1 Setting the Scene

1.1 Strange Beginnings

"It isn't German philosophy." So said the roboticist Rodney Brooks in and about his own on-the-barricades paper, *Intelligence without Representation* (Brooks 1991b), a paper that, as it landed on desks around the world, caused excitement and controversy in about equal measure—and that measure was a large one. In this widely read and much debated piece, Brooks targeted some of the deepest theoretical assumptions made by mainstream work in *artificial intelligence* (henceforth *AI*). Historically and philosophically, AI can reasonably be identified as the intellectual core of cognitive science (Boden 1990a), or at least as the source of many of cognitive science's most cherished concepts and models. So, in effect, what was under fire here was not merely the prevailing approach to reproducing intelligence in artifacts, but also the dominant scientific framework for explaining mind and cognition. No wonder people were upset.[1]

So what was the content of Brooks's argument? He rightly observed that most AI research had been concerned with the production of disembodied programs capable of performing feats of reasoning and inference in abstracted subdomains of human cognition (subdomains such as natural language processing, visual scene analysis, logical problem solving, hypothesis formation from data, and so on). This overwhelming tendency to concentrate on abstracted, disembodied reasoning and inference was, according to Brooks, a serious mistake. Indeed, his argument went, by sidelining the problem of how whole, physically embodied agents, including nonhuman animals, achieve successful real-time sensorimotor control in dynamic, sometimes unforgiving environments, mainstream AI had misled us as to the true character of intelligence. So Brooks proposed a rather different goal for AI as a discipline, namely the design and

construction of complete robots that, while embedded in dynamic real-world situations, are capable of integrating perception and action in real time so as to generate fast and fluid embodied adaptive behavior (Brooks 1991b; see also Brooks 1991a).

Given the astonishing psychological and behavioral complexity of adult humans, the demand that AI turn out whole, fully integrated agents led naturally to the thought that reproducing less sophisticated styles of perceptually guided action, such as those exhibited by insects, should be the immediate aim of the field—a necessary stepping stone on the way to fancier cognitive agents.[2] But what was on the cards here was not merely a methodological reorientation. One of Brooks's key claims (and it is a claim that I shall echo, explore, and develop in my own way as the argument of this book unfolds) was that once the physical embodiment and the world-embeddedness of the intelligent agent are taken seriously, the explanatory models on offer from mainstream AI—models that trade paradigmatically in the concept of representation—begin to look decidedly uncompelling. As it happens, and despite the inflammatory title of his paper, Brooks was in fact not advocating the rejection of all representation-based control. Rather, he was objecting to a certain version of the idea (see chapter 8 below). Still, battle lines were drawn, friends fell out, and things have never been quite the same since.

With this brief introduction to the context and the content of Brooks's seminal paper under our belts, we might wonder why he felt the need even to mention German philosophy, let alone claim that he wasn't doing it. The reason is that he was endeavoring to place some intellectual distance between himself and none other than Martin Heidegger, the heavy-duty German phenomenologist whose difficult and complex book *Being and Time* (1926) is widely acknowledged as one of the most important works in twentieth century philosophy. Heidegger had been praised in some quarters in and around AI as a thinker who understood more than perhaps anyone else about what it means for an agent to be embedded in the world, and as someone whose ideas could be used to generate telling critiques of standard approaches to AI (see, e.g., Agre 1988; Dreyfus 1991, 1992; Winograd and Flores 1986). So Brooks was registering the point that although there might conceivably be a connection of some sort between his message and Heidegger's, he had no desire to forge any such connection himself. In a sense (one that will become fully clear only much later in our story), it is from Brooks's distancing comment that this book takes its cue.

1.2 Muggles Like Us

"A what?" said Harry, interested.

"A Muggle," said Hagrid. "It's what we call non-magic folk like them. An' it's your bad luck you grew up in a family o' the biggest Muggles I ever laid eyes on."

—J. K. Rowling, *Harry Potter and the Philosopher's Stone*

The range of human activities is vast. It includes getting out of bed, taking a shower, brushing your teeth, getting dressed, doing up your shoelaces, making sandwiches, unlocking and opening the front door, walking to the railway station, ordering and paying for a ticket, locating the right train, getting on that train, finding a seat, getting off at the right stop, navigating the way to your office while avoiding slow-moving people and fast-moving cars, unlocking and opening your office door, sitting down at your desk, logging in to your computer, accessing the right file, typing . . . and that's only a fraction of the activities in which I've already engaged this morning. Later on I'll be taking part in a seminar, playing squash, marking some essays, cooking dinner, and engaging in lively communicative interactions (verbal and nonverbal) not only with other people, but also with Penny and Cindy, my two pet rats. The mention of Penny and Cindy here should remind us that nonhuman animals also often exhibit a diverse range of competences in their day-to-day behavior, competences such as hunting, fighting, avoiding predators, foraging, grooming, mate finding, and biting their owner's fingers. Animal communication too is often a subtle and sophisticated business.

