1

The Nature of Cognition

1.1 *Motivation for Studying Artificial Cognitive Systems*

When we set about building a machine or writing a software application, we usually have a clear idea of what we want it to do and the environment in which it will operate. To achieve reliable performance, we need to know about the operating conditions and the user's needs so that we can cater for them in the design. Normally, this isn't a problem. For example, it is straightforward to specify the software that controls a washing machine or tells you if the ball is out in a tennis match. But what do we do when the system we are designing has to work in conditions that aren't so well-defined, where we cannot guarantee that the information about the environment is reliable, possibly because the objects the system has to deal with might behave in an awkward or complicated way, or simply because unexpected things can happen?

Let's use an example to explain what we mean. Imagine we wanted to build a robot that could help someone do the laundry: load a washing machine with clothes from a laundry basket, match the clothes to the wash cycle, add the detergent and conditioner, start the wash, take the clothes out when the wash is finished, and hang them up to dry (see Figure 1.1). In a perfect world, the robot would also iron the clothes,[1] and put them back in the wardrobe. If someone had left a phone, a wallet, or something else in a pocket, the robot should either remove it before putting the garment in the wash or put the garment to

[1] The challenge of ironing clothes as a benchmark for robotics [1] was originally set by Maria Petrou [2]. It is a difficult task because clothes are flexible and unstructured, making them difficult to manipulate, and ironing requires careful use of a heavy tool and complex visual processing.

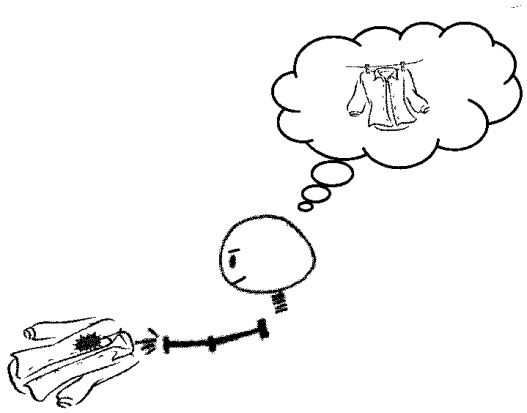

Figure 1.1: A cognitive robot would be able to see a dirty garment and figure out what needs to be done to wash and dry it.

one side to allow a human to deal with it later. This task is well beyond the capabilities of current robots[2] but it is something that humans do routinely. Why is this? It is because we have the ability to look at a situation, figure out what's needed to achieve some goal, anticipate the outcome, and take the appropriate actions, adapting them as necessary. We can determine which clothes are white (even if they are very dirty) and which are coloured, and wash them separately. Better still, we can also learn from experience and adapt our behaviour to get better at the job. If the whites are still dirty after being washed, we can apply some extra detergent and wash them again at a higher temperature. And best of all, we usually do this all on our own, autonomously, without any outside help (except maybe the first couple of times). Most people can work out how to operate a washing machine without reading the manual, we can all hang out damp clothes to dry without being told how to do it, and (almost) everyone can anticipate what will happen if you wash your smartphone.

We often refer to this human capacity for self-reliance, for being able to figure things out, for independent adaptive anticipatory action, as *cognition*. What we want is the ability to create machines and software systems with the same capacity, i.e., *artificial cognitive systems*. So, how do we do it? The first step would be to model cognition. And this first step is, unfortunately, where things get difficult because cognition means

[2] Some progress has been made recently in developing a robot that can fold clothes. For example, see the article "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding" by Jeremy Maitin-Shepard *et al.* [3] which describes how the PR2 robot built by Willow Garage [4] tackles the problem. However, the focus in this task is not so much the ill-defined nature of the job — how do you sort clothes into different batches for washing and, in the process, anticipate, adapt, and learn — as it is on the challenge of vision-directed manipulation of flexible materials.

different things to different people. The issue turns on two key concerns: (a) the purpose of cognition — the role it plays in humans and other species, and by extension, the role it should play in artificial systems — and (b) the mechanisms by which the cognitive system fulfils that purpose and achieves its cognitive ability. Regrettably, there's huge scope for disagreement here and one of the main goals of this book is to introduce you to the different perspectives on cognition, to explain the disagreements, and to tease out their differences. Without understanding these issues, it isn't possible to begin the challenging task of developing artificial cognitive systems. So, let's get started.

1.2 *Aspects of Modelling Cognitive Systems*

There are four aspects which we need to consider when modelling cognitive systems:[3] how much inspiration we take from natural systems, how faithful we try to be in copying them, how important we think the system's physical structure is, and how we separate the identification of cognitive capability from the way we eventually decide to implement it. Let's look at each of these in turn.

[3] For an alternative view that focusses on assessing the contributions made by particular models, especially computational and robotic models, see Anthony Morse's and Tom Ziemke's paper "On the role(s) of modelling in cognitive science" [5].

To replicate the cognitive capabilities we see in humans and some other species, we can either invent a completely new solution or draw inspiration from human psychology and neuroscience. Since the most powerful tools we have today are computers and sophisticated software, the first option will probably be some form of computational system. On the other hand, psychology and neuroscience reflect our understanding of biological life-forms and so we refer to the second option as a bio-inspired system. More often than not, we try to blend the two together. This balance of pure computation and bio-inspiration is the first aspect of modelling cognitive systems.

Unfortunately, there is an unavoidable complication with the bio-inspired approach: we first have to understand how the biological system works. In essence, this means we must come up with a model of the operation of the biological system and then use this model to inspire the design of the artificial system. Since biological systems are very complex, we need to choose the level

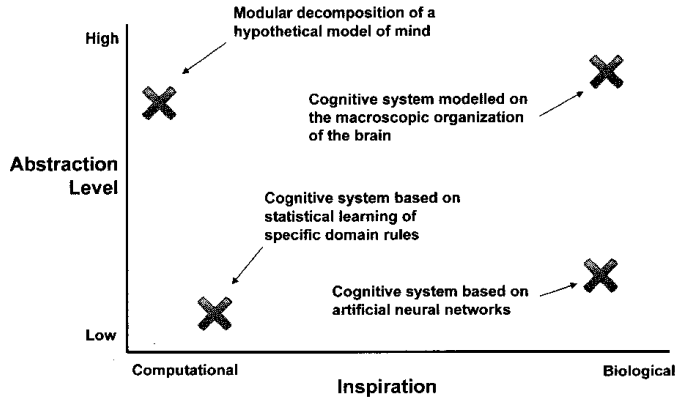

Figure 1.2: Attempts to build an artificial cognitive system can be positioned in a two-dimensional space, with one axis defining a spectrum running from purely computational techniques to techniques strongly inpired by biological models, and with another axis defining the level of abstraction of the biological model.

of abstraction at which we study them. For example, assuming for the moment that the centre of cognitive function is the brain (this might seem a very safe assumption to make but, as we'll see, there's a little more to it than this), then you might attempt to replicate cognitive capacity by emulating the brain at a very high level of abstraction, e.g. by studying the broad functions of different regions in the brain. Alternatively, you might opt for a low level of abstraction by trying to model the exact electrochemical way that the neurons in these regions actually operate. The choice of abstraction level plays an important role in any attempt to model a bio-inspired artificial cognitive system and must be made with care. That's the second aspect of modelling cognitive systems.

Taking both aspects together — bio-inspiration and level of abstraction — we can position the design of an artificial cognitive system in a two-dimensional space spanned by a computational / bio-inspired axis and an abstraction-level axis; see Figure 1.2. Most attempts today occupy a position not too far from the centre, and the trend is to move towards the biological side of the computational / bio-inspired spectrum and to cover several levels of abstraction.

In adopting a bio-inspired approach at any level of abstraction it would be a mistake to simply replicate brain mechanisms in complete isolation in an attempt to replicate cognition. Why? Because the brain and its associated cognitive capacity is the result

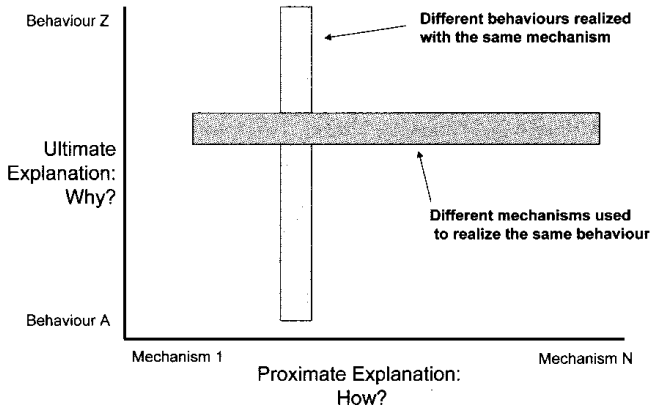

Figure 1.3: The ultimate-proximate distinction. Ultimate explanations deal with *why* a given behaviour exists in a system, while proximate explanations address the specific mechanisms by which these behaviours are realized. As shown here, different mechanisms could be used to achieve the same behaviour or different behaviours might be realized with the same mechanism. What's important is to understand that identifying the behaviours you want in a cognitive system and finding suitable mechanisms to realize them are two separate issues.

of evolution and the brain evolved for some purpose. Also, the brain and the body evolved together and so you can't divorce one from the other without running the risk of missing part of the overall picture. Furthermore, this brain-body evolution took place in particular environmental circumstances so that the cognitive capacity produced by the embodied brain supports the biological system in a specific ecological niche. Thus, a complete picture may really require you to adopt a perspective that views the brain and body as a complete system that operates in a specific environmental context. While the environment may be uncertain and unknown, it almost always has some in-built regularities which are exploited by brain-body system through its cognitive capacities in the context of the body's characteristics and peculiarities. In fact, the whole purpose of cognition in a biological system is to equip it to deal with this uncertainty and the unknown nature of the system's environment. This, then, is the third aspect of modelling cognitive systems: the extent to which the brain, body, and environment depend on one another.[4]

Finally, we must address the two concerns we raised in the opening section, i.e., the purpose of cognition and the mechanisms by which the cognitive system fulfils that purpose and achieves its cognitive ability. That is, in drawing on bio-inspiration, we need to factor in two complementary issues: what cognition is for and how it is achieved. Technically, this is known as the

[4] We return to the relationship between the brain, body, and environment in Chapter 5 on embodiment.

ultimate-proximate distinction in evolutionary psychology; see Figure 1.3. Ultimate explanations deal with questions concerned with *why* a given behaviour exists in a system or is selected through evolution, while proximate explanations address the specific mechanisms by which these behaviours are realized. To build a complete picture of cognition, we must address both explanations. We must also be careful not to get the two issues mixed up, as they very often are.[5] Thus, when we want to build machines which are able to work outside known operating conditions just like humans can — to replicate the cognitive characteristics of smart people — we must remember that this smartness may have arisen for reasons other than the ones in which it is being deployed in the current task-at-hand. Our brains and bodies certainly didn't evolve so that we could load and unload a washing machine with ease, but we're able to do it nonetheless. In attempting to use bio-inspired cognitive capabilites to perform utilitarian tasks, we may well be just piggy-backing on a deeper and quite possibly quite different functional capacity. The core problem then is to ensure that this *system* functional capacity matches the ones we need to get *our* job done. Understanding this, and keeping the complementary issues of the purpose and mechanisms of cognition distinct, allows us to keep to the forefront the important issue of how one can get an artificial cognitive system (and a biological one, too, for that matter) to do what we want it to do. If we are having trouble doing this, the problem may not be the operation of the specific (proximate) mechanisms of the cognitive model but the (ultimate) selection of the cognitive behaviours and their fitness for the given purpose in the context of the brain-body-mind relationship.

[5] The importance of the ultimate-proximate distinction is highlighted by Scott-Phillips *et al.* in a recent article [6]. This article also points out that ultimate and proximate explanations of phenomena are often confused with one another so we end up discussing proximate concerns when we really should be discussing ultimate ones. This is very often the case with artificial cognitive systems where there is a tendency to focus on the proximate issues of *how* cognitive mechanisms work, often neglecting the equally important issue of *what* purpose cognition is serving in the first place. These are two complementary views and both are needed. See [7] and [8] for more details on the ultimate-proximate distinction.

To sum up, in preparing ourselves to study artificial cognitive systems, we must keep in mind four important aspects when modelling cognitive systems:

1. The computational / bio-inspired spectrum;
2. The level of abstraction in the biological model;
3. The mutual dependence of brain, body, and environment;
4. The ultimate-proximate distinction (why *vs.* how).

Understanding the importance of these four aspects will help us make sense of the different traditions in cognitive science, artificial intelligence, and cybernetics (among other disciplines) and the relative emphasis they place on the mechanisms and the purpose of cognition. More importantly, it will ensure we are addressing the right questions in the right context in our efforts to design and build artificial cognitive systems.

1.3 *So, What Is Cognition Anyway?*

It should be clear from what we have said so far that in asking "what is cognition?" we are posing a badly-framed question: what cognition *is* depends on what cognition is *for* and *how* cognition is realized in physical systems — the ultimate and proximate aspects of cognition, respectively. In other words, the answer to the question depends on the context — on the relationship between brain, body, and environment — and is heavily coloured by which cognitive science tradition informs that answer. We devote all of Chapter 2 to these concerns. However, before diving into a deep discussion of these issues, we'll spend a little more time here setting the scene. In particular, we'll provide a generic characterization of cognition as a preliminary answer to the question "what is cognition?", mainly to identify the principal issues at stake in designing artificial cognitive systems and always mindful of the need to explain how a given system addresses the four aspects of modelling identified above. Now, let's cut to the chase and answer the question.

Cognition implies an ability to make inferences about events in the world around you. These events include those that involve the cognitive agent itself, its actions, and the consequences of those actions. To make these inferences, it helps to remember what happened in the past since knowing about past events helps to anticipate future ones.[6] Cognition, then, involves predicting the future based on memories of the past, perceptions of the present, and in particular anticipation of the behaviour[7] of the world around you and, especially, the effects of your actions in it. Notice we say actions, not movement of motions. Actions usually involve movement or motion but an action also involves

[6] We discuss the forward-looking role of memory in anticipating events in Chapter 7.

[7] Inanimate objects don't behave but animate ones do, as do inanimate objects being controlled by animate ones (e.g. cars in traffic). So agency, direct or indirect, is implied by behaviour.

something else. This is the *goal* of the action: the desired outcome, typically some change in the world. Since predictions are rarely perfect, a cognitive system must also learn by observing what does actually happen, assimilate it into its understanding, and then adapt the way it subsequently does things. This forms a continuous cycle of self-improvement in the system's ability to anticipate future events. The cycle of anticipation, assimilation, and adaptation supports — and is supported by — an on-going process of action and perception; see Figure 1.4.

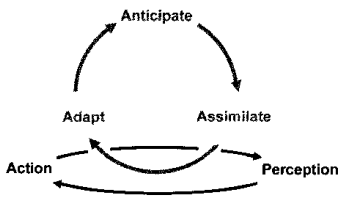

Figure 1.4: Cognition as a cycle of anticipation, assimilation, and adaptation: embedded in, contributing to, and benefitting from a continuous process of action and perception.

We are now ready for our preliminary definition.

> Cognition is the process by which an autonomous system perceives its environment, learns from experience, anticipates the outcome of events, acts to pursue goals, and adapts to changing circumstances.[8]

We will take this as our preliminary definition of cognition and, depending on the approach we are discussing, we will adjust it accordingly in later chapters.

