Adaptive Behavior

The Dynamics of Active Categorical Perception in an Evolved Model Agent

Randall D. Beer
Department of Electrical Engineering and Computer Science, and Department of Biology Case Western Reserve University

Notions of embodiment, situatedness, and dynamics are increasingly being debated in cognitive science. However, these debates are often carried out in the absence of concrete examples. In order to build intuition, this paper explores a model agent to illustrate how the perspective and tools of dynamical systems theory can be applied to the analysis of situated, embodied agents capable of minimally cognitive behavior. Specifically, we study a model agent whose "nervous system" was evolved using a genetic algorithm to catch circular objects and to avoid diamond-shaped ones. After characterizing the performance, behavioral strategy and psychophysics of the best-evolved agent, its dynamics are analyzed in some detail at three different levels: (1) the entire coupled brain/body/environment system; (2) the interaction between agent and environment that generates the observed coupled dynamics; (3) the underlying neuronal properties responsible for the agent dynamics. This analysis offers both explanatory insight and testable predictions. The paper concludes with discussions of the overall picture that emerges from this analysis, the challenges this picture poses to traditional notions of representation, and the utility of a research methodology involving the analysis of simpler idealized models of complete brain/body/environment systems.

Keywords dynamics · minimally-cognitive behavior · categorical perception

1 Introduction

"May you live in interesting times." This ancient Chinese curse has never been more appropriate than it is of cognitive science today. Once upon a time, life was simple. Cognition was computation, and everyone knew what that meant: formal manipulation of quasi-linguistic symbolic representations by syntactic rules. Then, in the early 1980s, connectionism began to muddy the theoretical waters by significantly expanding our notion of what a representation could be and the form a computation could take.

Now notions of situatedness, embodiment, and dynamics are threatening to once again revolutionize some of our most basic assumptions about what cognition is and how it should be understood. For a situated, embodied agent, taking action appropriate to both its immediate circumstances and its long-term goals is the primary concern, and cognition becomes only one resource among many in service of this objective. An agent's physical body, the structure of its environment, and its social context can play as important a role in the generation of its behavior as its brain. Indeed, in a very real sense, cognition can no longer be seen as limited to an agent's head, but can be distributed across a group of agents and artifacts.

Dynamical approaches emphasize the temporal dimension of cognition, and the way in which an

Correspondence to: Randall D. Beer, Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, Ohio 44106-7071, USA.
E-mail: beer@eecs.cwru.edu;
Tel: +1-216-368-2801; *Fax*: +1-216-368-2816.

(2003), Vol 11(4): 209–243.
[1059-7123(200312) 11:4; 209-243; 041345]

agent's behavior arises from the ongoing interaction between its brain, its body and its environment. On this view, the focus shifts from accurately representing an environment to continuously engaging that environment with a body so as to stabilize coordinated patterns of behavior that are adaptive for the agent. Rather than assigning representational content to neuronal states, the mathematical tools of dynamical systems theory are used to characterize the structure of the space of possible behavioral trajectories and the internal and external forces that shape the particular trajectory that unfolds. Indeed, a dynamical approach to situated action raises important questions about the very necessity of notions of representation and computation in cognitive theorizing.

Not surprisingly, such claims have met with a great deal of skepticism, and there is considerable controversy regarding the fundamental significance of situatedness, embodiment, and dynamics to cognitive science. Are these notions just minor corrections to the classical picture of cognition, or do they radically alter the entire enterprise of cognitive science? Unfortunately, as in the early days of connectionism, the initial debates have often been based more on competing intuitions than facts. Concrete dynamical models of a wide range of cognitive phenomena are under construction (including language, cognitive development, decision-making, perception, and action (for reviews, see Port & van Gelder (1995) and Beer (2000)). But in the meantime, one powerful way to improve our intuitions, clarify the key issues and sharpen the debate is through the careful study of simpler idealized models of *minimally cognitive behavior*, the simplest behavior that raises issues of genuine cognitive interest (Beer, 1996). Detailed dynamical analyses of situated and embodied model agents can help us to expand our theoretical imagination and to better ground the ongoing debates in concrete examples (Beer, 1997, 1998). What kinds of cognitive phenomena can dynamical notions account for? What kinds of mathematical tools can be brought to bear, and what sorts of insights do they provide? What is the nature of a dynamical explanation and how does it differ from more traditional modes of explanation in cognitive science? What are the advantages and disadvantages of such an explanation? We will explore these and other questions in the context of a detailed dynamical analysis of a particular model agent.

What sort of minimally cognitive behavior should we examine? Stably partitioning the world into objects with distinctive properties is one of the most fundamental activities that an agent must perform. Indeed, it has been argued that *categorical perception*, as this activity is often called, may provide a necessary substrate for higher cognition (Harnad, 1987), making it particularly appropriate for detailed study as a minimally cognitive behavior. In categorical perception, the continuous signals received by sense organs are sorted into discrete categories, whose members resemble one another more than they resemble members of other categories. The paradigmatic example of categorical perception is the perception of phoneme boundaries in continuous speech (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967), but other examples include color categorization in humans (Boynton, 1975) and species-specific sound categorization in monkeys, birds, frogs, toads and insects (Ehret, 1987; Wyttenbach, May, & Hoy, 1996). A variety of neural network models of categorical perception have been developed and analyzed (Harnad, Hanson, & Lubin, 1995; Tijsseling & Harnad, 1997; Damper & Harnad, 2000; Cangelosi, Greco, & Harnad, 2000).

In this paper, I present a dynamical analysis of a previously developed model agent that "visually" discriminates between objects (Beer, 1996). This agent is controlled by a dynamical "nervous system" that was evolved using a genetic algorithm to catch circular objects and avoid diamond-shaped objects. The intention here is not to propose a serious model of categorical perception, but rather to use this model agent to explore the implications of dynamical explanation for cognitive agents. For this reason, we will take the analysis of this agent very seriously, systematically examining in considerable depth each aspect of its behavior and the mechanisms underlying that behavior. While some may find the level of technical detail tedious, it is a basic premise of this work that a proper appreciation of the nature of dynamical explanation requires a thorough grounding in concrete examples. We must also address nontrivial issues involved in extending the techniques utilized for dynamical analysis of sensorimotor behavior to minimally cognitive behavior. Furthermore, the ability to analyze a single minimally cognitive agent is a crucial prerequisite for abstracting general principles from the analysis of many such agents. Think of this exercise, then, as a form of mental calisthenics, an intellectual warm-up for the dynamical analyses of a wider range of agents and behaviors.

The paper is organized as follows. The next section presents an overview of the conceptual framework within which the work described in this paper is carried out, one that attempts to do justice to the embodied, situated and dynamical nature of cognitive agents. This is followed by a review of our object discrimination experiments. I then summarize the overall performance and the behavioral strategy used by the best agent that was evolved. Next, I describe the results of psychophysical experiments that demonstrate categorical perception in the evolved agent, determine the object features that matter to the discrimination, and characterize the manner in which the decision to catch or avoid is made. I then analyze this evolved agent, first characterizing the dynamics of the entire agent-environment system, then decomposing this in terms of the interactions between the agent and its environment, and finally examining the neural implementation of the agent dynamics. The paper concludes with discussions of the overall picture that emerges from this analysis, the relationship of this picture to traditional notions of representation, and the utility of a research methodology based on analyzing simpler idealized models of complete brain/body/environment systems.

2 An Embodied, Situated, Dynamical Perspective on Cognition

What does it mean to take an embodied, situated or dynamical perspective on cognition, let alone all three simultaneously? There are probably as many answers to this question as there are advocates or critics of these various positions. Thus, I will make no attempt here to provide a critical survey of the ongoing debates or to develop and defend a general situated or dynamical position. Fortunately, this ground is being well-covered elsewhere (Clark, 1997). Rather, the more modest aim of this section is to merely provide a brief sketch of the general conceptual framework within which the particular analysis described in this paper will be carried out (Beer, 1995a, 1997).

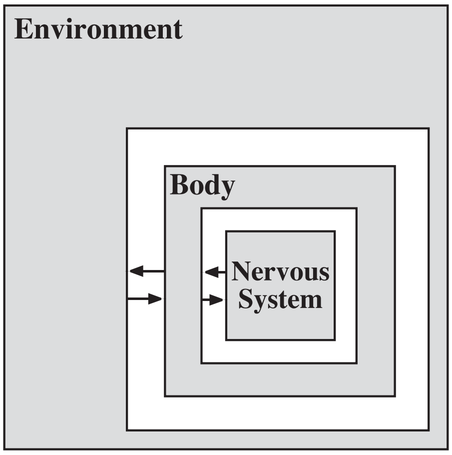

Figure 1 A dynamical perspective on a situated, embodied agent. The nervous system, body and environment of an agent are each conceptualized as dynamical systems that are in constant interaction. From this viewpoint, an agent's behavior arises from this interaction between these subsystems and cannot properly be attributed to any one component in isolation from the others.

As I intend them, embodiment and situatedness are closely related concepts. Both reject the notion that purely intellectual ratiocination is the defining feature of the cognitive. Rather, embodiment emphasizes the role of an agent's own body in its cognition, while situatedness emphasizes the role of an agent's immediate physical and social environment (Gibson, 1979; Suchman, 1987; Beer, 1990; Brooks, 1991a; Varela, Thompson, & Rosch, 1991; Ballard, 1991; Hutchins, 1995; Hendriks-Jansen, 1996; Clancey, 1997; Agre, 1997; Glenberg, 1997; Lakoff & Johnson, 1998; Arkin, 1998; Pfeifer & Scheier, 1999). On this view, situated action is the fundamental concern, and cognition is an elaboration of this more basic capacity, and one resource among many that can be brought to bear as an agent encounters its world. After all, nervous systems originally evolved to coordinate behavior, not to balance checkbooks. An embodied nervous system can utilize the natural biomechanics of its body, the geometry of its sensory surfaces, and its ability to actively control these surfaces to considerably simplify many problems. Likewise, a situated agent can utilize and manipulate the physical and functional organization of the space around it, as well as the social organization in which it exists, to offload problems to its environment. On this view, cognitive behavior arises from the interaction between a brain, a body and an environment (Figure 1), and cannot properly be studied as a product of the mind/brain alone (Beer, 1995a). Given that nervous systems and bodies co-evolve with one another and with their environment, the need for such a coupled perspective is hardly surprising. What is surprising is the extent to which such an embodied, situated perspective can alter our very conception of cognition, as well as the domains in which we seek to understand its underlying mechanisms.

By a dynamical perspective, I mean simply one in which the concepts and mathematical tools of dynamical systems theory are applied to the analysis of the temporal behavior of a cognitive system (Ashby, 1952; Skarda & Freeman, 1987; Thelen & Smith, 1994; Kelso, 1995; Elman, 1995; Beer, 1995 ; Port & van Gelder, 1995; van Gelder, 1998; Beer, 2000; Thelen, Schöner, Scheier, & Smith, 2001). A dynamical system is a mathematical object that unambiguously describes how the state of some system evolves over time. More formally, a dynamical system is a triple $\langle T, S, \phi_t \rangle$ consisting of an ordered time set T, a state space S, and an evolution operator $\phi_t : S \to S$ that transforms an initial state $x_0 \in S$ at time $t_0 \in T$ to another state $x_t \in S$ at time $t \in T$. The time set T may be continuous or discrete. The state space S may be numerical or symbolic, continuous or discrete, or a hybrid of the two, and it may be finite- or infinite-dimensional depending on the number of variables required to fully describe the state of the system. The evolution operator ϕ_t may be given explicitly or defined implicitly, it may be deterministic or stochastic and it may have inputs (non-autonomous) or not (autonomous). Sets of differential or difference equations, cellular automata, finite state machines and Turing machines are all examples of dynamical systems. Dynamical systems theory offers a wide variety of tools for analyzing the temporal behavior of such systems, including the identification of invariant sets (e.g., equilbrium points, limit cycles, etc.), a characterization of their local and global structure (e.g., stability), and their dependence on parameters (e.g., bifurcations) (Abraham & Shaw, 1992; Strogatz, 1994).

Thus, to apply the tools of dynamical systems theory to a cognitive process is to seek a set of state variables that uniquely characterize the behavior of the process and to uncover the laws by which these states evolve in time. Armed with such a dynamical model, the dynamicist then attempts to understand the organizing features of its dynamics and the response of those dynamics to inputs or changes in system parameters. Here the focus is on how the system's behavior unfolds over time and the various "forces" that shape this trajectory. The importance of the states that the system passes through lies not so much in any content that they may be assigned, but rather in their sensitivity to subsequent inputs and the future behavior that they make possible.

While embodied, situated, and dynamical perspectives can be and have been independently developed and defended, these viewpoints work much better as a unit. Taking a dynamical perspective on situated, embodied action, an agent's nervous system, its body and its environment are each described as dynamical systems, which are coupled (Figure 1). Because all of the individual components are described in the same mathematical language, it is much easier to approach questions involving their interaction. To say that an agent's behavior arises from the interaction between its brain, body, and environment, for example, simply means that it is a trajectory of the complete dynamical system formed by their coupling, and the explanatory focus shifts to the properties of this trajectory and the mechanisms by which it is generated and stabilized. This common language can be applied to cognitive processes, the neurophysiological processes that support it, non-cognitive human behavior, and the adaptive behavior of simpler animals, as well as the emergence of cognition in development and evolution. Thus, in dynamical language, a unified theoretical framework for cognitive science broadly construed begins to seem possible.

3 An Object Discrimination Task

In a previous series of experiments, model agents were evolved that could "visually" discriminate between circles and diamonds, catching the former while avoiding the latter. These experiments were designed to produce evolved examples of categorical perception, the best of which we will study in Sections 4–8. Complete details of these experiments can be found in an earlier paper (Beer, 1996).

The agent has a circular body with a diameter of 30, and can move horizontally as objects fall from above (Figure 2A). Horizontal velocity is proportional to the sum of opposing forces produced by two motors. The agent's "eye" consists of seven rays distributed over a visual angle of $\pi/6$. An intersection between a ray and an object causes an input to be injected into the corresponding sensory neuron, with the magnitude of the injected input inversely propor-

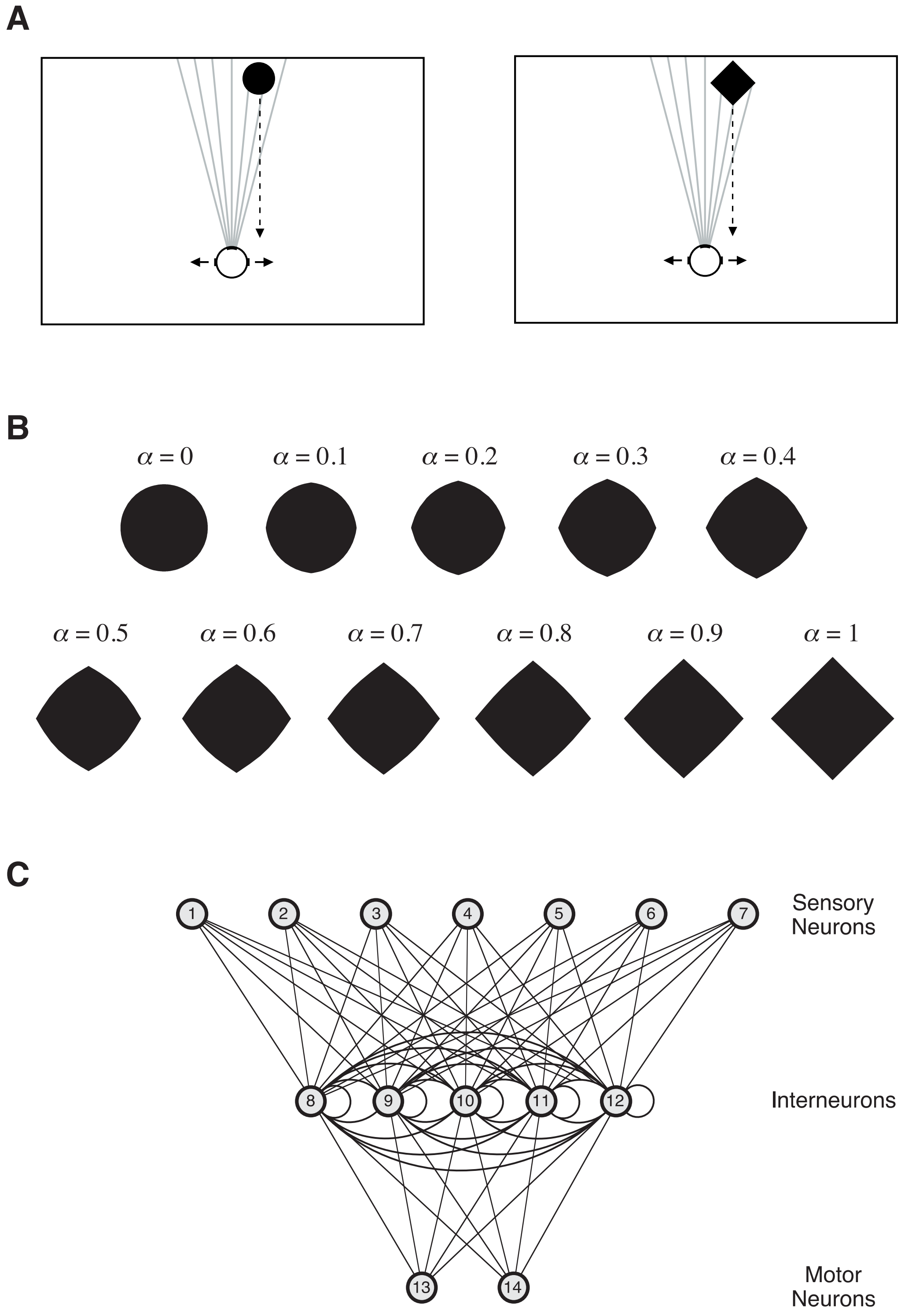

Figure 2 Basic setup for the categorical perception experiments. (A) The agent can move horizontally while objects fall from above. It uses an array of seven distance sensors to discriminate between circular (left) and diamond-shaped (right) objects, catching the former while avoiding the latter. (B) Examples of the class of hybrid objects, R_H (ϕ; α) as, α is varied from 0 (circle) to 1 (diamond). (C) The network architecture consists of seven sensory neurons fully connected to five fully interconnected interneurons, which are in turn fully connected to two motor neurons. All connections and neurons are bilaterally symmetric.

tional to the distance to the object along that ray. When the rays are at their maximum length, no input is injected, while the maximum input is injected for rays of zero length.

