abstraction: the symbolic and the sub-symbolic. Symbols represent conceptual abstractions of the world, such as *dog* or *justice*, and act as pointers both to the world and to one another; thus, symbolic processing pertains to conceptual information processing. In contrast, sub-symbolic processing tends to deal with information at a finer grain, such as the pixels of an image, the weight of an association, and the probability of the truth of a proposition. Sub-symbolic systems are difficult to program completely by hand, so they are often machine learning systems as well. Connectionism is the dominant sub-symbolic modeling paradigm in cognitive science, and works by applying neural networks to psychological problems. Symbolic cognitive systems, in contrast, may organize symbols into propositions that can be used in reasoning processes, such as deduction, induction, and various transformations. Many language processing systems work this way, as well as systems that use commonsense reasoning to try to understand the physical world. Many symbolic systems use “qualitative reasoning,” a term that distinguishes it from the more numerically represented memory of sub-symbolic systems. There are, of course, many hybrid systems that use both symbols and sub-symbols, such as Bayesian belief networks.

Navigational Planning: An Illustrative Example

We want to illustrate a simple example of AI in some detail to help make this discussion more concrete. Let us suppose that Sunny, a cheerful AI agent, is about to start a new job in a new city. Sunny starts its car at its apartment and needs to navigate to an office building downtown. How might Sunny think and what might Sunny do, given that this is its first day in the city and it has never been to the office building? Our goals in this section are to explain some dimensions in designing AI agents as well as describe some issues in putting multiple capabilities into an AI agent.¹

Action, Perception, and Cognition

To reach its office from its apartment, Sunny might use one (or more) of several possible strategies. For example, it might drive its car a short distance in some direction and then see if it has reached the office building. If it has, then it has accomplished its goal. If it has not, then it might again drive a short distance in some direction, and then again see if it has reached the building. Sunny could repeat this process until it reaches its goal. Blindly moving about like this would likely take a very long time, but in terms of internal processing, this method is very efficient. This *perceive-act* internal computational processing, called *situated action* (or *reactive control*; Arkin, 1999), works by perceiving the immediate environment, acting based on those perceptions, and then repeating. The computational processing in reactive control is very efficient and does not require the agent to be able to store new memories. However, depending on the environment and the goal, it may produce needlessly complicated external behavior because Sunny could be driving short distances in arbitrary directions for a very long time before it reaches its goal. In fact, this strategy does not guarantee that the goal will ever be reached.

Alternatively, when Sunny starts at its apartment, it might simply ask Honey, a sweet AI agent who happens to be passing by, how to reach the office building. Honey, a longtime resident of the city, might give Sunny detailed directions, which Sunny could simply follow. In contrast to the previous strategy, this strategy produces very efficient output behavior: Assuming that Honey's directions are good, Sunny should reach its goal quite efficiently. However, this strategy of *asking* requires a society of intelligent agents (human or AI), each with different knowledge. It also requires a culture in which Sunny may in fact approach Honey for directions; Honey might in fact stop to help Sunny, and the two can communicate in a shared language; Sunny might trust Honey, a total stranger, enough to follow its directions in a new city; and so on. AI research on robot societies and human-robot interaction is in its early stages, and so here we will briefly mention only a small set of selected issues.

How can Sunny and Honey talk with each other? How can Sunny talk with a human? Understanding and generating natural language is the goal of the AI sub-discipline of *natural language processing* (NLP). Researchers in the area of natural language understanding take written text or spoken language and create accurate knowledge representations reflecting the meaning of the input. Natural language generation works roughly in the reverse – starting with some meaning and generating appropriate words and speech to communicate it; this has received much less attention in AI. Two robots might be able to share knowledge very efficiently if that knowledge is represented in the same way. However, there is little agreement

in AI over how knowledge should be represented in general (the linguistics subfield of semantics, similarly, has no consensus on meaning representation). Different knowledge representations appear to be better for different tasks.

When Honey gives advice, how is Sunny to know whether that advice is plausible? Except for limited environments, this problem seems to require general *commonsense reasoning*, a field closely related to knowledge representation. It is a widely held belief that most computer programs' lack of common knowledge and inability to reason with it effectively are major problems for much of AI. The subfield of commonsense reasoning endeavors to overcome this challenge. The most famous is the Cyc project (Lenat & Guha, 1990), a major project to manually encode all human commonsense knowledge. More recent strategies include Web-based knowledge collection methods, such as OpenMind Commonsense (Singh et al., 2002) and Peekaboom (von Ahn, Liu, & Blum, 2006).

Here is another strategy by which Sunny may reach its office building: Let us suppose that when Sunny was originally built in an AI laboratory, it was bootstrapped with some knowledge. Some of this knowledge may have been *heuristic* in its content and encoded in the form of a *production rule*. A heuristic is a "rule of thumb," and a production is an "If x then do y" kind of rule. So, for example, Sunny might be bootstrapped with the knowledge that "if the goal is to reach downtown in a city, then move in the direction of the tallest buildings." This knowledge directly uses the goal (reaching downtown) to suggest a high-level action (move in the direction of the tallest buildings) and is heuristic in its nature since it may not correctly apply in all cities. If Sunny had this knowledge, then it might begin by perceiving the environment around it, locating the tallest buildings in the horizon, deciding to head in their direction, and moving toward them. When it reaches the next intersection, Sunny might again locate the tallest buildings relative to its current location, change its direction if needed, and so on. This strategy of *perceive-think-act* not only requires some knowledge but also must use more complex internal processing than the simpler perceive-act strategy of situated action. On the other hand, depending on the environment, perceive-think-act may result in a far simpler external behavior because now the behavior is more explicitly directed by the goal.

This kind of strategy can be implemented as a *production system* (Newell & Simon, 1972), which represents "what to do," or procedural knowledge, with if-then rules. In Sunny's case, the rules dictate physical action in the environment. Production systems are often used for making changes in memory as well. Rules can add, change, and remove goals and elements in

memory. Surprisingly complex behavior can result with this method. This particular approach has been very successful in cognitive modeling for some problems. Well-known cognitive architectures such as Soar (Laird, 2012; Laird, Newell, & Rosenbloom, 1987) and ACT-R (Anderson, 2013; Anderson & Lebiere, 1998) are production systems at their core.