While definitions are convenient, the problem with them is that they have to be continuously amended as we learn more about the thing they define.[9] So, with that in mind, we won't become too attached to the definition and we'll use it as a memory aid to remind us that cognition involved at least six attributes of autonomy, perception, learning, anticipation, action, and adaptation.

For many people, cognition is really an umbrella term that covers a collection of skills and capabilities possessed by an agent.[10] These include being able to do the following.

- Take on goals, formulate predictive strategies to achieve them, and put those strategies into effect;
- Operate with varying degrees of autonomy;
- Interact — cooperate, collaborate, communicate — with other agents;
- Read the intentions of other agents and anticipate their actions;
- Sense and interpret expected and unexpected events;

[8] These six attributes of cognition — autonomy, perception, learning, anticipation, action, adaptation — are taken from the author's definition of cognitive systems in the Springer *Encyclopedia of Computer Vision* [9]

[9] The Nobel laureate, Peter Medawar, has this to say about definitions: "My experience as a scientist has taught me that the comfort brought by a satisfying and well-worded definition is only short-lived, because it is certain to need modification and qualification as our experience and understanding increase; it is explanations and descriptions that are needed" [10]. Hopefully, you will find understandable explanations in the pages that follow.

[10] We frequently use the term *agent* in this book. It means any system that displays a cognitive capacity, whether it's a human, or (potentially, at least) a cognitive robot, or some other artificial cognitive entity. We will use *agent* interchangably with *artifical cognitive system.*

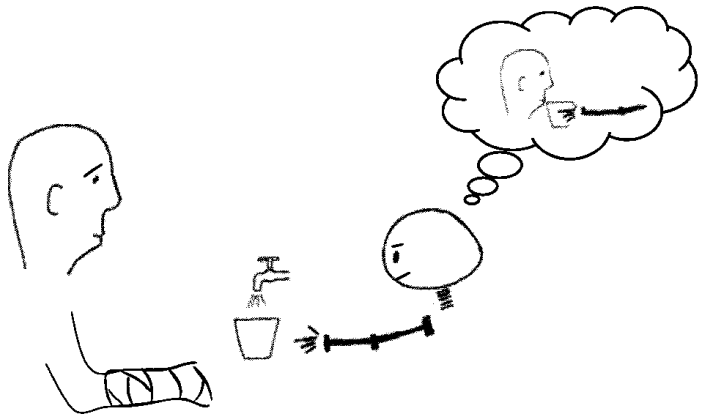

Figure 1.5: Another aspect of cognition: effective interaction. Here the robot anticipates someone's needs (see Chapter 9, Section 9.4 *Instrumental Helping*).

- Anticipate the need for actions and predict the outcome of its own actions and those of others;
- Select a course of action, carry it out, and then assess the outcome;
- Adapt to changing circumstances, in real-time, by adjusting current and anticipated actions;
- Learn from experience: adjust the way actions are selected and performed in the future;
- Notice when performance is degrading, identify the reason for the degradation, and take corrective action.

These capabilities focus on what the agent should do: its functional attributes. Equally important are the effectiveness and the quality of its operation: its non-functional characteristics (or, perhaps more accurately, its meta-functional characteristics): its dependability, reliability, usability, versatility, robustness, fault-tolerance, and safety, among others.[11]

These meta-functional characteristics are linked to the functional attributes through system capabilities that focus not on carrying out tasks but on maintaining the integrity of the agent.[12] Why are these capabilities relevant to artificial agents? They are relevant — and critically so — because artificial agents such as a robot that is deployed outside the carefully-configured environments typical of many factory floors have to deal with a

[11] The "non-" part of "non-functional" is misleading as it suggests a lesser value compared to functional characteristics whereas, in reality, these characteristics are equally important but complementary to functionality when designing a system. For that reason, we sometimes refer to them as meta-functional attributes; see [11] for a more extensive list and discussion of meta-functionional attributes.

[12] We will come back to the issue of maintaining integrity several times in this book, briefly in the next section, and more at length in the next chapter. For the moment, we will just remark that the processes by which integrity is maintained are known as *autonomic* processes.

world that is only partially known. It has to work with incomplete information, uncertainty, and change. The agent can only cope with this by exhibiting some degree of cognition. When you factor interaction with people into the requirements, cognition becomes even more important. Why? Because people are cognitive and they behave in a cognitive manner. Consequently, any agent that interacts with a human needs to be cognitive to some degree for that interaction to be useful or helpful. People have their own needs and goals and we would like our artificial agent to be able to anticipate these (see Figure 1.5). That's the job of cognition.

So, in summary, cognition is not to be seen as some module in the brain of a person or the software of a robot — a planning module or a reasoning module, for example — but as a system-wide process that integrates all of the capabilities of the agent to endow it with the six attributes we mentioned in our memory-aid definition: autonomy, perception, learning, anticipation, action, and adaptation.

1.3.1 *Why Autonomy?*

Notice that we included autonomy in our definition. We need to be careful about this. As we will see in Chapter 4, the concept of autonomy is a difficult one. It means different things to different people, ranging from the fairly innocent, such as being able to operate without too much help or assistance from others, to the more controversial, which sees cognition as one of the central processes by which advanced biological systems preserve their autonomy. From this perspective, cognitive development has two primary functions: (1) to increase the system's repertoire of effective actions, and (2) to extend the time-horizon of its ability to anticipate the need for and outcome of future actions.[13]

Without wishing to preempt the discussion in Chapter 4, because there is a tight relationship between cognition and autonomy — or not, depending on who you ask — we will pause here just a while to consider autonomy a little more.

From a biological perspective, autonomy is an organizational characteristic of living creatures that enables them to use their

[13] The increase of action capabilities and the extension anticipation capabilities as the primary focus of cognition is the central message conveyed in *A Roadmap for Cognitive Development in Humanoid Robots* [12], a multi-disciplinary book co-written by the author, Claes von Hofsten, and Luciano Fadiga.

own capacities to manage their interactions with the world in order to remain viable, i.e., to stay alive. To a very large extent, autonomy is concerned with the system maintaining itself: self-maintenance, for short.[14] This means that the system is entirely self-governing and self-regulating. It is not controlled by any outside agency and this allows it to stand apart from the rest of the environment and assert an identity of its own. That's not to say that the system isn't influenced by the world around it, but rather that these influences are brought about through interactions that must not threaten the autonomous operation of the system.[15]

If a system is autonomous, its most important goal is to preserve its autonomy. Indeed, it must act to preserve it since the world it inhabits that may not be very friendly. This is where cognition comes in. From this (biological) perspective, cognition *is* the process whereby an autonomous self-governing system acts effectively in the world in which it is embedded in order to maintain its autonomy.[16] To act effectively, the cognitive system must sense what is going on around it. However, in biological agents, the systems responsible for sensing and interpretation of sensory data, as well as those responsible for getting the motor systems ready to act, are actually quite slow and there is often a delay between when something happens and when an autonomous biological agent comprehends what has happened. This delay is called *latency* and it is often too great to allow the agent to act effectively: by time you have realized that a predator is about to attack, it may be too late to escape. This is one of the primary reasons a cognitive system must anticipate future events: so that it can prepare the actions it may need to take in advance of actually sensing that these actions are needed.

In addition to sensory latencies, there are also limitations imposed by the environment and the cognitive system's body. To perform an action, and specifically to accomplish the goal associated with an action, you need to have the relevant part of your body in a certain place at a certain time. It takes time to move, so, again, you need to be able to predict what might happen and prepare to act. For example, if you have to catch an object, you need to start moving your hand before the object arrives and

[14] The concepts of self-maintenance and recursive self-maintenance in self-organizing autonomous system was introduced by Mark Bickhard [13]. We will discuss them in more detail in Chapter 2. The key idea is that self-maintenant systems make active contributions to their own persistence but do not contribute to the maintenance of the conditions for persistence. On the other hand, recursive self-maintenant systems do contribute actively to the conditions for persistence.

[15] When an influence on a system isn't directly controlling it but nonetheless has some impact on the behaviour of the system, we refer to it as a *perturbation*.

[16] The idea of cognition being concerned with *effective action*, i.e. action that helps preserve the system's autonomy, is due primarily to Francisco Varela and Humberto Maturana [14]. These two scientists have had a major impact on the world of cognitive science through their work on biological autonomy and the organizational principles which underpin autonomous systems. Together, they provided the foundations for a new approach to cognitive science called *Enaction*. We will discuss enaction and enactive systems is more detail in Chapter 2.

sometimes even before it has been thrown. Also, the world in which the system is embedded is constantly changing and is outside the control of the system. Consequently, the sensory data which is available to the cognitive system may not only be late in arriving but critical information may also be missing. Filling in these gaps is another of the primary functions of a cognitive system. Paradoxically, it is also often the case that there is too much information for the system to deal with and it has to ignore some of it.[17]

Now, while these capabilities derive directly from the biological autonomy-preserving view of cognition, it should be fairly clear that they would also be of great use to artificial cognitive systems, whether they are autonomous or not. However, before moving on to the next section which elaborates a little more on the relationship between biological and artificial cognitive systems, it is worth noting that some people consider that cognition should involve even more than what we have discussed so far. For example, an artificial cognitive system might also be able to explain what it is doing and why it is doing it.[18] This would enable the system to identify potential problems which could appear when carrying out a task and to know when it needed new information in order to complete it. Taking this to the next level, a cognitive system would be able to view a problem or situation in several different ways and to look at alternative ways of tackling it. In a sense, this is similar to the attribute we discussed above about cognition involving an ability to anticipate the need for actions and their outcomes. The difference in this case is that the cognitive system is considering not just one but *many* possible sets of needs and outcomes. There is also a case to be made that cognition should involve a sense of self-reflection:[19] an ability on the part of the system to think about itself and its own thoughts. We see here cognition straying into the domain of consciousness. We won't say anything more in this book on that subject apart from remarking that computational modelling of consciousness is an active area of research in which the study of cognition plays an important part.

[17] The problem of ignoring information is related to two problems in cogitive science: the *Frame Problem* and *Attention*. We will take up these issues again later in the book.

[18] The ability not simply to act but to explain the reasons for an action was proposed by Ron Brachman in an article entitled "Systems that know what they're doing" [15].

[19] Self-reflection, often referred to as meta-cognition, is emphasized by some people, e.g. Aaron Sloman [16] and Ron Sun [17], as an important aspect of advanced cognition.

1.4 *Levels of Abstraction in Modelling Cognitive Systems*

All systems can be viewed at different levels of abstraction, successively removing specific details at higher levels and keeping just the general essence of what is important for a useful model of the system. For example, if we wanted to model a physical structure, such as a suspension bridge, we could do so by specifying each component of the bridge — the concrete foundations, the suspension cables, the cable anchors, the road surface, and the traffic that uses it — and the way they all fit together and influence one another. This approach models the problem at a very low level of abstraction, dealing directly with the materials from which the bridge will be built, and we would really only know after we built it whether or not the bridge will stay up. Alternatively, we could describe the forces at work in each member of the structure and analyze them to find out if they are strong enough to bear the required loads with an acceptable level of movement, typically as a function of different patterns of traffic flow, wind conditions, and tidal forces. This approach models the problem at a high level of abstraction and allows the architect to established whether or not his or her design is viable before it is constructed. For this type of physical system, the idea is usually to use an abstract model to validate the design and then realize it as a physical system. However, deciding on the best level of abstraction is not always straightforward. Other types of system — biological ones for example — don't yield easily to this top-down approach. When it comes to modelling cognitive systems, it will come as no surprise that there is some disagreement in the scientific community about what level of abstraction one should use and how they should relate to one another. We consider here two contrasting approaches to illustrate their differences and their relative merits in the context of modelling and designing artificial cognitive systems.

As part of his influential work on modelling the human visual system, David Marr[20] advocated a three-level hierarchy of abstraction;[21] see Figure 1.6. At the top level, there is the computational theory. Below this, there is the level of representation and algorithm. At the bottom, there is the hardware implementation.

[20] David Marr was a pioneer in the field of computer vision. He started out as a neuroscientist but shifted to computational modelling to try to establish a deeper understanding of the human visual system. His seminal book *Vision* [18] was published posthumously in 1982.

[21] Marr's three-level hierarchy is sometimes known as the *Levels of Understanding* framework.

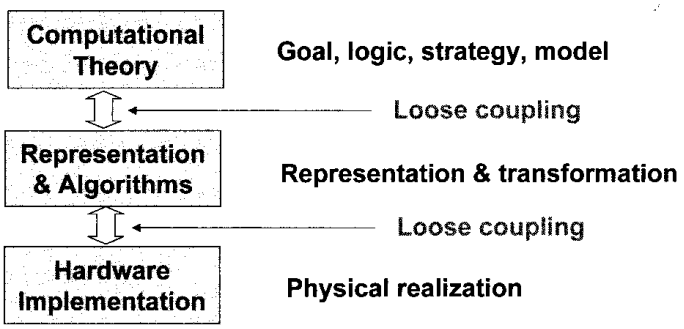

Figure 1.6: The three levels at which a system should be understood and modelled: the computational theory that formalizes the problem, the representational and algorithmic level that addresses the implementation of the theory, and the hardware level that physically realizes the system (after David Marr [18]). The computational theory is primary and the system should be understood and modelled first at this level of abstraction, although the representational and algorithmic level is often more intuitively accessible.

At the level of the computational theory, you need to answer questions such as "what is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it is carried out?" At the level of representation and algorithm, the questions are different: "how can this computational theory be applied? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?" Finally, the question at the level of hardware implementation is "how can the representation and algorithm be physically realized?" In other words, how can we build the physical system? Marr emphasized that these three levels are only loosely coupled: you can — and, according to Marr, you should — think about one level without necessarily paying any attention to those below it. Thus, you begin modelling at the computational level, ideally described in some mathematical formalism, moving on to representations and algorithms once the model is complete, and finally you can decide how to implement these representations and algorithms to realize the working system. Marr's point is that, although the algorithm and representation levels are more accessible, it is the computational or theoretical level that is critically important from an information processing perspective. In essence, he states that the problem can and should first be modelled at the abstract level of the computational theory without strong reference to the lower and less abstract levels.[22] Since many people believe that cognitive systems — both biological and artificial — are effectively information processors, Marr's hierarchy of abstraction is very useful.

Marr illustrated his argument succinctly by comparing the

[22] Tomaso Poggio recently proposed a revision of Marr's three-level hierarchy in which he advocates greater emphasis on the connections between the levels and an extension of the range of levels, adding *Learning and Development* on top of the computational theory level (specifically hierarchical learning), and *Evolution* on top of that [19]. Tomaso Poggio co-authored the original paper [20] on which David Marr based his more famous treatment in his 1982 book *Vision* [18].

problem of understanding vision (Marr's own goal) to the problem of understanding the mechanics of flight.

> "Trying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense"

Objects with different cross-sectional profiles give rise to different pressure patterns on the object when they move through a fluid such as air (or when a fluid flows around an object). If you choose the right cross-section then there is more pressure on the bottom than on the top, resulting in a lifting force that counters the force of gravity and allows the object to fly. It isn't until you know this that you can begin to understand the problem in a way that will yield a solution for your specific needs.

Of course, you eventually have to decide how to realize a computational model but this comes later. The point he was making is that you should decouple the different levels of abstraction and begin your analysis at the highest level, avoiding consideration of implementation issues until the computational or theoretical model is complete. When it is, it can then subsequently drive the decisions that need to be taken at the lower level when realizing the physical system.

