CHAPTER 1
Introduction

We will adopt the overall goal of artificial intelligence (AI) to be 'to build machines with minds, in the full and literal sense' as prescribed by the Canadian philosopher John Haugeland (1985).

Not to create machines with a clever imitation of human-like intelligence. Or machines that exhibit behaviours that would be considered intelligent if done by humans – but to build machines that reason.

This book focuses on search methods for problem solving. We expect the user to define the goals to be achieved and the domain description, including the moves available with the machine. The machine then finds a solution employing first principles methods based on search. A process of trial and error. The ability to explore different options is fundamental to thinking.

As we describe subsequently, such methods are just amongst the many in the armoury of an intelligent agent. Understanding and representing the world, learning from past experiences, and communicating with natural language are other equally important abilities, but beyond the scope of *this* book. We also do not assume that the agent has meta-level abilities of being self-aware and having goals of its own. While these have a philosophical value, our goal is to make machines do something useful, with as general a problem solving approach as possible.

This and other definitions of what AI is do not prescribe *how to test* if a machine is intelligent. In fact, there is no clear-cut universally accepted definition of intelligence. To put an end to the endless debates on machine intelligence that ensued, the brilliant scientist Alan Turing proposed a behavioural test.

1.1 Can Machines Think?

Ever since the possibility of building intelligent machines arose, there have been raging debates on whether machine intelligence is possible or not. All kinds of arguments have been put forth both for and against the possibility. It was perhaps to put an end to these arguments that Alan Turing (1950) proposed his famous *imitation game*, which we now call the *Turing Test*. The test is simply this: if a machine interacts with a human using text messages and can fool human judges a sufficiently large fraction of times that they are chatting with another human, then we can say that the machine has passed the test and is intelligent.

Since then, many programs have produced text based interactions that are convincingly human-like, for example, *ChatGPT*[1] being one of the latest. Advances in machine learning algorithms for building language models from large amounts of training data have enabled machines to churn out remarkably well structured impressive text. Humans are quite willing to believe that if it talks like a human, then it must think like a human. Even when the very first chat program *Eliza* threw back user sentences with an interrogative twist, its creator Edward Weizenbaum was shocked to discover that his secretary was confiding her personal woes to the program (Weizenbaum, 1966). Pamela McCorduck (2004) has observed in *Machines Who Think* that in medieval Europe people were willing to ascribe intelligence to mechanical toy statues that could nod or shake their head in response to a question.

Clearly relying on human impressions based on interaction in natural language is not the best way of determining whether a machine is intelligent or not. With more and more machines becoming good at generating text rivalling that produced by humans, a need is being felt for something that delves deeper and tests whether the machine is actually reasoning when answering questions.

Hector Levesque and colleagues have proposed a new test of intelligence which they call the *Winograd Schema Challenge*, after Terry Winograd who first suggested it (Levesque et al., 2012; Levesque, 2017). The idea is that the test cannot be answered by having a large language model or access to the internet but would need common sense knowledge about the world. The test subject is given a sentence that refers to two entities of the same kind and a pronoun that could refer to either one of them. The question is which one, and the task is called anaphora resolution. The ambiguity can easily be resolved by humans using common sense knowledge. The strategy is to have two variations of the sentence, each having a different word or a phrase that leads to different anaphora resolution. One of the versions is presented to the subject with a question about what the pronoun refers to. Guesswork on a series of such questions is only expected to produce about half the correct answers, whereas a knowledgeable (read intelligent) agent would do much better. The following is the example attributed to Winograd (1972).

- The town councillors refused to give the angry demonstrators a permit because they feared violence. Who feared violence?
  (a) The town councillors
  (b) The angry demonstrators
- The town councillors refused to give the angry demonstrators a permit because they advocated violence. Who advocated violence?
  (a) The town councillors
  (b) The angry demonstrators

In both cases, two options are given to the subject who has to choose one of the two. Here are two more examples of the Winograd Schema Challenge, with two sets of sentences, each one of which is presented and followed by a question.

- The trophy doesn't fit in the brown suitcase because it's too big. What is too big?
  (a) the trophy
  (b) the suitcase

[1] ChatGPT: Optimizing Language Models for Dialogue. https://openai.com/blog/chatgpt/, accessed December 2022.

- The trophy doesn't fit in the brown suitcase because it's too small. What is too small?
  (a) the trophy
  (b) the suitcase

The following sentence is from the First Winograd Challenge at the International Joint Conference on AI in 2016 (Davis et al., 2017).