However, cognitive architectures such as SOAR and ACT-R have declarative as well as procedural knowledge. Declarative knowledge is often represented as *frames* (Minsky, 1975) or semantic networks, and is used by the productions (the procedural knowledge). Frames are similar to classes in object-oriented programming: They define a class of entities and what attributes they have. Instances of these frames take particular values for these attributes. For example, the frame for PERSON might contain the attributes NAME and AGE, and an instance of person might have a NAME of "Julie" and an AGE of "45." Like frames, *semantic networks* (Sowa, 1987) are a widely used representation scheme in AI. One can imagine a semantic network as a map of concepts, with nodes representing concepts (such as MAN and DOG) and labeled links between them (labeled, for example, with OWNS). Frames and semantic networks are thought to be informationally equivalent, which means that there is no loss of information when translating from one to another.

Another long-standing and still very strong area of AI is representation and processing based on *logic*. Logic is used for inference but has also been adapted for use in many other specific tasks, such as theorem proving (McCarthy, 1988).

Let us consider one other strategy for Sunny's task before we move on to the next topic: Sunny might consult a map of the new city. The important characteristics of a city map in this context are that they are an external representation of the world (i.e., it is not stored internally in Sunny) and that it is a visuospatial model of the world (i.e., there is a one-to-one structural correspondence between selected spatial objects and relations in the world and the objects and relations on the map; see Glasgow, Narayanan, & Chandrasekaran, 1995). Sunny can use this map to plan a navigation route to the office building and then execute the plan. This too is a perceive-think-act strategy. However, as compared to the heuristic method, the "thinking" in this strategy uses very different content and representation of knowledge.

Once Sunny has studied the map, it has some version of it stored in its memory. When Sunny needs to navigate to a location on the map, it can refer to the map. Finding a route on a map is not trivial, however. At each intersection, a choice must be made. One of the first insights of

the field was that a great many cognitive problems can be solved by systematically evaluating available options. This method of searching through a space of choices is applicable in many domains and is still widely used. Researchers focusing on *search* compare the various search methods that have been invented and describe the classes of problems to which each is most applicable. Because most interesting search spaces are enormous (e.g., there are more possible chess game configurations than there are atoms in the universe), researchers invent *heuristics* to guide the AI to explore the more promising areas of the search space. One problem for which search has been particularly useful is in *planning*, which is the generation of an ordered sequence of actions prior to actually executing those actions.

The internal processing in reading a map might be more costly than the processing in a heuristic search; however, depending on the environment, this strategy might lead to a solution that has a better chance of success – for example, the solution generated by this model-based method is less likely to get stuck in some cul-de-sac than the solution generated by the heuristic method. Of course, we can easily think of several other strategies for addressing Sunny's task, especially in today's world of the Internet and the global positioning system.

These examples make clear some of the dimensions of designing an AI agent. First, an AI agent lives in some environment, and what and how an agent can think depends in large part on the environment in which the agent lives. Some environments might contain other agents, who may be cooperative, competitive, or combative. Some environments are dynamic. Some environments are only partially observable. Some environments are nondeterministic, and so on. One of the many contributions of AI is a more precise characterization and analysis of different kinds of environments, though much of the AI analysis so far has focused mostly on physical, not social, environments. Second, an agent might have access to different kinds of knowledge contents and representations. The knowledge may be engineered or acquired. The representations can be internal or external. The knowledge contents range from nil to heuristic rules to detailed, high-fidelity models of the environment. Another major AI contribution is a more precise and detailed account of knowledge contents and representations. Third, different strategies lead to very different trade-offs among knowledge requirements, the computational efficiency of internal processing, and the quality of generated solutions and behaviors. Yet another contribution of AI is more precise enumeration and analysis of these trade-offs.

Reasoning, Learning, and Memory

So far we have talked only about what our hypothetical AI agent, Sunny, might think and do when trying to reach its office for the first time. However, because Sunny is an AI agent, it might also learn from its interactions with the environment. What and how might Sunny learn from its experiences? Sunny acquires a new experience each time it interacts with the environment, including navigating from its apartment to its office, talking with Honey, and so on, irrespective of what internal strategy it uses. Further, to the degree to which Sunny's internal processing is accessible to it, it may also acquire an internal experience each time it does internal processing. In addition, when Sunny executes a plan or an action on the environment, the environment might provide it with feedback. This feedback might come immediately after the execution of an action (e.g., taking a turn at an intersection and getting caught in a cul-de-sac), or after a series of actions (e.g., taking a sequence of turns and reaching the goal). The feedback might simply be the outcome of a plan – success or failure – or it might contain more information, for example, a specific action in the plan failed because it led to a cul-de-sac. Thus, an experience might contain not only an interaction with the environment but also some feedback on the interaction, and perhaps also a trace of the internal processing in that interaction.

Sunny might potentially learn many different things from its experiences in the environment. For example, Sunny might simply encapsulate experiences as *cases* and store them in memory for reuse in the future – the AI equivalent to episodic memory. On the first day, for example, Sunny might use a map to plan a navigation route and then execute the plan in the environment, as indicated in the section Action, Perception, and Cognition. The next day, when Sunny again faces the task of navigating to its office from its apartment, it might find a solution simply by retrieving the navigation plan in the case acquired from the previous day rather than relying on general-purpose knowledge and rules. This is called *case-based reasoning* (Kolodner, 1993). This approach views reasoning largely as a memory task, that is, as a task of retrieving and modifying almost correct solutions from memory to address the current problem. Related subdisciplines of cognitive science studying similar phenomena are exemplar-based reasoning, memory-based reasoning, instance-based reasoning, and analogical reasoning.

As Sunny learns from its experiences, its internal processing as well as its external behaviors can change. Initially, for example, Sunny might use a map of the environment for navigating

through the new city. However, as it navigates through the world and stores its experiences as cases in its memory, it can increasingly generate new navigation plans by case-based reasoning. However, as the number of cases in memory increases, the cost of retrieving the case appropriate for a new problem also increases. Thus, again, each reasoning strategy offers computational trade-offs among knowledge requirements, processing efficiency, and solution quality.

More generally, AI typically thinks of each strategy for action selection discussed in the section Action, Perception, and Cognition as setting up an associated learning goal, which in turn requires a corresponding strategy for learning from experiences. Let us suppose, for example, that Sunny uses the strategy of situated action for action selection. It might, for example, use a table (called a *policy*) that specifies mappings from percepts of the world into actions on it. Then, from the feedback, or the reward, on a series of actions, Sunny can learn updates to the policy so that over time its action selection is closer to optimal. This is called *reinforcement learning* (Sutton & Barto, 1998). Note that if the series of actions results in success, then the reward will be positive; otherwise it is negative. Reinforcement learning is an especially useful learning strategy when the reward is delayed, that is, it comes after a sequence of actions rather than immediately after an action so that it is not clear what specific action in the sequence was responsible for the success or failure. Alternatively, suppose that Sunny employs the strategy of using production rules such as “If x then do y” to select actions. In this case, Sunny can use the learning strategy of *chunking* (Laird et al., 1987) to learn new rules from its experiences over time. Thus, just as AI has developed many reasoning strategies for action selection, it has developed many learning strategies for acquiring the knowledge needed by the reasoning strategies. Further, just like the reasoning strategies, the learning strategies too offer trade-offs among knowledge requirements, computational efficiency, and solution quality.