It seems to me that all the activities just listed, plus any others that involve behaving appropriately (e.g., adaptively, in a context-sensitive fashion) with respect to some (usually) external target state of affairs, should be counted as displays of intelligence and as outcomes of cognitive processing. In other words, I wish to join with many other thinkers in using psychological terms such as "intelligence" and "cognition" in a deliberately broad and nonanthropocentric manner. On this inclusive view, the cognitive umbrella should be opened wide enough to cover not only human-skewed examples of reflective thought and concept-based categorical perception, such as wondering what the weather's like in Paris now, mentally weighing up the pros and cons of moving to a new city, or identifying what's in the refrigerator by way of concepts such as "orange juice" and "milk carton," but also cases in which an agent coordinates sensing and movement, in real time, so as to generate fluid and flexible responses to incoming stimuli. We humans realize the latter phenomenon

when, say, playing squash or engaging in lively communicative interactions. In their own ways, other animals do the same when, say, tracking mates or escaping from predators. One might wonder why we should adopt this somewhat open-door policy as to what counts as falling within the domain of the cognitive. For me, one key justification is that it accurately reflects the diverse array of psychologically interesting behaviors and mechanisms that (owing partly to Brooks-style shifts toward embodiment and world embeddedness) are right now being investigated by researchers in cognitive science. On this point, I see no good reason for our philosophical map of the terrain to be at odds with contemporary scientific practice (for a related approach to the characterization of cognition, see van Gelder 1998b).

Three comments: first, nothing about the inclusive view prevents us from making useful theoretical distinctions between different classes of phenomena. Indeed, implicit in my setting up of that view, there is already one such crucial distinction (between what I shall later call *offline* and *online* styles of intelligence). Second, it is worth noting that using the term "cognition" so as to incorporate real-time action in the way just described does not require us to sacrifice the connection that is traditionally thought to hold between cognition and knowledge. It merely requires that the term "cognition" may be applied both to knowing *that* something is the case and to knowing *how* to perform some action. Third, in saying that a behavior or state is cognitive, one should be seen as making no a priori commitment to the specific character of the underlying mechanisms in play. In particular, cognitive behavior does not presuppose the presence of inner representations. Any such presence is something that would need to be established by further argument and evidence.

The issues just raised will reverberate throughout this book. For the present, however, let's move on. Whatever the correct account of mind, cognition, and intelligence is, it must, it seems, proceed from an intellectual marriage of philosophy and science, although exactly how the conceptual relations between these two sometimes very different partners should be understood remains a highly controversial issue. I will say more on this question in chapters 5 and 7. For the present, I want merely to impose a condition on the operation of philosophy in this arena. In the modern age, it seems to me that philosophical accounts of psychological phenomena have a duty to meet what I call the *Muggle constraint*. So what is that? In J. K. Rowling's Harry Potter books, there are two coexisting and intersecting worlds. The first is the magical realm, populated by wizards, witches, dragons, dementors, and the like. This is a realm in which, for

example, getting from A to B can be achieved by flying broomstick, flying carpet, or more dramatically, teleportation, and in which one object can be transformed into another by a transfiguration spell. The second world is the nonmagical realm, populated by Muggles—Muggles like us. Muggles, being nonmagical folk, are condemned to travel by boringly familiar (to us) planes, trains, and automobiles, and to operate without the manifest benefits of supernatural object-altering powers. Now, if you want an understanding of how Muggles work, you had better not appeal to anything magical. So one's explanation of some phenomenon meets the Muggle constraint just when it appeals only to entities, states, and processes that are wholly nonmagical in character. In other words, no spooky stuff allowed. But how are we to tell if the Muggle constraint is being met on some particular occasion? It seems clear that the most reliable check we have is to ask of some proposed explanation (philosophical or otherwise), "Is it consistent with natural science?" If the answer is "No," then that explanation fails to pass the test, and must be rejected.

It is useful to see the Muggle constraint as expressing a weak form of the philosophical position known as *naturalism*. Naturalism may be defined as the conjunction of two claims: (i) that physicalism is true, and (ii) that philosophy is continuous with natural science (see, e.g., Sterelny 1990). The stripe of one's naturalism will then be determined by how one fills in the details of (i) and (ii). In my book, physicalism amounts to the ontological claim that there is ultimately nothing but physical stuff. It does not impose the additional explanatory condition that every worldly phenomenon be ultimately explicable by physical laws. (This additional condition is imposed by, for example, Sterelny [1990], but not by, for example, Flanagan [1992].) My purely ontological species of physicalism is in tune with the fact that I read continuity with natural science in the weakest possible way, that is, as mere *consistency with* natural science, a reading that makes room, in principle, for multiple modes of explanation. Thus the view I advocate does not demand reductionist explanations of psychological phenomena, although it certainly allows for such explanations in specific cases. (For a pretty much equivalent conception of naturalism, see Elton's analysis of Dennett [Elton 2003].)

Although the kind of naturalism expressed by the Muggle constraint is somewhat restrained, it is not toothless. It still has the distinctively naturalistic consequence that (stated baldly) if philosophy and natural science clash (in the sense that philosophy demands the presence of some entity, state, or process that is judged to be inconsistent with natural science), then it is philosophy and not science that must give way.[3] Of course, good

philosophy shouldn't capitulate to bad natural science. Strictly speaking, then, the claim ought to be that if there is a clash between philosophy and some *final* natural science, then it is philosophy that should give way. Nevertheless, in practice, at any specific point in history, one has reason to be suspicious of any philosophical theory that conflicts with some seemingly well-supported scientific view, although there will often be room for negotiation. Later in this book I shall explore a more detailed account of the relations between philosophy and science that does justice to this general picture.

