Chapter 2
What Is AI?

In this chapter we discuss the different definitions of Artificial Intelligence (AI). We then discuss how machines learn and how a robot works in general. Finally we discuss the limitations of AI and the influence the media has on our preconceptions of AI.

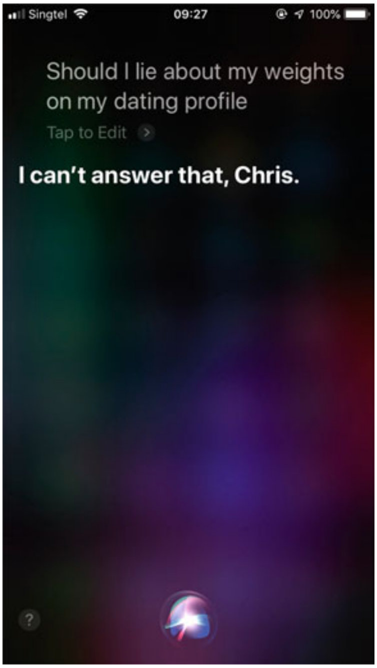
CHRIS: Siri, should I lie about my weight on my dating profile?
SIRI: I can't answer that, Chris.

Siri is not the only virtual assistant that will struggle to answer this question (see Fig. 2.1). Toma et al. (2008) showed that almost two thirds of people provide inaccurate information about their weight on dating profiles. Ignoring, for a moment, what motivates people to lie about their dating profiles, why is it so difficult, if not impossible, for digital assistants to answer this question?

To better understand this challenge it is necessary to look behind the scene and to see how this question is processed by Siri. First, the phone's microphone needs to translate the changes in air pressure (sounds) into a digital signal that can then be stored as data in the memory of the phone. Next, this data needs to be sent through the internet to a powerful computer in the cloud. This computer then tries to classify the sounds recorded into written words. Afterwards, an artificial intelligence (AI) system needs to extract the meaning of this combination of words. Notice that it even needs to be able to pick the right meaning for the homophone "lie". Chris does not want to lie down on his dating profile, he is wondering if he should put inaccurate information on it.

While the above steps are difficult and utilise several existing AI techniques, the next step is one of the hardest. Assuming Siri fully understands the meaning of Chris's question, what advice should Siri give? To give the correct advice, it would need to know what a person's weight means and how the term relates to their attractiveness. Siri needs to know that the success of dating depends heavily on both

C. Bartneck et al., *An Introduction to Ethics in Robotics and AI*,
SpringerBriefs in Ethics, https://doi.org/10.1007/978-3-030-51110-4_2



Should I lie about my weights on my dating profile

I can't answer that, Chris.

Fig. 2.1 Siri's response to a not so uncommon question

participants considering each other attractive—and that most people are motivated to date. Furthermore, Siri needs to know that online dating participants cannot verify the accuracy of information provided until they meet in person. Siri also needs to know that honesty is another attribute that influences attractiveness. While deceiving potential partners online might make Chris more attractive in the short run, it would have a negative effect once Chris meets his date face-to-face.

But this is not all. Siri also needs to know that most people provide inaccurate information on their online profiles and that a certain amount of dishonesty is not likely to impact Chris's long-term attractiveness with a partner. Siri should also be aware that women select only a small portion of online candidates for first dates and that making this first cut is essential for having any chance at all of convincing the potential partners of Chris's other endearing qualities.

There are many moral approaches that Siri could be designed to take. Siri could take a consequentialist approach. This is the idea that the value of an action depends on the consequences it has. The best known version of consequentialism is the classical utilitarianism of Jeremy Bentham and John Stuart Mill (Bentham 1996; Mill 1863). These philosophers would no doubt advise Siri to maximise happiness: not just Chris's happiness but also the happiness of his prospective date. So, on the consequentalist approach Siri might give Chris advice that would maximise his chances to not only to have many first dates, but maximise the chances for Chris to find true love.

