tivity (see text and Box 5.4) depends on the flow of actual environmental triggers, e.g., encountering a table and then switching into can-seeking mode. A more advanced twist on this strategy occurs when we actively structure our environments in ways designed to off-load control and action selection (as when we place reminders in select locations, or when we lay out the parts of a model plane in the correct order for assembly). This devolution of control to the local environment is a topic to which we shall return.

In sum, it remains unclear how best to press coordinated behavior from a "bag of tricks" style of cognitive organization. But preserving the gains and advantages that such a style of organization offers precludes the use of a central executive and a heavy duty, message-passing code. Instead, appropriate coordination must somehow emerge from the use of simpler forms of internal routing and signaling and (perhaps) from the structure of the environment itself.

5.3 Suggested Readings

For a *general introduction to the contemporary neuroscience of perception and action*, try M. Jeannerod *The Cognitive Neuroscience of Action* (Oxford, England: Blackwell, 1997). This covers work on reaching and grasping, and is an especially clear introduction to the interface between psychology and neuroscience. See also A. D. Milner and M. Goodale, *The Visual Brain in Action* (Oxford, England: Oxford University Press, 1995) for a clear but provocative story about vision and action. The review article by T. Decety and T. Giezes, "Neural mechanisms subserving the perception of human actions." *Trends in Cognitive Sciences*, 3(5), 172–178, 1999, is also a useful resource.

For a philosophically, computationally, and neuroscientifically informed discussion of the *questions about levels of analysis and explanation*, see P. S. Churchland and T. J. Sejnowski, *The Computational Brain* (Cambridge, MA: MIT Press, 1992), a dense but accessible treatment of contemporary computational neuroscience, with especially useful discussions of the issues concerning levels of analysis and levels of description, and P.S. Churchland, *Neurophilosophy* (Cambridge, MA: MIT Press, 1986), which also contains a useful and accessible primer on basic neuroscience and neuroanatomy.

The work on *interactive vision and change-blindness* is nicely described in P. S. Churchland, V. S. Ramachandran, and T. Sejnowski, "A critique of pure vision." In C. Koch and T. Davis (eds.), *Large-Scale Neuronal Theories of the Brain* (Cambridge, MA: MIT Press, 1994, pp. 23–60). See also the review articles by D. Simons and D. Levin, "Change blindness." *Trends in Cognitive Sciences*, 1, 261–267, 1997; and D. Ballard, "Animate vision." *Artificial Intelligence*, 48, 57–86, 1991. The latter is just about the perfect introduction to computational work on real-world, real-time vision.

For a nice review of the work on *real-world robotics*, see J. Dean, "Animats and what they can tell us." *Trends in Cognitive Science*, 2(2), 60–67, 1998. For a longer treatment, integrating themes in philosophy, robotics, and neuroscience, see A. Clark, *Being There: Putting Brain, Body and World Together Again* (Cambridge, MA: MIT Press, 1997).

And finally, the various essays in *Daedalus*, 127(2), 1998 (special issue on the brain) range over a variety of topics relating to the *current state of mind/brain research* and include useful general introductions to work on vision, sleep, consciousness, motor action, and lots more.

6 ROBOTS AND ARTIFICIAL LIFE

6.1 Sketches

In Chapter 5, we began to encounter our first examples of work in robotics—work that falls broadly within the field that has come to be known as artificial life. This work is characterized by three distinct, but interrelated themes:

1. An interest in complete but low-level systems (whole, relatively autonomous artificial organisms that must sense and act in realistic environments).
2. Recognition of the complex contributions of body, action, and environmental context to adaptive behavior.
3. Special attention to issues concerning emergence and collective effects.

In this sketch, I introduce these topics using two concrete examples: cricket phonotaxis and termite nest building.

The interest in complete but low-level systems is most famously illustrated by Rodney Brooks' work on mobile robots (mobots), and by robots such as Herbert, whom we already met in Chapter 5. But the idea of building such creatures goes back at least to the early 1950s when W. Grey Walter created a pair of cybernetic turtles named Elmer and Elsie. In 1978, the philosopher Daniel Dennett published a short piece called "Why Not the Whole Iguana" that likewise argued in favor of studying whole simple systems displaying integrated action, sensing, and planning routines (contrast this with the stress on isolated aspects of advanced cognition such as chess playing, story understanding, and medical diagnosis displayed by classical artificial intelligence—see Chapters 1 and 2). One powerful reason for such a switch, as we have noted before, is that biological solutions to these more advanced

problems may well be profoundly shaped by preexisting solutions to more basic problems of locomotion, sensing, and action selection. Moreover, the idea that it is fruitful to separate basic functions such as vision, planning, and action taking is itself open to doubt: these functions (as we also saw in the previous chapter) look to be quite intimately interrelated in naturally intelligent systems. As an example of the whole system approach in action, let us consider (partly by way of variety—Brooks' robots are a nice, but overused, example) Barbara Webb's recent work on cricket phonotaxis.

Female crickets are able to identify a male of the same species by his song, and are able to use the detected song as a signal allowing the female to find the male. The term "phonotaxis" names this capacity to detect and reliably move toward a specific sound or signal. The male cricket's song is produced by rubbing its wings together and consists in a carrier frequency (a simple tone) and a rhythm (the way the tone is broadcast in discrete bursts, separated by silence, as the wings open and close). The repetition rate of the bursts (or "syllables") is an important indicator of species, whereas the loudness of the song may help to pick out the most desirable male from a group. The female cricket must thus

1. hear and identify the song of her own species,
2. localize the source of the song, and
3. locomote toward it.