1.3 Three Kinds of Cognitive Science

For the naturalist of whatever strength, the field of cognitive science must occupy a pivotal place in our contemporary understanding of ourselves and other animals. So this is a book about cognitive science. More specifically, it's a book about the philosophical foundations of cognitive science, foundations that, if I am right, are entering a period of quite dramatic reconstruction.

Modern cognitive science was launched when the claim that cognitive processes are computational in character was annexed to the representational theory of mind. The latter doctrine (which goes back at least as far as Plato—the term "idea" was the precursor to the term "representation") is the view according to which mental states are, for the most part, conceived as inner representational states. Such representational states are understood as explaining the very possibility of psychologically interesting behavior. So how do we go about recognizing a mental (internal, inner) representation when we come across one? This question (which will exercise us at length in what is to come) remains far from settled. Viewed from one perspective, the situation is an embarrassing scandal. The idea that there are internal representations is a deep assumption of the most influential branches of philosophy of mind and cognitive science, and we really ought to know how to spot one. Moreover, the basic idea is surely straightforward enough, namely, that there exist, in the cognizer's mind, entities or structures (the representations) that *stand in for* (typically) external states of affairs. From another perspective, however, the shortfall in our current theoretical understanding is, perhaps, less surprising. As we shall see, representations are slippery characters that come in a veritable plethora of different forms. Moreover, although the issue of how to specify the meanings of the representations that we (allegedly) have has received library loads of philosophical attention, the question of under what circumstances it is

appropriate to engage in representational explanation at all remains curiously underexplored (Cummins 1996).

One problem that confronts the scientifically minded fan of the representational theory of mind is to explain, in a way that meets the Muggle constraint, how any purely physical system, such as a brain, might generate the kind of systematic and semantically coherent representational activity that, on this story, will constitute a mind. This is no stroll in the park. When the seventeenth-century philosopher John Locke, who was a science-friendly champion of representational thinking, wondered how our ideas of colors, smells, sounds, and so on resulted from purely material processes in our brains, he felt he had no option but to appeal to the extraordinary power of God to support the mysterious transition (Locke 1690). From a contemporary naturalistic perspective, that's simply throwing in the towel. But it does indicate one historical reason why the very idea of a representationalist cognitive science got a much needed leg up when human-built computers came on the scene, since any such computer precisely is an existence proof that a lump of the physical world can build and process representations in systematic and semantically coherent ways (cf. Fodor 1985). A computer can accomplish this impressive trick because its more familiar operations (semantically interpretable symbol manipulations according to the rules of the program) are hierarchically decomposed into much simpler operations (e.g., logical conjunction, register manipulation); and these simpler operations are implemented directly in the machine language. In effect, the machine is hardwired to carry out certain basic processes. The trick is then to set up the machine so that its physical state transitions track or mirror semantically coherent transitions (e.g., from "The televised football match starts in five minutes" to "I'll turn on the TV"), under some appropriate interpretation of the symbols concerned. Extending this picture to biological brains was irresistible. Hence we witness the rise of the computational theory of cognition, the position according to which the processes by which the intelligent agent's inner representational states are constructed, manipulated, and transformed are computational in character.

To guarantee that the computational theory of cognition has real explanatory cash value, one would at least need to say exactly what it is that makes a process a computational one. This seems as if it ought to be an easy job: look up the answer in any first-year undergraduate textbook on computer science. But there is a complication. What we require is an account of computation that is not only theoretically well grounded, but also duly sensitive to the particular way in which that term functions as

an explanatory primitive within cognitive science. Meeting this demand will take up much of chapter 4 below.

The representational theory of mind and the computational theory of cognitive processing are empirical hypotheses. However, they are empirical hypotheses whose truth has been pretty much assumed by just about everyone in cognitive science. So even though the actual details of the representations and computations concerned have remained a matter of some dispute, the overwhelming majority of cognitive scientists have at least been able to agree that if one is interested in mind, cognition, and intelligence, then one is interested in representational states and computational processes. Against this background, modern cognitive science has, for the bulk of its relatively short history, been divided into two camps—the *classical* (e.g., Fodor and Pylyshyn 1988; Newell and Simon 1976) and the *connectionist* (e.g., Rumelhart and McClelland 1986a; McClelland and Rumelhart 1986). As the argument of this book unfolds, I shall develop an analysis of the deep explanatory structures exhibited by most theorizing within these two approaches (including the accounts of representation and computation in play), and, as an important element in this analysis, I shall describe and discuss a number of models that each have produced. For the present, however, our task is less demanding: it is to orient ourselves adequately for what is to come, by way of a brief, high-level sweep over the intellectual landscape.[4]