Objects fall straight down from the top of the environment with a fixed vertical velocity of –3 and an initial horizontal offset from the agent in the range ±50. Circular objects had a diameter of 30 and diamonds had sides of length 30. Both for uniformity of notation and for later analysis, it is convenient to define a parameterized class of hybrid objects that continuously interpolate between a circle and a diamond, even though the agent was exposed only to circles and diamonds during evolution. This is most easily done in polar coordinates. If the circle is defined by $R_C(\phi) = 15$ and the diamond is defined by

$$R_D(\phi) = \begin{cases} \dfrac{15\sqrt{2}\sec\phi}{1+\tan\phi} & \text{if } 0 \le \phi \le \pi/2 \\ \dfrac{15\sqrt{2}\sec\pi-\phi}{1+\tan\pi-\phi} & \text{if } \pi/2 \le \phi \le \pi \\ \dfrac{-15\sqrt{2}\sec\phi}{1+\tan\phi} & \text{if } \pi \le \phi \le 3\pi/2 \\ \dfrac{-15\sqrt{2}\sec\pi-\phi}{1+\tan\pi-\phi} & \text{if } 3\pi/2 \le \phi \le 2\pi \end{cases}$$

then the hybrid objects are given by $R_H(\phi; \alpha) = (1 - \alpha) \times R_C(\phi) + \alpha R_D(\phi)$. Sample objects produced as α varing from 0 to 1 are shown in Figure 2B.

The agent's behavior is controlled by a continuous-time recurrent neural network (CTRNN) with the following state equation (Beer, 1995c)

$$\tau_i \dot{s}_i = -s_i + \sum_{j=1}^{N} w_{ji}\sigma(g_j(s_j + \theta_j)) + I_i \qquad i = 1, \ldots, N$$

where s is the state of each neuron, τ is its time constant, w_{ji} is the strength of the connection from the j^{th} to the i^{th} neuron, g is a gain, θ is a bias term, $\sigma(x) = 1/(1 + e^{-x})$ is the standard logistic activation function, and I represents an external input (e.g., from a sensor). It will sometimes be convenient to refer to the output of a neuron $o_i = \sigma(s_i + \theta_i)$ rather than its state. The network architecture was bilaterally symmetric in the connection weights, biases, and time constants. The architecture consisted of seven ray sensory neurons projecting to five fully interconnected interneurons, which in turn projected to two motor neurons controlling horizontal motion (Figure 2C). All sensory neurons shared the same gain, bias, and time constant, while all other neurons had a gain of 1.

Thus, the equations for the complete model take the form

$$\begin{aligned} \dot{y} &= -3 \\ \tau\dot{s}_i &= -s_i + I_i(x, y; \alpha) \qquad i = 1, \ldots, 7 \\ \tau_i\dot{s}_i &= -s_i + \sum_{j=1}^{7} w_{ji}\sigma(g(s_j + \theta)) \\ &\quad + \sum_{j=8}^{12} w_{ji}\sigma(s_j + \theta_j) \qquad i = 8, \ldots, 12 \\ \tau_i\dot{s}_i &= -s_i + \sum_{j=8}^{12} w_{ji}\sigma(s_j + \theta) \qquad i = 13, 14 \\ \dot{x} &= 5(\sigma(s_{13} + \theta_{13}) - \sigma(s_{14} + \theta_{14})) \end{aligned} \tag{1}$$

where x is the horizontal position of the object relative to the agent's midline, y is the vertical position of the object relative to the agent, and $I_i(x, y; \alpha)$ is the sensory input from the i^{th} ray due to a hybrid object with the given α at location (x, y) in agent-centered coordinates.

The connection weights, biases, time constants, and gain in the earlier equations were evolved. The evolutionary algorithm used was a real-valued population-based hill climber (Bäck, 1996). The performance measure to be maximized was $\sum_{j=1}^{24} p_i/24$, where $p_i = 1 - |d_i|$ for circles ($\alpha = 0$) and $p_i = |d_i|$ for diamonds ($\alpha = 1$), and d_i is the horizontal distance between the centers of the object and the agent when their vertical separation goes to zero on the i^{th} trial. A total of 24 evaluation trials were uniformly distributed over the range of horizontal offsets.

Fourteen circle/diamond discrimination experiments were run with minor variations in the protocol described earlier and with different random seeds. Seven of these achieved an average performance of > 90% on the discrimination task. Additional experiments that involved discriminating a circle from a line were also run. The results of all of these experiments have been previously reported (Beer, 1996). In the present paper, we focus on the detailed analysis of the best circle/diamond discrimination agent that was evolved over all of these experiments.

4 Performance and Behavior

The best evolved agent achieved a mean performance of 99.83% on the 24 evaluation trials after 80 generations of evolution, with a mean performance of 97.08% on 10,000 random trials whose initial horizontal position was drawn from a uniform distribution over the range ± 50. The dependence of performance on the initial horizontal position of an object is illustrated in Figure 3. Note that there are very narrow spikes of poor performance near the centerline for both circles and diamonds. In general, due to the bilateral symmetry of the task, objects whose initial positions are very near the centerline are difficult to handle. For example, a diamond that begins exactly centered over the agent must be caught, because there is nothing to break the bilateral symmetry of the body, environment and nervous system. There are also two broader valleys of poor performance for circles. These fall between the locations of the 24 evaluation trials used during evolution.

The basic strategy that this agent uses to distinguish between circles and diamonds utilizes active scanning of the object as it falls (Figure 4A). Initially, the agent foveates any object by moving so as to center the object within its field of view. Next, the agent scans back and forth over the object. For a circle, the amplitude of the scan decreases as the object nears until it is centered, while the agent makes a large avoidance movement for a diamond. Note that, although the agent responds more quickly to a diamond than a circle, the first half of the motion is qualitatively very similar for both circles and diamonds (Figure 4B). However, the circle and diamond motions differ qualitatively at the second reversal (Figure 4B, arrow), with the agent moving back toward a circle, but pausing and then moving further away from a diamond. This suggests that the second reversal may be a key event in the discrimination, and something that we would like to better understand.

Note how the motion trajectories group into two bilaterally symmetric bundles depending on which side of the agent the object appears initially. There is one exception to this. For circles, the centermost trial on each side crosses over to join the bundle characteristic of initial positions on the opposite side. This pair of central trajectories corresponds to the inner two "fingers" of high performance that are separated from the more peripheral trials by the low-performance valleys mentioned above (Figure 3A). Thus, it appears that the central two trials are treated somewhat differently from the remaining trials, and the low-performance valleys may represent a transition region between these two strategies.

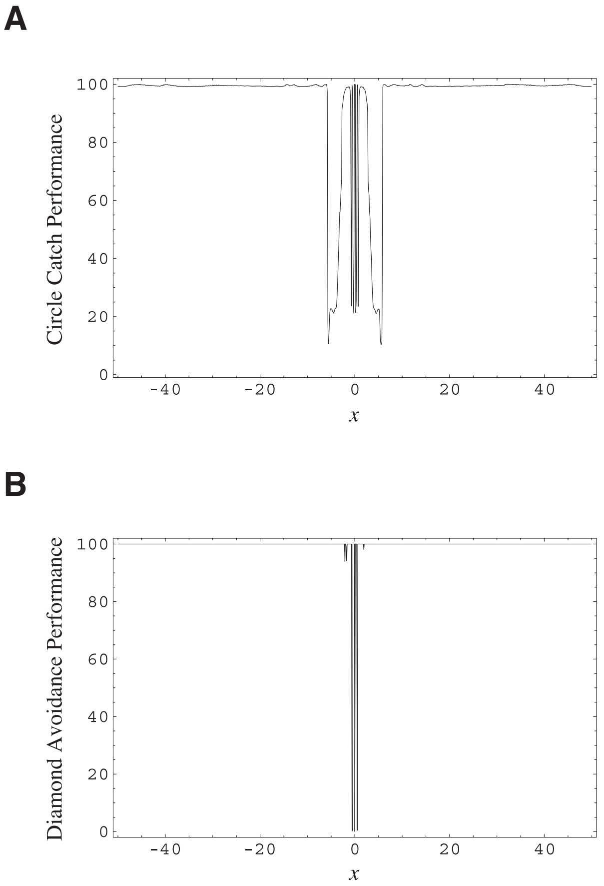

Figure 3 Generalization performance over initial horizontal position of the best-evolved agent. (A) Circle catch performance as a function of initial object location. (B) Diamond avoidance performance as a function of initial object location.

The agent's active responses to a falling object can be characterized by examining its reaction to objects held at fixed distances (Figure 5). The agent foveates any object placed slightly off center at large vertical distances (y = 200, 175). At intermediate distances (y = 150, 125, 100), the agent actively scans objects, with diamonds scanned at a higher frequency than circles. At closer distances (y = 75), a circle is

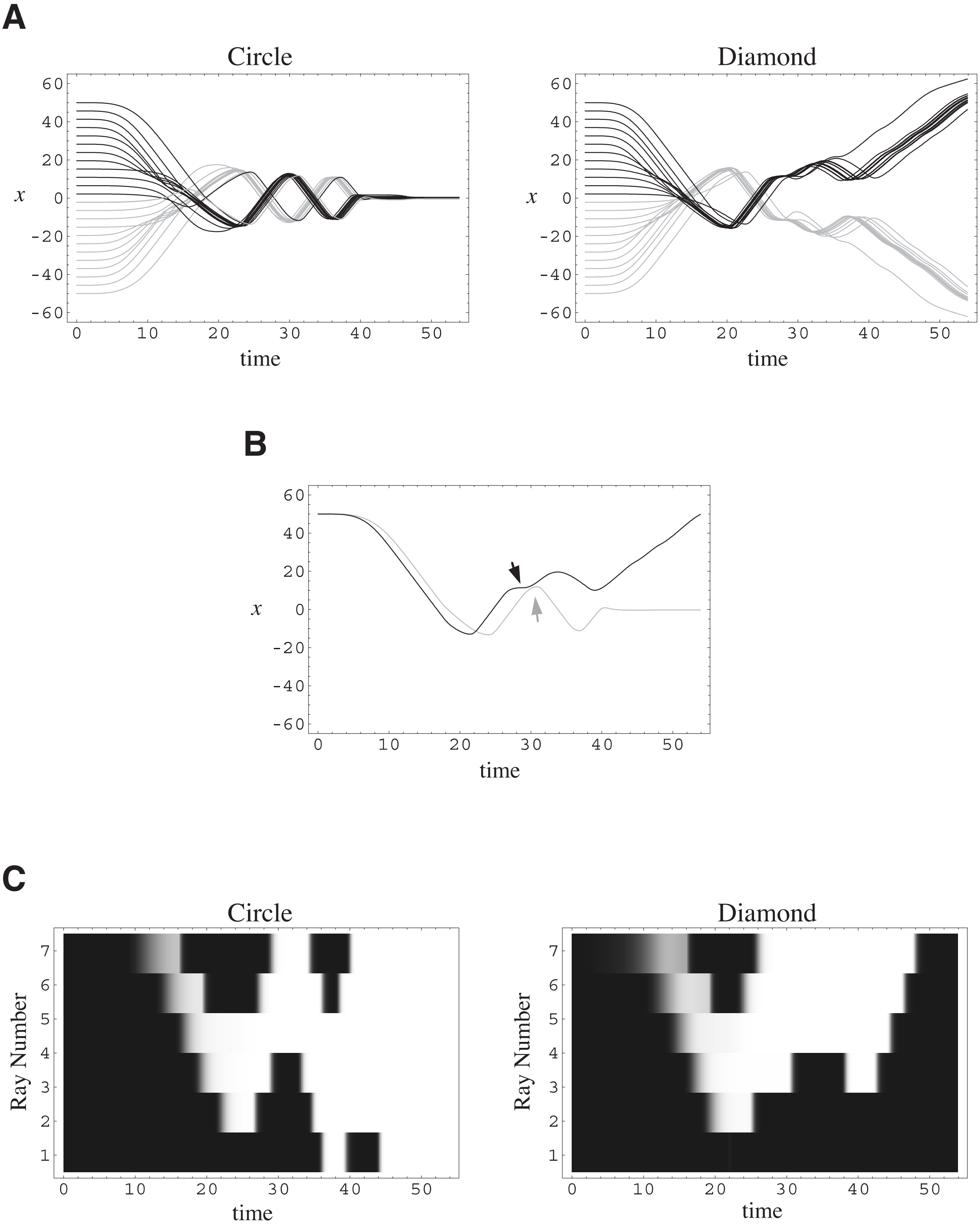

Figure 4 Behavior of the best evolved agent. (A) The agent's motion over time in object-centered coordinates for circles (left) and diamonds (right) beginning on the left side (gray) and the right side (black) of the agent. Note that the central most trajectory of the circle trials follows the path characteristic of trajectories from the opposite side. (B) A comparison of the motion for a single circle (gray) and diamond (black) trial from the same initial location. Arrows indicate the location of the second reversal in each trajectory. (C) What the agent "sees". The outputs of the seven ray sensors over time for the circle (left) and diamond (right) trials from part B are shown as density plots. The shading ranges from black for an output of 0 to white for an output of 1.