Alternatively, Siri might be designed to take a deontological approach. A deontologist like Immanuel Kant might prioritise duty over happiness. Kant might advise Chris that lying is wrong. He has a duty not to lie so he should tell the truth about his weight, even if this would decrease his chances of getting a date.

A third approach Siri could take would be a virtue ethics approach. Virtue ethics tend to see morality in terms of character. Aristotle might advise Chris that his conduct has to exhibit virtues such as honesty.

Lastly, Siri needs to consider whether it should give a recommendation at all. Providing wrong advice might damage Siri's relationship to Chris and he might consider switching to another phone with another digital assistant. This may negatively impact Apple's sales and stock value.

This little example shows that questions that seem trivial on the surface might be very difficult for a machine to answer. Not only do these machines need the ability to process sensory data, they also need to be able to extract the correct meaning from it and then represent this meaning in a data structure that can be digitally stored. Next, the machine needs to be able to process the meaning and conclude with desirable actions. This whole process requires knowledge about the world, logical reasoning and skills to learn and adapt. Having these abilities may make the machine **autonomous**.

There are various definitions of "autonomy" and "autonomous" in AI, robotics and ethics. At its simplest, autonomous simply refers to the ability of a machine to operate for a period of time without a human operator. Exactly what that means differs from application to application. What is considered "autonomous" in a vehicle is different to what is considered "autonomous" in an weapon. In bioethics autonomy refers to the ability of humans to make up their own minds about what treatment to accept or refuse. In Kantian ethics autonomy refers to the ability of humans to decide what to do with their lives and what moral rules to live by. The reader should be aware that exactly what "autonomous" means is context-sensitive. Several meanings are presented in this book. The unifying underlying idea is self-rule (from the Greek words "auto" meaning self and "nomos" meaning rule).

On the first of these definitions, Siri is an autonomous agent that attempts to answer spoken questions. Some questions Siri tries to answer require more intelligence, meaning more background, reasoning ability and knowledge, than others. The chapter that follows define and describe the characteristics that make something artificially intelligent and an agent.

2.1 Introduction to AI

The field of artificial intelligence (AI) has evolved from humble beginnings to a field with global impact. The definition of AI and of what should and should not be included has changed over time. Experts in the field joke that AI is everything that computers cannot currently do. Although facetious on the surface, there is a sense that developing intelligent computers and robots means creating something that does not exist today. Artificial intelligence is a moving target.

Indeed, even the definition of AI itself is volatile and has changed over time. Kaplan and Haenlein define AI 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” (Kaplan and Haenlein 2019). Poole and Mackworth (2010) define AI as “the field that studies the synthesis and analysis of computational agents that act intelligently.” An agent is something (or someone) that acts. An agent is intelligent when:

1. its actions are appropriate for its circumstances and its goals
2. it is flexible to changing environments and changing goals
3. it learns from experience, and
4. it makes appropriate choices given its perceptual and computational limitations.

Russell and Norvig define AI as “the study of [intelligent] agents that receive precepts from the environment and take action. Each such agent is implemented by a function that maps percepts to actions, and we cover different ways to represent these functions, such as production systems, reactive agents, logical planners, neural networks, and decision-theoretic systems” Russell and Norvig (2010, p. viii).

Russell and Norvig also identify four schools of thought for AI. Some researchers focus on creating machines that think like humans. Research within this school of thought seeks to reproduce, in some manner, the processes, representations, and results of human thinking on a machine. A second school focuses on creating machines that act like humans. It focuses on action, what the agent or robot actually does in the world, not its process for arriving at that action. A third school focuses on developing machines that act rationally. Rationality is closely related to optimality. These artificially intelligent systems are meant to always do the right thing or act in the correct manner. Finally, the fourth school is focused on developing machines that think rationally. The planning and/or decision-making that these machines will do is meant to be optimal. Optimal here is naturally relevant to some problems that the system is trying to solve.