- John took the water bottle out of the backpack so that it would be lighter.
- John took the water bottle out of the backpack so that it would be handy.

What does 'it' refer to? Again, two options are given to the subject who is asked to choose one.

The authors report that the Winograd Schema Test was preceded by a pronoun disambiguation test in a single sentence, with examples chosen from naturally occurring text. Only those programs that did well in the first test were allowed to advance to the Winograd Schema Test. Here is an example from their paper which has been taken from the story 'Sylvester and the Magic Pebble'.

- The donkey wished a wart on its hind leg would disappear, and it did.

What vanished? The important thing is that such problems can be solved only if the subject is well versed with sufficient common sense knowledge about the world and also the structure of language.

A question one might ask is why should a test of intelligence be language based? After all, intelligence manifests itself in other ways as well. Could one of these also be an indicator of intelligence?

One area that has been proposed is in the arts, where creativity is the driving force. Computer generated art has time and again come to the limelight. Many artworks by *AARON*, the drawing artist created by Harold Cohen (1928–2016), have been demonstrated at AI conferences over the years (Cohen, 2016). A slew of text-to-image AI systems including DALL-E, Midjourney, and Stable Diffusion have all been released for public use recently.

Erik Belgum and colleagues have proposed a Turing Test for musical intelligence (Belgum et al., 1989). In the fall of 1997, Douglas Hofstadter organized a series of five public symposia centred on the burning question 'Are Computers Approaching Human-Level Creativity?' at Indiana University. This fourth symposium was about a particular computer program, David Cope's *EMI* (Experiments in Musical Intelligence) as a composer of music in the style of various classical composers (Cope, 2004). A two-hour concert took place in which compositions written by EMI and compositions written by famous human composers were performed without identification, and the audience was asked to vote for which pieces they thought were human-composed and which were computer-composed. Subsequently, David Coco-Pope published an article written by a computer program *EWI* (Experiments in Written Intelligence) in the style of Hofstadter, grudgingly conceded by Hofstadter himself at the end of the article (Hofstadter, 2009).

After the 2011 spectacular win by IBM's program *Watson* in the game of *Jeopardy* over two players who were widely considered to be the best that the game had seen, the company unveiled a program *Chef Watson* with the following claim – 'In our application, a computationally creative computer can automatically design and discover culinary recipes that

are flavorful, healthy, and novel!'[2] The market is now abuzz with robots that can cook for you, for example, as reported in Cain (2022).

Recently, when DeepMind's *AlphaGo* program beat the reigning world champion Lee Sedol in the oriental game of *go*, the entire world sat up and took notice (Silver et al., 2016). This followed an equally impressive win almost twenty years earlier in 1997 when IBM's *Deep Blue* program beat the then world champion Garry Kasparov in the game of chess (Campbell et al., 2002). Both the games are two person board games in which programs can search game trees as described in Chapter 8. The challenge in these games is to search the huge trees that present themselves. While chess is played on an 8×8 board, *go* is played on a 19×19 board, which generates a much larger game tree. And yet a combination of machine learning and selective search proved invincible. Both these games are conceptually simple even though the search trees are large. In the author's opinion, only when a computer program can play the game of contract bridge at the level described in Ottlik and Kelsey (1983) can we legitimately stake a claim to have created an AI.

Meanwhile, one should perhaps take a cue from Alan Turing himself, move away from the bickering, and get on with the design and implementation of autonomous machines who[3] do useful things for us. In the summer of 1956, John McCarthy and Marvin Minsky had organized the Dartmouth Conference with the following stated goal – 'The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it' (McCorduck, 2004). *That* is the spirit of our quest for AI.

1.2 Problem Solving

Our quest is for a machine that is autonomous and whose behaviour is goal directed. Whatever it does, it should do autonomously. We imagine a scenario in which the machine is an agent to serve the goals given it to it by a *user*. Current applications take a short horizon view of achieving specific goals, though we can imagine a persistent agent engaging with the human over long periods, perhaps even the user's lifetime. We ignore the doomsday scenarios in which machines overcome and subjugate humans, though this idea has been fashionable amongst certain sections of science writers. That so-called *singularity* is not even on the horizon (Larson, 2021).

We want our machines to solve problems for us. Given a set of goals, the machine must engage with the world to achieve those goals. The goals may be short term or long term, and the world in which the problem solving agent operates may be changing, even in the simplest case when the agent is the only one acting and effecting the change. The agent must sense its environment, deliberate over its goals, and act in the domain. The agent must not just be reactive, operating in a hard-coded *stimulus–response* cycle, but should be able to act flexibly in a *sense–deliberate–act* cycle autonomously. A schematic of an autonomous agent is shown in Figure 1.1.