One crude but effective way to state, in very broad terms, the difference between classicism and connectionism is to say that whereas classicism used the abstract structure of human language as a model for the nature of mind, connectionism used the abstract structure of the biological brain. Human language (on one popular account anyway) is at root a finite storehouse of essentially arbitrary atomic symbols (words) that are combined into complex expressions (phrases, sentences, and so on) according to certain formal-syntactic rules (grammar). This formal-syntactic dimension of language is placed alongside a theory of semantics according to which each atomic symbol (each word) typically receives its meaning in a causal or denotational way, and each complex expression (each phrase, sentence, etc.) receives its meaning from the meanings of its constituent atomic symbols, plus its syntactic structure (as determined by the rules of the grammar). In short, human language features a combinatorial syntax and semantics. And, for the classical cognitive scientist, so it goes for our inner psychology. That too is based on a finite storehouse of essentially arbitrary atomic symbols. In this case, however, the symbols are our inner representations, conceived presemantically. In accordance with certain formal-

syntactic rules, these symbols may be combined into complex expressions. These expressions are our thoughts, also conceived presemantically. The meaning of each atomic symbol (each representation) is once again fixed in a causal or denotational way; and the meaning of each complex expression (each thought) is once again generated from the meanings of its constituent atomic symbols, plus its syntactic structure. In short thinking, like language, features a combinatorial syntax and semantics. Thus, Fodor famously speaks of our inner psychological system as a *language of thought* (1975).[5]

So how has classicism fared empirically? Here is a summary of what (I believe) the history books will say about classical AI, the intellectual core of the approach. The tools of classical AI are undoubtedly powerful weapons when one's target is, for instance, logic-based reasoning or problem solving in highly structured search spaces (for discussions of many key examples, see, e.g., Boden 1977). However, these heady heights of cognitive achievement are, in truth, psychological arenas in which most humans perform rather badly, and in which most other animals typically don't perform at all. This should immediately make us wary of any claim that classicism provides a general model for natural intelligence. Moreover, the word on the cognitive-scientific street (at least in the neighborhood where I live) is that classical systems have, by and large, failed to capture, in anything like a compelling way, specific styles of thinking at which most humans naturally excel. These include the flexible ability to generalize to novel cases on the basis of past experience, and the capacity to reason successfully (or, at least, sensibly) given incomplete or corrupt data. Attempts by classical AI to reproduce these styles of thinking have either looked suspiciously narrow in their domain of application, or met with a performance-damaging explosion in computational costs. In other words, classical systems have often seemed to be rigid where we are fluid, and brittle where we are robust. Into this cognitive breach stepped connectionism.

Roughly speaking, the term "connectionism" picks out research on a class of systems in which a (typically) large number of interconnected units process information in parallel.[6] In as much as the brain too is made up of a large number of interconnected units (neurons) that process information in parallel, connectionist networks are "neurally inspired," although usually at a massive level of abstraction. (This is an issue to which we shall return.) Each unit in a connectionist network has an activation level regulated by the activation levels of the other units to which it is connected, and, standardly, the effect of one unit on another is either

positive (if the connection is excitatory) or negative (if the connection is inhibitory). The strengths of these connections are known as the network's weights, and it is common to think of the network's "knowledge" as being stored in its set of weights. The values of these weights are (in most networks) modifiable, so, given some initial configuration, changes to the weights can be made that improve the performance of the network over time. In other words, within all sorts of limits imposed by the way the input is encoded, the specific structure of the network, and the weight-adjustment algorithm, the network may learn to carry out some desired input–output mapping. As we shall see in more detail later, most connectionist networks also exploit a distinctive kind of representation, so-called *distributed representation*, according to which a representation is conceived as a pattern of activation spread out across a group of processing units.

In the interests of historical accuracy, it is important to stress that what we now call connectionism can be traced back, in many ways, to the seminal work of McCulloch and Pitts (1943), work that set the stage for both classical AI and connectionism (Boden 1991). For years, work within the first wave of connectionism moved ever onward, although in a more reserved way than its then media-grabbing classical cousin (for important examples of early connectionism, see Hebb 1949; Rosenblatt 1962). There were some troubled times in the 1970s, following an influential critique by Minsky and Papert (1969). However, armed with some new tools (multilayered networks and the back-propagation learning rule), tools that were immune to the Minsky and Papert criticisms, connectionism bounced back, and, in the 1980s, two volumes of new studies awakened mass interest in the field (Rumelhart and McClelland 1986a; McClelland and Rumelhart 1986). One (perhaps, the) major reason why connectionism gripped the cognitive-scientific imagination of the 1980s was that connectionist networks seemed, to many cognitive theorists, to demonstrate precisely the sorts of intelligence-related capacities that were often missing from, or difficult to achieve in, classical architectures, capacities such as flexible generalization and the graceful degradation of performance in the face of restricted damage or noisy or inaccurate input information. As we noted above, such capacities appear to underlie the distinctive cognitive profile of biological thinkers. However, it wasn't merely the thought that connectionist networks exhibited these exciting properties that inspired devotion; it was the extra thought that they exhibited them as "natural" by-products of the basic processing architecture and form of representation that characterized the connectionist approach. In other words, what

the classicist had to pay dearly for—in a currency of computational time, effort, and complexity—the connectionist seemed to get for free.