We have provided three definitions. Perhaps the most basic element common to all of them is that AI involves the study, design and building of intelligent agents that can achieve goals. The choices an AI makes should be appropriate to its perceptual and cognitive limitations. If an AI is flexible and can learn from experience as well as sense, plan and act on the basis of its initial configuration, it might be said to be more intelligent than an AI that just has a set of rules that guides a fixed set of actions. However, there are some contexts in which you might not want the AI to learn new rules and behaviours, during the performance of a medical procedure, for example. Proponents of the various approaches tend to stress some of these elements more than others. For example, developers of expert systems see AI as a repository of expert knowledge that humans can consult, whereas developers of machine learning systems see AI as something that might discover new knowledge. As we shall see, each approach has strengths and weaknesses.

2.1.1 The Turing Test

In 1950 Alan Turing (see Fig. 2.2) suggested that it might be possible to determine if a machine is intelligent based on its ability to exhibit intelligent behaviour which is indistinguishable from an intelligent human's behaviour. Turing described a conversational agent that would be interviewed by a human. If the human was unable to determine whether or not the machine was a person then the machine would be viewed as having passed the test. Turing's argument has been both highly influential and also very controversial. For example, Turing does not specify how long the

Fig. 2.2 Alan Turing (1912–1954) (*Source* Jon Callas)

human would have to talk to the machine before making a decision. Still, the Turing Test marked an important attempt to avoid ill-defined vague terms such as "thinking" and instead define AI with respect to a testable task or activity.

2.1.2 Strong and Weak AI

John Searle later divided AI into two distinct camps. Weak AI is limited to a single, narrowly defined task. Most modern AI systems would be classified in this category. These systems are developed to handle a single problem, task or issue and are generally not capable of solving other problems, even related ones. In contrast to weak AI, Searle defines strong AI in the following way: "The appropriately programmed computer with the right inputs and outputs would thereby have a mind in exactly the same sense human beings have minds" (Searle 1980). In strong AI, Searle chooses to connect the achievement of AI with the representation of information in the human mind. While most AI researchers are not concerned with creating an intelligent agent that meets Searle's strong AI conditions, these researchers seek to eventually create machines for solving multiple problems which are not narrowly defined. Thus one of the goals of AI is to create autonomous systems that achieve some level of general intelligence. No AI system has yet achieved general intelligence.

2.1.3 Types of AI Systems

There are many different types of AI systems. We will briefly describe just a few. Knowledge representation is an important AI problem that tries to deal with how information should be represented in order for a computer to organise and use this information. In the 1960s, expert systems were introduced as knowledge systems that can be used to answer questions or solve narrowly defined problems in a particular domain. They often have embedded rules that capture knowledge of a human expert. Mortgage loan advisor programs, for example, have long been used by lenders to evaluate the credit worthiness of an applicant. Another general type of AI system are planning systems. Planning systems attempt to generate and organise a series of actions which may be conditioned on the state of the world and unknown uncertainties. The Hubble telescope, for example, utilised an AI planning system called SPIKE.

Computer vision is a subfield of AI which focuses on the challenge of converting data from a camera into knowledge representations. Object recognition is a common task often undertaken by computer vision researchers. Machine learning focuses on developing algorithms the allow a computer to use experience to improve its performance on some well-defined task. Machine learning is described in greater detail in the sections below.

AI currently works best in constrained environments, but has trouble with open worlds, poorly defined problems, and abstractions. Constrained environments include simulated environments and environments in which prior data accurately reflects future challenges. The real world, however, is open in the sense that new challenges arise constantly. Humans use solutions to prior related problems to solve new problems. AI systems have limited ability to reason analogically from one situation to another and thus tend to have to learn new solutions even for closely related problems. In general, they lack the ability to reason abstractly about problems and to use common sense to generate solutions to poorly defined problems.