Marr's dissociation of the different levels of abstraction is significant because it provides an elegant way to build a complex system by addressing it in sequential stages of decreasing abstraction. It is a very general approach and can be applied successfully to modelling, designing, and building many different systems that depend on the ability to process information. It also echoes the assumptions made by proponents of a particular paradigm of cognition — *cognitivism* — which we will meet in the next chapter.[23]

Not everyone agrees with Marr's approach, mainly because they think that the physical implementation has a direct role to play in understanding the computational theory. This is particularly so in the emergent paradigm of embodied cognition which we will meet in the next chapter, the embodiment reflecting the physical implementation. Scott Kelso,[24] makes a case for a com-

[23] The cognitivist approach to cognition proposes an abstract model of cognition which doesn't require you to consider the final realization. In other words, cognitivist models can be applied to any platform that supports the required computations and this platform could be a computer or a brain. See Chapter 2, Section 2.1, for more details.

[24] Over the last 25 years, Scott Kelso, the founder of the Center for Complex Systems and Brain Sciences at Florida Atlantic University, has developed a theory of *Coordination Dynamics*. This theory, grounded in the concepts of self-organization and the tools of coupled nonlinear dynamics, incorporates essential aspects of cognitive function, including anticipation, intention, attention, multimodal integration, and learning. His book, *Dynamic Patterns – The Self-Organization of Brain and Behaviour* [21], has influenced research in cognitive science world-wide.

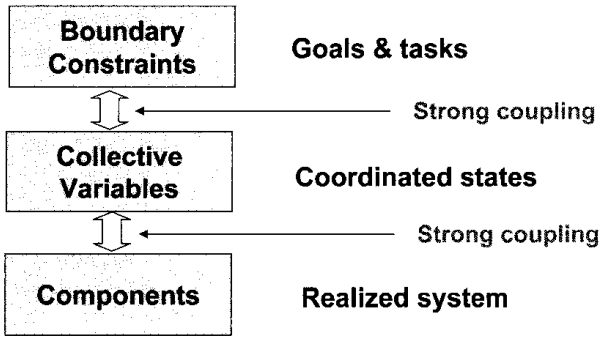

Figure 1.7: Another three levels at which a system should be modelled: a boundary constraint level that determines the task or goal, a collective variable level that characterizes coordinated states, and a component level which forms the realized system (after Scott Kelso [21]). All three levels are equally important and should be considered together.

pletely different way of modelling systems, especially non-linear dynamical types of systems that he believes may provide the true basis for cognition and brain dynamics. He argues that these types of system should be modelled at three distinct levels of abstraction, but at the same time. These three levels are a boundary constraint level, a collective variables level, and a components level. The boundary constraint level determines the goals of the system. The collective variable[25] level characterizes the behaviour of the system. The component level forms the realized physical system. Kelso's point is that the specification of these three levels of model abstraction are tightly coupled and mutually dependent. For example, the environmental context of the system often determines what behaviours are feasible and useful. At the same time, the properties of the physical system may simplify the necessary behaviour. Paraphrasing Rolf Pfeifer,[26] "morphology matters": the properties of the physical shape or the forced needed for required movements may actually simplify the computational problem. In other words, the realization of the system and its particular shape or morphology cannot be ignored and should not be abstracted away when modelling the system. This idea that you cannot model the system in isolation from either the system's environmental context or the system's ultimate physical realization is linked directly to the relationship between brain, body, and environment. We will meet it again later in the book when we discuss enaction in Chapter 2 and when we consider the issue of embodiment in Chapter 5.

The mutual dependence of system realization and system

[25] Collective variables, also referred to as order parameters, are so called because they are responsible for the system's overall collective behaviour. In dynamical systems theory, collective variables are a small subset of the system's many degrees of freedom but they govern the transitions between the states that the system can exhibit and hence its global behaviour.

[26] Rolf Pfeifer, University of Zurich, has long been a champion of the tight relationship between a system's embodiment and its cognitive behaviour, a relationship set out in his book *How the body shapes the way we think: A new view of intelligence* [22], co-authored by Josh Bongard.

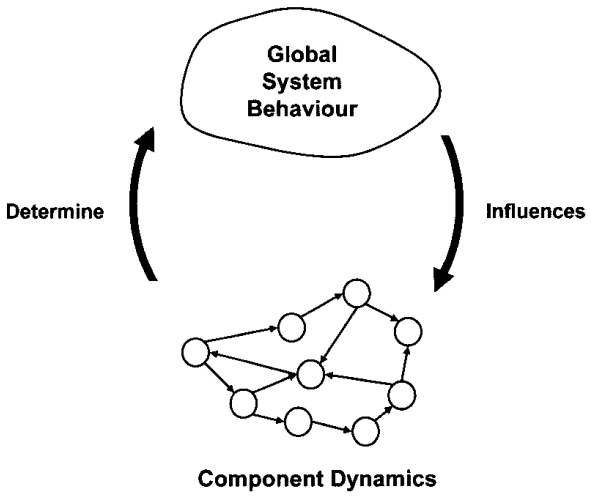

Figure 1.8: Circular causality — sometimes referred to as continuous reciprocal causation or recursive self-maintenance — refers to the situation where global system behaviour somehow influences the local behaviour of the system components and yet it is the local interaction between the components that determines the global behaviour. This phenomenon appears to be one of the pivotal mechanisms in autonomous cognitive systems.

modelling presents us with a difficulty, however. If we look carefully, we see a circularity, with everything depending on something else. It's not easy to see how you break into the modelling circle. This is one of the attractions of Marr's approach: there is a clear place to get started. This circularity crops up repeatedly in cognition and it does so in many forms. All we will say for the moment is that circular causality[27] — where global system behaviour somehow influences the local behaviour of the system components and yet it is the local interaction between the components that determines the global behaviour; see Figure 1.8 — appears to be one of the key mechanisms of cognition. We will return again to this point later in the book. For the moment, we'll simply remark that the two constrasting approaches to system modelling mirror two opposing paradigms of cognitive science. It is to these that we now turn in Chapter 2 to study the foundations that underpin our understanding of natural and artificial cognitive systems.

[27] Scott Kelso uses the term "circular causality" to describe the situation in dynamical systems where the cooperation of the individual parts of the system determine the global system behaviour which, in turn, governs the behaviour of these individual parts [21]. This is related to Andy Clark's concept of continuous reciprocal causation (CRC) [23] which "occurs when some system S is both continuously affecting and simultaneously being affected by, activity in some other system O" [24]. These ideas are also echoed in Mark Bickhard's concept of recursive self-maintenance [13]. We will say more about these matters in Chapter 4.

2

Paradigms of Cognitive Science

In Chapter 1, we were confronted with the tricky and unexpected problem of how to define cognition. We made some progress by identifying the main characteristics of a cognitive system — perception, learning, anticipation, action, adaptation, autonomy — and we introduced four aspects that must be borne in mind when modelling a cognitive system: (a) biological inspiration *vs.* computational theory, (b) the level of abstraction of the model, (c) the mutual dependence of brain, body, and environment, and (d) the ultimate-proximate distinction between what cognition is for and how it is achieved. However, we also remarked on the fact that there is more than one tradition of cognitive science so that any definition of cognition will be heavily coloured by the background against which the definition is set. In this chapter, we will take a detailed look at these various traditions. Our goal is to tease out their differences and get a good grasp of what each one stands for. Initially, it will seem that these traditions are polar opposites and, as we will see, they do differ in many ways. However, as we get to the end of the chapter, we will also recognize a certain resonance between them. This shouldn't surprise us: after all, each tradition occupies its own particular region of the space spanned by the ultimate and proximate dimensions which we discussed in Chapter 1 and it is almost inevitable that there will be some overlap, especially if that tradition is concerned with a general understanding of cognition.

Before we begin, it's important to appreciate that cognitive sci-

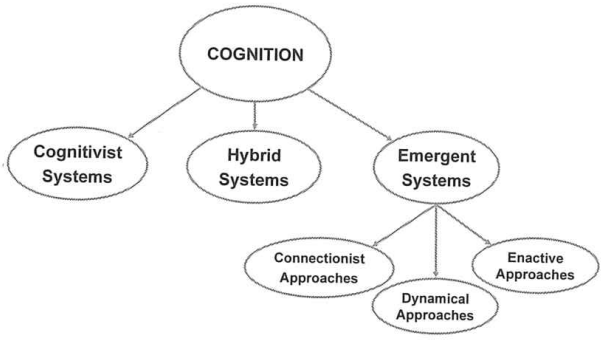

Figure 2.1: The cognitivist, emergent, and hybrid paradigms of cognition.

ence is a general umbrella term that embraces several disciplines, including neuroscience, cognitive psychology, linguistics, epistemology, and artificial intelligence, among others. Its primary goal is essentially to understand and explain the underlying processes of cognition: typically human cognition and ideally in a way that will yield a model of cognition that can then be replicated in artificial agents.

To a large extent, cognitive science has its origins in cybernetics which in the early 1940s to 1950s made the first attempts to formalize what had up to that point been purely psychological and philosophical treatments of cognition. Cybernetics was defined by Norbert Wiener as "the science of control and communication in the animal and the machine."[1] The intention of the early cyberneticians was to understand the mechanisms of cognition and to create a science of mind, based primarily on logic. Two examples of the application of cybernetics to cognition include the seminal paper by Warren S. McCulloch and Walter Pitts "A logical calculus immanent in nervous activity"[2] and W. Ross Ashby's book *Design for a Brain*.[3]

The first attempts in cybernetics to uncover the mechanisms of cognitive behaviour were subsequently taken up and developed into an approach referred to as *cognitivism*. This approach built on the logical foundations laid by the early cyberneticians and exploited the newly-invented computer as a literal metaphor for cognitive function and operation, using symbolic information processing as its core model of cognition. The cognitivist tradition continued to be the dominant approach over the next

[1] The word *cybernetics* has its roots in the Greek word κυβερνήτης or kybernētēs, meaning steersman. It was defined in Norbert Wiener's book *Cybernetics* [25], first published in 1948, as "the science of control and communication" (this was the sub-title of the book). W. Ross Ashby remarks in his book *An Introduction to Cybernetics* [26], published in 1956, that cybernetics is essentially "the art of steersmanship" and as such its themes are co-ordination, regulation, and control.

[2] As well as being a seminal work in cybernetics, the 1943 paper by Warren S. McCulloch and Walter Pitts, "A logical calculus immanent in nervous activity" [27], is also regarded as the foundation for artificial neural networks and connectionism [28].

[3] Complementing his influential book *Design for a Brain* [29, 30, 31], W. Ross Ashby's *Introduction to Cybernetics'* [26] is also a classic text.

30 or so years and, indeed, it was so pervasive and became so deeply embedded in our mind-set that it still holds sway today to a considerable extent.

Paradoxically, the early cybernetics period also paved the way for a completely different approach to cognitive science — the emergent systems approach — which recognized the importance of self-organization in cognitive processes. Initially, emergent systems developed almost under the radar — it was difficult to challenge the appeal of exciting new computer technology and the computational model of cognition — but it progressed nonetheless in parallel with the cognitivist tradition over the next fifty years and more, growing to embrace connectionism, dynamical systems theory, and enaction, all of which we will discuss in more detail later in the chapter.

In recent years, a third class — hybrid systems — has become popular, and understandably so since, as the name suggests, it attempts to combine the best from each of the cognitivist and emergent paradigms; see Figure 2.1.

In the sections that follow, we will take a closer look at all three traditions of cognitive science — cognitivist, emergent, and hybrid — to draw out the key assumptions on which they build their respective theories and to compare and contrast them on the basis of several fundamental characteristics that reflect different points in the ultimate-proximate space. Before we proceed to do this, it is worth noting that, although we have referred so far to the different *traditions* of cognitive science, the title of the chapter refers to the different *paradigms* of cognitive science. Is there any great significance to this? Well, in fact, there is. As we will see in what follows, and not withstanding the resonance that we mentioned above, the two traditions do make some fundamentally different assertions about the nature of cognition (i.e. its ultimate purpose) and its processes (i.e. its proximate mechanisms). In fact, these differences are so strong as to render the two approaches intrinsically incompatible and, hence, position them as two completely different paradigms. It isn't hard to see that this incompatibility is going to cause problems for hybrid approaches, but we'll get to that in due course. For the present, let's proceed with our discussion of the cognitivist and emergent

traditions of cognitive science, seeking out the issues on which they agree, but recognizing too those on which they do not, both in principle and in practice.[4]

2.1 *The Cognitivist Paradigm of Cognitive Science*

2.1.1 *An Overview of Cognitivism*

As we have seen, the initial attempt in cybernetics to create a science of cognition was followed by the development of an approach referred to as cognitivism. The birth of the cognitivist paradigm, and its sister discipline of Artificial Intelligence (AI), dates from a conference held at Dartmouth College, New Hampshire, in July and August 1956. It was attended by people such as John McCarthy, Marvin Minsky, Allen Newell, Herbert Simon, and Claude Shannon, all of whom had a very significant influence on the development of AI over the next half-century. The essential position of cognitivism is that cognition is achieved by computations performed on internal symbolic knowledge representations in a process whereby information about the world is taken in through the senses, filtered by perceptual processes to generate descriptions that abstract away irrelevant data, represented in symbolic form, and reasoned about to plan and execute mental and physical actions. The approach has also been labelled by many as the information processing or symbol manipulation approach to cognition.

For cognitivist systems, cognition is representational in a particular sense: it entails — requires — the manipulation of explicit symbols: localized abstract encapsulations of information that denote the state of the world external to the cognitive agent. The term 'denote' has particular significance here because it asserts an identity between the symbol used by the cognitive agent and the thing that it denotes. It is as if there is a one-to-one correspondence between the symbol in the agent's cognitive system and the state of the world to which it refers. For example, the clothes in a laundry basket can be represented by a set of symbols, often organized in a hierarchical manner, describing the identity of each item and its various characteristics: whether it is

[4] For an accessible summary of the different paradigms of cognition, refer to a paper entitled "Whence Perceptual Meaning? A Cartography of Current Ideas" [32] It was written by Francisco Varela, one of the founders of a branch of cognitive science called Enaction, and it is particularly instructive because, as well as contrasting the various views of cognition, it also traces them to their origins. This historical context helps highlight the different assumptions that underpin each approach and it shows how they have evolved over the past sixty years or so. Andy Clark's book *Mindware – An Introduction to the Philosophy of Cognitive Science* [33] also provides a useful introduction to the philosophical and scientific differences between the different paradigms of cognition.

heavily soiled or not, its colour, its recommended wash cycle and water temperature. Similarly, the washing machine can be represented by another symbol or set of symbols. These symbols can represent objects and events but they can also represent actions: things that can happen in the world. For example, symbols that represent sorting the clothes into bundles, one bundle for each different wash cycle, putting them into the washing machine, selecting the required wash cycle, and starting the wash. It is a very clear, neat, and convenient way to describe the state of the world in which the cognitive agent finds itself.