As one might expect, classicism fought back. To recall just two famous attempts to derail the connectionist bandwagon, Fodor and Pylyshyn (1988) argued that classicism can, but connectionism cannot, satisfactorily account for the nonnegotiable psychological property of systematicity, and Pinker and Prince (1988) published a stinging attack on one of connectionism's apparently big successes, namely Rumelhart and McClelland's past-tense acquisition network (Rumelhart and McClelland 1986b). But as interesting and important as these disputes are, they need not detain us here.[7] Indeed, in this book, it will not be the much publicized differences between classicism and connectionism that will come to exercise our attention. Rather, it will be certain deep, but typically overlooked, similarities. For although connectionism certainly represents an advance over classicism along certain important dimensions (e.g., biological sensitivity, adaptive flexibility), the potentially revolutionary contribution of connectionist-style thinking has typically been blunted by the fact that, at a more fundamental level of analysis than that of, say, combinatorially structured versus distributed representations, such thinking has left all the really deep explanatory principles adopted by classicism pretty much intact. So if we are searching for a sort of Kuhnian revolution in cognitive science, the second dawn of connectionism is not the place to look.

The stage is now set for our *third* kind of cognitive science. Following others, I shall call this new kid on the intellectual block *embodied–embedded cognitive science*. In its raw form, the embodied–embedded approach revolves around the thought that cognitive science needs to put cognition back in the brain, the brain back in the body, and the body back in the world. This is all very laudable as a general statement of intent, but it certainly does not constitute a specification of a research program, since it allows for wide-ranging interpretations of what exactly might be required of its adherents by way of theoretical commitments. So I intend to focus on, and stipulatively reserve the term "embodied–embedded cognitive science" for, what I take to be a central and distinctive theoretical tendency within the more nebulous movement. Conceived this way, the embodied–embedded approach is the offspring of four parallel claims: (1) that online intelligence (see below) is the primary kind of intelligence; (2) that online intelligence is typically generated through complex causal interactions in an extended brain–body–environment system; (3) that cognitive science should increase its level of biological sensitivity; and (4) that cognitive science should adopt a dynamical systems perspective. For now let's

take a quick look at these claims, with the promise that each will be explored in proper detail, with abundant references, in due course.

1 The primacy of online intelligence Here is a compelling, evolutionarily inspired thought: biological brains are, *first and foremost*, systems that have been designed for controlling action (see, e.g., Wheeler 1994; Clark 1997a; Wheeler and Clark 1999). If this is right, then the primary expression of biological intelligence, even in humans, consists not in doing math or logic, but in the capacity to exhibit what I shall call *online intelligence* (Wheeler and Clark 1999). We met this phenomenon earlier. A creature displays online intelligence just when it produces a suite of fluid and flexible real-time adaptive responses to incoming sensory stimuli. On this view, the natural home of biological intelligence turns out to be John Haugeland's fridge: "[W]hat's noteworthy about our refrigerator aptitudes is not just, or even mainly, that we can visually identify what's there, but rather the fact that we can, easily and reliably, reach around the milk and over the baked beans to lift out the orange juice—without spilling any of them" (Haugeland 1995/1998, p. 221). Other paradigmatic demonstrations of on-line intelligence, cases that have already featured in our story, include navigating a path through a dynamic world without bumping into things, escaping from a predator, and playing squash. The general distinction here is with *offline intelligence*, such as (again, to use previous examples) wondering what the weather's like in Paris now, or mentally weighing up the pros and cons of moving to a new city. Of course, as soon as one reflects on the space of possibilities before us, it becomes obvious there will be all sorts of hard-to-settle intermediate cases. But the recognition of this complexity doesn't, in and of itself, undermine the thought that there will be cognitive achievements that fall robustly into one category or the other; so the online–offline distinction remains, I think, clear enough and illuminating.

2 Online intelligence is generated through complex causal interactions in an extended brain–body–environment system Recent work in, for example, neuroscience, robotics, developmental psychology, and philosophy suggests that on-line intelligent action is grounded not in the activity of neural states and processes alone, but rather in complex causal interactions involving not only neural factors, but also additional factors located in the nonneural body and the environment. Given the predominant role that the brain is traditionally thought to play here, one might say that evolution, in the interests of adaptive efficiency, has been discovered to out-

source a certain amount of cognitive intelligence to the nonneural body and the environment. In chapters 8 and 9 we shall explicate this externalistic restructuring of the cognitive world—with its attendant (typically mild, but sometimes radical) downsizing of the contribution of the brain—in terms of what Andy Clark and I have called *nontrivial causal spread* (Wheeler and Clark 1999).