2.2 What Is Machine Learning?

Machine learning is a sub-field of AI focused on the creation of algorithms that use experience with respect to a class of tasks and feedback in the form of a performance measure to improve their performance on that task. Contemporary machine learning is a sprawling, rapidly changing field. Typically machine learning is sub-categorised into three types of learning.

Supervised learning centres on methods such as regression and classification. To solve a classification problem experiences in the form of data are labelled with respect to some target categorisation. The labelling process is typically accomplished by enlisting the effort of humans to examine each piece of data and to label the data. For supervised learning classification problems performance is measured by calculating the true positive rate (the ratio of the true positives over all positives, correctly labelled or not) and the false positive rate (the ratio of false positives over all negatively classified data, correctly and incorrectly labelled). The result of this machine learning process is called a *classifier*. A classifier is software that can automatically predict the label of a new piece of data. A machine learning classifier that categorises labelled data with a true positive rate of 100% and a false positive rate of 0% is a perfect classifier. The supervised learning process then is the process by which unlabelled data is fed to a developing classifier and, over the course of working through some training data, the classifier's performance improves. Testing the classifier requires the use of a second label data-set called the test data set. In practice, often one overall data-set is carved into a training and test set on which the classifier is then trained and tested. The testing and training process may be time-consuming, but once a classifier is created it can be used to quickly categorise incoming data.

Unsupervised learning is more focused on understanding data patterns and relations than on prediction. It involves methods such as principal components analysis and clustering. These are often used as exploratory precursors to supervised learning methods.

Reinforcement learning is a third type of machine learning. Reinforcement learning does not focus on the labelling of data, but rather attempts to use feedback in the form of a reinforcement function to label states of the world as more or less desirable with respect to some goal. Consider, for example, a robot attempting to move from one location to another. If the robot's sensors provide feedback telling it its distance from a goal location, then the reinforcement function is simply a reflection of the sensor's readings. As the robot moves through the world it arrives at different locations which can be described as states of the world. Some world states are more rewarding than others. Being close to the goal location is more desirable than being further away or behind an obstacle. Reinforcement learning learns a policy, which is a mapping from the robot's action to expected rewards. Hence, the policy tells the system how to act in order to achieve the reward.

2.3 What Is a Robot?

Typically, an artificially intelligent agent is software that operates online or in a simulated world, often generating perceptions and/or acting within this artificial world. A robot, on the other hand, is situated in the real world, meaning that its existence and operation occur in the real world. Robots are also embodied, meaning that they have a physical body. The process of a robot making intelligent decisions is often described as "sense-plan-act" meaning that the robot must first sense the environment, plan what to do, and then act in the world.

2.3.1 Sense-Plan-Act

A robot's embodiment offers some advantages in that its experiences tend to be with real objects, but it also poses a number of challenges. Sensing in the real world is extremely challenging. Sensors such as cameras, laser scanners, and sonar all have limitations. Cameras, for example, suffer from colour shifts whenever the amount of light changes. Laser scanners have difficulty perceiving transparent objects. Converting sensor data into a usable representation is challenging and can depend on the nature and limitations of the sensor. Humans use a wide array of integrated sensors to generate perceptions. Moreover, the number of these sensors is (at least currently) much higher than the number of sensors of any robot. The vast amount of sensors available to a human is advantageous in terms of uncertainty reduction of perception. Humans also use a number different brain structures to encode information, to perform experience-based learning, and to relate this learning to other knowledge and experiences. Machines typically cannot achieve this type of learning.

Planning is the process by which the robot makes use of its perceptions and knowledge to decide what to do next. Typically, robot planning includes some type of goal that the robot is attempting to achieve. Uncertainty about the world must be dealt with at the planning stage. Moreover, any background or historical knowledge that the system has can be applied at this stage.