Having this information about the world represented by such an explicit abstract symbolic knowledge is very useful for two reasons. First, it means that you can easily combine this knowledge by associating symbolic information about things and symbolic information about actions that can be performed on them and with them. These associations effectively form rules that describe the possible behaviours of the world and, similarly, the behaviours of the cognitive agent. This leads to the second reason why such a symbolic representational view of cognition is useful: the cognitive agent can then reason effectively about this knowledge to reach conclusions, make decisions, and execute actions. In other words, the agent can make inferences about the world around it and how it should behave in order to do something useful, i.e. to perform some task and achieve some goal. For example, if a particular item of clothing is heavily soiled but it is a delicate fabric, the agent can select a cool wash cycle with a pre-soak, rather than putting it into a hot water cycle (which will certainly clean the item of clothing but will probably also cause it to shrink and fade). Of course, the agent needs to know all this if it is to make the right decisions, but this doesn't present an insurmountable difficulty as long as someone or something can provide the requisite knowledge in the right form. In fact, it turns out to be relatively straightforward because of the denotational nature of the knowledge: other cognitive agents have the same representational framework and they can share this domain knowledge directly[5] with the cognitive robot doing the laundry. This is the power of the cognitivist perspective on cognition and knowledge.

[5] The idea of cognitive robots sharing knowledge is already a reality. For example, as a result of the RoboEarth initiative, robots are now able to share information on the internet: see the article by Markus Waibel and his colleagues "RoboEarth: A World-Wide Web for Robots" [34]. For more details, see Chapter 8, Section 8.6.1.

A particular feature of this shared symbolic knowledge — rules that describe the domain in which the cognitive agent is operating — is that it is even more abstract than the symbolic knowledge the agent has about its current environment: this domain knowledge describes things *in general*, in a way that isn't specific to the particular object the agent has in front of it or the actions it is currently peforming. For example, the knowledge that delicate coloured fabrics will fade and shrink in very hot water isn't specific to the bundle of laundry that the agent is sorting but it does apply to it nonetheless and, more to the point, it can be used to decide how to wash this particular shirt.

Now, let's consider for a moment the issue of where the agent's knowledge comes from. In most cognitivist approaches concerned with creating artificial cognitive systems, the symbolic knowledge is the descriptive product of a human designer. The descriptive aspect is important: the knowledge in question is effectively a description of how the designer — a third-party observer — sees or comprehends the cognitive agent and the world around it. So, why is this a problem? Well, it's not a problem if every agent's description is identical or, at the very least, compatible: if every agent sees and experiences the world the same way and, more to the point, generates a compatible symbolic representation of it. If this is the case — and it will be the case if the assertion which cognitivism makes that an agent's symbolic representation denotes the objects and events in the world is true — then the consequence is very significant because it means that these symbolic representations can be directly accessed, understood, and shared by the cognitive agent (including other people). Furthermore, it means that domain knowledge can be embedded directly in to, and extracted from, an artificial cognitive agent. This direct transferrability of knowledge between agents is one of cognitivism's key characteristics. Clearly, this makes cognitivism very powerful and extremely appealing, and the denotational attribute of symbolic knowledge is one of its cornerstones.

You may have noticed above a degree of uncertainty in the conditional way we expressed the compatibility of descriptive knowledge and its denotational quality ("if the assertion

... that an agent's symbolic representation denotes the objects and events in the world is true"). Cognitivism asserts that this is indeed the case but, as we will see in Section 2.2, the emergent paradigm of cognitive science takes strong issue with this position. Since this denotational aspect of knowledge and knowledge sharing (in effect, cognitivist epistemology) is so important, it will come as no surprise that there are some far-reaching implications. One of them concerns the manner in which cognitive computations are carried out and, specifically, the issue of whether or not the platform that supports the required symbolic computation is of any consequence. In fact, in the cognitivist paradigm, it isn't of any consequence: any physical platform that supports the performance of the required symbolic computation will suffice. In other words, a given cognitive system (technically, for a given cognitive architecture; but we'll wait until Chapter 3 to discuss this distinction) and its component knowledge (the content of the cognitive architecture) can exploit any machine that is capable of carrying out the required symbol manipulation. This could be a human brain or a digital computer. The principled separation of computational operation from the physical platform that supports these computations is known as *computational functionalism*.[6] Cognitivist cognitive systems are computationally functionalist systems.

Although the relationship between computational functionalism and the universal capability of cognitivist symbolic knowledge is intuitively clear — if every cognitive agent has the same world view and a compatible representational framework, then the physical support for the symbolic knowledge and associated computation is of secondary importance — they both have their roots in classical artificial intelligence, a topic to which we now turn. We will return to the cognitivist perspective on knowledge and representation in Chapter 8.

[6] The principled decoupling of the computational model of cognition from its instantiation as a physical system is referred to as *computational functionalism*. It has its roots in, for example, Allen Newell's and Herbert Simon's seminal paper "Computer Science as Empirical Enquiry: Symbols and Search" [35]. The chief point of computational functionalism is that the physical realization of the computational model is inconsequential to the model: any physical platform that supports the performance of the required symbolic computations will suffice, be it computer or human brain; also see Chapter 5, Section 5.2.

2.1.2 *Cognitivism and Artificial Intelligence*

As we mentioned at the outset, both cognitivist cognitive science and artificial intelligence share a common beginning and they developed together, building a strong symbiotic relationship over

a period of approximately thirty years.[7] Artificial intelligence then diverged somewhat from its roots, shifting its emphasis away from its orginal concern with human and artificial cognition and their shared principles to issues concerned more with practical expediency and purely computational algorithmic techniques such as statistical machine learning. However, the past few years have seen a return to its roots in cognitivist cognitive science, now under the banner of Artificial General Intelligence (to reflect the reassertion of the importance of non-specific approaches built on human-level cognitive foundations).[8] Since there is such a strong bond between cognitivism and classical artificial intelligence, it is worth spending some time discussing this relationship.

In particular, because it has been extraordinarily influential in shaping how we think about intelligence, natural as well as computational, we will discussAllen Newell's and Herbert Simon's "Physical Symbol System" approach to artificial intelligence.[9] As is often the case with seminal writing, the commentaries and interpretations of the original work frequently present it in a somewhat distorted fashion and some of the more subtle and deeper insights get lost. Despite our brief treatment, we will try to avoid this here.

Allen Newell and Herbert Simon, in their 1975 ACM Turing Award Lecture "Computer Science as Empirical Enquiry: Symbol and Search," present two hypotheses:

1. The *Physical Symbol System Hypothesis:* A physical symbol system has the necessary and sufficient means for general intelligent action.

2. The *Heuristic Search Hypothesis:* The solutions to problems are represented as symbol structures. A physical-symbol system exercises its intelligence in problem-solving by search, that is, by generating and progressively modifying symbol structures until it produces a solution structure.

The first hypothesis implies that any system that exhibits general intelligence is a physical symbol system and, furthermore, any physical symbol system of sufficient size can be configured somehow ("organized further," in the words of Newell and

[7] Some observers view AI less as a discipline that co-developed along with cognitivist cognitive science and more as the direct descendent of cognitivism. Consider the following statement by Walter Freeman and Rafael Núñez: "... the positivist and reductionist study of the mind gained an extraordinary popularity through a relatively recent doctrine called *Cognitivism*, a view that shaped the creation of a new field — *Cognitive Science* — and its most hard core offspring: Artificial Intelligence" (emphasis in the original) [36].

[8] The renewal of the cognitivist goals of classical artificial intellegence to understand and model human-level intelligence is typified by the topics addressed by the cognitive systems track of the AAAI conference [37], and the emergence of the discipline of *artificial general intelligence* promoted by, among others, the Artificial General Intelligence Society [38] and the Artificial General Intelligence Research Institute [39].

[9] Allen Newell and Herbert Simon were the recipients of the 1975 ACM Turing Award. Their Turing Award lecture "Computer Science as Empirical Enquiry: Symbol and Search" [35] proved to be extremely influential in the development of artificial intelligence and cognitivist cognitive science.

Simon) to exhibit general intelligence. This is a very strong assertion. It says two things: (a) that it is necessary for a system to be a physical symbol system if the system is to display general intelligence (or cognition), and (b) that being a physical symbol system of adequate size is sufficient to be an intelligent system — you don't need anything else.

The second hypothesis amounts to an assertion that symbol systems solve problems by heuristic search, *i.e.* "successive generation of potential solution structures" in an effective and efficient manner: "The task of intelligence, then, is to avert the ever-present threat of the exponential explosion of search." This hypothesis is sometimes caricatured by the statement that "All AI is search" but this is to unfairly misrepresent the essence of the second hypothesis. The point is that a physical symbol system must indeed search for solutions to the problem but it is intelligent because its search strategy is effective and efficient: it doesn't fall back into blind exhaustive search strategies that would have no hope of finding in a reasonable amount of time a solution to the kinds of problems that AI is interested in. Why? Because these are exactly the problems that defy simple exhaustive search techniques by virtue of the fact that the computational complexity of these brute-force solutions — the amount of time needed to solve them — increases exponentially with the size of the problem.[10] It is in this sense that the purpose of intelligence is to deal effectively with the danger of exponentially large search spaces.

A physical symbol system is essentially a machine that produces over time an evolving collection of symbol structures.[11] A symbol is a physical pattern and it can occur as a component of another type of entity called an expression or symbol structure: in other words, expressions or symbol structures are arrangements of symbols. As well as the symbol structures, the system also comprises processes that operate on expressions to produce other expressions: to process, create, modify, reproduce, and destroy them. An expression can *designate* an object and thereby the system can either affect the object itself or behave in ways that depend on the object. Alternatively, if the expression designates a process, then the system *interprets* the expression by

[10] Formally, we say that exponential complexity is of the order k^n, where k is a constant and n is the size of the problem. By contrast, we also have polynomial complexity: n^2, n^3, or n^k. The difference between these two classes is immense. Problems with exponential complexity solutions scale very badly and, for any reasonably-sized problem, are usually intractable, i.e. they can take days or years to solve, unless some clever — intelligent — solution strategy is used.

[11] For a succinct overview of symbol systems, see Stevan Harnad's seminal article "The Symbol Grounding Problem" [40].

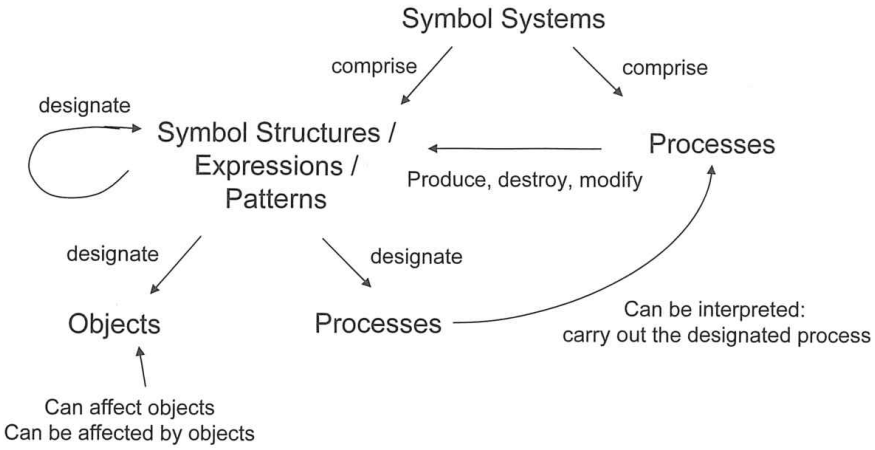

Figure 2.2: The essence of a physical symbol system [35].

carrying out the process (see Figure 2.2). In the words of Newell and Simon,

> "Symbol systems are collections of patterns and processes, the latter being capable of producing, destroying, and modifying the former. The most important properties of patterns is that they can designate objects, processes, or other patterns, and that when they designate processes, they can be interpreted. Interpretation means carrying out the designated process. The two most significant classes of symbol systems with which we are acquainted are human beings and computers."

There is an important if subtle point here. This explanation of a symbol system is much more general and powerful than the usual portrayal of symbol-manipulation systems in which symbols designate only objects, and in which case we have a system of processes that produces, destroys, and modifies symbols, and no more. Newell's and Simon's original view is considerably more sophisticated. There are two recursive aspects to it: processes can produce processes, and patterns can designate patterns (which, in turn, can be processes). These two recursive loops are closely linked. Not only can the system build ever more abstract representations and reason about those representation, but *it can modify itself* as a function of its processing and its symbolic representations. The essential point is that in Newell's and Simon's original vision, physical symbol systems are in prin-

ciple capable of development. This point is often lost in contemporary discussions of cognitivist AI systems. On the other hand, as we will see later, emergent approaches focus squarely on the need for development. This is one of the resonances between the cognitivist and emergent paradigms we mentioned, although one that isn't often picked up on because the developmental capacity that is intrinsic in principle to cognitivist system, by virtue of the physical symbol systems hypothesis, often doesn't get the recognition it deserves.

In order to be realized as a practical agent, symbol systems have to be executed on some computational platform. However, the behaviour of these realized systems depend on the details of the symbol system, its symbols, operations, and interpretations, and *not* on the particular form of the realization. This is something we have already met: the functionalist nature of cognitivist cognitive systems. The computational platform that supports the physical symbol system or the cognitive model is arbitrary to the extent that it doesn't play any part in the model itself. It may well influence how fast the processes run and the length of time it takes to produce a solution, but this is only a matter of timing and not outcome, which will be the same in every case.

Thus, the physical symbol system hypothesis asserts that a physical symbol system has the necessary and sufficient means for general intelligence. From what we have just said about symbol systems, it follows that intelligent systems, either natural or artificial ones, are effectively equivalent because the instantiation is actually inconsequential, at least in principle. It is evident that, to a very great extent, cognitivist systems and physical symbol systems are effectively identical with one another. Both share the same assumptions, and view cognition or intelligence in exactly the same way.

Shortly after Newell and Simon published their influential paper, Allen Newell defined intelligence as the degree to which a system approximates the ideal of a knowledge-level system.[12] This is a system which can bring to bear *all* its knowledge onto *every* problem it attempts to solve (or, equivalently, every goal it attempts to achieve). Perfect intelligence implies complete utilization of knowledge. It brings this knowledge to bear ac-

[12] In addition to his seminal 1975 Turing Award Lecture, Allen Newell made several subsequent landmark contributions to the establishment of practical cognitivist system, beginning perhaps in 1982 with his introduction of the concept of a knowledge-level system, the Maximum Rationality Hypothesis, and the principle of rationality [41], in the mid-1980s with the development of the Soar cognitive architecture for general intelligence (along with John Laird and Paul Rosenbloom) [42], and in 1990 the idea of a Unified Theory of Cognition [43].

cording to the Maximum Rationality Hypothesis expressed as the *principle of rationality* which was proposed by Allen Newell in 1982 as follows: "If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action." John Anderson later offered a slightly different perspective, referred to as *rational analysis*, in which the cognitive system optimizes the adaptation of the behaviour of the organism. Note that Anderson's principle considers optimality to be necessary for rationality, something that Newell's principle does not.[13] In essence, the principle of rationality formalizes the intuitive idea that an intelligent agent will never ignore something if it knows it will help achieve its goal and will always use as much of its knowledge as it can to guide its behaviour and successfully complete whatever task it is engaged in.

[13] For a good comparison of Newell's principle of rationality [41] and Anderson's rational analysis [44], refer to the University of Michigan's Survey of Cognitive and Agent Architectures [45].

As we might expect, the knowledge in such an artificial intelligence system, i.e. in a knowledge-level system, is represented by symbols. Symbols are abstract entities that may be instantiated and manipulated as *tokens*. Developing his physical symbol systems hypothesis, Newell characterizes a symbol system as follows.[14] It has:

[14] Newell's characterization of a symbol system can be found on a website at the University of Michigan dedicated to cognitive and agent architectures [45] .