This way of describing the problem may, however, be misleading, and for some increasingly familiar reasons. The hear–localize–locomote routine constitutes a neat task decomposition and identifies a sequence of subtasks that would plainly solve the problem. But it is again hostage to a nonbiological vision of single functionality and sequential flow. Webb, heavily inspired by what is known of real cricket anatomy and neurophysiology, describes the following alternative scenario, which was successfully implemented in a robot cricket.

The cricket's ears are on its forelegs and are joined by an inner tracheal tube that also opens to the world at two other points (called spiracles) on the body (see Figure 6.1). External sounds thus arrive at each ear via two routes: the direct external route (sound source to ear) and an indirect internal route (via the other ear, spiracles, and tracheal tube). The time taken to travel through the tube alters the phase of the "inner route" sound relative to the "outer route" sound on the side (ear) nearest to the sound source (since sound arriving at the ear *closer* to the external source will have traveled a much shorter distance than sound arriving at the same ear via the inner route). As a result, simple neural or electronic circuitry can be used to sum the out-of-phase sound waves, yielding a vibration of greater amplitude (heard as a louder sound) at the ear nearest the sound source. Orientation in the direction of the male is directly controlled by this effect. Each of the two in-

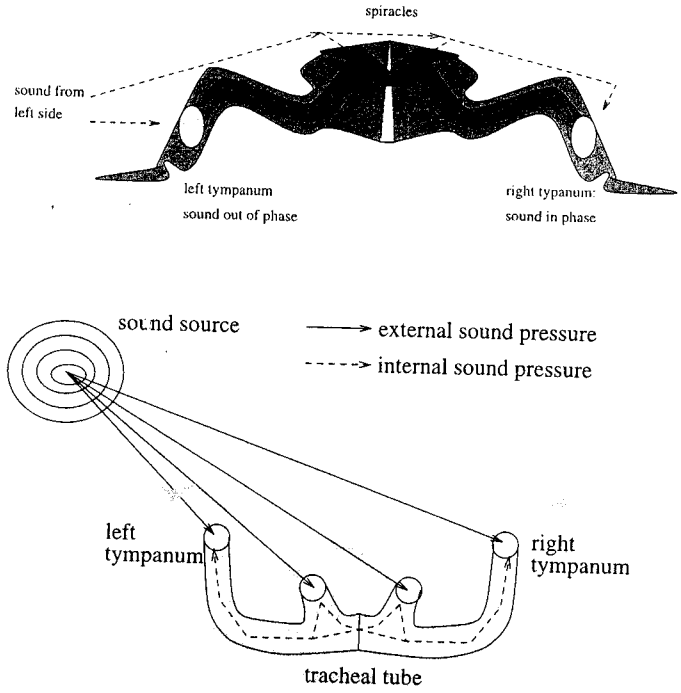

Figure 6.1 Cricket phonotaxis. The cricket's body channels sounds through an internal tracheal tube that connects the insect's ears to each other and to two openings, called spiracles, at the top of the body. Each ear is near a knee on a front leg. Because of the tube, sound reaches each ear in two ways: directly from the sound source, and indirectly, via the tube, from the spiracles and other ear. At the ear closer to the sound source, the sound that has traveled directly to the outside of the eardrum has traveled a shorter distance than the sound arriving through the tube at the inside of the eardrum. Because of this difference in distance, the sound arriving at one side of this eardrum is out of phase with respect to the sound arriving at the other side. At this eardrum the out-of-phase waves are summed, causing a vibration of greater amplitude, sensed as a louder sound. (Pictures courtesy of Barbara Webb.)

terneurons (one connected to each ear) fires when the input (vibration amplitude) reaches a critical level. But the one connected to the ear nearest the sound source will reach this threshold first. The cricket's nervous system is set up so as to reliably turn the cricket to the side on which the dedicated interneuron fires first. As a result, the insect responds, at the start of each burst of male song, by turning and moving in the direction of the sound (hence the importance of syllable *repetition* in attracting a mate). Notice, finally, that in this story the particularities of the tracheal tube are especially crucial to success. As Webb puts it:

> One of the fundamental principles of this system is that the cricket's tracheal tube transmits sounds of the desired calling song frequency, and the phase shifts in this transmission are suited to that particular wavelength. (Webb, 1996, p. 64)

The result is that the robot cricket (see Figure 6.2) does not possess any *general* mechanism for identifying the direction of sounds, nor does it need to *actively* discriminate the song of its own species from other songs. For other sounds are structurally incapable of generating the directional response. The robot cricket does not succeed by tailoring general purpose capacities (such as pattern recognition and sound localization) to the special case of mate detection: instead, it exploits *highly efficient but (indeed, because) special-purpose strategies.* It does not build a rich model of its environment and then apply some logicodeductive inference system to generate action plans. It does not even possess a central sensory information store capable of integrating multimodel inputs.