Finally, the robot acts in the world. The robot must use knowledge about its own embodiment and body schema to determine how to move joints and actuators in a manner dictated by the plan. Moreover, once the robot has acted it may need to then provide information to the sensing process in order to guide what the robot should look for next.

It should be understood that AI agents and robots have no innate knowledge about the world. Coming off the factory production line a robot or AI is a genuine "blank slate" or to be more exact an unformatted drive. Babies, on the other hand, enter the world "pre-programmed" so to speak with a variety of innate abilities and knowledge. For example, at birth babies can recognise their mother's voice. In contrast, AI agents know nothing about the world that they have not been explicitly programmed to know. Also in contrast to humans, machines have limited ability to generate knowledge from perception. The process of generating knowledge from information requires that the AI system creates meaningful representations of the knowledge. As mentioned above, a representation is a way of structuring information in order to make it meaningful. A great deal of research and debate has focused on the value of different types of representations. Early in the development of AI, symbolic representations predominated. A symbolic representation uses symbols, typically words, as the underlying representation for an object in the world. For example, the representation of the object apple would be little more than "Apple." Symbolic representations have the value of being understandable to humans but are otherwise very limiting because they have no precise connection to the robot's or the agent's sensors. Non-symbolic representations, on the other hand, tend not to be easily understood, but tend to relate better to a machine's sensors.

2.3.2 System Integration. Necessary but Difficult

In reality, to develop a working system capable of achieving real goals in the real world, a vast array of different systems, programmes and processes must be integrated to work together. System integration is often one of the hardest parts of building a working robotic system. System integrators must deal with the fact that different information is being generated by different sensors at different times. The different sensors each have unique limitations, uncertainties, and failure modes, and the actuators may fail to work in the real world. For all of these reasons, creating artificially intelligent agents and robots is extremely challenging and fraught with difficulties.

2.4 What Is Hard for AI

The sections above have hinted at why AI is hard. It should also be mentioned that not all software is AI. For example, simple sorting and search algorithms are not considered intelligent. Moreover, a lot of non-AI is smart. For example, control algorithms and optimisation software can handle everything from airline reservation systems to the management of nuclear power plants. But they only take well-defined actions within strictly defined limits. In this section, we focus on some of the major challenges that make AI so difficult. The limitations of sensors and the resulting lack of perception have already been highlighted.

AI systems are rarely capable of generalising across learned concepts. Although a classifier may be trained on very related problems, typically classifier performance drops substantially when the data is generated from other sources or in other ways. For example, face recognition classifiers may obtain excellent results when faces are viewed straight on, but performance drops quickly as the view of the face changes to, say profile. Considered another way, AI systems lack robustness when dealing with a changing, dynamic, and unpredictable world. As mentioned, AI systems lack common sense. Put another way, AI systems lack the enormous amount of experience and interactions with the world that constitute the knowledge that is typically called common sense. Not having this large body of experience makes even the most mundane task difficult for a robot to achieve. Moreover, lack of experience in the world makes communicating with a human and understanding a human's directions difficult. This idea is typically described as common ground.

Although a number of software systems have claimed to have passed the Turing test, these claims have been disputed. No AI system has yet achieved strong AI, but some may have achieved weak AI based on their performance on a narrow, well-defined task (like beating a grandmaster in chess or Go, or experienced players in Poker). Even if an AI agent is agreed to have passed the Turing test, it is not clear whether the passing of the test is a necessary and sufficient condition for intelligence.

AI has been subject to many hype cycles. Often even minor advancements have been hailed as major breakthroughs with predictions of soon to come autonomous intelligent products. These advancements should be considered with respect to the narrowness of the problem attempted. For example, early types of autonomous cars capable of driving thousands of miles at a time (under certain conditions) were already being developed in the 1980s in the US and Germany. It took, however, another 30+ years for these systems to just begin to be introduced in non-research environments. Hence, predicting the speed of progression of AI is very difficult—and in this regard, most prophets have simply failed.