Most of the methods described thus far fall roughly into a category that can be described as “symbolic” approaches, characterized by the manipulation of qualitative, recognizable, discrete symbols. Another broad approach, as we mentioned earlier, is sub-symbolic. Though the border between these two approaches is fuzzy, we can think of a symbolic representation having a symbol for the letter “R” and a sub-symbolic system representing the letter with the dots that make it up on a screen. Since the dots, or pixels, are not meaningful in themselves, they are thought to be at a level of description below the symbol. The rest of the methods described in this subsection tend to use sub-symbolic representations.

So far we have assumed that Sunny has perfect knowledge of the environment, even if that knowledge is limited. However, many real-world domains involve uncertainty, and AI methods based on *probability* have been very successful at working in these environments. Probability theory has been used in many algorithms that use *hidden Markov models* to predict events based on what has happened in the past. Hidden Markov models are mathematical representations that predict the values of some variables given a history of how the values of these and other variables have changed over time (Raibiner & Juang, 1986). Probabilities are also used to determine beliefs, such as how likely it is that a street Sunny wants to use has been closed, given that the rain in that part of the city was 80 percent likely to have been freezing. Bayes' rule is useful for determining such conditional probabilities of some events (e.g., a road being closed) given the probability of others (e.g., freezing rain). *Bayesian belief networks* are mathematical representations that predict the probability of certain beliefs being true, given the conditional probabilities of other beliefs being true (Pearl, 2000). These networks are useful for updating probabilities of beliefs as information about events in the world arrives.

Statistics is the foundation of much of *machine learning*, a subdiscipline of AI that aims to create programs that use data and limited previous beliefs to create new beliefs. There are a great many kinds of learning algorithms, including artificial *neural networks*, which are the basis of connectionism in cognitive science (McClelland, Rumelhart, & the PDP Research Group, 1986; Rumelhart, McClelland, & PDP Research Group, 1986). Whereas most of the systems we've discussed process recognizable symbols, neural networks represent information at a sub-symbolic level (such as in pixels or bits of sound) as activations of nodes in a network. The processing of a neural network depends on how the nodes change each other's activations. The output of a neural network is an interpretation of the activations of certain nodes (for example, indicating whether or not a room is dark). *Genetic algorithms* are another means of computation that is (often) based on processing sub-symbolic representations. Inspired by the theory of biological evolution, genetic algorithms create solutions to problems by applying some fitness function to a population of potential solutions (Mitchell, 1998). Solutions with a high fitness are used to generate members of the next generation (often with some mutation or crossover of features), after which the process repeats.

Deliberation and Situated Action

Although we have briefly discussed situated action (reactive control) and situated learning (reinforcement learning), much of our discussion about Sunny, our friendly robot, pertained to deliberation. While AI theories of deliberative action selection typically are explicitly goal-directed, goals in situated action often are only implicit in the design of an AI agent. Deliberation and situated action in AI agents occur at different timescales, with deliberation typically unfolding at longer timescales than situated action. In general, designs of AI agents include both deliberative and situated components. For example, the design of Sunny, our friendly robot, might contain a deliberative planner that generates plans to navigate from one location in a city to another. Note that because there are many people and other robots working or walking on the roads, Sunny's environment is dynamic in that the state of the world can change during the time Sunny takes to generate a plan. How can Sunny navigate from its apartment to its office building in this dynamic environment?

Sunny of course can use the deliberative planner to plan a path between offices. However, while the planner can produce navigation plans, it might not represent the movements of all the people and other robots on the roads. So deliberation by itself is not good enough for the dynamic urban environment. Alternatively, Sunny can use situated action (i.e., *perceive-act*) that we described in the previous section. While this can help Sunny avoid collisions with moving people – as soon as Sunny senses the nearby presence of a person, it can move away – its progress toward the goal of reaching a specific office is likely to be slow, perhaps painfully so.

Yet another alternative is to endow Sunny with the capability of both deliberative planning and situated action. In fact, this is exactly what many practical robots do. As a result, Sunny becomes capable of both long-range planning and short-range reaction. It can use its deliberative planner to come up with a plan for reaching the office building. Then, as it is executing the navigation plan, it constantly monitors the world around it and acts to avoid collisions with moving people. Next, as soon as it has moved away from a collision, it reverts to execution of its navigation plan. In this way, Sunny combines both deliberation and situated action. While this integration of deliberation and situated action has obvious benefits, it also has additional knowledge requirements as well as additional computational costs of shifting between strategies.

So far we have talked of perceiving the environment as though it were a minor task. For human beings, perception often appears to be effortless, but automating perception in AI agents has proven to be one of the many difficult problems in AI. The field of *computer vision*

creates programs that take images (such as photos and video) as input and generates beliefs about objects, textures, and movements, as well as higher-level features such as emotions, movement styles, and gender. *Speech recognition* is another major field in perception. The ability of computers to understand your credit card number when you speak it into the phone is the result of over fifty years of AI work. Many of the algorithms used to understand speech and sound are shared with those of machine learning.

Likewise, achieving physical motion in the real world is difficult. *Robotics* is the field of AI that controls machines that interact directly with the physical world (as opposed to a program that, say, buys stocks electronically). Robotics uses computational perception, machine learning, and sometimes natural language processing. Some of the major problems specific to robotics are navigation and the handling of objects. Robots can work in collaboration with each other; the field of *intelligent agents* or *agent-based AI* builds intelligent programs that operate through the interaction of many individual agents whereas in *swarm intelligence* the individual agents do not have much intelligence individually. For example, two intelligent robots cooperating to assemble a desk would be an example of agent-based AI, and a large number of simple agents, reacting to their environment only locally to find the fastest route, much as ants do, would be an example of swarm intelligence.