As a result, it is not at all obvious that the robot cricket uses anything worth calling internal representations. Various inner states do correspond to salient outer parameters, and certain inner variables to motor outputs. But Webb argues:

> It is not necessary to use this symbolic interpretation to explain how the system functions: the variables serve a mechanical function in connecting sensors to motors, a role epistemologically comparable to the function of the gears connecting the motors to the wheels. (Webb, 1994, p. 53)

In fact, understanding the behavior of the robot cricket requires attention to details that (from the standpoint of classical cognitive science) look much more like descriptions of implementation and environmental context than substantive features of an intelligent, inner control system. Key factors include, as noted, the fixed-length trachea and the discontinuity and repetition of the male song. The explanation of real-life cricket phonotaxis, if the Webb model is anywhere near correct,[1] involves a complex interaction among brain, body, and world, with no single component bearing the brunt of the problem-solving burden.

[1]The issue of biological plausibility has been addressed in two ways. First, by direct confrontation with cricket physiology and neuroanatomy (Webb, 1996) and second, by reimplementing the robotic solution so as to allow phonotaxis to real cricket song—a nice (though nonconclusive) test previously ruled out by details of size and component speed. Once reimplemented, the robot was indeed able to direct and locate real singing males (see Lund et al., 1997).

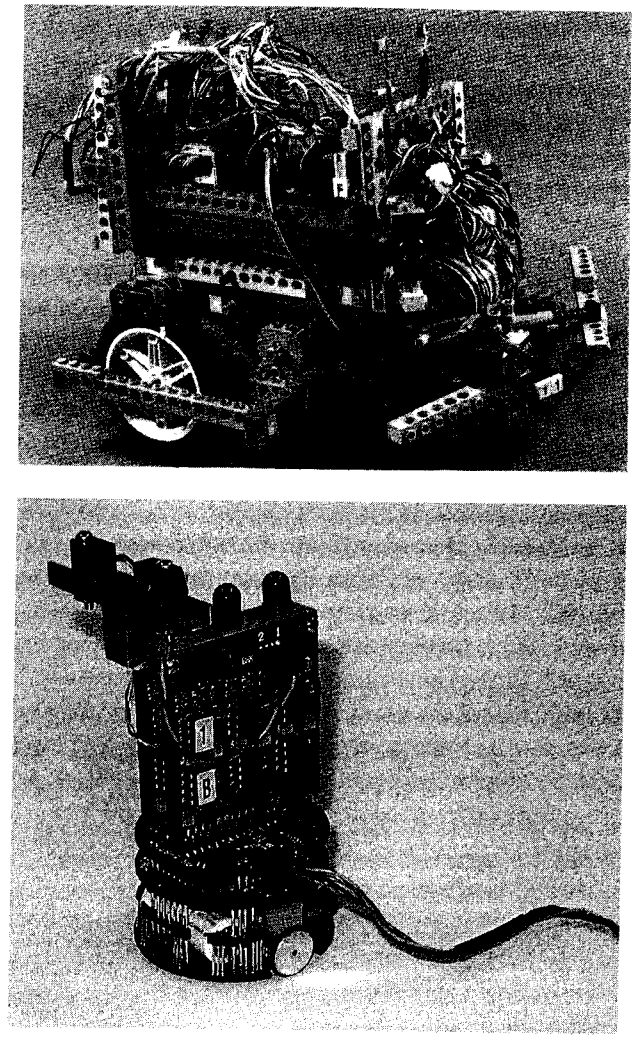

Figure 6.2 Two versions of the robot cricket: the original LEGO version and a newer version based on the Khepera robot platform. (Photos courtesy of Barbara Webb.)

One major strand of work in artificial life thus stresses the importance of real-time, real-world activity and the distribution of problem-solving contributions across body, brain, and local environment. Another strand, to which we now turn, stresses issues concerning emergence and collective effects in large ensembles. To get the flavor, consider Craig Reynolds groundbreaking work on flocking. Reynolds (1987) showed that the fluid and elegant flocking behavior of birds and other animals could be replicated (in computer animation) using a group of simulated agents (boids) each of which followed just three simple, local rules.

The rules were, roughly, to try to stay near a mass of other boids, to match your velocity to that of your neighbors, and to avoid getting too close to any one neighbor. When each boid followed these rules, patterns of on-screen activity ensued that quite closely resembled the flocking behavior of real birds, schooling fish, and other animals. Widely spaced boids immediately closed ranks, then group motion ensued with each boid making subtle speed and position adjustments as needed. And unexpectedly, when the mobile flock ran into an obstacle, it simply parted, washed around it, and reformed elegantly on the other side!

The boid work, although initially conceived as a simple tool for computer animation, clearly offered possible insight into the mechanisms of flocking in real animals. More importantly, for current purposes, it exemplified several themes that have since become central to work in artificial life. It showed that interesting collective effects can emerge as a result of the interactions between multiple simple agents following a few simple rules. It showed that the complexity and adaptability of such emergent behavior can often exceed our untutored expectations (witness the elegant obstacle-avoidance behavior). And it began to raise the question of what is real and what is mere simulation: the boids were not real animals, but the *flocking* behavior, it was later claimed (Langton, 1989, p. 33) was still an instance of *real flocking*. (We will return to this issue in the discussion.)