- *Memory* to contain the symbolic information;
- *Symbols* to provide a pattern to match or index other symbolic information;
- *Operations* to manipulate symbols;
- *Interpretations* to allow symbols to specify operations;
- *Capacities* for *composability*, so that the operators may produce any symbol structure; for *interpretability*, so that the symbol structures are able to encode any meaningful arrangement of operations; and sufficient *memory* to facilitate both of the above.

Newell suggests a progression of four bands of operation, depending on the timescale over which processing takes place, ranging from biological, cognitive, rational, to social. The typical execution time in the biological band is 10^{-4} to 10^{-2} seconds, the cognitive 10^{-1} to 10^{1} seconds, the rational 10^{2} to 10^{4} seconds,

and the social 10^5 to 10^7 seconds. The biological band corresponds to the neurophysiological make-up of the system. Newell identifies three layers in this band: the organelle, the neuron, and the neural circuit. Connectionist systems and artifical neural networks are often focussed exclusively on this band.

The cognitive band corresponds to the symbol level and its physical instantiation as a concrete architecture. The idea of a cognitive architecture is one of the most important topics in the study of cognitive systems (both biological and artificial) and artificial intelligence, and cognitive architectures play a key role in cognitivist cognitive science in particular. We devote all of the next chapter to this topic.

Newell identifies three layers in the cognitive band. First, there are deliberate acts that take a very short amount of time, typically 100ms. For example, reaching to grasp an object. Second, there are 'composed operations' which comprise sequences of deliberate acts. For example, reaching for an article of clothing, grasping it, picking it up, and putting it in a washing machine. These composed operations take on the order of a second. Third, there are complete actions that take up to ten seconds. For example, finding the washing powder tablets, opening the washing powder tray in the washing machine, inserting the tablet, and adding the fabric conditioner.

The rational band is concerned with actions that are typically characterized by tasks and require some reasoning. For example, doing the laundry. The social band extends activity to behaviours that occupy hours, days, or weeks, often involving interaction with other agents.

All knowledge is represented (symbolically) at the cognitive symbol level. All knowledge-level systems contain a symbol system. As we have already seen, this is the strong interpretation of the physical symbol system hypothesis: not only is a physical symbol system *sufficient* for general intelligence, it is also *necessary* for intelligence.

This section has summarized very briefly the close relationship between classical AI and cognitivism and, by extension, the new incarnation of classical AI in Artificial General Intelligence (AGI). It is impossible to overestimate the influence that AI

has had on cognitivist cognitive science and, to a slightly lesser extent, that of cognitivism on AI. This is not surprising when you consider that they were both born as disciplines at the same time by more or less the same people and that their goals were identical: to develop a comprehensive theory — inspired by the computational model of information processing — that moved forward the original agenda of the early cyberneticians to formalize the mechanisms that underpin cognition in animals and machines. AI may well have deviated from that goal in the last 20 or so years to pursue alternative strategies such as statistical machine learning but, as we mentioned above, there is now a large and growing body of people who are championing a return to the original goals of the founders of cognitivism and classical AI which, in the words of Pat Langley at Arizona State University, was to understand and reproduce in computational systems the full range of intelligent behaviour observed in humans.

With this important pillar of cognitive science now firmly established, let's move on to the second pillar that also grew out of the goals and aspirations of the early cyberneticians: emergent systems.

2.2 *The Emergent Paradigm of Cognitive Science*

The view of cognition taken by emergent approaches is very different to that taken by cognitivism. The ultimate goal of an emergent cognitive system is to maintain its own autonomy, and cognition is the process by which it accomplishes this. It does so through a process of continual self-organization whereby the agent interacts with the world around it but only in such a way as not to threaten its autonomy. In fact, the goal of cognition is to make sure that the agent's autonomy is not compromised but is continually enhanced to make its interactions increasingly more robust. In achieving this, the cognitive process determines what is real and meaningful for the system: the system constructs its reality — its world and the meaning of its perceptions and actions — as a result of its operation in that world. Consequently, the system's understanding of its world is inherently specific to the form of the system's embodiment and is dependent on the

system's history of interactions, i.e., its experiences.[15] This process of making sense of its environmental interactions is one of the foundations of a branch of cognitive science called *enaction*, about which we will say much more in Section 2.2.3. Cognition is also the means by which the system compensates for the immediate "here-and-now" nature of perception, allowing it to anticipate events that occur over longer timescales and prepare for interaction that may be necessary in the future. Thus, cognition is intrinsically linked with the ability of an agent to act prospectively: to deal with what might be, not just with what is.

[15] This mutual specification of the system's reality by the system and its environment is referred to as co-determination [14] and is related to the concept of radical constructivism [46] (see Chapter 8, Section 8.3.4).

Many emergent approaches also adhere to the principle that the primary mode of cognitive learning is through the acquisition of new anticipatory skills rather than knowledge, as is the case in cognitivism.[16] As a result, in contrast to cognitivism, emergent approaches are necessarily embodied and the physical form of the agent's body plays a pivotal role in the cognitive process. Emergent systems wholeheartedly embrace the idea of the interdependence between brain, body, and world that we mentioned in the previous chapter. Because of this, cognition in the emergent paradigm is often referred to as *embodied cognition*.[17] However, while the two terms might be synonymous, they are not equivalent. Embodied cognition focusses on the fact that the body and the brain, together, form the basis of a cognitive system and they do so in the context of a structured environmental niche to which the body is adapted. Emergent systems, as we will see, do so too but often they make even stronger assertions about the nature of cognition.

[16] Emergent approaches typically proceed on the basis that processes which guide action and improve the capacity to guide action form the root capacity of all intelligent systems [47].

[17] We discuss the issue of *embodied cognition* in detail in Chapter 5.

The emergent paradigm can be sub-divided into three approaches: connectionist systems, dynamical systems, and enactive systems (refer again to Figure 2.1). However, it would be wrong to suggest that these three approaches have nothing to do with one another. On the contrary, there are very important ways in which they overlap. The ultimate-proximate relationship again helps to clarify the distinctions we make between them. Both connectionist and dynamical systems are more concerned with proximate explanations of cognition, i.e. the mechanisms by which cognition is achieved in a system. Typically, connectionist systems correspond to models at a lower level of abstraction, dy-

namical systems to a higher level. Enaction, on the other hand, makes some strong claims about what cognition is for (as does the emergent paradigm in general) and why certain characteristics are important. Enaction does have a lot to say about how the process of cognition is effected, touching on the proximate explanations, but it does so at quite an abstract level. To date, explicit formal mechanistic, computational, or mathematical models of enaction remain goals for future research.[18] In contrast, connectionism and dynamical systems theory provide us with very detailed and well-understood formal techniques, mathematically and computationally, but the challenge of scaling them to a fully-fledged theory of cognition on a par with, say, the cognitivist Unified Theory of Cognition (a concept that was already mentioned above in Section 2.1.2 and that will be discussed more fully in Chapter 3) requires much more time and effort, not to mention some intellectual breakthroughs.

Many working on the area feel that the future of the emergent paradigm may lie in the unification of connectionist, dynamical, and enactive approaches, binding them together in a cohesive joint ultimate-proximate explanation of cognition. Indeed, as we will see shortly, connectionism and dynamical systems theory are best viewed as two complementary views of a common approach, the former dealing with microscopic aspects and the latter with macroscopic. On the other hand, others working in the field prefer the view that a marriage of cognitivist and emergent approaches is the best way forward, as exemplified by the hybrid systems approach about which we will say more in Section 2.3. There are other interesting perspectives too, such as the computational mechanics espoused by James Crutchfield who argues for a synthesis and extension of dynamical and information processing approaches.[19]

Bearing in mind these relationships, let us now proceed to examine the three emergent approaches, one by one, highlighting the areas where they overlap.

[18] *Enaction: Towards a New Paradigm for Cognitive Science* [48], edited by John Stewart, Olivier Gapenne, and Ezequiel Di Paolo, and published by MIT Press in 2010, provides an excellent snapshot of the current state of development of the enactive paradigm.

[19] James Crutchfield agrees with those who advocate a dynamical perspective on cognition, asserting that time is a critical element, and he makes the point that one of the advantages of dynamical systems approaches is that it renders the temporal aspects geometrically in a state space. Structures in this state space both generate and constrain behaviour and the emergence of spatio-temporal patterns. Dynamics, then, are certainly involved in cognition. However, he argues that dynamics *per se* are "not a substitute for information processing and computation in cognitive processes" but neither are the two approaches incompatible [49]. He holds that a synthesis of the two can be developed to provide an approach that does allow dynamical state space structures to support computation. He proposes *computational mechanics* as the way to tackle this synthesis of dynamics and computation.

2.2.1 *Connectionist Systems*

Connectionist systems rely on parallel processing of non-symbolic distributed activation patterns in networks of relatively simple processing elements. They use statistical properties rather than logical rules to analyze information and produce effective behaviour. In the following, we will summarize the main principles of connectionism, briefly tracing its history and highlighting the main developments that have led us to where we are today. Unfortunately, but inevitably, we will be forced into making use of many technical terms with little or no explanation: to do justice to connectionism and the related field of artificial neural networks would require a substantial textbook in its own right. All we can hope for here is to convey some sense of the essence of the connectionism, its relevance to cognitive science, and the way it differs from cognitivism. References to supplementary material are provided in the sidenote.[20]

The roots of connectionism reach back well before the computational era. Although the first use of connectionism for computer-based models dates from 1982,[21] the term connectionism had been used in psychology as early as 1932.[22] Indeed, connectionist principles are clearly evident in William James' nineteenth century model of associative memory,[23] a model that also anticipated mechanisms such as Hebbian learning, an influential unsupervised neural training process whereby the synaptic strength — the bond between connecting neurons — is increased if both the source and target neurons are active at the same time. The introduction to Donald Hebb's book *The Organization of Behaviour* published in 1949 also contains one of the first usages of the term connectionism.[24]

We have already noted that cognitivism has some of its roots in earlier work in cybernetics and in the seminal work by Warren McCulloch and Walter Pitts in particular.[25] They showed that any statement within propositional logic could be represented by a network of simple processing units, i.e. a connectionist system. They also showed that such nets have, in principle, the computational power of a Universal Turing Machine, the theoretical basis for all computation. Thus, McCulloch and Pitts managed

[20] David Medler's paper "A Brief History of Connectionism" [50] provides an overview of classical and contemporary approaches and a summary of the link between connectionism and cognitive science. For a selection of seminal papers on connectionism, see James Anderson's and Edward Rosenfeld's *Neurocomputing: Foundations of Research* [28] and *Neurocomputing 2: Directions of Research* [51]. Paul Smolensky reviews the field from a mathematical perspective [52, 53, 54, 55]. Michael Arbib's *Handbook of Brain Theory and Neural Networks* provides very accessible summaries of much of the relevant literature [56].

[21] The introduction of the term *connectionist models* in 1982 is usually attributed to Jerome Feldman and Dana Ballard in their paper "Connectionist Models and their Properties" [57].

[22] Edward Thorndike used the term connectionism to refer to an extended form of associationism in 1932 [58, 59].

[23] Anderson's and Rosenfeld's collection of papers [28] opens with Chapter XVI "Association" from William James's 1890 *Psychology, Briefer Course* [60].

[24] The introduction of Donald Hebb's book *The Organization of Behaviour* [61] can be found in Anderson's and Rosenfeld's collection of papers [28].

[25] See Sidenote 2 in this chapter.

the remarkable feat of contributing simultaneously to both the foundation of cognitivism and the foundation of connectionism.

The connectionist approach was advanced significantly in the late 1950s with the introduction of Frank Rosenblatt's perceptron and Oliver Selfridge's Pandemonium model of learning.[26] Rosenblatt showed that any pattern classification problem expressed in binary notation can be solved by a perceptron network, a simple network of elementary computing elements which do little more than sum the strength of suitably-weighted input signals or data streams, compare the result to some fixed threshold value, and, on the basis of the result, they either fire or not, producing a single output which then connected to other computing element in the network.

Although network learning advanced in 1960 with the introduction of the Widrow-Hoff rule (also called the delta rule) for supervised training in the *Adaline* neural model,[27] perceptron networks suffered from a severe problem: no learning algorithm existed to allow the adjustment of the weights of the connections between input units and hidden associative units in networks with more than two layers.

In 1969, Marvin Minsky and Seymour Papert caused something of a stir by showing that these perceptrons can only be trained to solve linearly separable problems and couldn't be trained to solve more general problems.[28] As a result, research on neural networks and connectionism suffered considerably.

With the apparent limitations of perceptions clouding work on network learning, research focussed more on memory and information retrieval and, in particular, on parallel models of associative memory.[29]

During this period, alternative connectionist models were also being developed, such as Stephen Grossberg's Adaptive Resonance Theory (ART)[30] and Teuvo Kohonen's self-organizing maps (SOM), often referred to simply as Kohonen networks.[31] ART addresses real-time supervised and unsupervised category learning, pattern classification, and prediction, while Kohonen networks exploit self-organization for unsupervised learning and can be used as either an auto-associative memory or a pattern classifier.

[26] The perceptron was introduced in 1958 by Frank Rosenblatt [62] while the *Pandemonium* learning model was developed by Oliver Selfridge in 1959 [63]; both are included in Anderson's and Rosenfeld's collection of seminal papers [28].

[27] Adaline — for Adaptive Linear — was introduced by Bernard Widrow and Marcian Hoff in 1960 [64].

[28] *Perceptrons: An Introduction to Computational Geometry* by Marvin Minsky and Seymour Papert [65] was published in 1969 and had a very strong negative influence on neural network research for over a decade. For a review of the book, see "No Harm Intended" by Jordan Pollack [66].

[29] For examples of connectionist work carried out in the 1970s and early 1980s, see Geoffrey Hinton's and James Anderson's book "Parallel Models of Associative Memory" [67].

[30] Adaptive Resonance Theory (ART) was introduced by Stephen Grossberg in 1976 and has evolved considerably since then. For a succinct summary, see the entry in *The Handbook of Brain Theory and Neural Networks* [68].

[31] Kohonen networks [69] produce topological maps in which proximate points in the input space are mapped by an unsupervised self-organizing learning process to an internal network state which preserves this topology.

Perceptron-like neural networks underwent a strong resurgence in the mid-1980s with the development of the parallel distributed processing (PDP) architecture,[32] in general, and with the introduction of the back-propagation algorithm by David Rumelhart, Geoffrey Hinton, and Ronald Williams, in particular.[33] The back-propagation learning algorithm, also known as the generalized delta rule or GDR since it is a generalization of the Widrow-Hoff delta rule for training Adaline units, overcame the limitation identified by Minsky and Papert by allowing the connection weights between the input units and the hidden units be modified, thereby enabling multi-layer perceptrons to learn solutions to problems that are not linearly separable. This was a major breakthrough in neural network and connectionist research. In cognitive science, PDP had a significant impact in promoting a move away from the sequential view of computational models of mind, towards a view of concurrently-operating networks of mutually-cooperating and competing units. PDP also played an important role in raising an awareness of the importance of the structure of the computing system on the computation, thereby challenging the functionalist doctrine of cognitivism and the principled divorce of computation from computational platform.

[32] David Rumelhart's and James McClelland's 1986 book *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* [70] had a major influence on connectionist models of cognition.

[33] Although the back-propagation learning rule made its great impact through the work of David Rumelhart *et al.* [71, 72], it had previously been derived independently by Paul Werbos [73], among others [50].