In all life forms, deliberation happens in the brain. Incoming sensory data is processed in the context of what the creature already knows. Our understanding of the animal brain is that it is a collection of a large number of very simple processing components called *neurons*. Each

[2] https://researcher.watson.ibm.com/researcher/view_group.php?id = 5077, accessed December 2022.

[3] In the style of *Machines Who Think* by Pamela McCorduck.

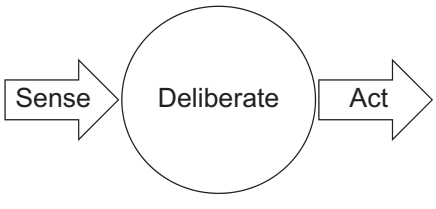

Figure 1.1 An autonomous agent operates in a three stage cycle. It receives input from its sensory mechanism, it deliberates over the inputs and its goals, and acts in the world.

neuron is connected to many other neurons and each connection has a weight that evolves with experience. This changing of weights is associated with the process of learning.

The neurons at the sensing end of the brain accept information coming in from various senses like sight, sound, smell, taste, and touch. The general model of processing is that once a neuron is activated, it sends a signal down its principal nerve called the axon, which distributes the signal to other connected neurons. The weights of the connections determine which connected neurons receive how much of the signal. Eventually the signals reach the neurons at the output end, sending signals down the motor neurons that activate muscles that produce sounds from the mouth and movement of the limbs.

Some simple creatures may be just reactive, recognizing food or prey and triggering appropriate actions, but as we move up the hierarchy, there may be more complex processing happening in the brain, involving memory (in Greek mythology, the dog Argos recognizes Odysseus at once when humans could not), planning (monkeys in Japan have been known to season their food with salt water), and reasoning (remember all those experiments with mice in mazes and Pavlov's dog). Whatever the cognitive capability of the creature, our view of their brains can be captured as shown in Figure 1.2.

Different life forms have differently sized brains relative to the sizes of their bodies. Earlier life forms had simple brains often referred to as the reptilian brain. In the 1960s, the American neuroscientist Paul MacLean (1990) formulated the *Triune Brain* model, which is based on

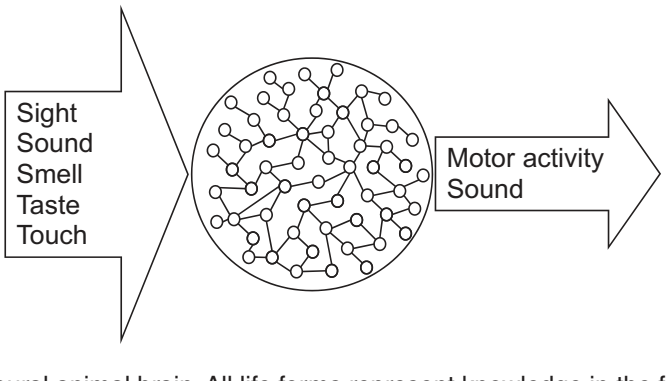

Figure 1.2 The neural animal brain. All life forms represent knowledge in the form of weights of connections between neurons in their brain and body. The numbers do not mean anything to us, and we say that the representation is sub-symbolic.

the division of the human brain into three distinct regions. MacLean's model suggests that the human brain is organized into a hierarchy, which itself is based on an evolutionary view of brain development. The three regions are as follows:

1. Reptilian or primal brain (basal ganglia) was the first to evolve and is the one in charge of our primal instincts.
2. Paleomammalian or emotional brain (limbic system) was the next to evolve and handles our emotions.
3. Neomammalian or rational brain (neocortex) which is responsible for what we call as thinking.

According to MacLean, the hierarchical organization of the human brain represents the gradual acquisition of the brain structures through evolution. The human brain, considered by many to be the most complex piece of matter in the universe, is made up of a cerebrum, the brain stem, and the cerebellum. The cerebrum is considered to be the seat of thought and, in humans, comprises two halves, each having an inner white core and an outer cerebral cortex made up of grey matter.

It is generally believed that the larger the brain, the greater the cognitive abilities of the owner.

1.3 Neural Networks

A neuron is a simple device that computes a simple function of the inputs it receives. Collections of interconnected neurons can do complex computations. Insights into animal brains have prompted many researchers to pursue the path of creating *artificial neural networks* (ANNs).