The boid research, however, really addresses only patterns emergent from agent–agent interaction. An equally important theme (and one also foregrounded in the kind of robotics work discussed earlier) concerns agent–environment interactions. Thus consider the way (real) termites build nests. The key principle behind termite nest building is the use of what have become known as "stigmergic" routines. In a stigmergic routine, repeated agent–environment interactions are used to control and guide a kind of collective construction process [the word derives from "stigma" (sign) and "ergon" (work) and suggests the use of work as a signal for more work—see Grasse (1959) and Beckers et al. (1994)]. A simple example is the termite's construction of the arches that structure the nests. Here, each termite deploys two basic strategies. First, they roll mud up into balls that are simultaneously impregnated—by the termite—with a chemical trace. Second, they pick up the balls and deposit them wherever the chemical trace is strongest. At first, this leads to random depositing. But once some impregnated mudballs are scattered about, these act as attractors for further deposits. As mudballs pile up, the attractive force increases and columns form. Some of these columns are, as luck would have it, fairly proximal to one another. In such cases, the drift of scent from a nearby column inclines the termites to deposit new mudballs on the side of the column nearest to the neighboring column. As this continues, so the columns gently incline together, eventually meeting in the center and creating an arch. Similar stigmergic routines then lead to the construction of cells, chambers, and tunnels. Recent computer-based simulations have replicated aspects of this process, using simple rules to underpin the piling of "wood chips" (Resnick, 1994, Chapter 3). And experiments using groups of small real-world robots have shown similar effects in laboratory settings (Beckers et al., 1994). The moral, once again, is that apparently complex problem solving need not always involve the use of heavy-duty individual reasoning engines, and that coordinated activity need not be controlled by a central plan or blueprint, nor by a designated "leader." In the termite studies just described no termite knows much at all: simply how to respond to an encountered feature of the local environment, such as the chemical trace in the mudballs. The collective activity is not even orchestrated by regular signaling or communication—instead, signals are channeled through the environmental structures, with one agent's work prompting another to respond according to some simple rule. (In Chapter 8, we will discuss some closely related ideas in the realm of advanced human problem solving).

In sum, work on artificial life aims to reconfigure the sciences of the mind by emphasizing the importance of factors other than rich, individual computation and cogitation. These factors include (1) the often unexpected ways in which multiple factors (neural, bodily, and environmental) may converge in natural problem solving, (2) the ability to support robust adaptive response without central planning or control, and (3) the general potency of simple rules and behavioral routines operating against a rich backdrop of other agents and environmental structure.

6.2 Discussion

A. THE ABSENT AND THE ABSTRACT

Work in artificial life and real-world robotics often has a rather radical flavor. This radicalism manifests itself as a principled antipathy toward (or at least agnosticism about) the invocation of internal representations, central planning, and rich inner models in cognitive scientific explanations of intelligent behavior.[2] Such radicalism looks, however, somewhat premature given the state of the art. For the notions of internal representation, inner world models and their ilk were introduced to help explain a range of behaviors significantly different from those studied by most roboticists: behaviors associated with what might reasonably[3] be called "advanced reason." Such behaviors involve, in particular:

1. The coordination of activity and choice with distal, imaginary, or counterfactual states of affairs.
2. The coordination of activity and choice with environmental parameters whose ambient physical manifestations are complex and unruly (e.g., open-endedly disjunctive—we will review examples below).

[2]See, e.g., Thelen and Smith (1994), Brooks (1991), van Gelder (1995), Keijzer (1998), and Beer (1995) among many others.

[3]This is not to downplay the difficulty or importance of basic sensorimotor routines. It is meant merely to conjure those distinctive skills by which some animals (notably humans) are able to maintain cognitive contact with distal, counterfactual, and abstract states of affairs.

It is these kinds of behavior, rather than locomotion, wall following, mate detection, and the like, for which the representationalist approach seems best suited.

Thus consider the first class of cases, the ones involving the coordination of activity and choice across some kind of physical disconnection.[4] Examples might include planning next year's family vacation, plotting the likely consequences of some imagined course of action, using mental imagery to count the number of windows in your London apartment while sitting at your desk in St. Louis, Missouri, or doing mental arithmetic. In all these cases, the objects of our cognitive activity are physically absent. By contrast, almost all[5] the cases invoked by the new roboticists involve behavior that is continuously driven and modified by the relevant environmental parameter—a light source, the physical terrain, the call of the male cricket, etc. Yet these kinds of problem domain, it seems clear, are simply not sufficiently "representation hungry" (Clark and Toribio, 1994) to be used as part of any *general* antirepresentationalist argument. This is why the best examples of representation-sparse real-world robotics strike us as rather poor examples of genuinely cognitive phenomena. Paradigmatically cognitive capacities involve the ability to generate appropriate action and choice despite physical disconnection. And this requires, prima facie, the use of some inner item or process whose role is to stand in for the missing environmental state of affairs and hence to support thought and action in the absence of on-going environmental input. Such inner stand-ins *are* internal representations, as traditionally understood.

The point here—to be clear—is *not* to argue that the capacity to coordinate action despite physical disconnection strictly implies the presence of anything like traditional internal representations. For it is certainly possible to *imagine* systems that achieve such coordination without the use of any stable and independently identifiable inner states whose role is to act as stand-ins or surrogates for the absent states of affairs [see Keijzer (1998) for a nice discussion]. The point is rather that it is dialectically unsound to argue *against* the representationalist by adducing cases where there is no physical disconnection. Such cases are interesting and informative. But they cannot speak directly against the representationalist vision.

Similar issues can be raised by focusing on our second class of cases. These involve not full-scale physical disconnection so much as what might be termed "attenuated presence." The issue here is related to a concern often voiced by Jerry Fodor, viz. that advanced reason involves selective response to nonnomic properties (see Box 6.1) of the stimulus–environment (see Fodor, 1986). Nomic properties are those that fall directly under physical laws. Thus detecting light intensity is detecting a nomic property. Humans (and other animals) are, however, capable of selective response to "nonnomic" properties such as "being a crumpled shirt"—a

[4]For an extended discussion of the themes of connection, and disconnection see Smith (1996).