3 An increased level of biological sensitivity Humans and other animals are biological systems. This is true, but what hangs on the fact? There is a strong tradition, in cognitive psychology and in philosophy of mind, according to which the details of the biological agent's biology are largely unimportant for distinctively psychological theorizing, entering the picture only as "implementation details" or as "contingent historical particulars." This venerable tradition is part and parcel of positions in which a physicalist ontology is allied with the claim that psychology requires "its own" explanatory language, one that is distinct from that of, say, neurobiology or biochemistry. Examples include traditional forms of functionalism and their offshoot, the orthodox computational theory of cognition. (More on this in chapters 2 and 3. See also Wheeler 1997.) To the fan of embodied-embedded cognitive science, this sidelining of biology is simply indefensible. Humans and animals are biological systems—*and that matters for cognitive science*. What is needed, therefore, is an increase in the biological sensitivity of our explanatory models. This can happen along a number of different dimensions. For example, one might argue that although mainstream connectionist networks represent an important step in the direction of neurally inspired processing architectures, such systems barely scratch the surface of the complex dynamical structures that, neuroscience increasingly reports, are present in real nervous systems. Alternatively, but harmoniously, one might build on the point that biology isn't exhausted by neurobiology. Since humans and animals are products of evolution, cognitive science ought also to be constrained by our scientific understanding of the general features exhibited by evolutionary systems (selection, adaptation, self-organization during morphogenesis, and so on). It is, of course, quite common to find Darwinian selection being wheeled in by naturalistic philosophers as a way of fixing representational content (see chapter 3); but typically that's about as far along this second dimension of biological sensitivity as cognitive theorists have managed to venture.[8]

4 A dynamical systems perspective As we have seen, the computational theory of cognition maintains that all cognitive processes are

computational processes. Our fourth and final embodied–embedded claim amounts to a rejection of this idea, in favor of the thought that cognitive processing is fundamentally a matter of state space evolution in certain kinds of dynamical system. In some ways this transition from the language of computation to the language of dynamics is the most controversial of the four claims that I have chosen to highlight. However, as I shall argue, once we nail down what a dynamical systems perspective ought to look like, and once claims 1–3 (above) are both developed systematically and understood in detail, there are good reasons to think that natural cognitive systems are, and should be explained as, dynamical systems.

Although embodied–embedded cognitive science is already open for business, it is, like many start-ups, a delicate success. Indeed, the fundamental conceptual profile of the research (just how different is it really?), and, relatedly, its scientific and philosophical implications (where does cognitive science go from here?), remain distinctly unclear. It is with respect to these two points that, in my view, generically Heideggerian thinking can make (indeed, in a sense to be determined, has already made) a crucial contribution. So the next step in the reconstruction of the cognitive world is, I suggest, a Heideggerian one. In fact, if I am right, Brooks was doing German philosophy after all. It is time for us to plot a course.

1.4 Where We Are Going

"Appearances can be deceptive" is a saying we teach to our children, in an attempt to prevent those gullible young minds from taking everything at face value. But it is a warning that is as useful to the student of cognitive theory as it is to the student of life. Here's why. Despite appearances, most research in cognitive science, that bastion of contemporary thought, is recognizably *Cartesian* in character. By this I mean that most cognitive-scientific theorizing bears the discernible stamp of a framework for psychological explanation developed by Descartes, that great philosopher and scientist of the seventeenth century. The Cartesian-ness to which I am referring here is elusive. Indeed, it is typically invisible to the external observer and even to the majority of working cognitive scientists, for it is buried away in the commitments, concepts, and explanatory principles that constitute the deep assumptions of the field. Nevertheless, in spite of the concealed nature of this Cartesian presence, its influence has been identified and described by (among others) Bickhard and Terveen (1996), Dennett (1991), Dreyfus (1991), Dreyfus and Dreyfus (1988), Fodor (1983), Harvey (1992), Haugeland (1995/1998), Lemmen (1998), Shanon (1993),

van Gelder (1992), Varela, Thompson, and Rosch (1991), and Wheeler (1995, 1996a, 1996b, 1997). For anyone interested in the philosophical foundations of cognitive science, this would be a good place to start.[9]

Given that so much has been said on the topic already, one might wonder whether anything remains to be done to establish that there is a Cartesian presence in cognitive science. The answer, I think, is "Yes." It seems to me that many (although not all) of the supporting analyses in the aforementioned literature turn on decontextualized, isolated features of Descartes's theory of mind, or appeal (explicitly or implicitly) to the sort of received interpretations of Descartes's views that, when examined closely, reveal themselves to be caricatures of the position that Descartes himself actually occupied. The appeal to such partial or potentially distorting evidence surely dilutes the plausibility of the analyses in question. That is why there is still a substantive contribution to be made. It is this observation that sets the agenda for the opening phase of our investigation proper.[10]

In what follows I shall use the term *orthodox cognitive science* to name the style of research that might be identified informally as "most cognitive science as we know it." My intention in using this term is to pick out not only classical cognitive science, but also most of the work carried out under the banner of connectionism. (Some decidedly *unorthodox* connectionist networks will be discussed in later chapters.) By orthodox cognitive science, then, I mean the first two kinds of cognitive science identified in the previous section. So here's the claim with which we shall begin our examination of the philosophical foundations of cognitive science: orthodox cognitive science is Cartesian in character. In order to defend this claim in a manner resistant to the sorts of worries (about caricatures and distortions) that I raised above, I begin (chapter 2) by extracting, from Descartes's philosophical and scientific writings, an integrated conceptual and explanatory framework for scientifically explaining mind, cognition, and intelligence. This framework, that I call *Cartesian psychology*, is defined by eight explanatory principles that capture the ways in which various crucial factors are located and played out in Descartes's own account of mind. These factors are the subject–object dichotomy, representations, general-purpose reason, the character of perception, the organizational structure of perceptually guided intelligent action, the body, the environment, and temporality.