The standard PDP model represents a static mapping between the input vectors as a consequence of the feed-forward configuration, i.e. a configuration in which data flows in just one direction through the network, from input to output. There is an alternative, however, in which the network has connections that loop back to form circuits, *i.e.* networks in which either the output or the hidden unit activation signals are fed back to the network as inputs. These are called recurrent neural networks. The recurrent pathways in the network introduce a dynamic behaviour into the network operation.[34] Perhaps the best known type of recurrent network is the Hopfield network. These are fully recurrent networks that act as an auto-associative memory or content-addressable memory.

[34] This recurrent feed-back has nothing to do with the feed-back of error signals (by, for example, back-propagation) to effect weight adjustment during learning.

As a brief aside, associative memory comes in two types: hetero-associative memory and auto-associative memory. Hetero-associative memory produces an output that is different in char-

acter from the input; the two are associated. Technically, the spaces to which the input and output vectors belong are different. For example, the input space might be an image of an object and the output might be a digitally-synthesized speech signal encoding a word or phrase describing the object's identity. On the other hand, auto-associative memory produces an output vector that belongs to the same space as the input vector. For example, a poorly-taken image of the object might produce — recall — a perfect image of the object taken previously.

Other recurrent networks include Elman networks (with recurrent connections from the hidden to the input units) and Jordan networks (with recurrent connections from the output to the input units). Boltzmann machines are variants of Hopfield nets that use stochastic rather than deterministic weight update procedures to avoid problems with the network becoming trapped in local minima during learning.[35]

Multi-layer perceptrons and other PDP connectionist networks typically use monotonic functions[36] such as hard-limiting threshold functions or sigmoid functions to trigger the activation of individual neurons. The use of non-monotonic activation functions, such as the Gaussian function, can offer computational advantages, *e.g.* faster and more reliable convergence on problems that are not linearly separable. Radial basis function (RBF) networks[37] use Gaussian functions but differ from multi-layer perceptrons in that the Gaussian function is used only for the hidden layer, with the input and output layers using linear activation functions.

Connectionist systems still continue to have a strong influence on cognitive science, either in a strictly PDP sense such as James McClelland's and Timothy Rogers' PDP approach to semantic cognition or in the guise of hybrid systems such as Paul Smolensky's and Geraldine Legendre's connectionist/symbolic computational architecture for cognition.[38]

With that all-too-brief overview of connectionism in mind, we can now see why connectionism, as a component of the emergent paradigm of cognitive science, is viewed as a viable and attractive alternative to cognitivism. Specifically, one of the original motivations for work on emergent systems was disaffection with

[35] For more information on Hopfield networks, Elman networks, Jordan networks, and Boltzmann machines, refer to [74], [75], [76], and [77], respectively.

[36] Monotonic functions grow in one direction only: monotonically-increasing functions only increase in value as the independent variable gets larger whereas monotonically-decreasing functions only decrease in value as the independent variable gets larger.

[37] For more details on radial basis function (RBF) networks, see for example [78].

[38] See [79] for details of James McClelland's and Timothy Rogers' PDP approach to semantic cognition and [80, 81] for details of Paul Smolensky's and Geraldine Legendre's connectionist/symbolic computational architecture for cognition.

the sequential, atemporal, and localized character of symbol-manipulation based cognitivism. Emergent systems, on the other hand, depend on parallel, real-time, and distributed architectures, just like natural biological systems do, and connectionist neural networks with their inherent capacity for learning are an obvious and appealing way to realize such systems. Of itself, however, this shift in emphasis isn't sufficient to constitute a new paradigm. While parallel distributed processing and real-time operation are certainly typical characteristics of connectionist systems, there must be more to it than this since modern cognitivist systems exhibit the very same attributes.[39] So, what are the key differentiating features? We defer answering this question until in Section 2.4, where we will compare and contrast the cognitivist and emergent paradigms on the basis of fourteen distinct characteristics. For now, we move on to consider dynamical systems approaches to emergent cognitive science.

[39] Walter Freeman and Rafael Núñez have argued that recent systems — what they term neo-cognitivist systems — exploit parallel and distributed computing in the form of artificial neural networks and associative memories but, nonetheless, still adhere to many of the original cognitivist assumptions [36]. A similar point is made by Timothy van Gelder and Robert Port [82].

2.2.2 *Dynamical Systems*

While connectionist systems focus on the pattern of activity that emerges from an adaptive network of relatively simple processing elements, dynamical systems theory models the behaviour of systems by using differential equations to capture how certain important variables that characterize the state of the system change with time. Dynamical systems theory is a very general approach and has been used to model many different types of systems in various domains such as biology, astronomy, ecology, economonics, physics, and many others.[40]

A dynamical system defines a particular pattern of behaviour. The system is characterized by a state vector and its time derivative, i.e. how it changes as time passes. This time derivative is determined by the state vector itself and also some other variables called control parameters. Usually, the dynamical equations also takes noise into account. To model a dynamical system, you need to identify the state variables and the control parameters, how noise will be modelled, and finally the exact form of the relationship which combines these and expresses them in terms of derivatives, i.e. how they change with time.

[40] For an intuitive introduction to dynamical systems theory, see Section 5.2 of Lawrence Shapiro's book *Embodied Cognition* [83]. For an overview of the way dynamical systems theory can be used to model cognitive behaviour, refer to Scott Kelso's book *Dynamic Patterns – The Self-Organization of Brain and Behaviour* [21].

The Nature of Dynamical Systems

In general, a dynamical system has several key attributes. First of all, it is a system. This may seem to be a bit obvious but it's important. It means that it comprises a large number of interacting components and therefore a large number of degrees of freedom.

Second, the system is dissipative, that is, it uses up or dissipates energy. This has an important consequence on the system behaviour. In particular, it means that the number of states that the system can reach reduces with time. Technically, we say that its phase space decreases in volume. The main upshot of this is that the system develops a preference for certain sets of states (again, technically, they are preferential sub-spaces in the complete space of possible states).

A dynamical system is also what is referred to as a non-equilibrium system. This just means that it never comes to rest. It doesn't mean that it can't exhibit stable behaviour — it can — but it does mean that it is unable to maintain its structure and carry out its function without external sources of energy, material, or information. In turn, this means that, at least from an energy, material, or information perspective, the system is open, i.e. stuff can enter and exit the system. In contrast, a closed system doesn't allow anything to cross the system boundary.

A dynamical system is also non-linear. This simply means that the equations that define the differential relationship between the state variables, the control parameters, and the noise components are combined together in a multiplicative manner and not simply by weighting and adding them together. Although non-linearity might appear to be a mathematical nicety (or, more likely, a mathematical complication) this non-linearity is extremely important because it provides the basis for complex behaviour — most of the world's interesting phenomena exhibit this hard-to-model characteristic of non-linearity — but, not only that, it also means that the dissipation is not uniform and that only a small number of the system's overall degrees of freedom contribute to its behaviour. In other words, when modelling the system, we need focus only on a small number of state variables instead of having to consider every single one (which would more or less make the task of modelling the system impossible). We refer to

these special variables by two different, but entirely equivalent, terms: *order parameters* and *collective variables*.[41] Which term you choose is largely a matter of convention or tradition.

Each collective variable plays a key role in defining the way the system's behaviour develops over time. In essence, the collective variables are the subset of the system variables that govern the system behaviour. The main consequence of the existence of these collective variables is that the system behaviour is characterized by a succession of relatively stable states: in each state the system is doing something specific and stays doing it until something happens to cause it to jump to the next relatively stable state. For this reason, we say that the states are meta-stable (stable but subject to change) and we call the local regions in the state space around them attractors (because once a behaviour gets close to one, it is attracted to stay in the vicinity of that behaviour until something significant disturbs it).

Being able to model the behaviour of the system — a system with very many variables and therefore a very high dimensional space of *possible* states — with a very small number of relevant variables — and therefore a very low dimensional space of *relevant* states — makes the modelling exercise practical and attractive, and it is one of the main characteristics that distinguish dynamical systems from connectionist systems.

[41] We already met the concept of a collective variable in Chapter 1 when we discussed Scott Kelso's ideas on the different levels of abstraction which need to be considered when modelling a system.

Dynamical Systems and Cognition

These are all very general characteristics of dynamical systems. So, what makes them suitable for modelling cognition? To answer this question, we need to understand the perspective that advocates of dynamical systems take on cognition. Esther Thelen and Helen Smith express it the following way:[42]

> Cognition is non-symbolic, nonrepresentational and all mental activity is emergent, situated, historical, and embodied.

To this we might add that it is socially constructed so that some aspects of cognition arise from the interaction between cognitive agents, again modelled as a dynamical process. It is clear that Thelen and Smith, along with many others who subscribe to the emergent paradigm, take issue with the symbolic nature of cognitivist models and with the representationalism it encapsulates.

[42] This quotation is taken from Esther Thelen's and Helen Smith's influential book *A Dynamic Systems Approach to the Development of Cognition and Action* [84].

Here we must exercise some caution in understanding their interpretation of symbolic representationalism and their assertion that this is not a true reflection of cognition.

There are two principle ways which proponents of dynamical systems models, and emergent models in general, object to symbol manipulation and representation. One is symbol manipulation in the literal sense that a computer program manipulates symbols. In other words, the objection is to the mechanism of symbolic processing: the rule-based shuffling of symbols in search of a state that satisfies the conditions defined by the goal of the system. Instead of this, proponents of dynamical systems and connectionism contend that cognitive behaviour arises as a natural consequence of the interaction of appropriately configured network of elementary components. That is, cognition is a behaviour that is a consequence of self-organization, i.e. a global pattern of activity that arises because of, and only because of, the dynamic interaction of these components. It is emergent in the sense that this global pattern of activity cannot be predicted from the local properties of the system components.[43]

The second aspect of the objections of proponents of dynamical systems concerns the issue of representation. This is a very hotly debated issue and there is considerable disagreement over exactly what different people mean by representation.[44] As we have seen, cognitivism hinges upon the direct denotation of an object or an event in the external world by a symbolic representation that is manipulated by the cognitive system. It is this strong denotational characterization that emergent systems people object to because it entails (i.e. it necessarily involves) a correspondence between the object as it appears and is represented by the symbol and the object as it is in the world. Furthermore, by virtue of the computational functionalism of cognition, these denoted symbolic object correspondences are shared by every cognitive agent, irrespective of the manner in which the system is realized in a cognitive agent: as a computer or as a brain. This is simultaneously the power of cognitivism and a great bone of contention among those who advocate an emergent position.

So, how do emergent systems manage without representations, as the quotation above suggests they do? Here again, we

[43] Strictly speaking, the pattern of activity that arises from self-organization can *in principle* be predicted from the properties of the components so in a sense emergence is a stronger — and more obscure — process which may, or may not, exploit self-organization; see the article "Self-Organizing Systems" by Scott Camazine in *Encyclopedia of Cognitive Science* [85].

[44] We discuss the troublesome issue of representation in some depth in Chapter 8.

need to be careful in our interpretation of the term representation. It is clearly evident from what we have discussed so far that connectionist and dynamical systems exhibit different states. Could these states not be interpreted as "representing" objects and events in the world and, if so, doesn't that contradict the anti-representational position articulated above? The answer to these two questions is a conditional "yes" and a cautious "no." Such states could be construed as a representation, but not in the sense that they *denote* the object or event in the cognitivist sense. Rather, it is a question of them being correlated in some way with these objects and events but they need not mean the same thing: it's a marriage of convenience, not one of absolute commitment.

We say that such a representation *connotes* the objects or events and, in so saying, we imply nothing at all about the nature of the object or the event except that the emergent system's state is correlated in some way with its occurence.[45] If this seems to be a very fine, almost pedantic, point, it's because it is. But it is a fundamentally important point nonetheless since it goes straight to the heart of one of the core differences between the cognitivist and emergent paradigms: the relationship between the state of the agent — cognitivist or emergent — and the world it interacts with.

Cognitivism asserts that the symbolic knowledge it represents about the world is a faithful counterpart of the world itself; emergent approaches make no such claim and, on the contrary, simply allows that the internal state reflects some regularity or lawfulness in the world which it doesn't know but which it can adapt to and exploit through its dynamically-determined behaviour. This helps explain what is meant in the quotation above by cognition being *situated*, *historical*, and *embodied*. A dynamical system must be embodied in some physical way in order to interact with the world and the exact form of that embodiment makes a difference to the manner in which the agent behaves (or can behave). Being situated means that the agent's cognitive understanding of the world around it emerges in the specific context of its local surroundings, not in any virtual or abstract sense. Furthermore, the history of these context-

[45] For a deep, if also very dense, discussion of the difference between denotation and connotation in the specific context of language, refer to Alexander Kravchenko's paper "Essential properties of language, or, why language is not a code" [86].

specific interactions have an effect on the way the dynamical system develops as it continually adjusts and adapts.[46]

[46] We discuss the issue of situated embodiment in some detail in Chapter 5.

Time

This brings us to a crucial aspect of dynamical systems which is rather obvious once we say it: it's about time.[47] To be dynamic means to change with time so it is clear that time must play a crucial part in any dynamical system. With emergent systems in general, and dynamical systems in particular, cognitive processes unfold over time. More significantly, they do so not just in an arbitrary sequence of steps, where the actual time taken to complete each step doesn't have any influence on the outcome of that step, but in real-time — in lock-step, synchronously — with events as they unfold in the world around the agent. So, time, and timing, is at the very heart of cognition and this is one of the reasons why dynamical systems theory may be an appropriate way to model it.

[47] *It's about Time: An Overview of the Dynamical Approach to Cognition* is the title of a book by Robert Port and Timothy van Gelder which is devoted to discussing the importance of time in cognition and arguing the case for a dynamical approach to modelling cognition.

The synchronicity of a dynamical cognitive agent with the events in its environment has two unexpected and, from the perspective of artificial cognitive systems, somewhat unwelcome consequences. First, it places a strong limitation on the rate at which the development of the cognitive agent can proceed. Specifically, it is constrained by the rate at which events in the world unfold and not on the speed at which internal changes can occur in the agent.[48] Biological cognitive systems have a learning cycle measured in weeks, months, and years. While it might be possible to collapse it into minutes and hours for an artificial system because of increases in the rate of internal adaptation and change, it cannot be reduced below the time-scale of the interaction. Second, taken together with the requirement for embodiment, we see that the historical and situated nature of the systems means that we cannot by-pass the developmental process: development is an integral part of cognition, at least in the emergent paradigm of cognitive science.

[48] Terry Winograd and Fernando Flores explain in their book *Computers and Cognition* [87] the impact of real-time interaction between a cognitive system and its environment on the rate at which the system can develop.

Dynamical systems and connectionism

We have already mentioned that there is a natural relationship between dynamical systems and connectionist systems. To a

significant extent, you can consider them to be complementary ways of describing cognitive systems, with dynamical systems focussing on macroscopic behaviour and connectionist systems focussing on microscopic behaviour.[49] Connectionist systems themselves are, after all, dynamical systems with temporal properties and structures such as attractors, instabilities, and transitions. Typically, however, connectionist systems describe the dynamics in a high dimensional space of computing element activation and network connection strengths. On the other hand, dynamical systems theory describes the dynamics in a low dimensional space because a small number of state variables are capable of capturing the behaviour of the system as a whole.[50] Much of the power of dynamical perspectives comes from this higher-level abstraction of the dynamics[51] and, as we have already noted above, this is the key advantage of the dynamical systems formulation of system dynamics: it collapses a very high-dimensional system defined by the complete set of system variables onto a low-dimensional space defined by the collective variables.