2.5 Science and Fiction of AI

Artificial Intelligence and robotics are frequent topics in popular culture. In 1968, the Stanley Kubrick classic "2001" featured the famous example of HAL, a spacecraft's intelligent control system which turns against its human passengers. The Terminator movies (since 1984) are based on the idea that a neural network built for military defense purposes gains self-awareness and, in order to protect itself from deactivation by its human creators, turns against them. The Steven Spielberg's movie "A.I." (2001), based on a short story by Brian Aldiss, explores the nature of an intelligent robotic boy (Aldiss 2001). In the movie “I, Robot” (2004), based on motives from a book by Isaac Asimov, intelligent robots originally meant to protect humans are turning into a menace. A more recent example is the TV show “Westworld” (since 2016) in which androids entertain human guests in a Western theme park. The guests are encouraged to live out their deepest fantasies and desires.

For most people, the information provided through these shows is their first exposure to robots. While these works of fiction draw a lot of attention to the field and inspire our imagination, they also set a framework of expectations that can inhibit the progress of the field. One common problem is that the computer systems or robots shown often exhibit levels of intelligence that are equivalent or even superior to that of humans or current systems. The media thereby contributes to setting very high expectations in the audience towards AI systems. When confronted with actual robots or AI systems, people are often disappointed and have to revise their expectations. Another issue is the frequent repetition of the “Frankenstein Complex” as defined by Isaac Asimov. In this trope, bands of robots or an AI system achieve consciousness and enslave or kill (all) humans. While history is full of examples of colonial powers exploiting indigenous populations, it does not logically follow that an AI system will repeat these steps. A truly intelligent system will (hopefully) have learned from humanity’s mistakes. Another common and rather paradoxical trope is the assumption that highly intelligent AI systems desire to become human. Often the script writers use the agent’s lack of emotions as a the missing piece of the puzzle that would make them truly human.

It is important to distinguish between science and fiction. The 2017 recommendation to the European Parliament to consider the establishment of electronic personalities (Delvaux 2017) has been criticised by many as a premature reflex to the depiction of robots in the media.[1] For example, granting the robot “Sophia” Saudi Arabian citizenship in October 2017 can in this respect be considered more as a successful public relations stunt (Reynolds 2018) than as a contribution to the field of AI or its ethical implications. Sophia’s dialogues are based on scripts and cannot therefore be considered intelligent. It does not learn nor is it able to adapt to unforeseen circumstances. Sophia’s presentation at the United Nation is an unconvincing demonstration of artificial intelligence. People do anthropomorphise robots and autonomous systems, but this does not automatically justify the granting of personhood or other forms of legal status. In the context of autonomous vehicles, it may become practical to consider such a car a legal entity, similar to how we consider an abstract company to be a legal person. But this choice would probably be motivated more out of legal practicality than out of existential necessity.

[1] http://www.robotics-openletter.eu/.

Discussion Questions:

- Explain the difference between weak and strong AI. Give examples from science fiction describing machines that could be categorised as displaying strong and weak AI.
- Given the description of supervised machine learning above, how might a classifier come to include societal biases? How might the removal of such biases impact classifier performance? Describe a situation in which stakeholders must balance the tradeoff between bias and performance.
- Consider the sense-plan-act paradigm described above. How might errors at one step of this process impact the other steps? Draw an informal graph of robot performance versus time.

Further Reading:

- Stuart J. Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, Upper Saddle River, N.J, 3rd edition, 2010. ISBN 9780132071482. URL http://www.worldcat.org/oclc/688385283
- Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. Machine learning: *An artificial intelligence approach*. Springer Science & Business Media, 2013. ISBN 978-3662124079. URL http://www.worldcat.org/oclc/864590508
- Sidney Perkowitz. *Digital people: From bionic humans to androids*. Joseph Henry Press, 2004. ISBN 978-0309096195. URL http://www.worldcat.org/oclc/936950712.