[5]A notable exception is Lynne Stein's work on imagination and situated agency. See Stein (1994) and comments in Clark (1999b).

Box 6.1

NONNOMIC PROPERTIES

Nomic properties are properties of an object such that possession of the properties causes the object to fall under specific scientific laws. The physical and chemical properties of a Picasso are thus nomic, whereas the property of "being admired by many" is not. The property of being worth a million dollars is likewise nonnomic, as is the property (according to Fodor—see text) of being a crumpled shirt. The parts of the physical universe that are, indeed, crumpled shirts are (of course) fully bound by physical laws. But such laws apply to them *not* because they are crumpled shirts (or even shirts) but because they, e.g., weigh 2 pounds or have such and such a mass, etc. For a nice discussion of the issues arising from Fodor's suggestion that selective response to nonnomic properties is the cash value of the use of mental representations, see Antony and Levine (1991) and Fodor's reply in the same volume.

property that (unlike, e.g., the shirt's mass) does not characterize the object in a way capable of figuring in physical laws. Ditto for "being a genuine dollar bill" or "being a labour victory in the 1996 election." Fodor's (1986, p. 14) view was that "selective response to [such] non-nomic properties is the great evolutionary problem that mental representation was invented to solve."

The nomic/nonnomic distinction does not, however, fully serve Fodor's purposes. For it is clearly possible to respond selectively to "nonnomic" properties such as "shirtness" (we do it all the time). If this is to be physically explicable, there must be *some* kind of (perhaps complex and multifaceted) lawful relation linking our reliable selective responses to shirt-presenting circumstances. The real issue, as Fodor (1991, p. 257) more recently acknowledges, is not whether shirt detection falls under laws, but "that there is no non-inferential way of detecting shirtness."

The deep issue, as Fodor now sees it, thus concerns what we might call "simple sensory transducability." To track a property such as "being a shirt" we seem to need to use an indirect route—we directly track a complex of other features that cumulatively signifies shirthood. No one could build a simple sensory transducer (where a transducer is loosely conceived as a device that takes sensory input and converts it into a different form or signal used for further processing) that (even roughly) itself isolated all and only those energy patterns that signify the presence of shirts. Instead, you need to detect the obtaining of properties such as "is shirt shaped," "could be worn by a human," etc. and then (or so Fodor insists) *infer* the presence of a shirt. It is the presence of inferred representations and the associated

capacity to go beyond simple, direct transduction that Fodor (1991, p. 257) now sees as the source of a "principled distinction" between very simple minds (such as that of a paramecium) and the minds of advanced reasoners (such as ourselves).

I think there is something in this. There certainly seems to be a large gap between systems that track directly transducible environmental features (such as the presence of sugar or the male cricket song) and ones that can respond to more arcane features, such as the carrying out of a charitable action or the presence of a crumpled shirt. Prima facie, the obvious way to support selective response to evermore arcane features is to detect the presence of multiple other features and to develop deeper inner resources that covary with the obtaining of such multiple simple features: complex feature detectors, in short. But internal states developed to serve such a purpose would, at least on the face of it, seem to count as internal representations in good standing.

The proper conclusion here, once again, is not that it is simply inconceivable that coordination with what is absent, counterfactual, nonexistent, or not directly transducible is *impossible* without deploying inner states worth treating as internal representations. Rather, it is that existing demonstrations of representation-free or representation-sparse problem solving should not be seen as directly arguing for the possibility of a more general antirepresentationalism. For the problem domains being negotiated are not, in general, the kind most characteristic of advanced "representation-hungry" reason.

All this, to be sure, invites a number of interesting (and sometimes potent) replies. This discussion continues in Chapters 7 and 8.

B. EMERGENCE[6]

The artificial life literature gives special prominence to the notions of emergence and collective effects. But the notion of emergence is itself still ill understood. Nor can it be simply identified with the notion of a collective effect, for not every collective effect amounts intuitively to a case of emergence, nor does every case of emergence seem (again, intuitively) to involve a collective effect. Thus consider the way a collection of small identical weights (billiard balls perhaps) may collectively cause a balance-beam to tip over onto one side. This is a collective effect all right (it needs, let us imagine, at least 30 billiard balls to tip the scale). But we seem to gain nothing by labeling the episode as one of "emergent toppling." Or consider, by contrast, the case of the simple robot described in Hallam and Malcolm (1994). This robot follows walls encountered to the right by means of an inbuilt bias to move to the right, and a right-side sensor, contact activated, that causes it to veer slightly to the left. When these two biases are well calibrated, the robot will follow the wall by a kind of "veer and bounce" routine. The resultant behavior is described as "emergent wall following," yet the number of factors and forces involved seems

[6]This section owes a lot to discussions with Pim Haselager and Pete Mandik.

too low, and the factors too diverse, to count this as a collective effect of the kind mentioned in our earlier sketch.