The complementary nature of dynamical systems and connectionist descriptions is reflected in the approach to modelling that we met in Chapter 1 in which systems are modelled simultaneously at three distinct levels of abstraction: a boundary constraint level that determines the task or goals (initial conditions, non-specific conditions), a collective variables level which characterize coordinated states, and a component level which forms the realized system (*e.g.* nonlinearly coupled oscillators or neural networks). This complementary perspective of dynamical systems theory and connectionism enables the investigation of the emergent dynamical properties of connectionist systems in terms of attractors, meta-stability, and state transition, all of which arise from the underlying mechanistic dynamics. It also offers the possibility of implementing dynamical systems theory models with connectionist architectures.

The strength of the dynamical systems approach

Those who advocate the use of dynamical systems theory to model cognition point to the fact that they provide you directly

[49] The intimate relationship between connectionism and dynamical systems is teased out in the book *Toward a New Grand Theory of Development? Connectionism and Dynamic Systems Theory Re-Considered*, edited by John Spencer, Michael Thomas, and James McClelland [88].

[50] Gregor Schöner argues that it is possible for a dynamical system model to capture the behaviour of the system using a small number of variables because the macroscopic states of high-dimensional dynamics and their long-term evolution are captured by the dynamics in that part of the space where instabilities occur: the low-dimensional Center-Manifold [89].

[51] There is a useful overview of the dynamical perspective on neural networks in a book *Mathematical perspectives on neural networks* edited by Paul Smolensky, Michael Mozer, and David Rumelhart [90]. The same book also provides useful overviews from computational and statistical viewpoints.

with many of the characteristics inherent in natural cognitive systems such as multistability, adaptability, pattern formation and recognition, intentionality, and learning. These are all achieved purely as a function of dynamical laws and the self-organization of the system that these laws enable. They require no recourse to symbolic representations, especially representations that are the result of human design: there is just an ongoing process of dynamic change and formation of meta-stable patterns of system activity. They also argue that dynamical systems models allow for the development of higher order cognitive functions, such as intentionality and learning, in a relatively straight-forward manner, at least in principle. For example, intentionality — purposive or goal-directed behaviour — can be achieved by superimposing a function that encapsulates the intention onto the equations that define the dynamical system. Similarly, learning can be viewed as a modification of existing behavioural patterns by introducing changes that allow new meta-stable behaviours to emerge, i.e. by developing new attractors in the state space. These changes don't just add the extra meta-stable patterns but in doing so they also may have an effect on existing attractors and existing behaviour. Thus, learning changes the whole system as a matter of course.

While dynamical models can account for several non-trivial behaviours that require sensorimotor learning and the integration of visual sensing and motoric control (e.g. the perception of affordances,[52] perception of time to contact,[53] and figure-ground bi-stability[54] the feasibility of realizing higher-order cognitive faculties has not yet been demonstrated. It appears that dynamical systems theory (and connectionism) needs to be embedded in a larger emergent context. This makes sense when you consider dynamical systems theory and connectionism from the ultimate-proximate perspective we discussed in Chapter 1: both focus more on the proximate aspects of modelling methodology and mechanisms of cognition than on the ultimate concern of what cognition is for, an issue to which we now turn.

[52] The concept of *affordance* is due to the influential psychologist J. J. Gibson [91]. It refers to the potential use to which an object can be put, *as perceived by an observer* of the object. This perceived potential depends on the skills the observer possesses. Thus, the affordance is dependent both on the object itself and the perception and action capabilities of the observing agent.

[53] The *time-to-contact* refers to the time remaining before an agent or a part of the agent's body will make contact with something in the agent's environment. It is often inferred from optical flow (a measure of how each point in the visual field of an agent is moving) and is essential to many behaviours; e.g., as Scott Kelso illustrates in his book *Dynamic Patterns*, gannets use optical flow to determine when to fold their wings before entering the water as they dive for fish [21].

[54] *Figure-ground bi-stability* refers to the way we alternately see one shape (the figure) or another (the background, formed by the complement of the figure), but never both at the same time; see Wolfgang Köhler's *Dynamics in Psychology* [92] for more details.

2.2.3 *Enactive Systems*

We now focus on an increasingly-important approach in cognitive science: enaction.[55] The principal idea of enaction is that a cognitive system develops it own understanding of the world around it through its interactions with the environment. Thus, enactive cognitive system operate autonomously and generate their own models of how the world works.

[55] Section 2.2.3 is based directly on a study by the author of enaction as a framework for development in cognitive robotics [93]. The paper contains additional technical details relating to enactive systems which are not strictly required here. Readers who are interested in delving more deeply into enaction are encouraged to refer to this paper as well as to the original literature [14, 32, 48, 87, 94, 95, 96, 97, 98]. The book *The Embodied Mind* by Francisco Varela, Evan Thompson, and Eleanor Rosch [98] would make a good starting point, followed perhaps by the book *Enaction: Toward a New Paradigm for Cognitive Science* by John Stewart, Olivier Gapenne, and Ezequiel Di Paolo [48] for a contemporary perspective on Enaction.

The five aspects of enaction

When dealing with enactive systems, there are five key elements to consider. These are:

1. Autonomy
2. Embodiment
3. Emergence
4. Experience
5. Sense-making

We have already encountered the first four of these elements.

The issue of autonomy was introduced in Chapter 1, Section 1.3, where we noted the link between cognition and autonomy, particularly from the perspective of biological systems. We take up this important issue again later in the book and devote all of Chapter 4 to unwrapping the somewhat complex relationship between the two topics.

Similarly, in this chapter we have already met the concept of embodiment and the related concept of embodied cognition. Again, a full chapter is dedicated to embodiment later in the book (Chapter 5) reflecting is importance in contemporary cognitive science.

Emergence is, of course, the topic of the current section and we have already discussed the relationship between emergence and self-organization (see Section 2.2.2). Emergence refers to the phenomenon whereby the behaviour we call cognition arises from the dynamic interplay between the components of the system and between the components and the system as a whole. We return to this issue in Chapter 4, Section 4.3.5.

Experience is the fourth element of enaction and, as we noted in the introduction to this section, it is simply the cognitive system's history of interaction with the world around it: the actions it takes in the environment in which it is embedded and the actions arising in the environment which impinge on the cognitive system. In enactive systems, these interactions don't control the system, otherwise it wouldn't be autonomous and, notwithstanding what we said in Chapter 1 about having to be cautious in approaching the relationship between cognition and autonomy, enactive systems are by definition autonomous. Even so, these interactions can and do trigger changes in the state of the system. The changes that can be triggered are *structurally determined*: they depend on the system structure, *i.e.* the embodiment of the self-organizational principles that make the system autonomous.[56] This structure is also referred to as the system's *phylogeny*: the innate capabilities of an autonomous system with which it is equipped at the outset (when it is born, in the case of a biological system) and which form the basis for its continued existence. The experience of the system — its history of interactions — involving *structural coupling*[57] between the system and its environment in an ongoing process of mutual perturbation is referred to as its *ontogeny*.

Finally, we come to the fifth and, arguably, the most important element of enaction: sense-making. This term refers to the relationship between the knowledge encapsulated by a cognitive system and the interactions which gave rise to it. In particular, it refers to the idea that this emergent knowledge is generated by the system itself and that it captures some regularity or lawfulness in the interactions of the system, *i.e.* its experience. However, the sense it makes is dependent on the way in which it can interact: its own actions and its perceptions of the environment's action on it. Since these perceptions and actions are the result of an emergent dynamic process that is first and foremost concerned with maintaining the autonomy and operational identity of the system, these perceptions and actions are unique to the system itself and the resultant knowledge makes sense only insofar as it contributes to the maintenance of the system's autonomy. This ties in neatly with the view of cognition as a pro-

[56] The founders of the enactive approach use the term *structural determination* to denote the dependence of a system's space of viable environmentally-triggered changes on the structure and its internal dynamics [14, 98]. The interactions of this structurally-determined system with the environment in which it is embedded are referred to as *structural coupling*: a process of mutual perturbations of the system and environment that facilitate the on-going operational identity of the system and its autonomous self-maintenance. Furthermore, the process of structural coupling produces a congruence between the system and its environment. For this reason, we say that the system and the environment are *co-determined*. The concepts of structural determination and structural coupling of autopoietic systems [14] are similar to Scott Kelso's circular causality of action and perception [21] and the organizational principles inherent in Mark Bickhard's self-maintenant systems [13]. The concept of enactive development has its roots in the structural coupling of organizationally-closed systems which have a central nervous system and is mirrored in Bickhard's concept of recursive self-maintenance [13].

[57] *Structural coupling*: see Sidenote 56 above.

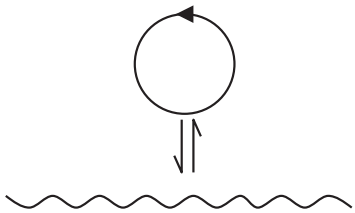

Figure 2.3: Maturana and Varela's ideogram to denote a structurally-determined organizationally-closed system. The arrow circle denotes the autonomy and self-organization of the system, the rippled line the environment, and the bi-directional half-lines the mutual perturbation — structural coupling — between the two.

cess that anticipates events and increases the space of actions in which a system can engage.

By making sense of its experience, the cognitive system is constructing a model that has some predictive value, exactly because it captures some regularity or lawfulness in its interactions. This self-generated model of the system's experience lends the system greater flexibility in how it interacts in the future. In other words, it endows the system with a larger repertoire of possible actions that allow richer interactions, increased perceptual capacity, and the possibility of constructing even better models that encapsulate knowledge with even greater predictive power. And so it goes, in a virtuous circle. Note that this sense-making and the resultant knowledge says nothing at all about what is really out there in the environment. It doesn't have to: all it has to do is make sense for the continued existence and autonomy of the cognitive system.

Sense-making is actually the source of the term enaction. In making sense of its experience, the cognitive system is somehow bringing out through its actions — enacting — what is important for the continued existence of the system. This enaction is effected by the system as it is embedded in its environment, but as an autonomous entity distinct from the environment, through an emergent process of making sense of its experience. To a large extent, this process of sense-making is exactly what we mean by cognition (in the emergent paradigm, at least).

Enaction and Development

The founders of the enactive approach, Humberto Maturana and Francisco Varela, introduced a diagrammatic way of conveying

the self-organizing and self-maintaining autonomous nature of an enactive system, perturbing and being perturbed by its environment: see Figure 2.3.[58] The arrowed circle denotes the autonomy and self-organization of the system, the rippled line the environment, and the bi-directional half-arrows the mutual perturbation.

[58] The Maturana and Varela ideograms depicting self-organizing, self-maintaining, developmental systems appear in their book *The Tree of Knowledge — The Biological Roots of Human Understanding* [14].

We remarked above that the process of sense-making forms a virtuous circle in that the self-generated model of the system's experience provides a larger repertoire of possible actions, richer interactions, increased perceptual capacity, and potentially better self-generated models, and so on. Recall also our earlier remarks that the cognitive system's knowledge is represented by the state of the system. When this state is embodied in the system's central nervous system, the system has much greater plasticity in two senses: (a) the nervous system can accommodate a much larger space of possible associations between system-environment interactions, and (b) it can accommodate a much larger space of potential actions. Consequently, the process of cognition involves the system modifying its own state, specifically its central nervous system, as it enhances its predictive capacity and its action capabilities. This is exactly what we mean by development. This generative (*i.e.* self-constructed) autonomous learning and development is one of the hallmarks of the enactive approach.

Development is the cognitive process of establishing and enlarging the possible space of mutually-consistent couplings in which a system can engage (or, perhaps more appropriately, which it can withstand without compromising its autonomy). The space of perceptual possibilities is founded not on an absolute objective environment, but on the space of possible actions that the system can engage in while still maintaining the consistency of the coupling with the environment. These environmental perturbations don't control the system since they are not components of the system (and, by definition, don't play a part in the self-organization) but they do play a part in the ontogenetic development of the system. Through this ontogenetic development, the cognitive system develops its own epistemology, *i.e.* its own system-specific history- and context-dependent knowledge of its

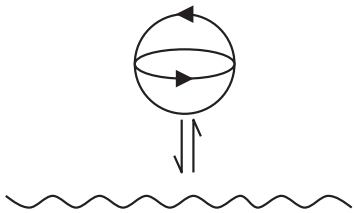

Figure 2.4: Maturana and Varela's ideogram to denote a structurally-determined organizationally-closed autonomous system *with a central nervous system*. This system is capable of development by means of self-modification of its nervous system, so that it can accommodate a much larger space of effective system action.

world, knowledge that has meaning exactly because it captures the consistency and invariance that emerges from the dynamic self-organization in the face of environmental coupling. Again, it comes down to the preservation of autonomy, but this time doing so in an ever-increasing space of autonomy-preserving couplings.

This process of development is achieved through self-modification by virtue of the presence of a central nervous system: not only does environment perturb the system (and *vice versa*) but the system also perturbs itself and the central nervous system adapts as a result. Consequently, the system can develop to accommodate a much larger space of effective system action. This is captured in a second ideogram of Maturana and Varela (see Figure 2.4) which adds a second arrow circle to the ideogram to depict the process of development through self-perturbation and self-modification. In essence, development *is* autonomous self-modification and requires the existence of a viable phylogeny, including a nervous system, and a suitable ontogeny.

Knowledge and Interaction

Let us now move on to discuss in a little more detail the nature of the knowledge that an enactive cognitive system constructs. This knowledge is built on sensorimotor associations, achieved initially by exploration of what the world offers. However, this is only the beginning. The enactive system uses the knowledge gained to form new knowledge which is then subjected to empirical validation to see whether or not it is warranted. After all, we, as enactive beings, imagine many things but not everything we imagine is valid in the sense that it is plausible or

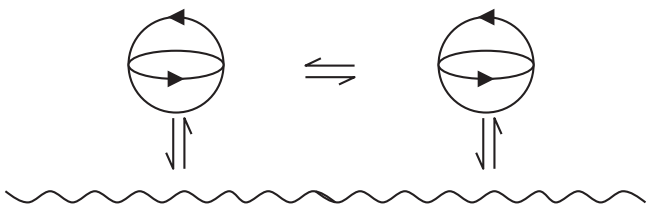

Figure 2.5: Maturana and Varela's ideogram to denote the development engendered by interaction between cognitive systems

corresponds well with reality. This brings us to one of the key issues in cognition: internal simulation, i.e. the ability to rehearse a train of imagined perceptions and actions, and assess the likely outcome in memory. This internal simulation is used to build on sensorimotor knowledge and accelerate development. Internal simulation thus provides the key characteristic of cognition: the ability to predict future events, to reconstruct or explain observed events (constructing a causal chain leading to that event), or to imagine new events.[59] Naturally, there is a need to focus on (re-)grounding predicted, explained, or imagined events in experience so that the system can *do* something new and interact with the environment in a new way. If the cognitive system wishes or needs to share this knowledge with other cognitive systems or communicate with other cognitive systems, it will only be possible if they have shared a common history of experiences and if they have a similar phylogeny and a compatible ontogeny. In essence, *the meaning of the knowledge that is shared is negotiated and agreed by consensus through interaction.*

[59] For more details on the nature of internal simulation, see [99, 100, 101]. We return to this topic in Sections 5.8 and 7.5.