An ANN is a computational model that can be trained to perform certain tasks by repeatedly showing a stimulus and the expected response. It is best suited for the classification task. The earliest neural network was the *perceptron* (McCulloch and Pitts, 1943; Rosenblatt, 1958) which had one layer of neurons and could serve as a binary linear classifier. That is, whenever two classes in some space could be separated by a line or a plane in appropriate dimensions, the perceptron could be trained to learn the position and the orientation of the separator. Research in this area suffered a setback when Minsky and Papert (1969) showed its limitations – it could only classify linearly separable classes. For example, one cannot draw a line to separate the shaded circles, representing data from class A, from the unshaded ones, representing data from class B, as shown in Figure 1.3.

The work in neural networks was revived with the publication of the *Backpropagation* algorithm. In the mid-1980s, Rumelhart, Hinton, and Williams (1986) showed that a *multi-layer perceptron* could be trained to learn any non-linear classifier. They also popularized the Backpropagation algorithm that fed the error at the output layer back via the hidden layer, adjusting the weights of the connections (McClelland and Rumelhart, 1986a, 1986b).

Figure 1.4 shows the schematic diagram of a typical feedforward neural network. On the left is the input layer where the discretized input is fed in, activating nodes in the layer. Then, activation spreads from the left to the output layer on the right via the nodes in the hidden layer. In Figure 1.4, there are five output nodes, which could stand for five class labels. In the simplest case, when the network has learned to classify the input, which could be an image, one output

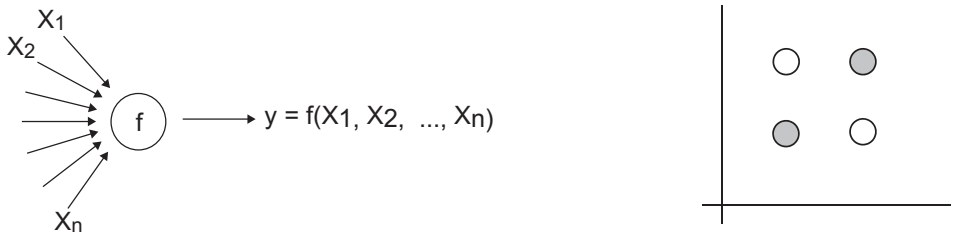

Figure 1.3 A neuron is a simple processing device that receives signals and generates an impulse as shown on the left. On the right is an example of a classification problem in which no line can be drawn to separate the shaded circle from the unshaded ones.

node is activated, indicating the class label. What the neural network has learnt is the *association* between the pattern of activation in the input layer and the class label at the output layer.

The fact that the input may be an image of a scene is only in the mind of the user, as is the name given to the class label. In the figure, the names are five animals, but the neural network has no idea that one is talking of animals, or a particular animal like a horse or a bear. It just *knows* which label to activate with a given image.

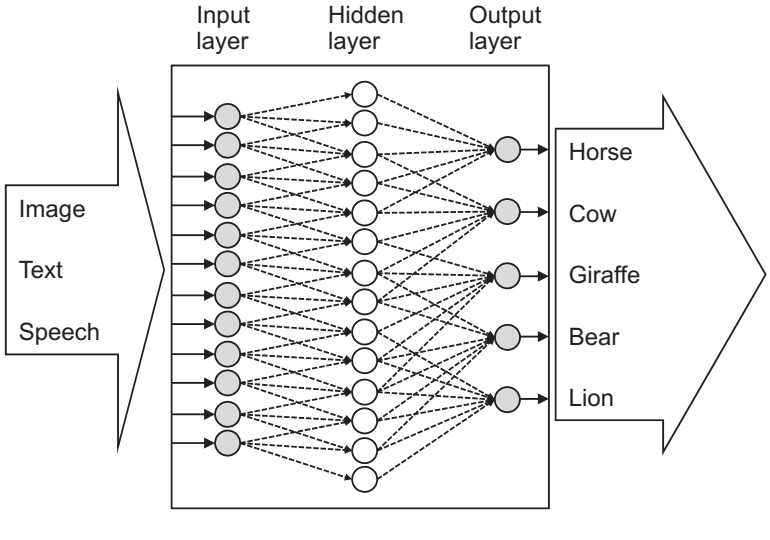

Figure 1.4 A feedforward artificial neural network learns a function from the input space to the output classes. Learning happens via the adjustment of edge weights. The labels of the output classes are meaningful only to the user.