Relatedly, we need to find an account of emergence that is neither so liberal as to allow just about everything to count as an instance of emergence (a fate that surely robs the notion of explanatory and descriptive interest), nor so strict as to effectively rule out any phenomenon that can be given a scientific explanation (we do not want to insist that only currently unexplained phenomena should count as emergent, for that again robs the notion of immediate scientific interest). Rather it should pick out a distinctive way in which basic factors and forces may conspire to yield some property, event, or pattern. The literature contains a number of such suggestions, each of which cuts the emergent/nonemergent cake in somewhat different ways. As a brief rehearsal of some prominent contenders, consider the following.

1. Emergence as Collective Self-Organization. This is the notion most strongly suggested by the earlier examples of flocking, termite nest building, etc. As a clinically pure example, consider the behavior of cooking oil heated in a pan. As the heat is applied it increases the temperature difference between the oil at the top (cooler) and at the bottom (hotter). Soon, there appears a kind of rolling motion known as a convection roll. The hotter, less dense oil rises, to be replaced by the cooler oil, which then gets hotter and rises, and so on. Of such a process, Kelso (1995, pp. 7–8) writes:

> The resulting convection rolls are what physicists call a collective or cooperative effect, which arises without any external instructions. The temperature gradient is called a control parameter [but it does not] prescribe or contain the code for the emerging pattern. . . . Such spontaneous pattern formation is exactly what we mean by self-organization: the system organized itself, but there is no 'self', no agent inside the system doing the organizing.

The proximal cause of the appearance of convection rolls is the application of heat. But the *explanation* of the rolls has more to do with the properties of an interacting mass of simple components (molecules) that, under certain conditions (viz. the application of heat), feed and maintain themselves in a specific patterned cycle. This cycle involves a kind of "circular causation" in which the activity of the simple components leads to a larger pattern, which then *enslaves* those same components, locking them into the cycle of rising and falling. (Think of the way the motion of a few individuals can start a crowd moving in one direction: the initial motion induces a process of *positive feedback* as more and more individuals then influence their own neighbors to move in the same direction, until the whole crowd moves as a coherent mass.)

Such collective effects, with circular causation and positive feedback, can be usefully understood using the notion of a "collective variable"—a variable whose

changing value reflects the interactive result of the activities of multiple systemic elements. Examples include the temperature and pressure of a gas, the rate of acceleration and direction of motion of the crowd, the amplitude of the convection rolls, and so on. Dynamic systems theory (which we will introduce in the next chapter) specializes in plotting the values of such collective variables as systemic behavior unfolds over time, and in plotting the relations between the collective variables and any control parameters (such as the temperature gradient in the oil). An emergent phenomenon, according to our first account, is thus any interesting behavior that arises as a direct result of multiple, self-organizing (via positive feedback and circular causation) interactions occurring in a system of simple elements.

Problems? This story works well for systems comprising large numbers of essentially identical elements obeying simple rules. It thus covers flocking, termite nest building, convection rolls, etc. But it is less clearly applicable to systems comprising relatively few and more heterogeneous elements (such as the robot cricket and the bounce and veer wall follower).

2. Emergence as Unprogrammed Functionality. By contrast, the idea of emergence as something like "unprogrammed functionality" is tailor-made for the problem cases just mentioned. In such cases we observe adaptively valuable behavior arising as a result of the interactions between simple on-board circuitry and bodily and environmental structure. Such behaviors (wall following, cricket phonotaxis) are not supported by explicit programming or by any fully "agent-side" endowment. Instead, they arise as a kind of *side-effect* of some iterated sequence of agent–world interactions. The point is not that such behaviors are necessarily unexpected or undesigned—canny roboticists may well set out to achieve their goals by orchestrating just such interactions. It is, rather, that the behavior is not subserved by an internal state encoding either the goals (follow walls, find males, etc.) or how to achieve them. Such behaviors thus depend on what Steels (1994) calls "uncontrolled variables"—they are behaviors that can only be very *indirectly* manipulated, since they depend not on central or explicit control structures but on iterated agent–environment interactions.

Problems? As you might guess, this story works well for the cases just mentioned. But it seems less clearly applicable to cases of collective self-organization. For cases of the latter kind clearly do allow for a form of direct control by the manipulation of a single parameter (such as the heat applied to the cooking oil).

3. Emergence as Interactive Complexity. I think we can do some justice to *both* the proceeding accounts by understanding emergent phenomena as the effects, patterns, or capacities made available by a certain class of complex interactions between systemic components. Roughly, the idea is to depict emergence as the process by which complex, cyclic interactions give rise to stable and salient patterns of systemic behavior. By stressing the complexity of the interactions we allow emergence

Box 6.2

NONLINEAR INTERACTIONS

A nonlinear interaction is one in which the value of x does not increase proportionally to the value of y. Instead, x may (for example) remain at zero until y reaches a critical value and then increase unevenly with an increase in the value of y. The behavior of a standard connectionist unit (see Chapter 4) is nonlinear since the output is not simply the weighted sum of the inputs but may involve a threshold, step function, or other nonlinearity. A typical example is a unit having a sigmoid activation function, in which certain input values (high positive or negative values, for example) yield a sharp response, causing the unit to output 0 (for high negative input) or 1 (for high positive input). But for certain intermediate input values (mildly positive or mildly negative ones), such a unit gives a more subtly gradated response, gradually increasing the strength of the output signal according to the current input. See Figure 6.3.