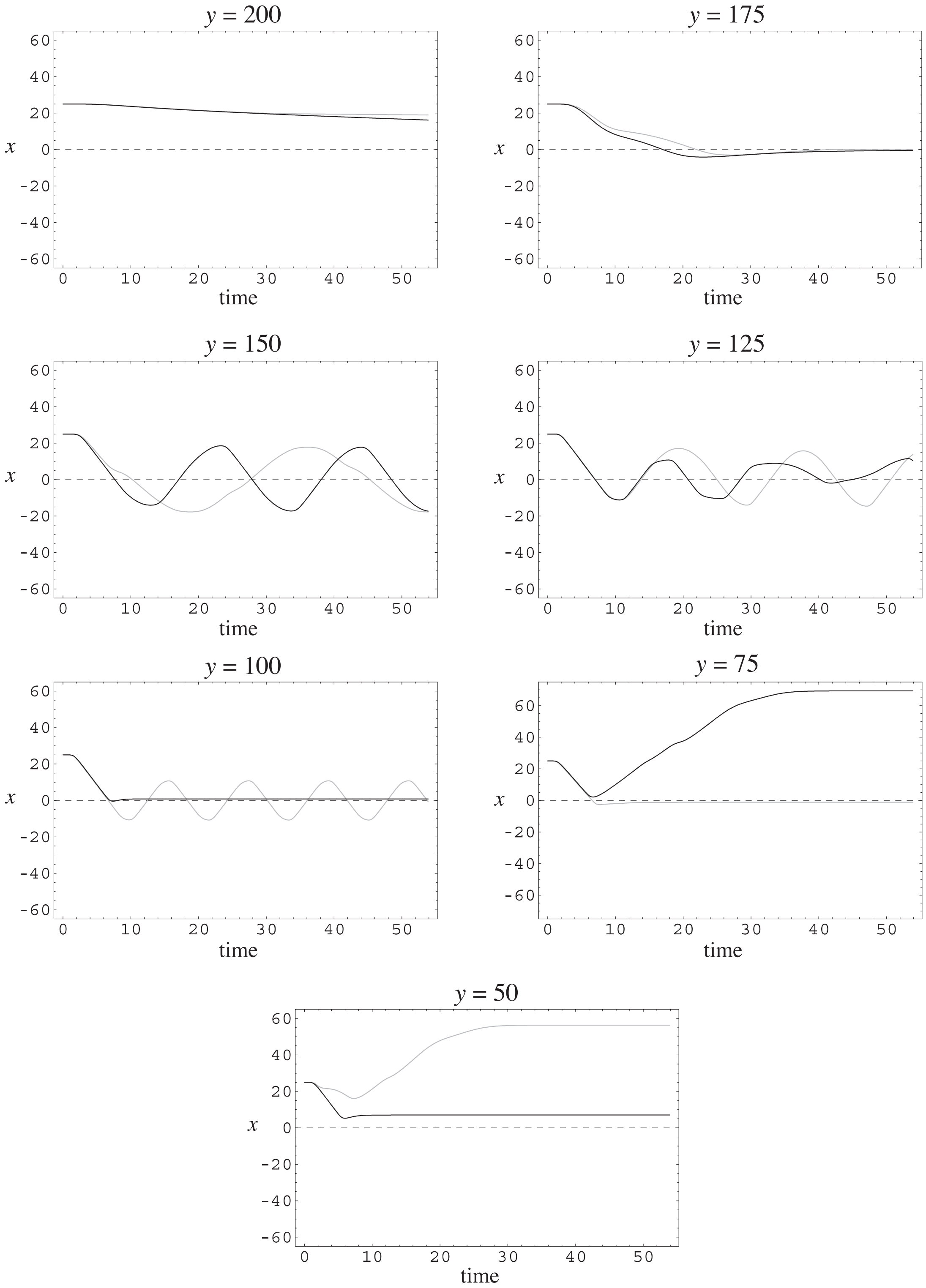

Figure 5 Active perception in the best evolved agent. The agent's motion over time in response to a circle (gray) or diamond (black) fixed at a range of vertical distances is plotted.

centered and a diamond is avoided. Finally, at even closer distances ($y = 50$), a diamond is centered and a circle is avoided. Of course, this is not a situation that the agent would normally find itself in, because by this point a circle would normally have already been centered and a diamond would normally have already been avoided.

The agent's strategy of actively scanning falling objects is an interesting one. Given the coarse spatial resolution of the agent's seven rays, it is likely that this scanning accentuates the small differences between a circle and a diamond. To see just how small these differences are, it is useful to visualize circles and diamonds from the agent's perspective (Figure 4C). While the difference between a circle and a diamond is obvious by the end of the interaction (because a caught circle fills the field of view while an avoided diamond leaves the field of view empty), it would be quite a challenge to perform this discrimination using only static snapshots of the visual array near the beginning of the behavior. Active control of body position and orientation has become an important theme in vision research, where it appears to considerably simplify otherwise intractable problems (Ballard, 1991; Churchland, Ramachandran, & Sejnsowski, 1994).

5 Psychophysics

5.1 Labeling and Discrimination

The two major defining characteristics of categorical perception are *labeling* and *discrimination* (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970; Ehret, 1987). Labeling requires that stimuli from a continuum are partitioned into distinct classes, whereas discrimination requires that stimuli are better distinguished between classes than within a class. For this agent, classification is manifested in its approach or avoidance behavior. In order to demonstrate labeling, the mean catch performance was plotted as the α parameter of a hybrid object smoothly varied from 0 (circle) to 1 (diamond). The resulting curve is sigmoidal in shape, with the category boundary occurring at approximately $\alpha = 0.3$ (Figure 6A). Below this value, objects are classified as circles and caught. Above this value, objects are classified as diamonds and avoided (i.e., their catch performance is 0). In order to demonstrate discrimination, the mean difference in catch perform-

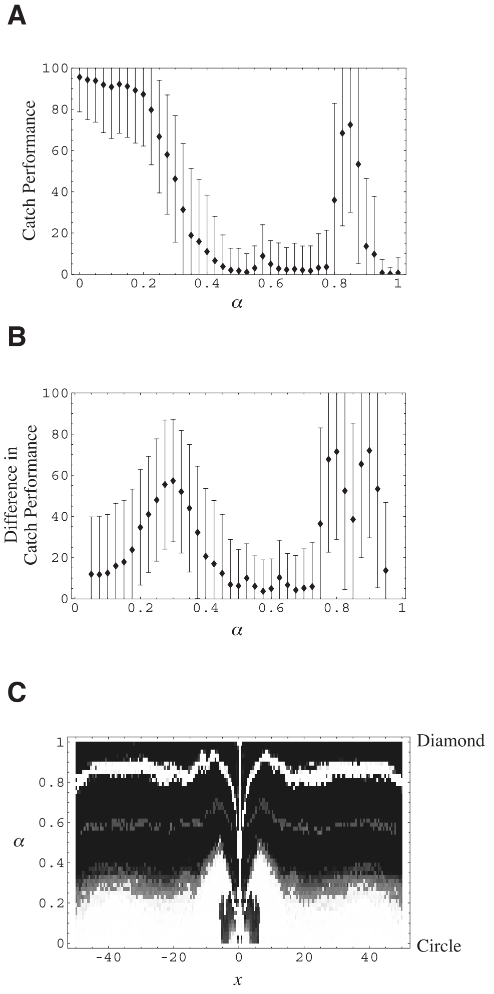

Figure 6 Demonstration of labeling and discrimination in the best evolved agent. (A) Average catch performance as a function of α shows a sigmoidal shape, with a transition from catching to avoidance behavior at about $\alpha = 0.3$. Note that there is also an anomalous peak of relatively high catch performance at about $\alpha = 0.85$. Each point represents mean value and standard deviation of 1000 trials, using initial horizontal offsets drawn from a uniform distribution over ± 50. (B) Average difference in catch performance for α values that differ by 0.1 as a function of α is bell shaped, with a peak discrimination at about $\alpha = 0.3$. Note that there is a double anomalous peak of high discrimination centered at about $\alpha = 0.85$. Each point represents mean value and standard deviation of 1000 trials. (C) A density plot illustrating the structure of the data underlying the plot in part A. The mean catch performance is plotted as a function of the initial horizontal position and the α value of an object, with black representing a catch performance of 0 and white representing a catch performance of 100%. The plot in part A can be recovered by averaging over x, while the plots in Figure 3 represent $\alpha = 0$ and $\alpha = 1$ slices through this plot.

ance for hybrid objects separated by a $\Delta\alpha$ of 0.1 was plotted as a function of α. The resulting curve is bell-shaped, with a peak discrimination occurring near the category boundary at $\alpha = 0.3$ (Figure 6B).

Note that both the labeling and discrimination plots are marred by an anomalous peak of high catch performance around $\alpha = 0.85$ (near-diamond shapes). Recall that the agent was only evolved to discriminate circles ($\alpha = 0$) and diamonds ($\alpha = 1$); it was never exposed to objects with intermediate values of α during evolution. Thus, the sigmoidal labeling curve and the bell-shaped discrimination curve were interpolated by the evolved network, and it is hardly surprising that mistakes might appear at intermediate values to which the agent was never exposed during evolution. As will be demonstrated later, this anomalous peak provides an important clue to the organization of the agent-environment dynamics that has evolved.

Finally, note that there is a fair amount of variability in both the labeling and discrimination plots. We already know that there are two valleys of avoidance behavior near the centerline for circles (Figure 3A) and that there are extremely narrow spikes of catch behavior near the centerline for diamonds (Figure 3B). Such misclassifications clearly contribute to variability in the mean performance statistics. In order to obtain a more detailed picture of the structure underlying this variability for general α, we can plot catch performance as a function of the initial horizontal offset of an object as α is varied from 0 to 1 (Figure 6C). In this plot, light bands correspond to catching behavior and dark bands correspond to avoidance behavior. Both the category boundary around $\alpha = 0.3$ and the anomalous peak at about $\alpha = 0.85$ have a very similar structure, with central and peripheral horizontal offsets switching behavior first as α is increased, followed by horizontal offsets at about ± 20, and finally intermediate offsets. A faint peak near $\alpha = 0.6$ also exhibits a very similar structure (this small peak can also be seen in Figure 6A).

5.2 What Features Matter to the Discrimination?

What object features does the evolved agent use to make its categorization? Since the size of the circle is such that it can be inscribed within the diamond, the two objects differ only in the four corners shown in Figure 7A, only three of which are visible to the agent as the object falls. A coarse determination of the features important to discrimination can be made by examining the agent's response to the eight possible combinations of these three features. As the Figure shows, the presence of a single corner on either side of the object is sufficient to trigger avoidance, while the presence or absence of the corner facing the agent appears to make very little difference to the agent's response. This suggests that object width is the major determining factor in discrimination.

In order to test this hypothesis, we can examine the agent's response to a variety of different objects whose width varies from 30 (the width of the circle) to $30\sqrt{2} \approx 42.43$ (the width of the diamond) (Figure 7B). In all cases, the resulting labeling curves are qualitatively similar to Figure 6A (shown as a solid black curve in Figure 7B for reference), with the category boundary occurring at a width of 33.1 (SD 0.6) and the anomalous peak occurring at a mean width of 40.3 (SD 1.0). Note that, as would be expected, the labeling curve of hybrid objects is more similar to that for circles (dark gray curve) for α values closer to 0 and it is more similar to that for diamonds (dotted curve) for α values closer to 1. The location of the category boundary varies in a consistent way with type of object, suggesting that the agent's behavior is also somewhat sensitive to the amount that an object protrudes toward the agent.

5.3 When is the Decision Made?

At what point does the agent make its decision to catch or avoid an object? We can probe this question by switching the identity of the object at different points in time and asking when the agent has committed itself to a response. Figure 8A shows how the final catch performance at a single horizontal offset varies as a circle is switched to a diamond at later and later times (left) and how the final avoidance performance at a single horizontal offset varies as a diamond is switched to a circle at later and later times (right). For example, a high value in the circle plots (left) means that, even though the circle is switched to a diamond at that point, the agent still responds as if it were a circle and was thus already committed to catching it. However, the multiple sharp changes in performance that occur in both plots as the time of the switch is varied demonstrate that the agent's behavior remains quite sensitive to object identity until very late in the

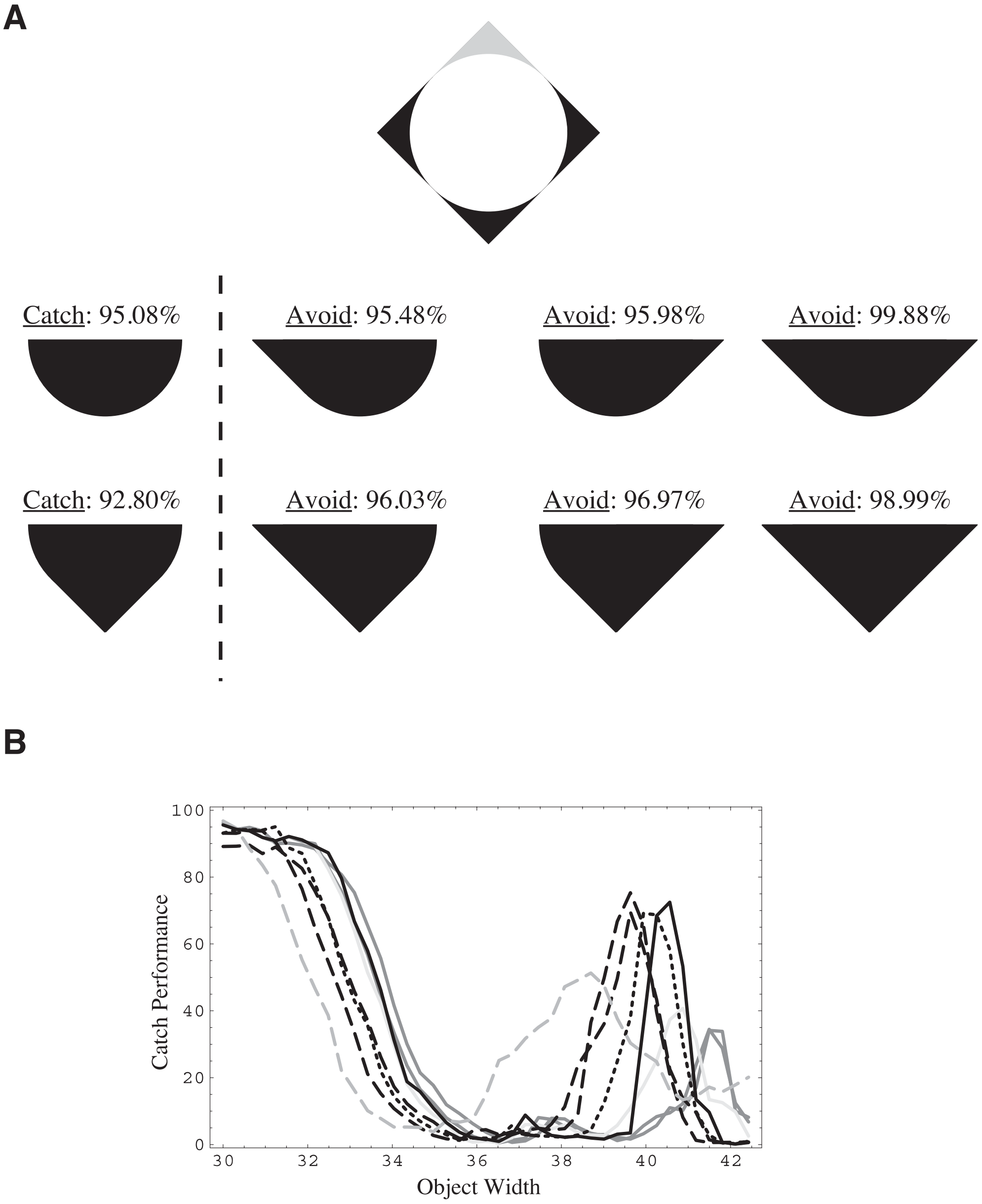

Figure 7 A demonstration that object width is the primary feature used by the best evolved agent to discriminate circles from diamonds. (A) The agent's performance on objects possessing all possible combinations of the three discrete features (shown in black at top) that distinguish a circle from a diamond. Note that the agent only catches objects with no left or right corners; the presence of either one or both left/right corners leads to avoidance. (B) Variation in catch performance with object width for a variety of different objects, including a horizontal line (dashed gray curve), wedges whose height is the same as the circle and the diamond (dashed black curves), diamonds (dotted black curve), the hybrid objects, R_H (ϕ; α) (solid black curve), ellipses whose height is the same as the circle (solid light gray curve), and circles and ellipses whose height is the same as a diamond (solid dark gray curve). Note that, as their width is varied, all objects produce a qualitatively similar response in the agent.

interaction. The plots for other horizontal offsets exhibit the same general pattern (Figure 8B), and some features of this pattern correlate closely with the agent's behavior (Figure 8C). For example, the brief peak in catch performance for circles at about $t = 33$ and the sharp rise in avoidance performance for diamonds at about $t = 29$ coincide with the second reversal.