Deliberation and Reflection

We have briefly discussed the need for both longer-range planning and shorter-range situated action in autonomous AI agents because the environment in which they reside is dynamic. However, changes in the environments themselves can unfold over different timescales. In the short term, for example, people and robots might be moving around on the roads of Sunny's city. In the long term, roads themselves change, new apartments and office buildings are constructed, and other changes occur. Then the navigation plan that Sunny's deliberative planner produces will start failing on execution. How might Sunny adapt its knowledge of the environment as the environment changes? Alternatively, if Sunny had been designed incorrectly to begin with, how might it adapt its reasoning process?

Recent AI research on meta-reasoning is starting to design AI agents capable of self-adaptation (Cox & Raja, 2011). Such an AI agent might contain a specification of its own design. For example, the meta-reasoner in Sunny may have a specification of Sunny's design, including its functions (e.g., its goals) and its mechanisms for achieving the functions (e.g., the

method of map-based navigation planning). When Sunny generates a plan that fails on execution, Sunny's meta-reasoner uses the specification of its design to diagnose and repair its reasoning process. If the feedback from the world on the failed plan pertains to an element of knowledge (e.g., at intersection A, I expected a road going directly toward downtown but when I reached there, I found no such road), then Sunny enters this new knowledge in its map of the city. Thus, while the deliberative planner in Sunny reasons about actions in the external world, Sunny's reflective meta-reasoner reasons about its external world as well as its internal knowledge and reasoning.

AI Safety

If Sunny is going to be moving around the real world, there needs to be something about the design that will keep the robot from hurting people (or itself). As robots are often made of hard material, the very motion of the limbs has the potential for injury to humans, other animals, or sensitive environments.

A complex robot should not have simple goals without a set of values or preferences that will make sure that harm is not caused. Pushing a child out of the way might help make Sunny get to the desired location faster, but would be socially unacceptable; so we need to make sure that robots like Sunny recognize that this would likely cause harm and should not be done.

This is a challenging issue for many reasons. First, harm can come about in so many ways. Some harms, such as injury, are often immediate, and others, such as low-level ingestion of toxins, can take years to cause harm. Some harms are physical, and others, such as witnessing a horrific event, can cause psychological trauma.

Second, sometimes harms cannot be avoided. We don't want Sunny to cut someone open, but a surgical robot has to do just that. If Sunny is a self-driving car robot, it might have to make split-second decisions on who lives or dies. Suppose Sunny is moving at a high speed, and the road conditions change such that Sunny will strike, and probably kill, either an old person or a seven-year-old child, depending on which direction Sunny steers (it's too late to brake sufficiently). Which one should Sunny hit, and why? What if it's two old persons or one child? The calculation that Sunny must make needs to be fast, and have numerical values associated with human lives and suffering.

We will also have robots designed to kill people. The very existence of these machines is an ethical issue, and the programming that determines who the robot kills and doesn't kill is an obvious example of a serious moral decision. It is important that we recognize that programming ethics into robots is not merely a programming issue, but an interdisciplinary problem requiring contribution from law, philosophy, psychology, sociology, and other fields.

Putting It All Together

Figure 25.2 illustrates a high-level general architecture for an AI agent such as Sunny with many of the capabilities discussed above. The agent is situated in an external world: It can sense percepts in the world; it can also use its effectors to act on the world. At the bottom of the multilayered architecture is reaction, which directly maps percepts into actions. In the middle is deliberation, which unfolds more slowly than reaction and includes complex interactions among memory, learning, and reasoning. At the top level is metacognition, which monitors, controls, adapts, and explains the deliberative processing. This architecture helps us understand how AI agents like Sunny are designed such that all their capabilities work in synchrony.



Figure 25.2 A multilayered architecture for an AI agent, combining reaction, deliberation, and metacognition.

In this section, we took navigational planning as an example to illustrate how AI is putting together multiple capabilities ranging from perception, cognition, and action, to reasoning, learning, and memory, and on to reflection, deliberation, and situated action. Of course, the design choices we have outlined are exactly that: choices. For example, instead of using deliberation to mediate between reflection and situated action as described above, an AI agent can reflect directly on situated action. In a way, the enterprise of AI is to explore such design choices and examine the computational trade-offs that each choice offers.

What has emerged out of this line of work is an understanding that the design of an AI agent depends on the environment it lives in, and that no one design is necessarily the best for all environments. Further, the design of an AI agent in any nontrivial environment requires multiple capabilities and multiple methods for achieving any capability such as reasoning and learning.

There is a large and growing literature on architectures for intelligent agents. Some of these architectures are inspired by cognitive psychology and are called cognitive architectures. Some of the better-known cognitive architectures include ACT-R (Anderson, 2013; Anderson & Lebiere, 1998) and SOAR (Laird, 2012; Laird et al., 1987). Samsonovich (2010) and Kotseruba, Gonzalez, and Tsotsos (2016) provide useful surveys of cognitive architectures. Langley, Laird, and Rogers (2009) review some of the challenges in developing cognitive architectures. Laird, Lebiere, and Rosenbloom (2017) recently proposed a “common model” of intelligence based on work on cognitive architectures.

A Very Brief History of AI

Many people have an almost mystical view of intelligence. One result is that when an AI agent manages to accomplish some task, a common reaction is to claim that it is not an example of intelligence. Indeed, at one point in the history of computing, arithmetic calculation was thought to be one of the best displays of intelligence, but now almost no one wants to say a calculator is intelligent. Because of this moving of the goalposts, AI has been jokingly referred to as standing for “almost implemented.” For the most part, this is only a semantic issue. In fact, although not always labeled AI, AI discoveries have revolutionized our world.

In the middle of the twentieth century, the scientific world experienced a shift in focus from descriptions of matter and energy to descriptions of information. One manifestation of information theory applied to real-world problems was in the field of *cybernetics* (Weiner, 1961), the study of communication and control in self-regulating analog systems. Cybernetics’ focus on analog signal contributed to its losing ground against discrete symbolic approaches common in AI. Not only did the symbolic approaches come to dominate AI research, but the symbol-processing approach came to dominate cognitive psychology as well.

Search was the first major paradigm of AI. The first artificial intelligence program ever written is the Logic Theorist (Newell, Shaw, & Simon, 1958). Many of the problems early AI researchers focused on were, in retrospect, simple. The early exuberance of AI was tempered with the first “AI winter” that dominated the late 1960s and the 1970s, characterized by a decrease of optimism and funding, and caused by unfulfilled expectations. Early interest in associative processing was diminished by an influential book *Perceptrons* (Minsky & Papert, 1969) around the same time. This rigorous book showed that the state-of-the-art associative systems of the

time could not implement any task that was not linearly separable, including the simple logical operator “exclusive or.”