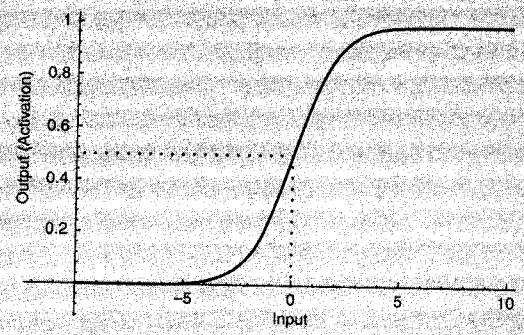

Figure 6.3 Nonlinear response in a connectionist unit. Notice that the unit responds with 0 to all high negative input values, with 1 to all high positives, with 0.5 when the input is zero, and with subtly gradated responses for all intermediate values. (Adapted from Elman et al., 1996, Fig. 2.2, p. 53.)

to come (obtain) in degrees. Phenomena that depend on repeated linear interactions with only simple kinds of feedback loop (e.g., a strict temporal sequence in which x affects y which then affects x) will count as, at best, only weakly emergent. In such cases it is usually unclear whether talk of emergence is explanatorily useful. By contrast, phenomena that depend on multiple, nonlinear (see Box 6.2), temporally asynchronous, positive feedback involving interactions will count as strongly emergent. Bounce-and-veer wall following is thus a case of weak emergence, whereas the convection roll example, when fully described, turns out to be

Box 6.3

A Case in Which Prediction Requires Simulation

Consider the decimal expansion of $\sqrt{2} - 1$. This (see Franklin, 1995, p. 282) defines an irrational number. The resulting sequence is unpredictable except by direct step-by-step calculation. To find the next digit you must always calculate the preceding digit. By contrast, some functions rapidly converge to a fixed point or a repeating pattern. In these cases (e.g., the infinite sequence .33333 recurring) we can predict the nth number in the sequence without calculating $n - 1$ and applying a rule. Such sequences afford short-cuts to prediction. Mathematical chaos represents a kind of middle option—sequences of unfolding that exhibit real *local* patterning but that resist long-term prediction (see Stewart, 1989).

a classic case of strong emergence (see Kelso, 1995, pp. 5–9). Emergent phenomena, thus defined, will typically reward understanding in terms of the changing values of a collective variable—a variable (see above) that tracks the pattern resulting from the interactions of multiple factors and forces. Such factors and forces may be wholly internal to the system or may include selected elements of the external environment.

4. Emergence as Uncompressible Unfolding. Finally (and for the sake of completeness), I should note another (and I think quite different) sense of emergence represented in the recent literature. This is the idea of emergent phenomena as those phenomena for which *prediction* requires *simulation*—and especially those in which *predication* of some macrostate P requires simulation of the complex interactions of the realizing microstates M_1–M_n. (See Box 6.3 for an example.) Bedau (1996, p. 344) thus defines a systemic feature or state as emergent if and only if you can predict it, in detail, *only* by modeling all the interactions that give rise to it. In such cases, there is no substitute for actual simulation if we want to predict, in detail, the shape of the macroscopic unfolding.

Problems? This definition of emergence strikes me as overly restrictive. For example, even in cases involving multiple, complex, nonlinear, and cyclic interactions, it will often be possible to model systemic unfolding by simulating only a *subset* of actual interactions. Convection roll formation, for example, succumbs to an analysis that (by exploiting collective variables) allows us to predict how the patterns (given a set of initial conditions) will form and unfold over time. Bedau's proposal, in effect, restricts the notion of emergence to phenomena that resist all such attempts at low-dimensional modeling. My intuition, by contrast, is that emergent phenomena are often *precisely* those phenomena in which complex interactions yield robust, salient patterns capable of supporting prediction and explanation, i.e., that lend themselves to various forms of low-dimensional projection.

C. LIFE AND MIND[7]

Work in artificial life also raises some fundamental questions concerning the very idea of life and the relationship between life and mind. On the very idea of life, the challenge is direct and simple: could life be actually instantiated (rather than simply modeled) in artificial media such as robots or computer-based ecosystems? Consider, for example, the virtual ecosystem named Tierra (Ray, 1991, 1994). Here, digital organisms (each one a kind of small program) compete for CPU time. The, "organisms" can reproduce (copy) and are subject to change via random mutations and occasionally incorrect copying. The system is implemented in the memory of the computer and the "organisms" (code fragments or "codelets") compete, change, and evolve. After a while, Ray would stop the simulation and analyze the resultant population. He found a succession of successful (often unexpected) survival strategies, each one exploiting some characteristic weakness in the proceeding dominant strategy. Some codelets would learn to exploit (piggyback on) the instructions embodied in other organisms' code, as "virtual parasites." Later still, codelets evolved capable of diverting the CPU time of these parasites onto themselves, thus parasitizing the parasites, and so on. The following question then arises: Are these *merely* virtual, simulated organisms or is this a *real* ecosystem populated by *real* organisms "living" in the unusual niche of digital computer memory? Ray himself is adamant that, at the very least, such systems can genuinely support several properties characteristic of life—such as real self-replication, real evolution, real flocking, and so on (see Ray, 1994, p. 181).