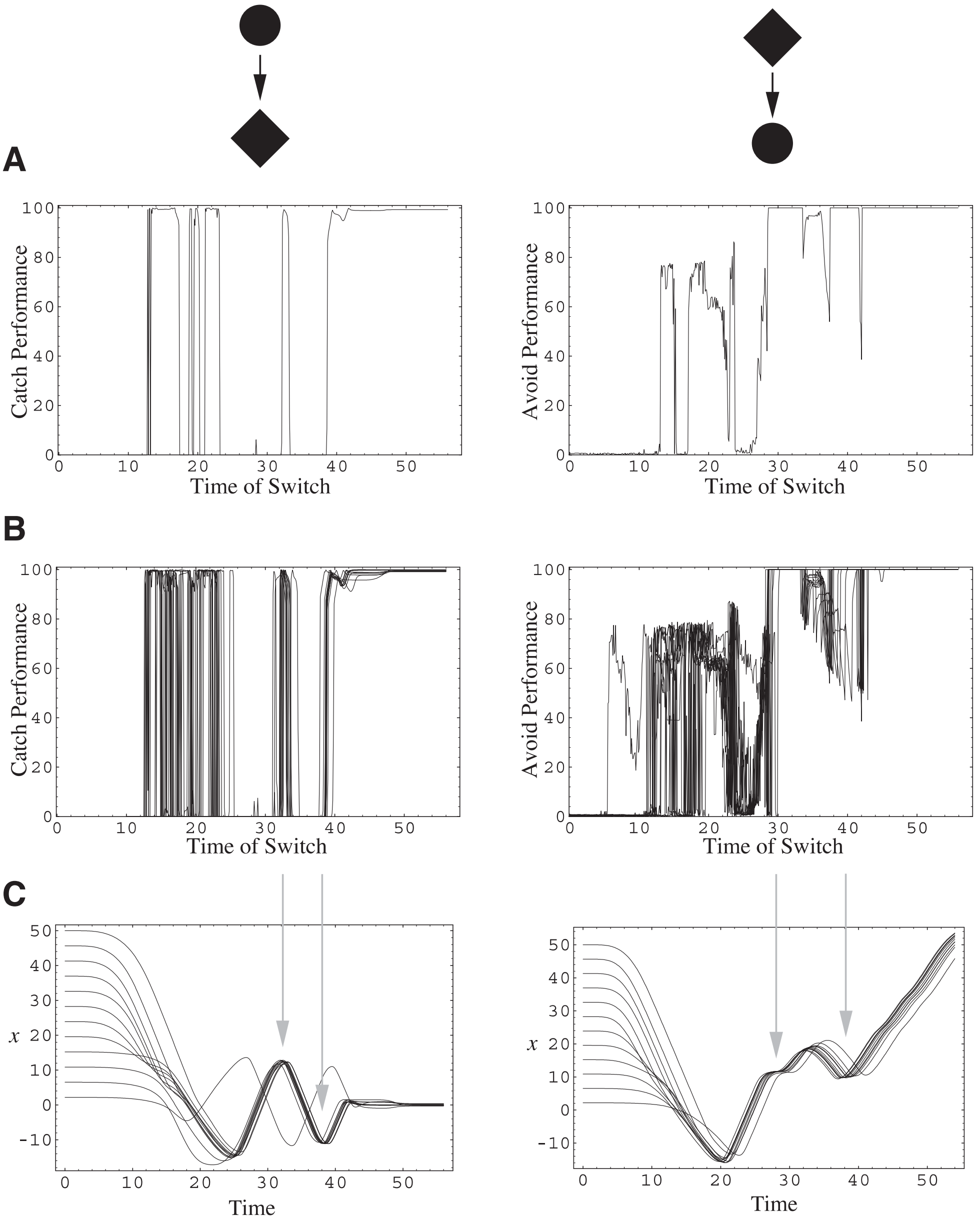

Figure 8 The effect of switching object identity at different times on the final decision to catch or avoid an object in the best evolved agent. (A) Catch performance as a function of the time of switch from a circle to a diamond (left) and avoidance performance as a function of the time of switch from a diamond to a circle (right) for a single initial horizontal offset. (B) Catch (left) and avoidance (right) performance as a function of time of switch for twelve different initial horizontal positions. In both cases, note that the agent's sensitivity to switches in object identity appears to vary considerably throughout the interaction. (C) The normal behavioral trajectories for the initial horizontal offsets used in (B). Note the correspondence between features of the behavioral trajectories and features of the performance plots in (B).

Taken at face value, these results suggest that the "decision" is repeatedly made and unmade as the agent and the object interact. For example, the agent seems to be completely uncommitted during the initial foveation of a circle, but is then largely committed to catching during the first scan (although there are times when this commitment wavers). The agent is then uncommitted during subsequent scans, but committed to catching at each reversal. By the final reversal near $t = 38$, the agent is largely committed to catching the

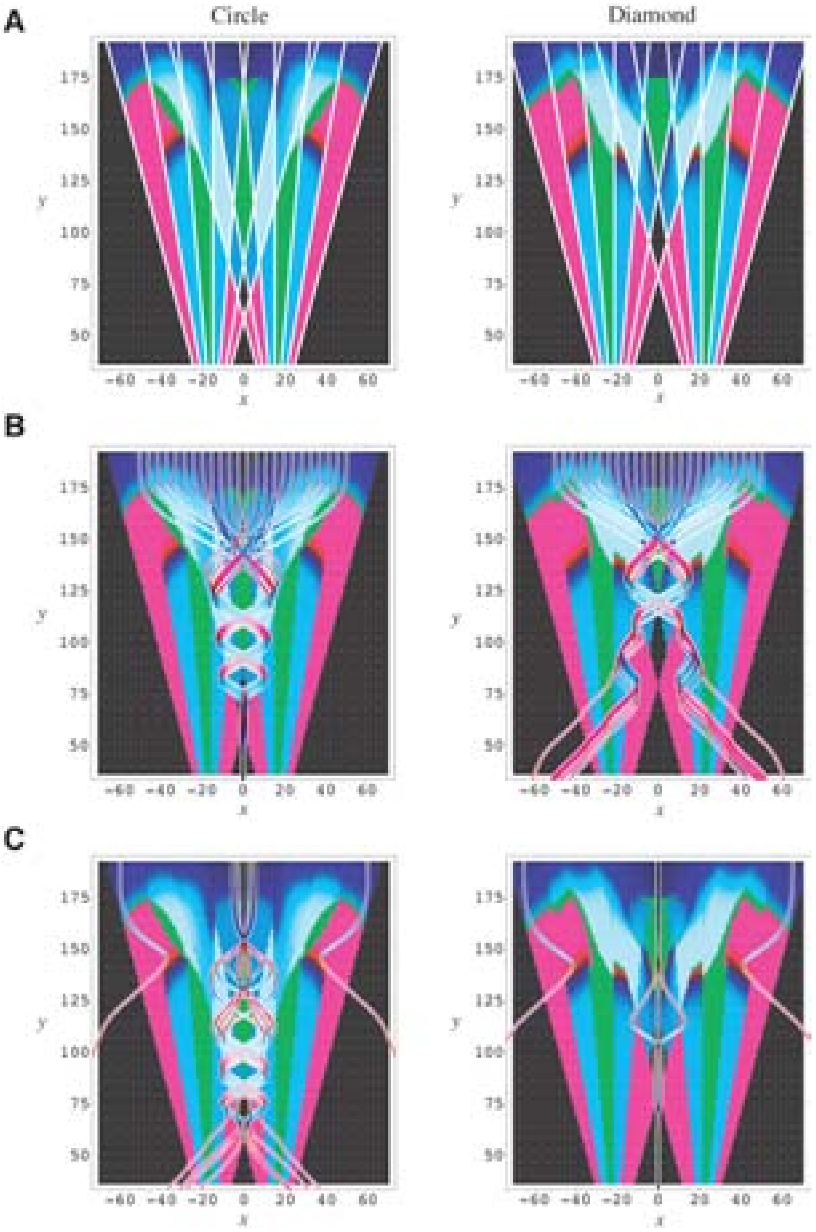

Plate 1 The dynamics of the best evolved agent. (A) Steady-state horizontal velocity fields for circles (left) and diamonds (right). Each point represents the long-term horizontal velocity that the agent would eventually adopt if an object were fixed at that location in its field of view. These fields correspond to an $\dot{x}$ projection of the coordinates of the equilibrium points of the evolved continuous-time recurrent neural network as a function of x and y. Regions in which $\dot{x}$ is directed toward the midline (centering regions) are colored blue, whereas those in which $\dot{x}$ is directed away from the midline (avoidance regions) are colored red. The magnitude of $\dot{x}$ is represented by the intensity of a color, with $\dot{x} = 0$ represented as black. Green regions are those in which multiple stable equilibrium points exist. The straight white lines indicate boundaries across which the object intersects different numbers of rays. These ray tangencies divide the visual field into disjoint regions of smooth dynamics. (B) The fields from part A with the movement trajectories from Figure 4A superimposed. The trajectories are color-coded with instantaneous horizontal velocity using the same scheme as in part A. If the interaction were frozen at any point along a trajectory, the color of that point would eventually approach the color of the corresponding point of the steady-state field. However, since both the agent and the object are moving, the instantaneous velocity lags behind the steady-state velocity. (C) Some additional example trajectories. The inner trajectories are misclassifications taken from the regions of low performance in Figure 3. The outer trajectories demonstrate that all sufficiently peripheral objects are avoided.

object. The pattern is less clear-cut for diamonds. However, by the time of the pause near $t = 27$, the agent is fairly committed to avoiding the object (although there are times during the final scan when this commitment wavers). Thus, it is not until the expression of the decision is nearly complete that a stable decision can really be said to have been made. These results suggest that the generation of a catch or avoidance response is a complex, temporally extended process, and it makes little sense to talk about the agent having made a decision before its behavioral expression of that decision. A similar view of decisions as temporally extended processes shaped by both sensory signals and neuronal state has begun to emerge from the decision-making literature (Townsend & Busemeyer, 1995; Platt, 2002).

6 The Dynamics of the Coupled Agent–Environment System

Let us now turn to an analysis of the dynamics underlying the behavioral and psychophysical phenomena described in the previous two sections. The first step in such an analysis is to characterize the dynamical structure of the entire coupled system formed by the evolved nervous system, the agent's body and the falling object (Figure 1). There are a number of questions that we would like to answer about the dynamics of this agent–environment system. How is its dynamics structured so that almost all circle trajectories result in centering, while almost all diamond trajectories lead to avoidance? How does this structure change as α increases so that the agent's response changes from catch to avoid, resulting in a category boundary between circles and diamonds at about $\alpha = 0.3$? What causes the anomalous peak at about $\alpha = 0.85$?

From the equations describing the coupled agent–environment system (equation 1), we can make several general observations about its dynamics. First, the system is 16-dimensional, because 16 separate differential equations are required to uniquely describe its motion. Second, this system has no closed limit sets such as equilibrium points or limit cycles because the $\dot{y} = -3$ equation means that y is always decreasing. Third, discontinuities are introduced into the dynamics through the ray sensory neurons (neurons one through seven) because $I_i(x, y; \alpha)$ is a discontinuous function of x and y. There are locations where an infinitesimal change in position can cause an object that was not intersecting a ray to suddenly intersect it, or vice versa. These ray tangencies divide the state space into discrete cells of smooth dynamics which depend on the size and shape of the object (see Plate 1), making the agent a hybrid dynamical system (Johnson, 1994). One final general observation that we can make is that the coupled agent-environment system is actually a one-parameter family of systems, because changing the object from a circle to a diamond amounts to changing the α parameter from 0 to 1.

In order to say anything more specific about this coupled dynamics, we must examine its detailed structure. While it is not possible to directly visualize the trajectories of such a high-dimensional system, carefully chosen projections can provide important insights into this structure. By examining various projections of the dynamics, it was discovered that combinations of x, y and an interneuron are most useful. The three-dimensional projection (x, y, o_9) leads to particularly clear visualizations of the dynamics. Interestingly, this projection includes an environmental state variable (the vertical position of the object), a body state variable (the horizontal position of the body relative to the object), and a neuronal state variable (the output of interneuron 9).

In the (x, y, o_9) projection, the trajectories begin as a line of initial states and move to decreasing y values as the object falls (Figure 9). When the leading edge of the object reaches the leading edge of the agent, a fitness value is assigned based on their horizontal separation. By adjusting network parameters, the evolutionary algorithm has been asked to structure the dynamics of the entire coupled system so that the line of initial states is concentrated near $x = 0$ by the end of the trial for circles, but is dispersed to large positive or negative values of x for diamonds.

Figure 9 shows the structure of the 24 evaluation trials as α is varied from 0 to 1, with the trajectories shaded according to final catch performance so that it is easier to distinguish those that are caught from those that are avoided. Note that the motion trajectories shown in Figure 4a are simply the projection of the trajectories shown in Figure 9 onto the (x, y) plane. As expected, there are no closed limit sets and the trajectories merely flow from left to right. However, they develop quite a bit of structure as the interaction unfolds. In particular, note how the line of initial states is separated into two distinct bundles of trajec-

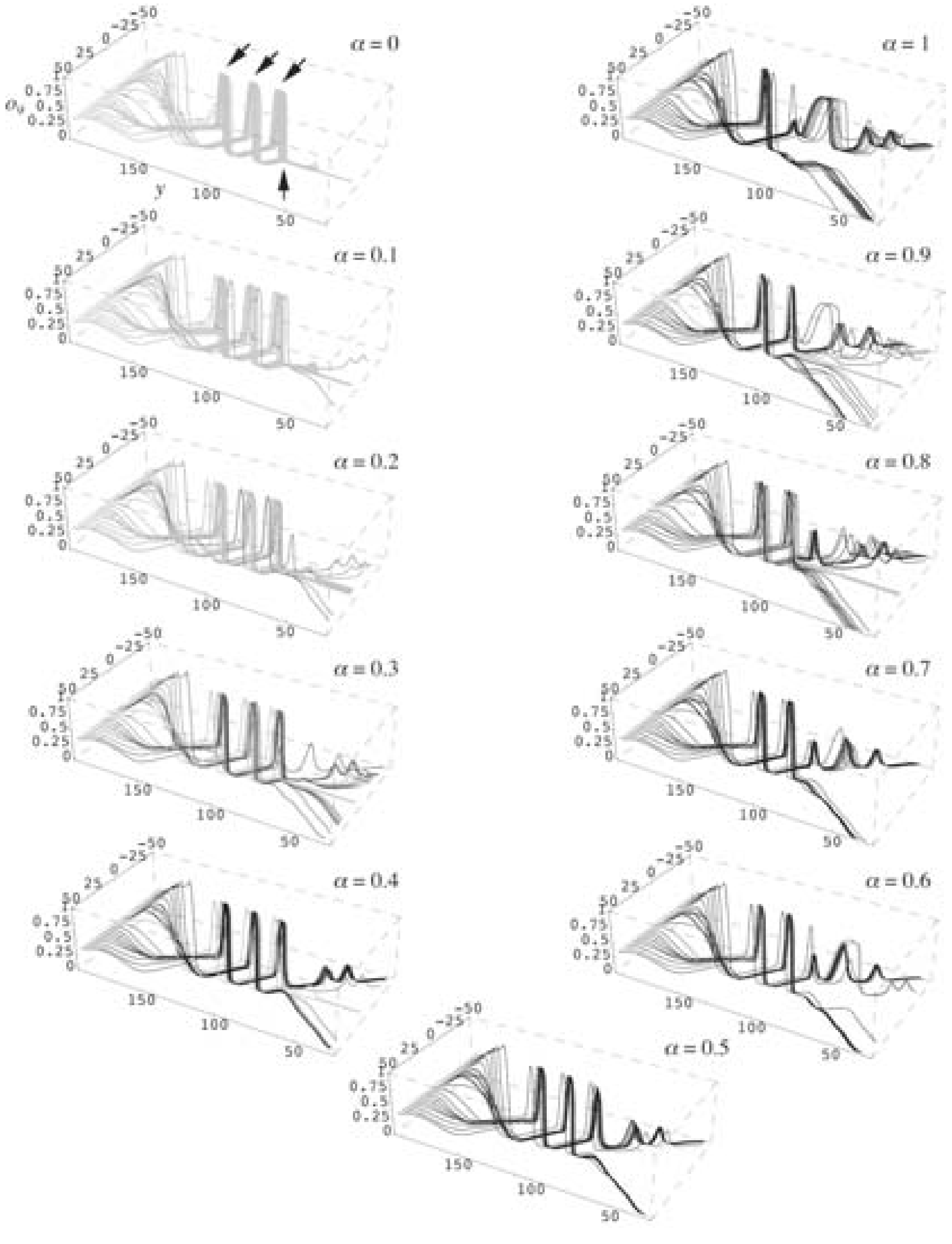

Figure 9 The dynamics of the coupled agent–environment system. The (x, y, o_9) projections of twenty-four trajectories of the system are shown. The plots are arranged in a counter-clockwise direction as α varies from 0 (circle) to 1 (diamond). The trajectories are shaded according to catch performance, with lighter shades representing high catch performance and darker shades representing low catch performance. Note that the trajectories are organized into two bundles that wind around one another. For small values of α, these bundles collide on the axis of bilateral symmetry and the objects are caught. As α increases, the bundles progressively unwind. As can be seen in the $\alpha = 0.8$ and $\alpha = 0.9$ plots, the anomalous peak in Figure 6 corresponds to the two bundles colliding at their third crossing.