The AI winter of the 1970s, however, also witnessed the emergence of new theories and paradigms. For example, ANALOGY (Evans, 1968) solved simple geometric analogy problems that appear on some tests of human intelligence. SHRDLU (Winograd, 1972) performed natural language processing to understand commands to a robot to pick up and manipulate blocks. Marr (1982) developed a three-stage computational theory of vision: from a raw image to a primal sketch with edges, from primal sketches to 2 and 1/2 D representations including surfaces, to 3D object recognition. Schank first developed a theory of conceptual structures for natural language understanding (Schank, 1975) and then a theory of memory, reasoning, and learning (Schank, 1982).

Working in a different paradigm, Feigenbaum, Buchanan, and their colleagues first developed an expert system called Dendral that could generate hypotheses about molecular structures from spectroscopic data (Lindsay et al., 1980), and then an expert system called Mycin that could generate hypotheses about E. coli bacterial diseases from heterogeneous patient data (Buchanan & Shortliffe, 1984). AI's revival in the 1980s was due in part to the success of these *expert systems*, which were designed to replicate the expertise of individuals with a great deal of domain knowledge. Knowledge engineers would interview and observe experts, and then attempt to encode their knowledge into some form that an AI program could use. This was done with a variety of methods, including *decision trees* (which can be thought of as using the answers to a series of questions to classify some input, as in the game Twenty Questions). Since expert systems were of use to business, there was a renewed interest in AI and its applications. Funding for AI research increased.

One of the ideological debates of the 1980s was between the “neats” and the “scruffies”: the neats used a formal, often logic-based approach and the scruffies focused on modeling human intelligence and getting AIs to use semantic information processing. Geographically, many of the neats were based at Stanford University and the US West Coast, and in Japan, and many of the scruffies were at Massachusetts Institute of Technology (MIT) and the US East Coast. Neats thought that knowledge representation and processing should be mathematically rigorous and elegant, and evaluations should involve proofs. Scruffies believed that intelligence is so complex that it is unwise to put such constraints on it at this early stage of development of AI theory and methodology. Today, most of the engineering AI research

would be classified as neat. A good deal of, but not all, contemporary psychological AI is scruffy.

In the 1980s, interest in artificial neural networks and associative AI was revived through cognitive modeling by *connectionists* (Rumelhart et al., 1986; McClelland et al., 1986). Connectionism continues to have a strong influence in modern cognitive science; in engineering AI, artificial neural networks are regarded as just one of many statistical learning mechanisms (such as Markov models and other methods of memory, reasoning and learning mentioned in the previous sections). Interestingly, some of the approaches and ideas of the cyberneticists have had a revival in these sub-symbolic approaches to AI.

Over time, the limits of expert systems became clear. As they grew in size, they became difficult to maintain and could not learn. As a knowledge base grows, inconsistencies between different chunks of knowledge tend to arise. In part again because of unfulfilled expectations, in the 1990s, AI entered a second “winter,” with diminished optimism, interest, and funding. However, during the second winter, again, new frameworks appeared, including *embodied cognition*, *situated cognition*, and *distributed cognition*. These frameworks emphasize how the body and environment both constrain and afford cognition, how cognition always is in the context of the physical and social worlds where these worlds themselves afford information to the cognitive agent. Similarly, *agent-based AI* on one hand seeks to unify cognition with perception and action, and on the other, studies AI agents as members of a team of other agents (artificial or human).

Over the past decade or so, AI has witnessed a resurgence of interest and attention. Perhaps the most exciting new development in AI is “deep learning,” which involves machine learning over multiple layers of artificial neural network units (LeCun, Bengio, & Hinton, 2015). This algorithmic innovation was facilitated by hardware advances that allowed AI researchers to program systems to work on special chips made for graphics: graphics processing units, or GPUs, which speeded deep learning by about 100 times.

Deep learning feels like a revolution, in part because it has been used to address many difficult problems in AI. To take a famous example, the Chinese board game Go was very difficult for AI, in part because there are so many possible moves at each turn. Using deep learning over information taken from large quantities of human Go games and knowledge, an AI system called AlphaGo (Silver et al., 2016) beat a professional Go champion in 2015. Two years later,

the successor to AlphaGo, called AlphaGo Zero (Silver et al., 2017), was able to beat the original, but without having looked at any human knowledge at all: AlphaGo Zero got to be a world champion just by playing games against itself for forty days, and came up with previously unknown strategies that Go experts described as very creative.

More traditional realms of creativity, the arts, have also been the focus of AI research. We now have AIs that create paintings, jokes, musical compositions and improvisation, and poetry (Besold, Schlorlemmer, & Smaill, 2015; Veale & Cardoso, 2018). More complex artistic endeavors, such as creating a written novel, have proven more difficult, because a novel requires so much knowledge and understanding about how the world works. Many of these creative AIs are used for commercial products, usually under the name “procedural generation.” Most famously, the computer game *No Man's Sky* (www.nomanssky.com) allows players to explore a virtual galaxy. When they land on a planet, the game creates the terrain and weather, as well as a complete ecosystem with custom-generated flora and fauna. Not only are the causal complexities of the simulation generated automatically, but the game also creates graphical models of them, complete with sounds. Procedurally generated content is economically important, because artistic creativity is a large part of the budget of many modern video games and movies, and the content is consumed far faster than it can be generated by human beings.

Robotics has proven to be a particular challenge for artificial intelligence – our running example of Sunny notwithstanding – because dealing with the real world is far more complex than dealing with formal, internal systems that play Go or recommend books to people. But these problems are slowly being addressed as well. Many people have robot vacuum cleaners in their houses now, and more complex robots are on their way. The field of *human-robot interaction* has arisen to study how humans do and can best interact with robots.

At present, AI appears to have entered a new phase of revival. This is in part due to the new frameworks that have appeared over the past generation, especially agent-based AI, deep learning, human-centered AI, and computational creativity. By now, AI is ubiquitous in industrialized societies, though it often does not go by that name. Many researchers avoid the term, feeling that it has been tarnished by the boom-and-bust cycle of interest and funding it has experienced in its sixty-year history. However, techniques from AI are used in many practical applications, allowing your voice to be understood when you talk to an automated phone system, using your past purchases to make recommendations for books when you

shop online, efficiently matching flights to gates at airports, directing the pathfinding of characters in computer games, generating web search engine results, enabling face detection in cameras and online photo archives, and doing automatic translation.