This knowledge is not explicit or symbolic in the network. It is buried in the weights of the edges from nodes in one layer to the next one. These weights are instrumental in directing the activation from the input layer to the relevant output layer node. Nowhere in the network is there any indication that one is looking at a giraffe or a lion. Such representations of knowledge are often called *sub-symbolic* in contrast with the explicit symbolic representations we humans commonly use.

Neural networks learn what they learn by a process of *training*. The most common form of training is called *supervised learning*, in which a user presents input patterns to the network, and for each input shows what the output label should be. Every time an input pattern is presented, the network makes its own decision of what the activation value of the class label is. For example, if a bear is shown to the network, it might compute the output values as [0.2, 0.1, 0.0, 0.4, 0.3] when the expected output is [0, 0, 0, 1, 0], indicating that it is the fourth node (the bear). The error in the actual output defines a loss function that BACKPROP (as it is also known) aims to minimize. Most variations of the algorithm compute the gradient of the loss function with respect to the weights and do a small change in the edge weights in each cycle to reduce the loss. This can be viewed as *gradient descent*, an algorithm we look at later in the book.

The other forms of learning that are popular are *unsupervised* learning in which algorithms can learn to identify clusters in data, and *reinforcement* learning in which feedback from the world is used to adjust relevant weights. Reinforcement learning has achieved great success in game playing programs that learn how to play by playing hundreds of thousands of games against themselves, learning how to evaluate board positions from the outcomes of the games.

1.3.1 Deep neural networks

In principle the three layered network could learn any function, but in practice it was hard to do so, requiring a large number of neurons in the hidden layer. Geoffrey Hinton persevered with neural networks and showed in 2012 that deep networks with many hidden layers can achieve phenomenal success in computer vision – recognizing thousands of *types* of objects. Alex Krizhevsky in collaboration with Ilya Sutskever and his PhD advisor Geoffrey Hinton implemented the convolutional neural network (CNN) named *AlexNet* (Krizhevsky et al., 2017). This program did phenomenally well on the task of image recognition in the ImageNet Large Scale Visual Recognition Challenge in 2012. The network achieved a top-5 error of 15.3%, much better than the runner up. Their paper claimed that the *depth of the model* was essential for its high performance. Since then, neural networks with many layers have been doing very well in pattern recognition tasks. In 2015, a deep CNN with over 100 layers from Microsoft Research Asia outperformed AlexNet. Even though many layers make them computationally expensive, the use of graphics processing units (GPUs) during training has made them feasible. Figure 1.5 shows a schematic of a deep neural network.

The development of newer architectures and newer algorithms was instrumental in the spurt of interest in deep neural networks. Equally responsible perhaps was the explosion in the amount of data available on the internet, for example, the millions of images with captions uploaded by users, along with rapid advances in the computing power available. In 2018, three scientists, Geoffrey Hinton, Yann LeCun, and Yoshua Bengio, were jointly awarded the Turing Award for their work in this area. Deep networks got further impetus with the availability

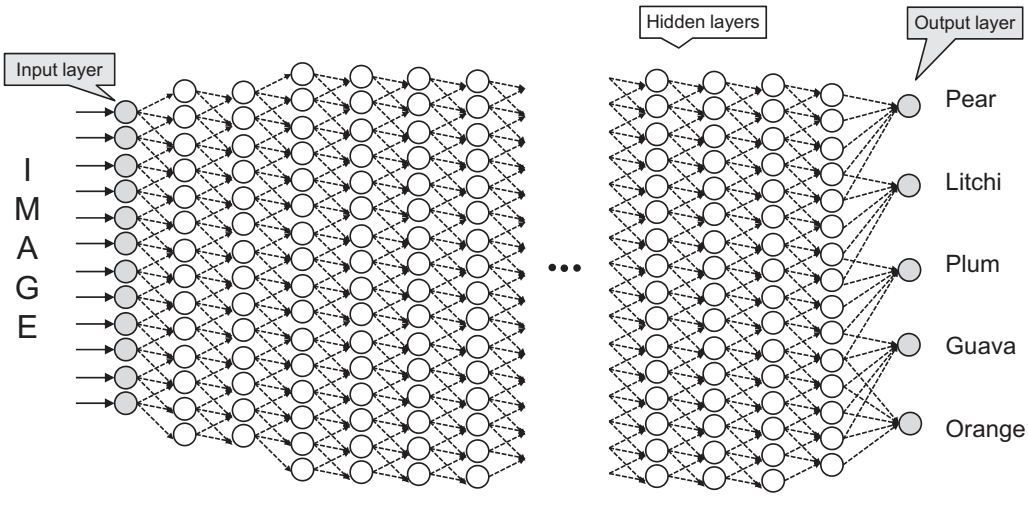

Figure 1.5 The schematic diagram of a deep neural network.

of open source software like *Tensorflow*[4] from Google that makes the task of implementing machine learning models easier for researchers.