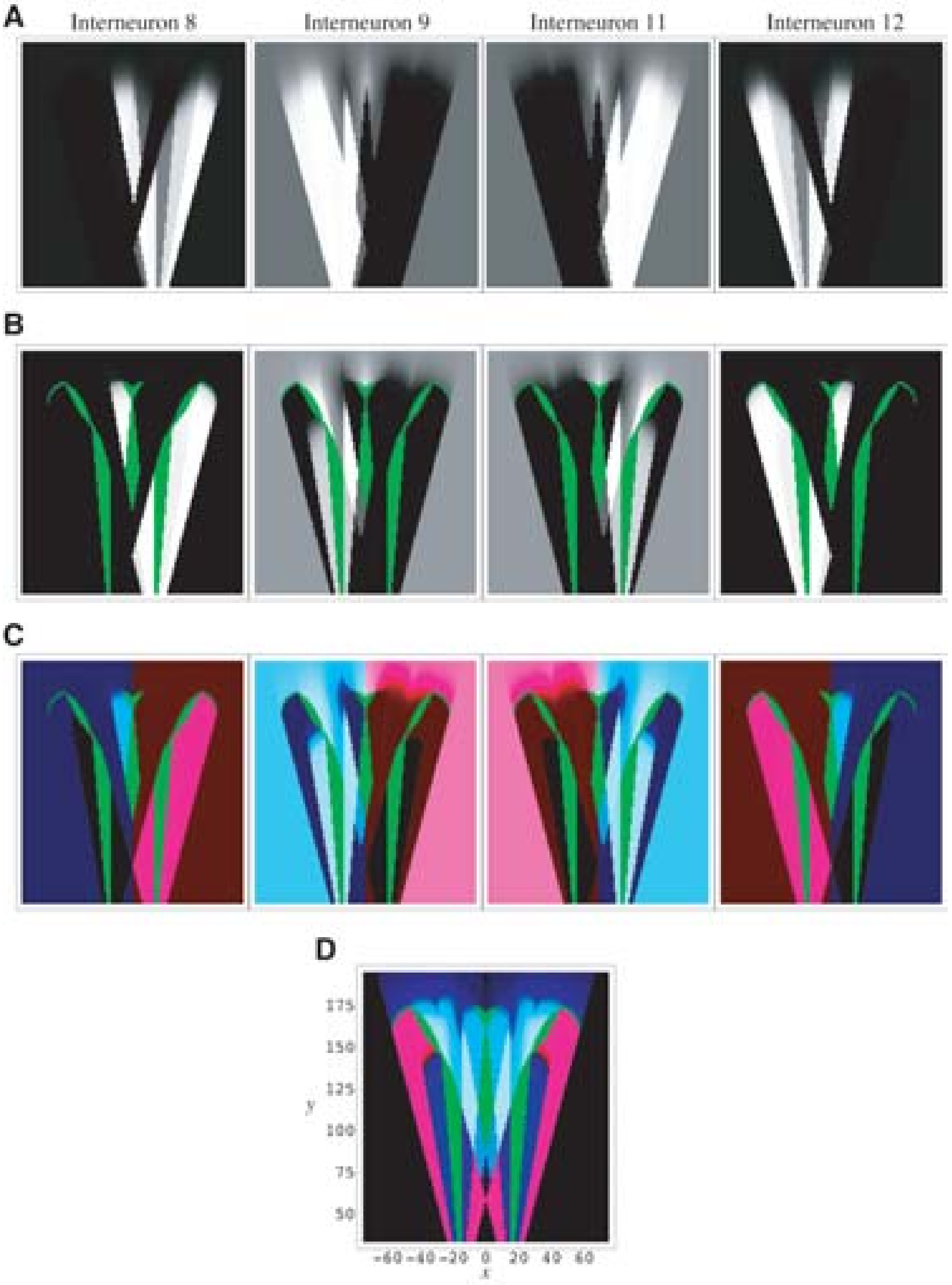

Plate 2 The circle steady-state response fields of the bilateral pairs of interneurons 8/12 and 9/11. (A). A density plot of the feedforward response fields of each interneuron in response to a circle placed at each point in the agent's field of view. Here only the feedforward connections from the sensory neurons to the interneurons are considered. White regions correspond to an output of 1 while black regions correspond to an output of 0. (B) A density plot of the interneuron response fields when the recurrent interactions are also taken into consideration. Note that intensity of the feedforward response fields is modified in places and that regions of multistability appear (green). (C) The steady-state horizontal velocity fields of each interneuron, color-coded in the same way as in Plate 1. These fields show the steady-state horizontal velocity produced by each interneuron in response to a circle placed at each point in the field of view. (D) The net steady-state horizontal velocity field formed by summing the four fields in Part C. Note that this summed response field closely matches the steady-state horizontal velocity field shown in Plate 1A (left).

tories that wind around each other several times. Each time the agent scans to the right, the output of interneuron 9 develops a peak (three arrows at top of $\alpha = 0$ plot) that allows the bundle to cross over the other bundle's corresponding scan to the left. Note that the trajectory bundle resulting from objects that initially appeared to the right of the agent exhibits two peaks, while the bundle resulting from objects that initially appeared to the left exhibits only one. Of course, due to bilateral symmetry, plots of the (x, y, o_{11}) projection would be mirror images of these, with the peaks occurring during scans to the left and the order and number of peaks for objects initially on the left of the agent reversed as compared with objects initially on the right.

How do these plots help us to answer the questions raised at the beginning of this section? Let us begin by examining the circle trials ($\alpha = 0$). After crossing over one another three times, the two circle trajectory bundles appear to "collide" and merge at the trajectory of bilateral symmetry rather than crossing a fourth time (arrow at bottom). Note that one must be careful in drawing such a conclusion from a single lower-dimensional projection. However, examination of other projections verifies that the two trajectory bundles do indeed interpenetrate at this point (since the agent dynamics is non-autonomous, distinct behavioral trajectories can pass through the same neuronal state). From the point of this collision on, both bundles follow the trajectory of bilateral symmetry (the trajectory that the dynamics follows in response to an object dropped at the midline), resulting in accurate catches. As α increases, the trajectories appear to unwind. While they retain their basic shape, they shift to larger values of y so that the peaks occur earlier in time. As the bundles shift earlier in time with increasing α, more and more trajectories miss the final collision and instead diverge from the trajectory of bilateral symmetry until, by $\alpha = 0.5$, the two bundles do not intersect at all. As α continues to increase, the two trajectory bundles continue to shift earlier in time until they begin to collide at the third crossing at $\alpha = 0.8$, resulting in catches that produce the anomalous peak. As α increases even further, the bundles shift to the point where, by $\alpha = 1$, they cross only twice before diverging.

These observations immediately suggest a prediction. From our psychophysical analysis, we know that the most important effect of increasing α is to increase the width of the object. Assuming that the two trajectory bundles continue to unwind with increasing object size, we would expect that two additional collisions should occur as the bundles intersect at the second and then the first crossing. This reasoning predicts that two additional anomalous peaks in catch performance should be observed for objects wider than the $\alpha = 1$ diamond. As shown in Figure 10A, this is indeed the case. The catch performance is plotted for diamonds ranging in width from 30 to 100. From 30 to 42.4, the plot is a compressed version of the dotted black line in Figure 7B, with the anomalous peak appearing as a narrow spike. From a width of about 47.7 to about 54.5, there is a second short anomalous peak. Finally, beginning around a width of 59, there is a final slow increase in catch performance. Figure 10B shows how a collision of the two trajectory bundles at the second crossing is responsible for the second short peak ($w = 51.2$) and a collision at the first crossing is responsible for final broad peak ($w = 66$), as predicted. Note that, between these two anomalous peaks, the bundles have unwound past the second crossing ($w = 54.8$). Beyond the start of the last anomalous peak, the trajectory bundles have unwound completely ($w = 89.6$).

7 The Analysis of Agent–Environment Interaction

The characterization of the overall dynamical structure of the coupled agent–environment system in the previous section certainly provided some explanatory and even predictive insight into this agent's behavior. However, it could be argued that this characterization is primarily descriptive in nature, detailing what the structure is without explaining how it arises. The second step in a dynamical analysis is to examine how this structure emerges from the interaction between the agent and its environment. What produces the active scans? What causes the two trajectory bundles to develop and to have the structure that they do? Why do circle trajectories move back toward the midline at the second reversal, while diamond trajectories merely pause and then continue moving away? What happens when the two bundles collide at the trajectory of bilateral symmetry? What causes the two bundles to unwind with increasing α? Why is object width the determining factor in discrimination? What underlies

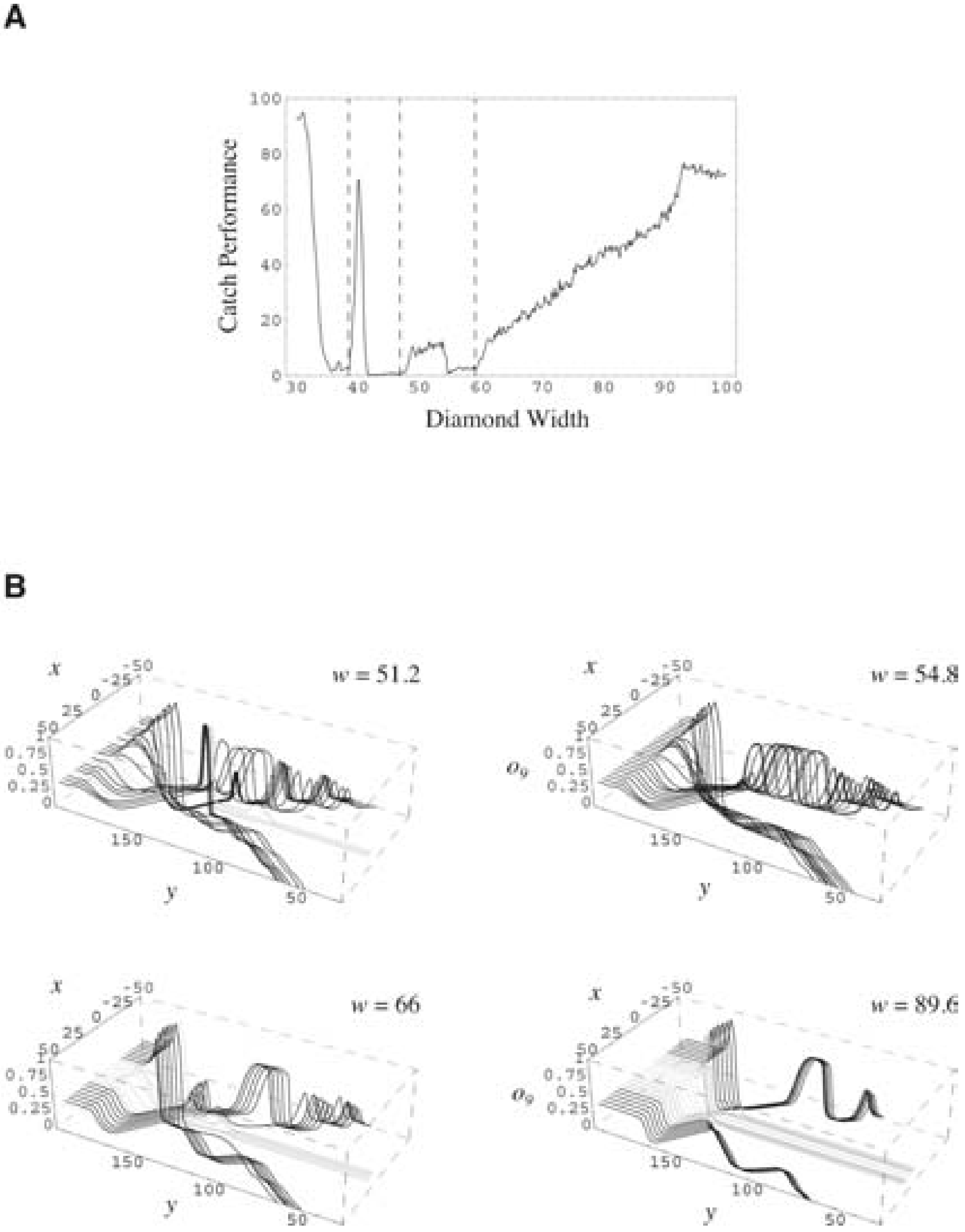

Figure 10 Confirming a prediction of additional peaks as object width increases. (A) A plot of catch performance for a diamond whose width varies from the width of a circle to more than twice the width of a normal diamond. Note that additional anomalous peaks occur. The dashed vertical lines correspond to the predicted widths at which anomalous peaks should begin based on the reasoning described in the text. (B) Trajectory plots for diamond widths within the third peak ($w = 51.2$), between the third and fourth peak ($w = 54.8$), and at two different points within the fourth peak ($w = 66$ and $w = 89.6$). The format is the same as in Figure 9. Note that the two bundles continue to unwind with increasing object width and that the two new anomalous peaks correspond to collisions occurring at the second or first crossing.

the misclassifications that occasionally occur? Why is the moment of decision to catch or avoid so difficult to pin down?

7.1 Steady-State Effects

Once again, we have many different choices as to how to decompose the dynamics of the agent–environment system and how to visualize it. Clearly, whatever decomposition we choose should emphasize the factors underlying movement, since the agent's decision is expressed in its motion over time and it is this behavior that we want to explain. Thus, we will decompose the agent–environment dynamics into: (1) the effect that the relative positions of the object and the agent have on the agent's motion; (2) the effect that the agent's motion has on the relative positions of the object and the agent.

Imagine a particular object fixed at some (x, y) location in the agent's field of view. This specifies a constant set of sensory inputs to the evolved neural circuit as described by the $I_i(x, y; \alpha)$ terms in equation 1. For constant inputs, the neural circuit is an autonomous dynamical system. It turns out that the only limit sets exhibited by this circuit are equilibrium points, whose positions and stabilities we can calculate. Although each stable equilibrium point is 14-dimensional, the only quantity that directly affects the agent's motion is the horizontal velocity $\dot{x}$, which is determined by o_{13} and o_{14} as described in equation 1. If we wait long enough after fixing the object in the agent's field of view, the state of the neural circuit will be found very near one of its stable equilibrium points. Thus, for each point in the agent's field of view, we can plot a representation of the steady-state velocity that would be produced by an object located at that point (Plate 1A). These points are color-coded according to whether the steady-state $\dot{x}$ is directed toward (blue) or away (red) from the centerline. The intensity of the color represents the magnitude of $\dot{x}$, with an $\dot{x}$ of zero colored black. Green points indicate object locations where the neural circuit has more than one stable equilibrium point. In this case, the long-term horizontal velocity is not uniquely determined by object position alone, but depends on the system's state. Note the straight, sharp borders between regions of different color that radiate at several angles from the bottom of the plots. These correspond to the ray tangencies described earlier (white lines in Plate 1A), which separate the regions of smooth dynamics where the object intersects different subsets of rays.