Having spelled out Cartesian psychology as a well-supported interpretation of the historical Descartes, I use it to underwrite a case for the target claim that orthodox cognitive science is Cartesian in character. To do this,

I argue (during chapter 3 and part of chapter 4) that each of the eight principles of Cartesian psychology is either (i) an assumption made by orthodox cognitive science before its empirical work begins, or (ii) an essential feature of key examples of that empirical work. One particular aspect of this analysis is worth highlighting here. Understanding the temporal character of orthodox cognitive-scientific explanation requires us to get clear about how the concept of computation is played out within the genre. To achieve this I lay out and defend a version of the view that computational systems are properly conceived as a subset of dynamical systems. It is from this vantage point that the temporal character of orthodox cognitive science becomes visible, and from which I work out what I think is the most plausible version of the idea that dynamical systems theory may provide the primary explanatory language for cognitive science (chapter 4).

It is at this point in the proceedings that the second phase of our investigation begins. It seems that we are in the midst of an anti-Cartesian turn in cognitive science. The first hints of this nascent transformation in the field are to be found in certain key examples of dynamical systems research (discussed in chapter 4). However, these are scattered points of pressure on the Cartesian hegemony. Going beyond Cartesianism in cognitive science requires a more fundamental reconstruction in the philosophical foundations of the discipline. It is in this context that I turn to Heidegger's radically non-Cartesian analysis of everyday cognition, and argue that the oppositions between it and the corresponding Cartesian analysis can help us to articulate the philosophical foundations of a genuinely non-Cartesian cognitive science.[11]

Crucially, the pivotal use that I make of Heidegger's work should not be heard as high-handed preaching on the part of a philosopher, telling science how it ought to be done. This is because in my view, embodied-embedded cognitive science has *already*, although in a largely implicit way, taken up a conceptual profile that reflects a distinctively Heideggerian approach to psychological phenomena. The philosophical task before us now (one that, as we shall see, is itself Heideggerian in character) is to articulate, amplify, and clarify that profile. If my analysis here is sound, it is closing time for those Euroskeptics in mainstream philosophy of cognitive science who think that continental philosophy has nothing of interest, certainly not of a positive nature, to say to cognitive science. But it is also closing time for those continental philosophers who claim that thinkers such as Heidegger have, in effect, presented arguments against the very idea of a cognitive science, concluding that any science of cognition must

be, in some way, radically misguided, necessarily incomplete, or even simply impossible (usually, the story goes, because the Muggle constraint cannot be met for mind and cognition).

At the outset, let me state for the record that I am of course not the first person to exploit Heidegger's philosophy to positive ends in cognitive science. For example, Dreyfus has occasionally used Heideggerian ideas to generate suggestions about how cognitive science might develop (see, e.g., Dreyfus 1992, introduction), although in truth it must be said that the more famous critical dimension of his ongoing engagement with the field (see below and chapter 7) has always dominated his writings. Heidegger's influence is also manifest in Agre's pioneering attempt to fuse a phenomenology of everyday behavior with an approach to AI that takes seriously the dynamics of agent–environment interactions (see, e.g., Agre 1988), and in Winograd and Flores's influential theory of human–computer interaction (1986). In addition, some writers have made very occasional comments to the effect that certain sorts of mechanisms or approaches may be suggestive of, or at least compatible with, a Heideggerian view. Here one might note the odd remark about connectionism by Dreyfus and Dreyfus (1988), about self-organizing systems by Varela, Thompson, and Rosch (1991), and about dynamical systems by van Gelder (1992). More generally, if we open our eyes a little wider, there are a number of instances in which theorists in and around cognitive science have allowed continental philosophy to shape their theorizing. For example, Dreyfus (2002), Lemmen (1998), Kelly (2000), and Hilditch (1995) all find lessons in the work of Merleau-Ponty, and Varela, Thompson, and Rosch (1991) and Tani (2002) have a similar experience with the work of Husserl. Finally, Haugeland's philosophical account of the essentially embodied and embedded nature of mind, an account that contains a discussion of Brooks-style robotics, also exhibits the marks of continental exposure (Haugeland 1995/1998).

It is not part of my project here to explore, in any comprehensive way, these prior episodes in which cognitive theorists have drawn positively on Heideggerian or, more generally, continental insights, in the vicinity of cognitive science. That would be a different book. As one would expect, there are points of contact between my position and these outbreaks of cognitive Europhilia, and there are points of divergence. Some of these will be explored in what follows, as demanded by the unfolding of my argument. My own engagement with Heideggerian philosophy is, I think, distinctive in a variety of critically important ways, not least because my interpretation of Heidegger contains certain nonstandard aspects,

especially in my accounts of (i) Heidegger on science in general, and (ii) Heidegger on the sciences of human agency. Moreover, as we shall see, I believe that the connections between Heideggerian philosophy (as I understand it) and embodied–embedded cognitive science emerge most clearly when one attempts to solve a number of conceptual problems posed by that new form of cognitive science.