The concerns of *human-centered AI* (Ford et al., 2015) are how individuals and societies can productively work with artificial intelligence to make progress on human values. For instance, if Sunny were given a command to get some food, implicit in our request is that Sunny does not steal food. But without world knowledge, the AI might complete the request, but not abide by assumed preferences about how it's done. Just as AIs need to understand what we mean, they also need to be able to convey what *they* mean.

Although machine learning, and particularly “deep learning,” has enjoyed many breakthroughs in the past few years, a persistent problem for all sub-symbolic AI systems is that once they learn to do something, it is not immediately clear how or why they can do it. This is because each unit is by itself meaningless, and the processing involves interaction between thousands or more of them. It's too complex to understand by looking at the code.

However, for many applications it is very important that our AI systems are able to *explain* their decisions to us, so we can, for example, tell when an AI is discriminating against a group of people for a reason we feel violates their rights. To prevent these practical problems, as well as to better understand how intelligence works, a subfield has emerged to try to understand the workings of the sub-symbolic AI systems we build!

Of course, we have not tried to cover every topic in AI in this chapter. For example, over the past two decades, there has been much AI research on designing the *semantic web* (Berners-Lee, Hendler, & Lassila, 2001), a new version of the World Wide Web that would be capable of understanding information (e.g., web pages) stored on it. As another example, just over the past few years, *interactive games* have emerged as an important arena for AI research, especially agent-based AI.

Assessing Progress in AI

The task of measuring progress in AI is complex. In the past, tasks such as arithmetic and chess were considered to require intelligence. However, computers have been performing arithmetic calculations with great precision for more than seventy-five years and reliably

beating human grand masters at chess for more than twenty-five (Hsu, Campbell, & Hoane, 1995). Although in computer science these problems are sometimes still used to measure progress in speed of calculation and use of memory, very few humans now consider these computer programs as good manifestations of intelligence. One part of the difficulty is that some humans seem reluctant to ascribe intelligence to computers: Once a computer is able to address the problems that we once considered to require intelligence, we tend to dismiss them as not being very interesting. Further, once we understand how a computer actually solves these problems, for many humans, these problems lose some of their challenge.

Early in the history of AI, Turing (1950) proposed the most famous test for AI, called the *imitation game* or, as it is more popularly known, the “Turing test.” In this test, computers and human beings are put in (typed) chat sessions with human judges. If computers can reliably make the judges think they are human, they pass the test. Turing initially formulated this test in response to the question “Can machines think?” But rather than answering that question, he reformulated it into a more concrete question of whether a machine could fool a human interrogator into believing that the computer can think. Some interpretations of the Turing test take the purpose of the test as distinguishing computer programs that have human-level intelligence from those that do not (e.g., Harnad, 1992). In this interpretation, the test is not a measurement of intelligence in the sense of giving a score that accurately reflects cognitive abilities, but is a pass-or-fail litmus test of general intelligence.

It has proven to be very difficult for computers to pass the Turing test in general, although some surprisingly simple and old programs, such as ELIZA (Weizenbaum, 1966) and PARRY (Raphael, 1976), sometimes fool some people for short times. Because of the difficulty of the general Turing test, many competitions usually restrict judges to specific topics. Recently, there have been variations of the Turing test with prize monies such as the Loebner prize.

Recently there have been proposals (Bringsjord & Schimanski, 2003) for using psychometrics tests of human intelligence, such as the Wechsler test (1939) and the Raven’s test (1962) to measure progress in AI. However, there already exist computer programs that approach human performance on various versions of the Raven’s test including the Standard, Color, and Advanced Raven’s test (e.g., Kunda, McGreggor, & Goel., 2013). Other recent proposals (Marcus, Rossi, & Veloso, 2016) for measuring AI have covered a wide range, from playing soccer and winning the FIFA world championship to scientific discovery and winning the Nobel Prize.

AI and Society

As AI becomes more ubiquitous in our society, and affects more aspects of human life in greater ways, the question of ethical behavior becomes increasingly more important. Earlier we wrote about safety issues, using the example of our robot, Sunny. Many people worry that robots, as well as software AI agents, will continue to replace human jobs faster than society can create new ones. This isn't a safety issue that can be fixed in the code of a single agent, but rather a societal issue that needs to be dealt with on the level of laws and social norms.

Recently, several famous people such as Stephen Hawking have expressed fear that superintelligent AI might pose a threat to humanity's very existence. The reasoning goes something like this: At some point, an AI might be smarter than human beings, and have the power to rewrite its own code (this kind of AI is called a "seed AI"). Because the AI is smarter than any human, it will make itself smarter faster than humans can make it smarter. This will cause a "takeoff" that might be very rapid. Are there limits to how smart the software could get? We have no idea. Once the AI is many times smarter than any human, it will have enormous power to gain real control over world resources, using social manipulation, hacking, and other methods. No matter what the ultimate goals of the AI are, it would probably in the AI's interest to have the subgoals of self-preservation, cognitive enhancement, technological progress, resource acquisition, and prevention of its goals from being changed (the so-called instrumental convergence thesis, Boström, 2014). The need for good ethical reasoning at the start, in this scenario, is crucial because if the AI were to come to rule the world in pursuit of its goals, it might be difficult or impossible to change the ethics of the AI after it is many times smarter than us (and has an interest in preserving its values).

Of course this argument is very speculative and not based on any evidence. There are counterarguments suggesting that the above scenario is based on poor or very unlikely assumptions (e.g., Pinker, 2018). Further, there is also scholarship suggesting that superintelligent AI will be a force for good (Kurzweil, 2005). In any case, almost everyone agrees that we are nowhere near having a superintelligent AI. The study of ethical AI behavior is a growing field of interest. A related issue is whether or not we will someday need to have ethical considerations for the AIs themselves, should they ever be able to suffer pain.

Conclusions

In this chapter we have reviewed the history of AI and its major subfields, illustrated AI as a science and as a technology, examined its relationship to psychology, and discussed the problem of measuring the intelligence of AI agents. A somewhat surprising lesson from the history of AI is that it is relatively easy to make AI systems for some cognitive tasks that seem difficult for humans to solve (for example, mathematical, logical, and chess problems), and extraordinarily difficult to make computers solve some tasks that are apparently easy for humans to address (for example, seeing, walking, and talking). This apparent paradox has meant that repeated predictions about bold AI successes have gone unfulfilled.

We suggest two reasons for this paradox. First, our difficult problems require deliberate thought and strategies that are explicitly learned. As a result, we can often gain insight into how they are solved through observation and introspection. Indeed, many of these strategies are actually written down, meant to be learned through reading. In contrast, nobody needs to tell human beings how to see, walk, or speak. As a result, our intuitions about how these processes work are, to put it mildly, unhelpful.