Superimposed on these steady-state velocity fields are the trajectories of motion of the object through the agent's field of view for the 24 trials used during evolution (Plate 1B). These trajectories are color-coded according to their instantaneous horizontal velocity using the same color scheme described earlier. If the interaction were frozen at any point along a trajectory, then the instantaneous velocity would approach the steady-state value over time as the internal state of the neural circuit approached the corresponding stable equilibrium point. However, since both the agent and the object are moving, the instantaneous velocity lags behind the steady-state value. As we shall see, this subtle interplay between sensory input and internal state is crucial to accurate discrimination.

How can these plots be used to understand the agent's behavior? Consider the circle trials shown in Plate 1B (left). The most striking feature of this plot is how the blue centering regions repeatedly turn the trajectory bundles back toward the midline until they become trapped in the central black regions of no horizontal movement. Initially, all trajectories begin in dark-blue regions corresponding to equilibrium points that attract the agent's velocity toward the center, slowly at first, and then with increasing velocity as the regions brighten. The trajectories overshoot the midline due to the lag mentioned above between the actual internal state of the agent and the changing equilibrium points that are attracting it (it takes some time for the neurons to respond to a sudden change in sensory input). This lag is also the reason that the brief passage of the outermost trajectories through a red region near the top of the plot has little effect on them. Note that the trajectories turn red as soon as they cross the midline. This is because, although a motion of the object toward the right is a centering movement when the object is on the left side of the midline, it becomes an avoidance movement when the object is on the right side of the midline and vice versa. As these now red trajectories pass through blue centering regions on the opposite side of the midline, they eventually turn blue and move back toward the center. This overshoot and turning back toward the midline occurs twice more. Thus, the active scans are produced by the interplay between the steady-state horizontal velocity field and a kind of "neuronal inertia" that causes the agent to repeatedly overshoot the midline. If this inertia is not allowed to build up (e.g., by decreasing the time constants of the interneurons), then the agent's discrimination performance drops sharply.

The last time the trajectory bundles approach the midline, they pass briefly through a red avoidance region before moving into the central black region and becoming trapped. The black region corresponds to object positions for which there exists a single stable equilibrium point whose $\dot{x}$ projection is 0. If the agent had sufficient neuronal inertia at this point, it would pass right through this region. In fact, there is a slight overshoot of the midline. However, this overshoot is not sufficient to carry the agent out of the black region. As the object continues to fall, the trajectories briefly intersect the corners of the central red regions, causing a small transient movement that actually serves to further center them (we will examine such transient effects shortly). Thus, what happens when the two trajectory bundles collide is that they run out of neuronal inertia in the upper black diamond-shaped region of the steady-state horizontal velocity field, leaving neither sensory inputs nor internal state to initiate further horizontal motion.

Now consider the diamond trials shown in Plate 1B (right). Although the details are somewhat different, the upper half of the steady-state horizontal velocity field and the initial motion of the trajectories is similar to that for circles. The most striking difference from the circle trials is that the central red regions of the steady-state horizontal velocity field in the lower half of the plot (and the black regions separating them) are much larger for diamonds than for circles, and penetrate much higher into the agent's field of view. These red avoidance regions have the effect of pushing the trajectory bundles away from the midline after the first scan, so that they never pass into the central black regions. Whereas the circle trajectories pass through only blue regions at the second reversal, turning them back toward the midline, the diamond trajectories pass through the top of the expanded central red region, causing only a brief pause rather than a reversal. The diamond trajectories then enter a blue centering region that pushes them back toward the midline, but the large central red region again pushes them away, resulting in a final avoidance. Because the size and location of the central regions are determined by the width of the object (compare the left and right sides of Plate 1A), it is easy to see how wider objects unwind the trajectory bundles (Figure 9) as the expanding red avoidance regions push the trajectories away from the midline earlier and earlier in the interaction.

Using this decomposition of the agent-environment dynamics, we can account for the shape of the catch performance plot (Figure 6A) as follows. As the width w of an object increases, the regions of smooth dynamics (delimited by the white lines in Plate 1A) enlarge and shift upwards (compare the left and right sides of the Figure), causing the central black region to intersect each of the four midline crossings in turn. Catch peaks begin whenever the trajectory bundles pass through a sufficiently wide portion of the black region that they become trapped. At $w = 30$, the original peak occurs when the mean vertical offset of the fourth midline crossing falls approximately 13.6 below the top of the black region. For a diamond, the ray tangencies occur at $y = m_i(x \pm w/2)$, where m_i is the slope of the i^{th} ray. By solving for the ray tangency intersections that define the upper central black region, we find that its top occurs at $y = m_6 w/2$. If we solve for the object widths at which the other three midline crossings occur 13.6 units below the top of the central black region, we obtain quantitative predictions for where the three anomalous peaks should begin. As the dashed lines in Figure 10A show, these predictions are fairly accurate. Catch peaks end when the red avoidance region surrounding the lower portion of the central black diamond become sufficiently large that the trajectories are pushed away before they ever enter the black region. This happens only three times because, for sufficiently wide objects, the central lower black triangle-shaped region becomes so large that no trajectories can avoid becoming trapped within it.

This decomposition of the agent–environment interaction can also explain a number of other features of the agent's behavior. Misclassifications occur when the agent's internal state gets sufficiently out of phase with the steady-state velocity field. For example, in the valleys of low circle performance (Figure 3A), the scan is delayed to the point where these trajectories pass through enough of the red avoidance region to carry them out of the black region that normally traps circle trajectories (Plate 1C, left). Similarly, in the narrow spikes of low diamond performance (Figure 3B), the initial scan is delayed long enough that the trajectories become trapped in the central black region and they are caught (Plate 1C, right). The different patterns of active scanning at different vertical distances (Figure 5) can be explained by the differences in the steady-state horizontal velocity fields for circles and diamonds at different vertical distances along the midline. The complexity of the decision plots (Figure 8) is due to the fact that switching object identity corresponds to replacing the steady-state horizontal velocity field for one object with that for the other. Depending on the position of the object and the internal state of the agent at the time of the switch, the subsequent interaction of the agent with the new velocity field can sometimes cause the object to be caught and sometimes to be avoided. Finally, the outer red avoidance regions observed in both circle and diamond velocity fields predict that the agent will avoid all sufficiently peripheral objects. As shown by the peripheral trajectories in Plate 1C, this is indeed the case.

7.2 Transient Effects

The steady-state horizontal velocity fields provide some qualitative insight into how the interaction between this agent and its environment generates the observed categorization behavior, and even provide some quantitative predictive power. However, they do not fully explain the fine details of the behavior. For example, note that some trajectories pass into the central black region in all cases shown in Plate 1B and C. Why is it that some become trapped there (Plate 1B, left and Plate 1C, right) while others do not (Plate 1B, right and Plate 1C, left)? In order to answer such questions, we must look in more detail at the transient structure of the agent dynamics when the object passes into the central black region of the steady-state velocity fields, taking into account the neuronal inertia provided by the lag of the CTRNN state.

Figure 11 illustrates the transient effects underlying the agent's behavior for two trajectories that become trapped in the black region (C1 and D2) and two trajectories that do not (C2 and D1). The actual trajectory of the second largest principal component of the CTRNN state is plotted against $\dot{x}$ (black curves) from the time an object enters the black region until the time it exits or the trial ends for one example of each of the four cases. At each point along a trajectory (black dots), the agent is following a particular instantaneous trajectory (gray curves) toward the equilibrium point currently attracting it (gray points). Note that although the vertical loca-

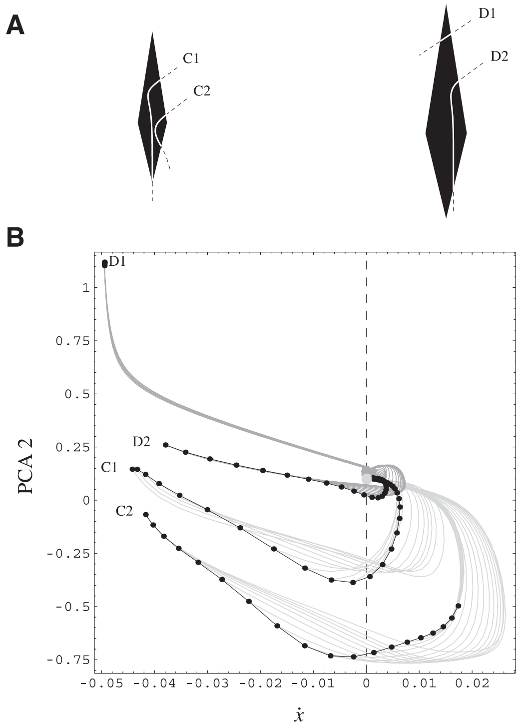

Figure 11 The role of transient effects in the dynamics of the best evolved agent for two circle (C) and two diamond (D) trajectories. (A) The passage of each of the four trajectories through the central black diamond-shaped region in the steady-state horizontal velocity field for circles (left) and diamonds (right) is shown. C1 is a normal circle catch from Plate 1B (left). C2 is a mistaken circle avoid from Plate 1C (left). D1 is a normal diamond avoid from Plate 1B (right). D2 is a mistaken diamond catch from Plate 1C (right). (B) The second principal component of continuous-time recurrent neural network (CTRNN) output is plotted against horizontal velocity as the agent enters the central black diamond for the four cases shown in part A. PCA 2 can be considered a depiction of the internal state or "neuronal inertia" of the CTRNN. PCA 2 was chosen over PCA 1 because the former led to less overlap of trajectories in this plot. The labeled points show the state of the agent at the time the object enters the black region and the sequence of black dots show how this state evolves over time until the agent leaves the black region. The gray curves show how the state would evolve if the interaction were frozen at selected points along the trajectory. Note that the two circle trajectories (C1 and C2, light gray) are ultimately headed toward an $\dot{x} = 0$ equilibrium point (light gray point), and the two diamond trajectories (D1 and D2, dark gray) are ultimately headed toward a slightly different $\dot{x} = 0$ equilibrium point (dark gray point). However, due to differences in internal state, only C1 and D2 actually reach the equilibrium points and the agent stops horizontal motion, whereas C2 and D1 never reach the equilibrium points before the object passes out of the black region.

tion of the equilibrium point is slightly different for circles (light gray) and diamonds (dark gray), both occur at $\dot{x} = 0$, which is why this region is colored black in the steady-state velocity fields. If the object remained in this region long enough, the black trajectories would eventually reach the equilibrium points at $\dot{x} = 0$ and the agent would stop moving in all four cases. However, because the CTRNN dynamics is parameterized by the (x, y) location of the object in the agent's field of view and both the object and agent are moving, the CTRNN dynamics is constantly changing, causing the state to follow a slightly different instantaneous trajectory at each time step (gray curves).

With this background, we are ready to explain why some trajectories pass through the central black region of the steady-state velocity fields while others become trapped there. For all four trials, the state begins at $\dot{x} < 0$, so the object is moving to the left in the agent's field of view. For a normal circle trial (C1), the state briefly crosses into the $\dot{x} < 0$ region, leading to a small rightward movement of the object, and then reaches the equilibrium point. This ceases horizontal motion of the agent and results in a catch. For the case in which a circle is mistakenly avoided (C2), a qualitatively similar trajectory is followed. However, due to the different CTRNN state and the different sequence of relative positions of the object and the agent, larger positive values of $\dot{x}$ are reached and a much larger rightward movement is generated, causing the object to reverse direction and move out of the black region before the state reaches the equilibrium point. For a normal diamond trial (D1), the time spent in the black region is so brief that the state barely has a chance to move toward the equilibrium point. Finally, for the case in which a diamond is mistakenly caught (D2), the CTRNN state and sequence of object positions are such that the motion toward the equilibrium point is even faster and more direct than for a normal circle trial. Thus, the particular trajectory that the agent follows depends on the interplay between the internal neural state and the changing (x, y) location of the object in the agent's field of view.

8 Neural Implementation of Agent Dynamics

The final step in a dynamical analysis of this agent is to examine how the agent dynamics described in the previous section is actually instantiated in the evolved neural circuit. How do the specific neuronal and synaptic parameters in this network produce the observed steady-state horizontal velocity fields? What neural properties give rise to the central black region so essential to catching circles? What neuronal interactions underlie the transient effects involved in overshooting the midline during scans and in trapping trajectory bundles in the central black region for circles? We will use a graphical illustration of the evolved circuit architecture (Figure 12) to present answers to all of these questions.

One observation that follows immediately from Figure 12 is that interneuron 10 has a high threshold, inhibits itself, and is either strongly inhibited or only very weakly excited by the sensory neurons. This suggests that this neuron will remain off under most conditions and thus may not play any significant role in the circuit's operation. Lesion studies largely confirm this. When interneuron 10 is removed, diamond avoidance performance only drops from 99.38 to 99.18% ($N = 1000$). However, circle catch performance drops from 95.08 to 88.77% due to a small but crucial effect that interneuron 10 has at one point in the behavior. Nevertheless, this difference is small enough that we will not consider interneuron 10 further in our analysis.

8.1 Steady-State Effects

Our first task is to account for the structure of the steady-state horizontal velocity fields (Plate 1). This is most easily done in three steps. First, we will examine the effect of the feedforward sensory connections in the circuit on the interneuron response fields (Plate 2A). Then we will consider the contribution of the recurrent interactions between the interneurons (Plate 2B). Finally, we will examine the consequences of the interneuron responses on the agent's motion (Plate 2C) and compare the combined effect of movement responses to the observed steady-state horizontal velocity fields (Plate 1). Our analysis will focus on the circle fields. The mechanisms underlying the diamond fields are very similar and we have already explained the differences in size and shape between the circle fields and the diamond fields in terms of how the locations of the lines of ray tangency differ for the two classes of objects.

The feedforward response fields for the bilateral interneuron pairs 8/12 and 9/11 are shown in Plate 2A.