More recently, *generative neural networks* have been successfully deployed for language generation and even creating paintings, for example, from OpenAI.[5] Generative models embody a form of unsupervised learning from large amounts of data, and are then trained to *generate data* like the one the algorithms were trained on. After having been fed with millions of images and text and their associated captions, they have now learnt to generate similar pictures or stories from similar text commands. Programs like *ChatGPT, Imagen,* and *DALL-E* have created quite a flurry amongst many users on the internet.

Deep neural networks are very good at the task of pattern recognition. Qualitatively, they are no different from the earlier networks, but in terms of performance they are far superior. The main task they are very good at is classification, a task that some researchers have commented is accomplished 'in the blink of an eye' by all life forms (Darwiche, 2018). The question one might ask is what after that?

Both in the case of generative models and deep neural network based classification, one must remember that the programs are throwing back at us whatever data has been fed to them. They do not *understand* what they are writing or drawing even though there is some correlation between the input query or command and the output generated.

For understanding and acting upon such perceived data, one needs to create models of the world to reason with. This is best done by explicit symbolic representations, which have the added benefit that they can contribute to explanations.

[4] https://developers.google.com/machine-learning/crash-course/first-steps-with-tensorflow/toolkit, accessed December 2022.

[5] https://openai.com/blog/generative-models/, accessed December 2022.

1.4 Symbolic AI

Imagine that you are the coach of a football team watching the game when your team is down two–nil. You have been watching the game for the best part of seventy minutes. The need of the hour is to make a couple of changes.[6] You pull out two players who are playing under their par and send in two substitutes.

Can a neural network take such a decision? Clearly not. The knowledge of the neural network is a kind of long term memory arrived at by training on many past examples. The neural network does not have the apparatus to represent the world around it in a dynamic scenario. What an agent also needs is short term memory that represents the current problem, and facilitates reasoning about the situation and planning of actions. We will talk about short term and long term memory in a little more detail in Chapter 7. But now we introduce the main idea at the core of this book – symbolic reasoning.

1.4.1 Symbols, language, and knowledge

Arguably, humankind broke away from the rest of the animal world with the development of language. The ability to give *names* to concepts combined with a shared understanding of what words mean not only has been a boon for communication (remember the boy who shouted wolf?) but has also provided a basis for representing complex concepts.

The core of *language*, whether spoken or written, is the *symbol*. A symbol is a perceptible something that stands for something else. The study of how signs and symbols are created and how they are interpreted is called *semiotics*. We are all familiar with road signs indicating the presence of schools, crossings, restaurants, U-turns, and so on. Most commercial activities are promoted using logos of companies which too stand for the company. Biosemiotics is the study of how complex behaviour emerges when simple systems interact with each other through signs. The pheromone trails left by ants for other ants to follow and the waggle dance of the honey bees to convey the location of food source to fellow bees are examples. The key feature is the use of words, behaviours, and shapes, collectively known as *semiosis*, as a means of transmitting meaningful information encoded in symbols and decoded by the receiver.

Human languages have evolved to describe what we see and perceive. The simplest kinds of names were probably just atomic but gradually we learnt to combine words to devise compound names, for example, *der liegestuhl* (the lounge chair) in German and *Himalaya* (the abode of snow) in Sanskrit. But at the simplest atomic level, a word, whether a noun, adjective, adverb, or verb, simply stands for something. Once many years ago a curious four-year-old had asked me: 'Why is a (ceiling) fan called a fan, and not something else?' The answer perhaps is that words acquire meaning via wide social agreement and also derive from related words. Words vary over regions, sometimes gradually and sometimes abruptly. The English word 'potato' corresponds to 'patata' in Sindhi, and 'batata' in Marathi. But some languages have a radically different name, 'alu' or 'aloo', for it. A look at the names of numbers across different languages also reveals a remarkable similarity that is unlikely to be sheer coincidence. Most languages have names starting with 's' for the equivalent of the number six, which is *sechs*

[6] Even as I write this, France has scored two goals in two minutes to draw level with Argentina in the FIFA World Cup final of 2022.