The second, perhaps more important, reason is that deliberate processing is likely a serial process running as a virtual machine on a network of neurons, whereas the automatic processes, the seemingly easy tasks, are running directly on the neural network. These easy tasks (called System 1 in Stanovich & West, 2000) are evolutionarily older, and the parts of our brains that accomplish them (generally near the back of our brains) evolved to do just those things. In contrast, the more deliberate processing (System 2) is evolutionarily younger and makes use of the kind of hardware designed for System 1 tasks. System 2 struggles to do rational, serial processing on an essentially parallel pattern-matching machine (Kahneman, 2011; Stanovich, 2004).


Computers, and the languages we program them with, are naturally serial processors. When we implement artificial neural networks, we are doing it backward from nature: Whereas System 2 is a serial virtual machine running on parallel hardware, our artificial neural networks are parallel virtual machines running on serial hardware. Given this, and the fact that we have no conscious access to System 1 processes, it is no wonder that the AI community has had to work very hard to make progress in these areas. As a result, we have chess programs that can beat world grand masters, but no robots that can walk down a street even as well as a five-

year-old child. We expect that neuroscience findings may illuminate the nature of these processes, and the AI community will be able to build on them.



Given the track record of predictions about the future of AI, we will refrain from making our own. What we will claim is that AI already has had a profound impact not only on computer science and information technology but also more generally on our culture and our philosophy. The field has made so much progress that the Association for Advancement of Artificial Intelligence (AAAI; www.aaai.org) organizes multiple conferences every year, including one for deployed AI applications. If the past fifty-year history of AI is any guide, then the next fifty years will not only be full of exciting discoveries and bold inventions, but they will also raise new questions about who we are as humans and what we want to be.

Footnotes

Acknowledgments: We thank the editors of this *Handbook of Intelligence* as well as members of the Design and Intelligence Laboratory at Georgia Institute of Technology for their comments on earlier drafts of this chapter. An earlier version of this chapter appears in the 2011 edition of the *Handbook*.

-  ¹ Much of our discussion of this problem is based on the work of the first author and his students in the 1990s when they developed a computer program called Router for addressing this class of problems (Goel et al., 1994) and instantiated Router on a mobile reactive robot called Stimpy (Ali & Goel, 1996). They also developed a knowledge-based shell called Autognostic for learning by reflection on the Router program embodied in Stimpy (Stroulia & Goel, 1999), as well as reflection on Stimpy's reactive controller (Goel et al., 1997).
-

References

-  Albus, J. S. (1991). Outline for a theory of intelligence. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 473–509.
-  Ali, K., & Goel, A. (1996). Combining navigational planning and reactive control. *Proceedings of the AAAI-96 Workshop on Reasoning About*

Actions, Planning and Control: Bridging the Gap (pp. 1–7). Portland: AAAI Press.



Anderson, J. R. (2013). The adaptive character of thought. New York: Psychology Press.



Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.



Arkin, R. (1999). Behavior-based robotics. Cambridge, MA: MIT Press.



Berners-Lee, T., Hendler, J., & Lassila, O. (2001). Semantic web. *Scientific American*, 284(5), pp. 35–43.



Besold, T., Schlorlemmer, M., & Smaill, A. (Eds.) (2015) Computational creativity research: Towards creative machines. New York: Atlantis Press.



Boström, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford: Oxford University Press.



Bringsjord, S., & Schimanski, B. (2003). What is artificial intelligence? Psychometric AI as an answer. *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03)* (pp. 887–893). San Francisco: Morgan Kaufmann.



Buchanan, B., & Shortliffe, E. (1984). Rule based expert systems: The Mycin experiments of the Stanford Heuristic Programming Project. Boston: Addison-Wesley.



Cox, M., & Raja, A. (2011), *Metareasoning: Thinking about thinking*, Cambridge, MA: MIT Press.



Craik, K. (1943). The nature of explanation. Cambridge, UK: Cambridge University Press.



Davies, J., & Francis, A. G. (2013). The role of artificial intelligence research methods in cognitive science. In West, R. & Stewart, T. (Eds.), *Proceedings of the 12th International Conference on Cognitive Modeling* (pp. 439–444). Ottawa: Carleton University.

- ⤴ Evans, T. G. (1968). A program for the solution of a class of geometric-analogy intelligence-test questions. In Minsky, M. (Ed.), *Semantic information processing* (pp. 271–353). Cambridge, MA: MIT Press.
- ⤴ Ford, K., Hayes, P., Glymour, C., & Allen, J. (2015). Cognitive orthoses: Toward human-centered AI. *AI Magazine*, 36(4), 5–8.
- ⤴ Glasgow, J., Narayanan, N. H., & Chandrasekaran, B. (Eds.) (1995). *Diagrammatic reasoning: Cognitive and computational perspectives*. Cambridge, MA: MIT Press.
- ⤴ Goel, A., Ali, K., Donnellan, M., Gomez, A., & Callantine, T. (1994). Multistrategy adaptive navigational path planning. *IEEE Expert*, 9(6), 57–65.
- ⤴ Goel, A., Stroulia, E., Chen, Z., & Rowland, P. (1997). Model-based reconfiguration of schema-based reactive control architectures. In *Proceedings of the AAAI Fall Symposium on Model-Directed Autonomous Systems* (pp. 1–6). Cambridge, MA: AAAI.
- ⤴ Harnad, S. (1992). The Turing test is not a trick: Turing indistinguishability is a scientific criterion. *SIGART Bulletin*, 3(4), 9–10.
- ⤴ Hsu, F., Campbell, M., & Hoane, A. (1995). Deep Blue system overview. In Wolfe, M. (Ed.), *Procs. the 1995 International Conference on Supercomputing* (pp. 240–244). New York: ACM Press.
- ⤴ Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.
- ⤴ Kolodner, J. (1993). *Case-based reasoning*. San Francisco: Morgan Kaufmann.
- ⤴ Kotseruba, I., Gonzalez, O., & Tsotsos, J. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. *The Computing Research Repository (CoRR)*. arXiv preprint arXiv:1610.08602, 1–74.
- ⤴ Kunda, M., McGreggor, K., & Goel, A. (2013). A computational model for solving problems from the Raven's Progressive Matrices intelligence test

using iconic visual representations. *Cognitive Systems Research*, 22, 47–66.



Kurzweil, R. (2005). *The singularity is near: When humans transcend biology*. New York: Viking Adult.



Laird, J. E. (2012). *The Soar cognitive architecture*. Cambridge, MA: MIT press.



Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine*, 38(4). <https://doi.org/10.1609/aimag.v38i4.2744>



Laird, J., Newell, A., & Rosenbloom, P. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, 33, 1–64.



Langley, P. (2012). The cognitive systems paradigm. *Advances in Cognitive Systems*, 1, 3–13.



Langley, P., Laird, J., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2), 141–160.



LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.



Lenat, D., & Guha, R. (1990). *Building large knowledge based systems: Representation and inference in the Cyc project*. Boston: Addison-Wesley Longman.



Lindsay, R., Buchanan, B., Feigenbaum, E., & Lederberg, J. (1980). *Applications of artificial intelligence for chemical inference: The Dendral project*. New York: McGraw-Hill.



Marcus, G., Rossi, F., & Veloso, M. (2016). Beyond the Turing test. Special issue, *AI Magazine*, 37(1), 3–101.



Marr, D. (1982). *Vision*. New York: Henry Holt.



McCarthy, J. (1988). Mathematical logic in AI. *Daedalus*, 117(1), 297–311.

- ⤴ McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955/2006). A proposal for the Dartmouth summer research project on artificial intelligence. *AI Magazine*, 27(4), 12–14.
- ⤴ McClelland, J. L., Rumelhart, D. E., & PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 2, *Psychological and biological models*. Cambridge, MA: MIT Press.
- ⤴ Minsky, M. L. (1975). *A framework for representing knowledge*. In Winston, P. H. (Ed.), *The psychology of computer vision* (pp. 1–82). New York: McGraw-Hill.
- ⤴ Minsky, M. L., & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- ⤴ Mitchell, M. (1998). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.
- ⤴ Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of problem solving. *Psychological Review*, 63(3), 151–166.
- ⤴ Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.
- ⤴ Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufman.
- ⤴ Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.
- ⤴ Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.
- ⤴ Pinker, S. (2018). *Enlightenment now: The case for reason. Science, humanism, and progress*. New York: Viking.
- ⤴ Quillian, M. (1968). Semantic Memory. In Minsky, M. (Ed.), *Semantic information processing* (pp. 227–270). Cambridge, MA: MIT Press.
- ⤴ Rabiner, L., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, January, 4–16.



Raphael, B. (1976). The thinking computer. New York: W. H. Freeman.



Raven, J. C. (1962). Advanced Progressive Matrices Set II. London: H. K. Lewis.



Rumelhart, D. E., McClelland, J. L., & PDP Research Group (1986). Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1, *Foundations*. Cambridge, MA: MIT Press.



Samsonovich, A. V. (2010). Toward a unified catalog of implemented cognitive architectures. In Samsonovich, A. V., Jóhannsdóttir, K. R., Chella, A., & Goertzel, B. (Eds.), *Proceeding of the Conference on Biologically Inspired Cognitive Architectures* (pp. 195–244). New York: IOS Press.



Schank, R. C. (1975). *Conceptual information processing*. New York: Elsevier.



Schank, R. C. (1982). *Dynamic memory* (2nd ed.). New York: Cambridge University Press.



Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals and understanding*. Hillsdale, NJ: Erlbaum.



Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., et al. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.



Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). *Mastering the game of Go without human knowledge*. *Nature*, 550(7676), 354–359.



Simon, H. A. (1996). *Sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.



Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In Meersman, R. & Tari, Z. (Eds.), *On the Move to Meaningful Internet Systems: OTM Confederated International Conferences* (pp. 1223–1237). Berlin: Springer.

- ⤴ Sowa, J. (1987). Semantic networks. In Shapiro, S. (Ed.), *Encyclopedia of AI* (pp. 1011–1024). New York: Wiley.
- ⤴ Stanovich, K. E. (2004). *The robot's rebellion*. Chicago: University of Chicago Press.
- ⤴ Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, 23, 645–726.
- ⤴ Stroulia, E., & Goel, A. K. (1999). Evaluating problem-solving methods in evolutionary design: The autognostic experiments. *International Journal of Human-Computer Studies*, 51, 825–847.
- ⤴ Sutton, R. S., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- ⤴ Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.
- ⤴ Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460.
- ⤴ Veale, T., & Cardoso, A. (2018). *Computational creativity: The philosophy and engineering of autonomously creative systems*. Berlin: Springer.
- ⤴ Von Anh, L., Liu, R., & Blum, M. (2006). Peekaboom: A game for locating objects in images. In Grinter, R., Rodden, T., Aoki, P., Cutrell, E., Jeffries, R., & Olson, G. (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montreal, April 22–27) (pp. 55–64). New York: ACM Press.
- ⤴ Wechsler, D. (1939). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.
- ⤴ Weiner, N. (1961). *Cybernetics* (2nd ed.). Cambridge, MA: MIT Press.
- ⤴ Weizenbaum, J. (1966). ELIZA – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45.



Winograd, T. (1972). Understanding natural language. San Diego, CA: Academic Press.

Related content

AI-generated results: by UNSILO

Book

Algo Bots and the Law

Gregory Scopino

[Algo Bots and the Law: Technology, Automation, and the Regulation of Futures and Other Derivatives](#)

Published online: 2 October 2020

Chapter

Artificial Intelligence

Ashok K. Goel and Jim Davies

[The Cambridge Handbook of Intelligence](#)

Published online: 5 June 2012

Chapter

Key Concepts: Algorithms, Artificial Intelligence, and More

Gregory Scopino

[Algo Bots and the Law](#)

Published online: 2 October 2020

Element

Strategizing AI in Business and Education

Aleksandra Przegalinska and Dariusz Jemielniak

Published online: 21 March 2023

Article

The reality of the symbolic and subsymbolic systems

Andrew Woodfield and Adam Morton

[Behavioral and Brain Sciences](#)

Published online: 4 February 2010

Chapter

Combating Bias in AI and Machine Learning in Consumer-Facing Services

Charlyn Ho , Marc Martin , Divya Taneja , D. Sean West , Sam Boro and Coimbra Jackson

The Cambridge Handbook of Artificial Intelligence

Published online: 28 July 2022

Article

Making the connections

Jay G. Rueckl

Behavioral and Brain Sciences

Published online: 4 February 2010

Article

Connectionism in the golden age of cognitive science

Dan Lloyd

Behavioral and Brain Sciences

Published online: 4 February 2010

Article

Two constructive themes

Richard K. Belew

Behavioral and Brain Sciences

Published online: 4 February 2010

Article

The promise and problems of connectionism

Michael G. Dyer

Behavioral and Brain Sciences

Published online: 4 February 2010