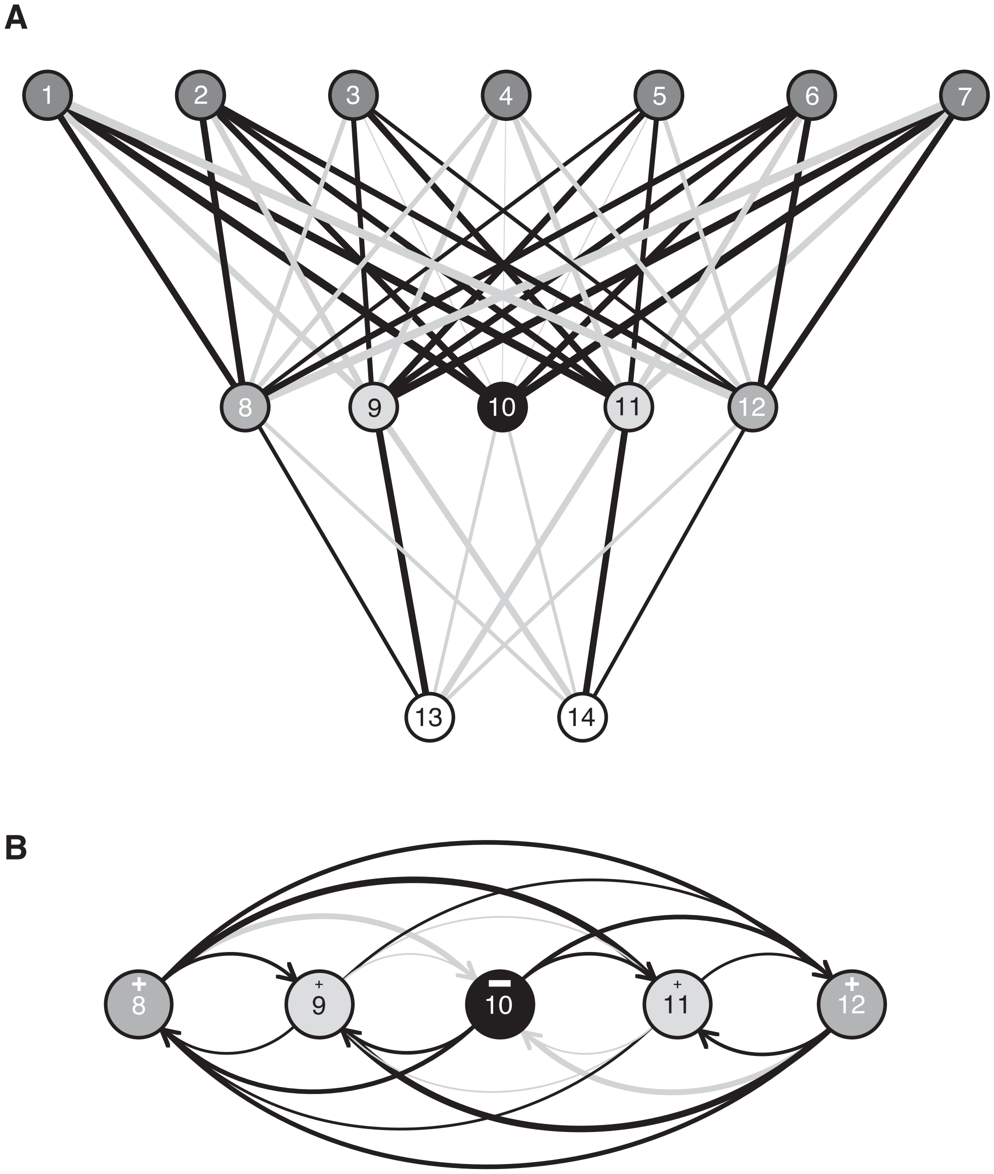

Figure 12 A graphical depiction of the evolved continuous-time recurrent neural network (CTRNN) of the best agent. The feedforward architecture is shown in (A), while the recurrent interactions between the interneurons are shown in (B). Neurons are shaded according to their bias, with higher threshold neurons (which require more excitation to activate) shaded darker. Excitatory connections are shaded gray and inhibitory connections are shaded black, with the width of the line proportional to the strength of the connection. The symbol above the neuron number in part B reflects the sign of the self-connection and the size of the symbol represents its strength. Thus, interneuron 10 has a high threshold and strongly inhibits itself. It is strongly inhibited by sensory neurons 1, 2, 6 and 7, very weakly excited by sensory neurons 3–5, strongly excited by interneurons 8 and 12, and very weakly excited by interneurons 9 and 11. In turn, interneuron 10 provides modest inhibition to the other interneurons and modest excitation to motor neurons 13 and 14.

These fields represent the output of each neuron in response to a circle at each point in the visual field. Note that interneurons 8/12 primarily respond to objects on the same side, whereas interneurons 9/11 primarily respond to objects on the opposite side. These features are relatively easy to explain. Both the sensory neurons and the interneurons are intrinsically off. The left side of interneuron 8's response field is black due to the inhibition it receives from sensory neurons 1 and 2, while the right half of interneuron 9's response field is black due to the strong inhibition from sensory neurons 5, 6 and 7. The remainder of the fields are produced by the pattern of excitation and inhibition from adjacent rays in a distributed fashion. Of course, the response fields and underlying synaptic mechanisms for interneurons 12 and 11 are simply the mirror images of those of interneurons 8 and 9 respectively. Most importantly, note the central black regions in all four response fields. These are due to the fact that inhibition dominates excitation when the central five or all seven sensory neurons are active. This is ultimately the mechanism responsible for the central black region in the horizontal steady-state velocity fields that traps circle trajectories.

The recurrent interactions between the interneurons modify the feedforward response fields as shown in Plate 2B. The most striking difference is that regions of multistability appear (green) in which the network can adopt different steady-state patterns of activation in response to the same input depending on the state it is in when the input occurs. The central region of multistability is primarily due to the mutual inhibition between interneurons 8 and 12, while the peripheral regions of multistability are produced by the mutual inhibition between interneurons 8 and 11 and between interneurons 12 and 9. In addition, the responses of interneurons 8 and 12 are both stronger than the feedforward responses alone due to self-excitation, while the responses of interneurons 9 and 11 are weaker due to inhibition from inteneurons 12 and 8 respectively.

Finally, each of the steady-state interneuron response fields produce a steady-state horizontal velocity field via their projections to the motor neurons (Plate 2C). These fields are color-coded in the same way as those in Plate 1. Note that all four interneurons produce motion away from the side they are on. In all cases, this is accomplished by inhibiting the motor neuron on the same side while exciting the motor neuron on the opposite side. Simply summing these individual horizontal velocity fields produces an estimate of their combined effect on the agent's motion (Plate 2D). Of course, this estimate is only approximate due to the nonlinear nature of the neural activation function (because of the saturation in $\sigma(x)$, $\sigma(x + y) \neq \sigma(x) + \sigma(y)$ in general). However, this summed horizontal velocity field corresponds quite closely in both structure and intensity to the observed velocity field (compare Plate 2D with Plate 1A, left). Thus, the step-by-step reconstruction of the agent's steady-state horizontal velocity field illustrated in Plate 2 can largely account for this important feature of its dynamics.

8.2 Transient Effects

As in the previous section, the steady-state responses of the interneurons to an object and the effects of those responses on motion do not give the full story. Transient effects also play an essential role. In order to illustrate the importance of these effects, I will examine the neural activity underlying the termination of active scanning and the final catching of circles (this was also explored in Figure 11). Because they cannot respond instantly to a change in input, all neurons contribute to some extent to the neuronal inertia that produces scanning, as shown by the fact that modifying the time constants of any of them lead to deficits in behavior. However, because all sensory signals must pass through the interneurons to affect behavior, we will focus on the transient responses of the interneurons during circle trials here.

Figure 13A shows the outputs of interneurons 8, 9, 11 and 12 along circle catch trajectories superimposed over the interneurons' feedforward response fields (from Plate 2A). Note that, for the most part, interneurons are only active along specific segments of the trajectory bundles and generally only one interneuron is active at a time (although there is significant overlap in places). This is presumably due to the almost universal mutual inhibition between the four interneurons. For example, for objects that begin to the right of the agent, interneuronal activation occurs in the following sequence: 11, 12, 9, 8, 11, 9, 12. We will focus on interneuron 9, which is active during left-to-right scans (Note that these bright regions of activity correspond to the peaks in the $\alpha = 0$ plot in Figure 9).

How are scans terminated to produce a catch? In the last section, we saw that whether or not a trajectory is trapped by the $\dot{x} = 0$ equilibrium point corresponding to the central black region depends on the system's state when the black region is entered (Figure 11). However, what *neuronal* mechanism underlies this behavior? Since the motor neurons require differential activation to produce movement and interneuron 9 is the only interneuron active during left-to-right scans, it stands to reason that it and its

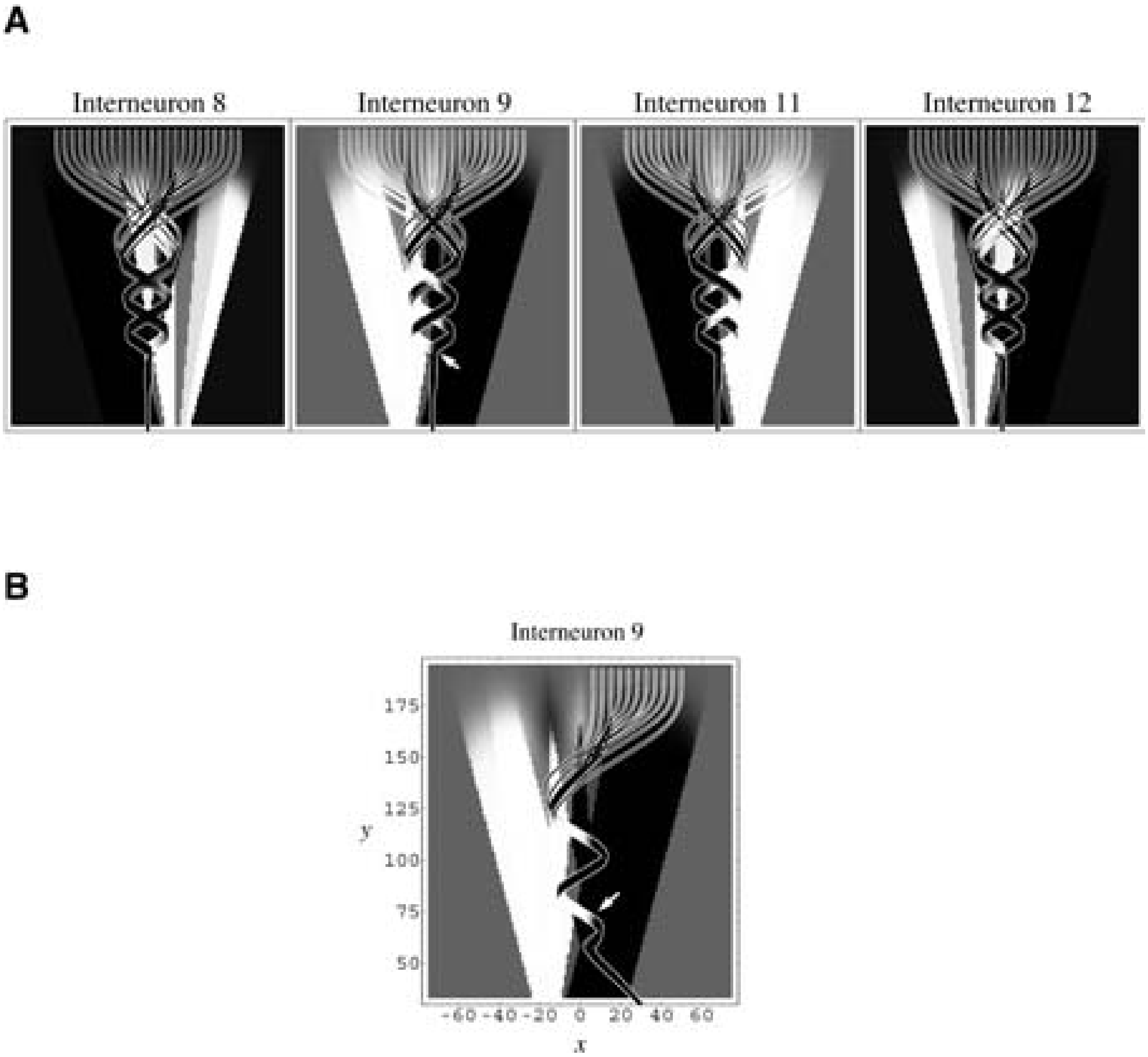

Figure 13 Transient behavior of the interneurons during circle catch trials. (A) Motion trajectories are superimposed on the feedforward interneuron response fields from Plate 2A. The trajectories are shaded according to the output of each interneuron, with black corresponding to an output of 0 and white corresponding to an output of 1. Note that, in most cases, only one interneuron is active at a time along each trajectory bundle. An important exception to this is that interneurons 9/12 and 8/11 are simultaneously active on the final scan before catching. (B). Interneuron 9 activity in response to circle trials beginning on the right of the agent when the connection from interneuron 12 to interneuron 9 is transiently lesioned as soon as the trajectories enter the black region. Note that the activity in interneuron 9 now lasts longer (compare white arrow here to white arrow in the interneuron 9 plot in part A), causing the trajectories to overshoot the central black region, and these circles are now avoided.

bilateral partner interneuron 11 are the principal producers of these scans. As can be seen in Figure 12, activity in interneuron 9 is primarily due to excitation from sensory neurons 1, 2 and 4, which produces the bright regions of excitation on the left side of its visual field. Normally, activity in interneuron 9 is terminated by inhibition from sensory neurons 5, 6 and 7 as the trajectory bundles pass into the right half of the visual field. However, notice that the activity during the final left-to-right scan is terminated early (Figure 13A, interneuron 9, white arrow). What causes this early termination? Unlike in previous left-to-right scans, interneuron 12 becomes active as the bundles approach the midline on the final scan. Since interneuron 12 strongly inhibits interneuron 9, its activity removes the source of left-to-right movement. Interneuron 12's activity also serves to terminate the left-to-right motion itself, because interneuron 12 excites and inhibits the opposite motor neurons from interneuron 9.

If this explanation is correct, it would predict that transiently lesioning the connection from interneuron 12 to 9 just as the trajectories pass into the black region of inhibition during the final scan should allow the trajectory bundles to pass through this region instead of becoming trapped there. As shown in Figure 13B (white arrow), this is indeed the case. Note that the length of time that interneuron 9 is active during the final scan is now comparable to that of earlier scans, causing the agent to move past the object and eventually produce an avoidance movement very much like that seen for diamonds (see Plate 1B, right), except that it is executed one scan later than usual.

9 Discussion

9.1 Summary

This paper has utilized a simpler idealized model agent to illustrate in some detail how the perspective and tools of dynamical systems theory can be applied to the analysis of embodied, situated agents capable of minimally cognitive behavior. A 14-neuron continuous-time recurrent neural network was evolved to allow an agent to visually discriminate between circular and diamond-shaped objects in its environment by catching the former and avoiding the latter. The best evolved agent used a strategy in which it foveated and actively scanned any object before eventually catching or avoiding it. It was also shown that this agent exhibited labeling and discrimination behavior characteristic of categorical perception, that object width was the primary determining factor in discrimination, and that the decision to catch or avoid was a temporally extended process rather than a discrete event.

The majority of the paper then focused on analyzing the dynamics of this evolved agent at three different levels: (1) the entire coupled brain/body/environment system; (2) the interaction between agent and environment that generates the observed coupled dynamics; (3) the underlying neuronal properties responsible for the agent dynamics. We found that the coupled dynamics was structured into a pair of bundles that twisted around one another, and that these two bundles unwound with increasing object width. Objects were caught whenever the two bundles collided at the axis of bilateral symmetry. These collisions underlay the category boundary between circles and diamonds, as well as producing an anomalous peak of catch performance. This structure also predicted two additional anomalous peaks that were subsequently observed. We were then able to explain these features of the coupled dynamics in terms of the interaction between the steady-state horizontal velocity fields of the agent dynamics and the neuronal inertia caused by the lag of CTRNN state in response to changing sensory inputs. Specifically, circle trajectories repeatedly overshot the midline and were then turned back toward it by blue centering regions until they became trapped in central black regions, whereas larger red avoidance regions ultimately pushed diamond trajectories away from the midline. Finally, we demonstrated how the horizontal steady-state fields were instantiated in the evolved CTRNN, as well as the neuronal interactions that underlay the trapping of circle trajectories in the central black region. This analysis successfully predicted that transiently lesioning a particular connection at a crucial point in time would cause circles to be avoided rather than caught.

9.2 The Dynamics of Embodied, Situated Cognition

The primary technical goal of this paper has been to show how dynamical analysis can be applied to a minimally cognitive agent. One very important question that remains to be addressed is the extent to which the specific result obtained here will generalize, both to other agents evolved on the same object discrimination task and to other related tasks. Such a generalization has successfully been performed for evolved neural circuits for walking (Chiel, Beer, & Gallagher, 1999; Beer, Chiel, & Gallagher, 1999). A key step in that generalization was a detailed study of a large set of high-fitness walkers that identified the range of possible solutions and their similarities and differences. Such a generalization study remains to be performed for the object discrimination task. Nevertheless, we can at least begin to sketch the general picture that is suggested by our analysis.

First and foremost, this analysis has illustrated what it means to say that an agent's behavior is a property only of the coupled brain/body/environment system, and cannot be attributed to any one subsystem. In a very real sense, the evolved CTRNN does not "know" the difference between circles and diamonds. It is only when embodied in its particular body and situated within the environment in which it evolved that this distinction arises over time through the interaction of these subsystems. For example, if the object ceases to fall, the agent produces stereotyped responses depending on the vertical distance to the object (Figure 5). The vertical motion of the object as it falls and the horizontal motion of the object through the agent's field of view as the agent moves combine with neural activity to produce a temporally extended approach or avoidance response that is more like the gradual accumulation of small differences than it is a discrete decision. Even though only CTRNN parameters have been modified by the evolutionary algorithm, *it is a property of the dynamics of the entire coupled system that has been selected for* (specifically, the horizontal separation between the agent and object at the end of a trial).