When there are two or more cognitive agents involved, interaction is a shared activity in which the actions of each agent influence the actions of the others, resulting in a mutually constructed pattern of shared behaviour. Again, Humberto Maturana and Francisco Varela introduce a succinct diagrammatic way of of conveying this coupling between cognitive agent and the development it engenders: see Figure 2.5.[60] Thus, explicit meaning is not necessary for something to be communicated in an interaction, it is simply necessary that the agents are engaged in a mutual sequence of actions. Meaning emerges through shared consensual experience mediated by interaction.

[60] Such mutually-constructed patterns of complementary behaviour is also emphasized in Andy Clark's notion of joint action [102].

SUMMARY

To recap: enaction involves two complementary processes: (a) phylogenetically-dependent structural determination, *i.e.* the preservation of autonomy by a process of self-organization which determines the relevance and meaning of the system's interactions, and (b) ontogenesis, *i.e.* the increase in the system's predictive capacity and the enlargement of its action repertoire through a process of model construction by which the system develops its understanding of the world in which is it embedded. Ontogenesis results in development: the generation of new couplings effected by the self-modification of the system's own state, specifically its central nervous system. This complementarity of structural determination — phylogeny — and development — ontogeny — is crucial.

Cognition is the result of a developmental process through which the system becomes progressively more skilled and acquires the ability to understand events, contexts, and actions, initially dealing with immediate situations and increasingly acquiring a predictive or prospective capability. Prediction, or anticipation, is one of the two hallmarks of cognition, the second being the ability to learn new knowledge by making sense of its interactions with the world around it and, in the process, enlarging its repertoire of effective actions. Both anticipation and sense-making are the direct result of the developmental process. This dependency on exploration and development is one of the reasons why an artificial cognitive system requires a rich sensory-motor interface with its environment and why embodiment plays such a pivotal role.

2.3 *Hybrid Systems*

Cognitivist and emergent paradigms of cognitive science clearly have very different outlooks on cognition and they each have their own particular strengths and weaknesses. Thus, it would seem to be a good idea to combine them in a hybrid system that tries to get the benefits of both without the disadvantages of either. This is what many people try to do. Typically, hybrid systems exploit symbolic knowledge to represent the agent's

world and logical rule-based systems to reason with this knowledge to pursue tasks and achieve goals. At the same time, they typically use emergent models of perception and action to explore the world and construct this knowledge. While hybrid systems still use symbolic representations, the key idea is that they are constructed by the system itself as it interacts with and explores the world. So, instead of a designer programming in all the necessary knowledge, objects and events in the world can be represented by observed correspondences between sensed perceptions, agent actions, and sensed outcomes. Thus, just like an emergent system, a hybrid system's ability to understand the external world is dependent on its ability to flexibly interact with it. Interaction becomes an organizing mechanism that establishes a learned association between perception and action.

2.4 *A Comparison of Cognitivist and Emergent Approaches*

Although cognitivist and emergent approaches are often contrasted purely on the basis of their use of symbolic representation — or not, as the case may be — it would be a mistake to think that this is the only issue on which they differ and, equally, it would be wrong to assume that the distinction is as black-and-white as it is sometimes presented. As we have seen, symbols have a place in both paradigms; the real issue is whether these symbols denote things in the real world or simply connote them from the agent's perspective. In fact, we can contrast the cognitivist and emergent paradigms in many different ways. The following are fourteen characteristics that have proven to be useful in drawing out the finer distinctions between the two paradigms.[61]

1. Computational operation
2. Representational framework
3. Semantic grounding
4. Temporal constraints
5. Inter-agent epistemology
6. Embodiment
7. Perception

[61] These fourteen characteristics are based on the twelve proposed by the author, Giorgio Metta, and Giulio Sandini in a paper entitled "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents" [103]. They have been augmented here by adding two more: the role of cognition and the underlying philosophy. The subsequent discussion is also an extended version of the commentary in [103].

8. Action
9. Anticipation
10. Adaptation
11. Motivation
12. Autonomy
13. The role of cognition
14. Philosophical foundations

Let us look at each of these in turn. In doing so, we sometimes have to refer to concepts that are covered later in the book. The relevant chapters or sections or end-notes are indicated in the end-notes.

Computational operation: cognitivist systems use rule-based manipulation of symbol tokens, typically but not necessarily in a sequential manner. On the other hand, emergent systems exploit processes of self-organization, self-production, self-maintenance, and development, through the concurrent interaction of a network of distributed interacting components.

Representational framework: cognitivist systems use patterns of symbol tokens that denote events in the external world. These often describe how the designer sees the relationship between the representation and the real-world, the assumption being that all agents see the world in the same way. The representations of emergent systems are global system states encoded in the dynamic organization of the system's distributed network of components.

Semantic grounding: semantic representations reflect the way that a particular cognitive agent sees the world. Cognitivist systems ground symbolic representations by identifying percepts with symbols, either by design or by learned association. These representations are accessible to direct human interpretation. In contrast, emergent systems ground representations by autonomy-preserving anticipatory and adaptive skill construction. These representations only have meaning insofar as they contribute to the continued viability of the system and are inaccessible to direct human interpretation.

Temporal constraints: cognitivist systems operate atemporally in the sense that time is not an essential element of the computation. It is just a measure of how long it takes to get the result and these results won't change with the amount of time taken. However, emergent systems are entrained by external events and timing is an intrinsic aspect of how they operate. The timing of the system's behaviour relative to the world's behaviour is crucial. This also limits the speed with which they can learn and develop.

Inter-agent epistemology: for cognitivist systems, an absolute shared epistemology, i.e. framework of knowledge, between agents is guaranteed by virtue of their positivist stance on reality; that is, each agent is embedded in an environment, the structure and semantics of which are independent of the system's cognition. This contrasts strongly with emergent systems for which epistemology is the subjective agent-specific outcome of a history of shared consensual experiences among phylogentically-compatible agents. This position reflects the phenomenological stance on reality taken by emergent systems, in general, and enaction, in particular.

Embodiment: cognitivist systems do not need to be embodied, in principle, by virtue of their roots in computational functionalism which holds that cognition is independent of the physical platform in which it is implemented. Again, in contrast, emergent systems are necessarily embodied and the physical realization of the cognitive system plays a direct constitutive role in the cognitive process.

Perception: in cognitivist systems, perception provides an interface between the absolute external world and the symbolic representation of that world. The role of perception is to abstract faithful spatio-temporal representations of the external world from sensory data. In emergent systems, perception is an agent-specific interpretation of the way the environment perturbs the agent and is, at least to some extent, dependent on the embodiment of the system.

Action: in cognitivist systems, actions are causal consequences of symbolic processing of internal representations, usually carried out when pursuing some task. In emergent systems, actions

are the way the agent perturbs the environment, typically to maintain the viability of the system. In both cases, actions are directed by the goals these actions are intended to fulfil.

Anticipation: in cognitivist systems, anticipation typically takes the form of planning using some form of procedural or probabilistic reasoning with some prior model. Anticipation in the emergent paradigm takes the form of the cognitive system visiting some subset of the states in its self-constructed perception-action state space but without commiting to the associated actions.

Adaptation: for cognitivism, adaptation ususally implies the acquisition of new knowledge. In emergent systems, adaptation entails a structural alteration or re-organization to effect a new set of dynamics. Adaptation can take the form of either learning or development; Chapter 6 explains the difference.

Motivation: in cognitivist systems, motives provide the criteria which are used to select a goal and the associated actions. In emergent systems, motives encapsulate the implicit value system that modulate the system dynamics of self-maintenance and self-development, impinging on perception (through attention), action (through action selection), and adaptation (through the mechanisms that govern change), such as enlarging the space of viable interaction.

Autonomy: the cognitivist paradigm does not require the cognitive agent to be autonomous but the emergent paradigm does. This is because in the emergent paradigm cognition is the process whereby an autonomous system becomes viable and effective through a spectrum of homeostatic processes of self-regulation. Chapter 4 explains the concept of homeostasis and exands further on the different nuances of autonomy.

Role of cognition: in the cognitivist paradigm, cognition is the rational process by which goals are achieved by reasoning with symbolic knowledge representations of the world in which the agent operates. This contrasts again with the emergent paradigm, in which cognition is the dynamic process by which the system acts to maintain its identity and organizational coherence in the face of environmental perturbation. Cognition entails system development to improve its anticipatory capabilities and

The Cognitivist Paradigm vs. the Emergent Paradigm

Characteristic	Cognitivist	Emergent
Computational Operation	Syntactic manipulation of symbols	Concurrent self-organization of a network
Representational Framework	Patterns of symbol tokens	Global system states
Semantic Grounding	Percept-symbol association	Skill construction
Temporal Constraints	Atemporal	Synchronous real-time entrainment
Inter-agent epistemology	Agent-independent	Agent-dependent
Embodiment	No role implied: functionalist	Direct constitutive role: non-functionalist
Perception	Abstract symbolic representations	Perturbation by the environment
Action	Causal result of symbol manipulation	Perturbation by the system
Anticipation	Procedural or probabilistic reasoning	Traverse of perception-action state space
Adaptation	Learn new knowledge	Develop new dynamics
Motivation	Criteria for goal selection	Increase space of interaction
Autonomy	Not entailed	Cognition entails autonomy
Role of Cognition	Rational goal-achievement	Self-maintenance and self-development
Philosophical Foundation	Positivism	Phenomenology

Table 2.1: A comparison of cognitivist and emergent paradigms of cognition; refer to the text for a full explanation (adapted from [103] and extended).

extend its space of autonomy-preserving actions.

Philosophical foundations: the cognitivist paradigm is grounded in positivism, whereas the emergent paradigm is grounded in phenomenology.[62]

Table 2.1 presents a synopsis of these key issues.

2.5 *Which Paradigm Should We Choose?*

The cognitivist, emergent, and hybrid paradigms each have their proponents and their critics, their attractions and their challenges, their strong points and their weak points. However, it is crucial to appreciate that each paradigm is not equally well developed as a science and so it isn't possible to make any definitive judgement on their long-term prospects. At the same time, it is important to recognize that while the arguments in favour of emergent systems are very compelling, the current capabilities of cognitivist systems are more advanced. At present, you can do far more with a cognitivist system than an emergent one

[62] For a discussion of the positivist roots of cognitivism, see "Restoring to Cognition the Forgotten Primacy of Action, Intention and Emotion" by Walter Freeman and Rafael Núñez [36]. The paper "Enactive Artificial Intelligence: Investigating the systemic organization of life and mind" by Tom Froese and Tom Ziemke [104] discusses the phenomenological leanings of enaction. A paper by the author and Dermot Furlong [105], "Philosophical Foundations of Enactive AI," provides an overview of the philosophical traditions of AI and cognitive science.

(from the perspective of artificial cognitive systems, at any rate). With that in mind, we wrap up this chapter by looking briefly at some of their respective strengths and weaknesses, and how they might be resolved.

According to some, cognitivist systems suffer from three problems:[63] the symbol grounding problem (the need to give symbolic representations some real-world meaning; see Chapter 8, Section 8.4), the frame problem (the problem of knowing what does and does not change as a result of actions in the the world),[64] and the combinatorial explosion problem (the problem of handling the large and possibly intractable number of new relations between elements of a representation when something changes in that representation as a consequence of some action; see Sidenote 10 in this chapter). These problems are put forward as reasons why cognitivist models have difficulties in creating systems that exhibit robust sensori-motor interactions in complex, noisy, dynamic environments, and why they also have difficulties modelling the higher-order cognitive abilities such as generalization, creativity, and learning. A common criticism of cognitivist systems is that they are are poor at functioning effectively outside narrow, well-defined problem domains, typically because they depend so much on knowledge that is provided by others and that depends very often on implicit assumptions about the way things are in the world in which they are operating. However, setting aside one's scientific and philosophical convictions, this criticism of cognitivism is unduly harsh because the alternative emergent systems don't perform particularly well at present (except, perhaps, in principle).

Emergent systems should in theory be much less brittle because they emerge — and develop — through mutual specification and co-determination with the environment. However, our ability to build artificial cognitive systems based on these principles is very limited at present. To date, dynamical systems theory has provided more of a general modelling framework rather than a model of cognition and has so far been employed more as an analysis tool than as a tool for the design and synthesis of cognitive systems. The extent to which this will change, and the speed with which it will do so, is uncertain.

[63] For more details on the problems associated with cognitivism, see Wayne Christensen's and Cliff Hooker's paper "Representation and the Meaning of Life" [106].

[64] In the cognitivist paradigm, the frame problem has been expressed in slightly different but essentially equivalent terms: how can one build a program capable of inferring the effects of an action without reasoning explicitly about all its perhaps very many non-effects? [107].

Hybrid approaches appear to offer the best of both worlds: the adaptability of emergent systems (because they populate their representational frameworks through learning and experience) and also the advanced starting point of cognitivist systems (because the representational invariances and representational frameworks don't have to be learned but are designed in). However, it is unclear how well one can combine what are ultimately highly antagonistic underlying philosophies. Opinion is divided, with arguments both for and against.[65] One possible way forward is the development of a form of *dynamic computationalism* in which dynamical elements form part of an information-processing system.[66]

Clearly, there are some fundamental differences these two general paradigms — for example, the principled body-independent nature of cognitivist systems vs. the body-dependence of emergent developmental systems, and the manner in which cognitivist systems often preempt development by embedding externally-derived domain knowledge and processing structures — but the gap between the two shows some signs of narrowing. This is mainly due to (i) a fairly recent recognition on the part of proponents of the cognitivist paradigm of the important role played by action and perception in the realization of a cognitive system; (ii) a move away from the view that internal symbolic representations are the only valid form of representation; and (iii) a weakening of the dependence on embedded pre-programmed knowledge and the attendant increased use of machine learning and statistical frameworks both for tuning system parameters and the acquisition of new knowledge.

Cognitivist systems still have some way to go to address the issue of true ontogenetic development with all that it entails for autonomy, embodiment, architecture plasticity, and agent-centred construction of knowledge, mediated by exploratory and social motivations and innate value systems. Nevertheless, to some extent they are moving closer together in the ultimate dimension of the ultimate-proximate space, if not in the proximate dimension. This shift is the source of the inter-paradigm resonances we mentioned in this chapter and the previous one. However, since fundamental differences remain it is highly un-

[65] For the case in favour of hybrid systems, see e.g. [33, 49, 108]; for the case against, see e.g. [106]

[66] Apart from offering a way out of the cognitivist-emergent stand-off through *dynamic computationalism*, Andy Clark's book *Mindware – An Introduction to the Philosophy of Cognitive Science* [33] provides a good introduction to the foundational assumptions upon which both paradigms are based. Regarding dynamic computationalism, James Crutchfield, whilst agreeing that dynamics are certainly involved in cognition, argues that dynamics *per se* are "not a substitute for information processing and computation in cognitive processes" [49]. He puts forward the idea that a synthesis of the two can be developed to provide an approach that does allow dynamical state space structures to support computation and he proposes *computational mechanics* as the way to tackle this synthesis of dynamics and computation.

likely they will ever fully coalesce. This puts hybrid systems in a difficult position. For them to be a real solution to the cognitivist/emergent dilemma, they need to overcome the deep-seated differences discussed in the previous section.

Let us close this chapter with a reminder that, to date, no one has designed and implemented a complete cognitive system. So, on balance, the jury is still out on which paradigm to choose as the best model of an artificial cognitive system, especially given that both fields continue to evolve. Nonetheless, we need to move forward and make some choices if we are to realize an artifical cognitive system. For cognitive science, this process of realization begins with the specification of what is known as the cognitive architecture, the subject of the next chapter.