One debate, then, concerns the effective definition of life itself, and perhaps of various properties such as self-replication. In this vein, Bedau (1996, p. 338) urges a definition of life as "supple adaptation"—the capacity to respond appropriately, in an indefinite variety of ways, to an unpredictable (from the perspective of the organism) variety of contingencies. Such a definition [unlike, for example, one focused on the metabolization of matter into energy—see Schrödinger (1969) and Boden (1999)] clearly allows events and processes subsisting in electronic and other media to count as instances of life properly so-called. Other authors focus on still other properties and features, such as autopoiesis (autopoietic systems actively create and maintain their own boundaries, within which complex circular interactions support the continued production of essential chemicals and materials—see Varela, Maturana, and Uribe, 1974), autocatalysis (sets of

[7]Thanks to Brian Keeley for insisting on the importance of these topics, and for helping me to think about them.

elements—chemical or computational—that catalyze their own production from available resources—see Kauffman, 1995), self-reproduction, genetics, and metabolization (Crick, 1981), and so on. A very real possibility—also mentioned by Bedau (1996)—is that "life" is a so-called cluster concept, involving multiple typical features none of which is individually necessary for a system to count as alive, and multiple different subsets of which could be sufficient.

There is also a debate about the relations between life and mind. One way to resist the worry (see Section A) that these simple, life-like systems tell us little about really *cognitive* phenomena is to hold that life and mind share deep organizational features and that the project of understanding mind is thus continuous with the project of understanding life itself. The position is nicely expressed by Godfrey-Smith (1996a, p. 320) in his description[8] of the thesis of "strong continuity":

> Life and mind have a common abstract pattern or set of basic organizational properties. The functional[9] properties characteristic of mind are an enriched version of the functional properties that are fundamental to life in general. Mind is literally life-*like*.

This, as Godfrey-Smith notes, is a deep claim about the phenomenon of mind itself. It thus goes beyond the more methodological claim that the scientific investigation of mind should proceed by looking at whole, embodied life-forms, and asserts that the central *characteristics* of mind are, in large part, those of life in general. This is not to imply, of course, that life and mind are exactly equivalent—just that if we understood the deep organizing principles of life in general, we would have come a very long way in the project of understanding mind. In more concrete terms, the thesis of strong continuity would be true if, for example, the basic concepts needed to understand the organization of life turned out to be self-organization, collective dynamics, circular causal processes, autopoiesis, etc., and if *those very same concepts and constructs* turned out to be central to a proper scientific understanding of mind. A specific—and currently quite popular—version of the strong continuity thesis is thus the idea that the concepts and constructs of dynamic systems theory will turn out to be the best tools for a science of mind, and will simultaneously reveal the fundamental organizational similarity of processes operating across multiple physical, evolutionary, and temporal scales. The danger, of course, is that by stressing unity and similarity we may lose sight of what is special and distinctive. Mind may indeed participate in many of the dynamic processes characteristic of life. But what about our old friends, the fundamentally reason-based transitions and the grasp of the absent and the abstract characteristic of advanced cognition?

Balancing these explanatory needs (the need to see continuity in nature and the need to appreciate the mental as somehow *special*) is perhaps the hardest part of recent cognitive scientific attempts to naturalize the mind.

[8]As far as I can tell, Godfrey-Smith remains agnostic on the truth of the strong continuity thesis. He merely presents it as one of several possible positions and relates it to certain trends in the history of ideas. See Godfrey-Smith (1996a,b).

[9]It may be that Godfrey-Smith overplays the role of functional description here. Recall our discussions of function versus implementation in Chapters 1 through 6. For a version of strong continuity without the functional emphasis, see Wheeler (1997).

6.3 Suggested Readings

Useful general introductions to work in robotics and artificial life include S. Levy, *Artificial Life* (London: Cape, 1992), a journalistic but solid introduction to the history and practice of artificial life, and S. Franklin, *Artificial Minds* (Cambridge, MA: MIT Press, 1995). C. Langton (ed.), *Artificial Life: An Overview* (Cambridge, MA: MIT Press, 1995) reprints the first three issues of the journal *Artificial Life* and includes excellent, specially commissioned overview articles covering robotics, collective effects, evolutionary simulations, and more. It includes one of Ray's papers on the Tierra project, as well as excellent introductory overviews by (among other) Luc Steels, Pattie Maes, and Mitchel Resnick.

For an excellent treatment of the issues concerning *emergence and collective effects*, the reader is strongly encouraged to look at M. Resnick, *Turtles, Termites and Traffic Jams* (Cambridge, MA: MIT Press, 1994). This is a delightful, simulation-based introduction to the emergence of complex effects from the interaction of simple rules. Software is available on the web.

For the *philosophical issues concerning emergence, representation, and the relation of life to mind*, see various essays in M. Boden (ed.), *The Philosophy of Artificial Life* (Oxford, England: Oxford University Press, 1996), especially the papers by Langton, Wheeler, Kirsh, and Boden. A. Clark, *Being There: Putting Brain, Body and World Together Again* (Cambridge, MA: MIT Press, 1997) is an extended treatment of many of the core issues.

For work on *real-world robotics and the importance of physical implementation*, see H. Chiel and R. Beer "The brain has a body." *Trends in Neuroscience*, 20, 553–557, 1997. This is an excellent short summary of evidence in favor of treating the nervous system, body, and environment as a unified system. R. McClamrock, *Existential Cognition: Computational Minds in the World* (Chicago: University of Chicago Press, 1995) is a well-executed philosophical argument for viewing the mind as essentially environmentally embedded, and B. Webb "A Cricket Robot." *Scientific American*, 275, 62–67, 1996, is a user-friendly account of the work on the robot cricket.

Volumes of conference proceedings probably offer the best view of the actual practice of artificial life. See, e.g., *Artificial Life I–VII* (and counting) published by MIT Press, Cambridge, MA.