I begin the Heideggerian phase of our investigation (in chapters 5 and 6) by developing and defending an interpretation of certain key elements from division 1 of Heidegger's *Being and Time* (1926). (Division 2 of this imposing text deals with a range of, as one might say, "spiritual" concerns, such as anxiety, guilt, and death. These are far beyond the scope of the present work.) My exegetical strategy will not be to spell out Heidegger's framework as a set of explicitly stated explanatory principles; that is, I shall not present that framework in a manner that structurally mirrors Cartesian psychology. In my view, the relationships in play are just too subtle for that tactic to be useful. However, the systematic differences between the two perspectives will be brought out as my interpretation of Heidegger unfolds. That interpretation revolves around three (what I call) *modes of encounter*. For Heidegger, these characterize the different ways in which agents may engage with entities, and he identifies them in terms of the crucial, famous, and much-discussed phenomenological categories of the *ready-to-hand* and the *present-at-hand*, plus the less famous and regularly ignored, but equally crucial, phenomenological category of the *un-ready-to-hand*. As I shall present the view, these three modes of encounter provide the backbone of Heidegger's approach to mind, cognition, and intelligence. Moreover, they propel us toward Heidegger's account of the agent as being essentially and in the first instance world embedded, where a world is to be understood as a holistic network of contexts in which things show up as meaningful. The radically anti-Cartesian character of Heidegger's thought emerges from this analysis.

As a rule I recommend steering clear of life's converts, followers, and uncritical devotees, and I certainly don't think one should approach Heidegger in an atmosphere of hands-off reverence. Indeed, at the end of chapter 6, I argue that the "pure" Heideggerian story faces severe difficulties over the status of animals, although the situation can be rescued with a little naturalistic tinkering. More significantly, at the beginning of chapter 7, I argue that Heidegger's head-on philosophical critique of Descartes fails dismally to show that a Cartesian metaphysics must be false. This means, of course, that Heidegger's own official response to Descartes cannot be used *directly* to undermine Cartesian cognitive science. Having

pinpointed this problem for the Heideggerian critic of orthodox cognitive science, I turn, for a potential solution, to what is arguably the frontline example of a largely Heideggerian approach in this area, namely Dreyfus's critique of orthodox AI (see, e.g., Dreyfus and Dreyfus 1988; Dreyfus 1991, 1992; Wrathall and Malpas 2000). Any fresh attempt to apply Heideggerian ideas to cognitive science has a duty to locate itself in relation to Dreyfus's work. In chapter 7 I do just that. I explain what I believe is going on in Dreyfus's (in my view) all-too-often misunderstood arguments, and I present reasons for thinking that those arguments, even when correctly understood, still fall short of their target.

To find a way out of this impasse, I suggest a shift in emphasis for the fan of Heideggerian thinking. This is a shift away from critique and toward the claim (previewed above) that there is evidence of a newly emerging paradigm in cognitive science, one that is not only generating compelling empirical work but is also usefully interpreted as having a distinctively Heideggerian conceptual profile. Of course, the overwhelming bulk of this work, in embodied–embedded cognitive science, is produced not as part of an *explicitly* Heideggerian research program, but rather under direct pressure to meet certain pressing explanatory challenges in the empirical arena. So, as in fact the Heideggerian would predict (see chapters 5 and 7), there is a philosophical job to be done here in identifying, amplifying, and clarifying the underlying philosophical foundations of the work. That's the job I take on next. In chapters 8, 9, and 10, I explore the underlying conceptual shape of embodied–embedded cognitive science. Much of the discussion focuses on recent research in AI-oriented robotics, especially evolutionary robotics. Among other things, we will find ourselves propelled headlong into a complex debate over the nature and status of representation as an explanatory primitive in cognitive science, and forced to take a stand on the equally difficult issue of to what extent cognition really is computation. Along the way we shall find abundant evidence that the conceptual profile of embodied–embedded cognitive science is plausibly and illuminatingly understood as being Heideggerian (and thus, non-Cartesian) in form.

Before we finally get down to business, a note about method: throughout my engagement with Heideggerian ideas, I have tried to work at an interface where analytic philosophy, continental philosophy, and cognitive science may meet in a mutually profitable way. This is a perilous task, so let me issue a couple of advance warnings. My aim is to present Heidegger's ideas in a form accessible and comprehensible to someone who has no previous knowledge of contemporary continental philosophy.

I trust that this restaging of Heidegger's thought does not distort its content, but I apologize in advance to any Heidegger scholars out there who conclude otherwise. From the other side of the tracks, some empirically minded cognitive scientists might find themselves being put off by some of the metaphysical issues that, on occasion, are discussed. I ask such readers to stay with me. Although I have constrained my coverage of *Being and Time* to target the crucial passages and ideas that matter most to our larger project, it seems to me that any understanding of those passages and ideas would at best be incomplete without some appreciation of the wider philosophical questions raised by the work. Having said that, I certainly don't want to appear apologetic for choosing to engage in a quite detailed exegetical treatment of Heidegger, or indeed of Descartes. As Marx and Engels famously and astutely pointed out, those who don't learn from history are doomed to repeat its mistakes. Learning from history—in this case from the work of dead philosophers—requires a proper appreciation of what exactly was done there.

So that, then, is where we are headed. We are ready to embark on a long journey, one that begins back in the seventeenth century with a landmark event in our philosophical and scientific pursuit of the mind—the birth of Cartesian dualism.