A second general theme that has emerged from the analysis of this model agent is that, if we seek to go beyond mere characterization of the coupled dynamics to an explanation of its generation, we must decompose the coupled brain/body/environment system into its components and study how their interactions give rise to the observed behavior. The main decomposition strategy that we have pursued here has been to separate the agent from its environment and ask how the structure of the environment affects the agent's behavior and how the agent's behavior affects the structure of its environment. Specifically, we characterized the steady-state structure of the agent's movement field as a function of the object's relative position in the agent's field of view (Plate 1). This kind of steady-state decomposition can be made mathematically rigorous when the time scale of the agent dynamics is much faster than the time scale of the sensory inputs from the environment. Even when this condition is not strictly satisfied, a great deal of qualitative insight can often be extracted from such a steady-state decomposition.

When the timescales of the agent significantly overlap those of its environment, as they do here and in many biological systems, then one cannot ignore the transient effects that arise from the lag between actual agent state and the equilibrium state. For example, this lag produces the differences shown in Plate 1 between the instantaneous horizontal velocity along a trajectory and the steady-state horizontal velocity given by the field. As the agent's state evolves from its current point, the agent's resulting actions and the environment's own dynamical evolution change the sensory input that the agent receives and modify the subsequent trajectories that are available to it. In this way, both the agent's dynamics and that of its environment continually shape the unfolding behavioral trajectory, as well as the future sensitivity of that trajectory to subsequent sensory input. As with any feedback system, it can be very difficult to disentangle cause and effect. Effects "play through the system" to become causes. Indeed, the entire discrimination behavior takes place on an extended transient of the coupled system. This certainly does not mean that the contribution of an agent's dynamics to its behavior cannot be understood. However, these results do suggest that we may need to modify our expectations regarding the demands we can reasonably place upon a componential explanation, stemming from the simple fact that the agent and its environment are really just two components of a single larger system.

The analysis described in this paper also has important implications for our understanding of perception. Traditionally, perception has been conceived as a process whereby an agent is informed about the state of its environment. But, as we have seen above, a dynamical agent follows a trajectory specified by its own dynamical laws and its current state. By parametrically shaping those laws, sensory inputs can certainly influence an agent's behavior. They cannot, however, in general steer it along any desired path or place it into a state uniquely corresponding to a given stimulus. Because the behavioral consequences of a given sensory input can differ greatly depending on the internal state of the agent when it occurs, ongoing neural activity sets a context for perceptual processing (Arieli, Sterkin, Grinvald, & Aertsen, 1996). For example, although the steady-state behavior specified by the perception of an object approaching the midline is the same in all four cases shown in Figure 11 (namely, to stop moving), the actual behavior differs due to differences in internal state.

Thus, it is perhaps most accurate to view perception as a *perturbation* to an agent, with the resulting behavior determined by the structure of the agent's dynamics and its internal state (Maturana & Varela, 1980). On this view, evolution structures an agent's dynamics so that its behavior is appropriately sensitive or insensitive to the different classes of perturbation that it may encounter (indeed, the agent creates these classes by its differential sensitivity). If the agent had been evolved to catch both circles and diamonds rather than to discriminate between them, then its dynamics would have had to suppress rather than amplify the behavioral effects of the small sensory differences between them. Of course, given the tight coupling between action and perception, a dynamical agent need not sit by passively while the environment shoves it around. Rather, it can modify its behavior so as to actively structure the perturbations it receives, making active perception a natural consequence of this view.

Dynamical systems theory provides a powerful set of mathematical and computational tools for characterizing the dynamics of a brain/body/environment system, only a few of which we have utilized here. However, there are clearly many challenges facing a dynamical approach to cognition. Even in the relatively simple agent considered in this paper, it is still quite impossible to fully visualize or characterize its complete 16-dimensional dynamics. Fortunately, such a complete characterization is usually unnecessary. The key is to focus only on the essential degrees of freedom, since only a tiny fraction of the full state space is actually visited during typical behavioral interactions. For example, note that the actual categorization behavior of the agent analyzed in this paper occurs in the neighborhood of the trajectory bundles shown in Figure 9. Appropriate projections allowed us to gain considerable insight into the operation of this evolved agent. Furthermore, many of the quantitative tools of dynamical systems theory do not require visualization for their application. Finally, it is likely that new tools will need to be developed, particularly in the areas of systematic reduction of dimensionality and the analysis of the non-autonomous dynamical subsystems that arise when a coupled brain/body/environment system is decomposed (Abarbanel & Rabinovich, 2001).

9.3 Stalking the Wily Representation

How does the concept of representation, the cornerstone of a computational theory of mind, fit into a situated/embodied/dynamical explanatory scheme? It is difficult to imagine a more controversial question in the foundations of cognitive science than this one (Carello, Turvey, Kugler, & Shaw, 1984; Brooks, 1991b; Bickhard & Terveen, 1995; Beer, 1995b; Clark & Toribio, 1995; Harvey, 1996; Ramsey, 1997; Bechtel, 1998; Clark & Grush; 1999; Wheeler & Clark, 1999; Markman & Dietrich, 2000; Keijzer, 2001). Much of the debate hinges on exactly what one is or is not willing to call a representation. Taken literally, the word "representation" means re-presentation, to present again. Thus, a sketch may re-present the essentials of a scene to someone who did not see it directly. Traditionally, *internal* representations are invoked as mediators between the outside world and internal cognitive processes. The basic idea is that these internal representations stand in for (re-present) external things in a way that both carries a coherent semantic interpretation and plays a direct causal role in the cognitive machinery.

Because representation is such a ubiquitous and fundamental concept in cognitive science, an ongoing critical assessment of representational notions is especially important. Thus, at the risk of inflaming what is already a sometimes acrimonious debate, it is worth understanding the challenges that dynamical analyses are beginning to raise for traditional representational notions. It is highly unlikely that any one of these challenges could ever definitively disprove the representation hypothesis. "Representation" is far too malleable a label for that. But minimally-cognitive agents do provide concrete examples within which we can explore the utility of these various notions and, by stretching our intuitions about the kinds of mechanisms that can account for sophisticated non-reactive behavior, they enlarge the intellectual playing field on which these debates are carried out. Remember that a skeptical stance is supposed to be a good thing in science. Someone who runs around yelling that the sky is falling is a trouble-maker. A skeptic just wants to know: How do you know that the sky *won't* fall today? Do *you* know what's keeping the sky up? What exactly do we *mean* by "the sky" anyway?

Consider the idea of an internal representation "standing in for" environmental entities that are not present in the agent's immediate environment (e.g., absent, hypothetical or fictitious objects, events, actions, or states of affairs), guiding behavior in their stead (Haugeland, 1991; Clark, 1997). Historically, this aspect of internal representation was one of the primary theoretical concepts by which cognitive approaches sought to overcome the inability of the stimulus/response framework of their behaviorist predecessors to account for many features of human behavior (Gardner, 1985). Unlike reactive agents, the behavior of an agent possessing internal representations can transcend the stimuli in its immediate environment. Even today, the limitations of reactive systems are often used to criticize situated approaches more generally.

Strictly speaking, however, *internal state* is the only requirement for such non-reactive behavior (Beer, 1995b). If an agent has internal state, then it can respond to the same environmental situation in different ways depending on that state, which in turn can depend upon its past experiences in that environment. Such a dynamical agent can also initiate and carry out sequences of actions even in the absence of environmental triggers and feedback, and it can organize its behavior in anticipation of future events. As an agent's internal state becomes more and more complex, the relationship between perception and action can become increasingly indirect. For these reasons, merely by virtue of its internal state, *any* dynamical agent will produce non-reactive behavior in general, as long as the timescale of its dynamics is sufficiently slower than that of its environment. Yet few would be willing to equate representation with internal state, since internal state is a property of physical systems in general. Finally, note that this observation cuts both ways. Not only does internal state allow an agent to respond differently to the same sensory stimulus, but different sensory stimuli can place a dynamical agent into the *same* internal state. This means that internal states are not necessarily uniquely associated with external events; the state of the body and environment must also be taken into account to fully disambiguate the trajectory.

Another common claim of representationalism that is challenged by dynamical analysis is the idea that the units of cognitive machinery will be individually semantically intelligible. The intuition here is that the objects and relations appearing in our conceptualization of an agent's environment and behavior must align in some sufficiently straightforward way with actual mechanisms internal to the agent. For example, since we conceptualize the object discrimination task in terms of a "decision" to "approach" "circles" (which are "smooth") and "avoid" "diamonds" (which have "sharp" "corners"), we expect to find representations of such things within the agent, as well as processes that manipulate and transform these representations. Indeed, a story is generally considered to be representational insofar as it "depicts whole systems of identifiable inner states (local or distributed) or processes (temporal sequences of such states) as having the function of bearing specific types of information" (Clark, 1997, p. 147).

There is no question that the patterns of activity of the interneurons play a key role in the operation of the agent analyzed here, but is this role best understood as a representational one? We have found no evidence for circle detectors or corner detectors within this agent's evolved nervous system, and even though we have shown that the agent's *behavior* is sensitive to object width, there is no evidence that any combination of interneuronal activity plays the role of re-presenting that width. Neither does there appear to be any straightforward neural correlate of the agent's "decision". Rather, the interneurons are active at multiple times during both circle and diamond trials, and in multiple contexts (Figure 13). Whatever "meaning" this interneuronal activity has lies in the way it shapes the agent's ongoing behavior, and in the way it influences the sensitivity of that behavior to subsequent sensory perturbations, not in coding particular features of the falling objects. For each possible internal state and sensory input of the agent, its nervous system defines how its behavior should change. Indeed, such a vectorfield of change is one common definition of what it means to be a dynamical system. This is wise dynamics to be sure, even dynamics for a purpose. But re-presentation?

Another common concern in many representational theories of mind is the possibility of misrepresentation (see e.g., Cummins, 1989), in which a putative internal representation fails to properly track its external referent in some circumstances. Since any discussion of mis-representation must presuppose a notion of re-presentation, there is little we can say about misrepresentation in the context of the evolved agent. However, the related issue of classification error due to misbehavior is certainly relevant here, e.g., the mistakenly avoided circles and caught diamonds in Plate 1C and the avoided circles in Figure 13B. But these misclassifications are probably best thought of as errors in dynamics that lead to a loss of synchronization between the agent and its environment.

Like juggling, a successful categorization is dynamically assembled and maintained. While a coordinated juggling pattern can be destroyed by a perturbed ball or a mistimed toss, a coordinated classification behavior can be destroyed by a perturbed object or neuron. But the fact that perturbing a specific element leads to a particular misbehavior can at best be taken as evidence of a dynamical role for that element in the behavior. It does not demonstrate a re-presentational role.

Since the object is always present in the agent's field of view, it might be argued that this categorical perception task is not sufficiently "representation-hungry" to fully engage questions regarding the utility of representational thinking (Clark & Toribio, 1995). There is some truth to this criticism, and it is certainly the case that categorical perception may fall below some cognitive scientists' threshold for cognitively interesting behavior. However, even if cognitive interest is equated with the extent to which a behavior is internally rather than externally driven, this task does at least begin to engage the issues. As I have repeatedly emphasized, the object alone does not drive the agent's behavior; rather, the object's effect on behavior is influenced by the agent's evolving internal state as the interaction proceeds. Thus, even in this simple task, there is a certain amount of causal disconnection between the object and the agent's behavior. However, it is important to further explore how the ideas developed here carry over to even more offline tasks. Toward that end, ongoing work is aimed at extending the approach described in this paper to tasks involving short-term memory and selective attention, in which the agent's behavior must depend in increasingly more sophisticated ways on information that is not available in its immediate environment (Slocum, Downey, & Beer, 2000).

It is becoming widely recognized that traditional notions of representation may need to be modified in order to do justice to the embodied, situated and dynamical facts emerging from recent work. For example, proposals are on the table to generalize the notion of representation to include analog, context-dependent, action-oriented, distributed, and dynamical features. This raises the question of how appropriate the term "representation" will continue to be. We are certainly free to redefine the term "representation" to mean anything we like, but some care must be taken in generalizing this concept. If "representation" comes to mean nothing more than correlated internal state and "computation" comes to mean nothing more than systematicity of behavior, then a computational theory of mind loses its empirical force (Beer, 1995b).

It may seem somewhat pointless to argue over a word. But the exemplar for computational theories of cognition is still quasilinguistic symbols syntactically manipulated, and by its connotations and its history the term "representation" carries with it preconceptions that may ultimately prove to be inappropriate for understanding the mechanisms of cognitive behavior. Thus, a little terminological hygiene seems to be in order. If we can reliably identify semantically intelligible internal configurations that can usefully be described as re-presenting an agent's internal goals and external environment, then "representation" it is. But if the mechanisms underlying situated, embodied cognitive behavior look nothing like representations, and they act nothing like re-presentations, then they are *not* representations, and continuing to call them representations is just going to confuse everyone.

9.4 Frictionless Brains

The research methodology that has been pursued in this paper is somewhat unusual, to say the least. Rather than studying the cognitive behavior of an existing animal through experiment and/or modeling, I have *invented* an artificial agent that engages in a cognitively interesting behavior and then analyzed the mechanisms underlying this behavior. While such approaches are quite common in adaptive behavior and artificial life research, they are still sufficiently novel in cognitive science and neuroscience that it is worth briefly articulating the motivation behind this methodology (Meyer, 1996; Beer, 1997; Parisi, 1997; Dean, 1998; Webb, 2000).

The integrated perspective illustrated in Figure 1 raises special difficulties that any scientific approach to the neural basis of behavior must address. Studying any one component of a brain/body/environment system is difficult enough, but studying all three components and their interactions in any animal is currently beyond our experimental capabilities. Not only must one be able to control, measure and manipulate the properties of a large fraction of an animal's nervous system, but also the relevant aspects of its body and environment. Perhaps even more importantly, even if the required data were available, we currently lack the theoretical concepts and mathematical tools necessary for an integrated understanding of complete agent–environment systems. These experimental and theoretical difficulties are only exacerbated by the complexity of cognitive processes.

The early theoretical development of a field typically involves the careful study of simpler idealized models that capture the essential conceptual features of the phenomena of interest. Such model systems have a long history in physics. For example, it was not until Galileo's consideration of such idealized situations as frictionless planes that theoretical physics in the modern sense of the word really began. The power of such an idealization is that it simultaneously makes clear a deep principle of motion (acceleration, not velocity, is proportional to force) and provides a well-defined way in which the complicating effects of friction can be understood (as an external force acting on the system). While testable quantitative predictions can sometimes be derived from simpler idealized models (e.g., when the friction coefficient of a physical surface is very low), their significance typically lies in the qualitative understanding they provide rather than their ability to account for specific empirical data.

Thus, in order to begin to flesh out the form that a dynamical perspective on situated, embodied agents might take, we need to undertake a careful study of simpler idealized models of complete brain/body/environment systems, an endeavor that has been termed *computational neuroethology* (Beer, 1990; Cliff, 1991). Instead of frictionless planes, we need frictionless *brains* (and bodies and environments). Only in idealized model agents can we presently hope to study the dynamics of a complete agent–environment system and thus clarify the fundamental nature of such systems. And if we are interested in cognitive phenomena, then we must focus on model agents that exhibit at least minimally cognitive behavior. Model agents can help us to formulate the necessary theoretical concepts, mathematical and computational tools, and experimental designs. They can serve not only as intuition pumps (Dennett, 1980), but also as theory pumps and mathematics pumps and even experiment pumps that can pave the way for more empirically driven analysis. They can also help to ground the ongoing debates about the proper place of embodied, situated, and dynamical notions in cognitive science by providing concrete examples to build intuition and sharpen the discussion.

In the work described here, evolutionary algorithms (Mitchell, 1996) are used to evolve dynamical nervous systems for situated model agents (Beer & Gallagher, 1992; Cliff , Harvey, & Husband, 1993; Nolfi & Floreano, 2000). There are, of course, many different ways to construct simpler idealized model agents that generate minimally cognitive behavior, and I certainly do not wish to claim that this is the only or best approach. However, there are certain advantages to this particular methodology that are worth mentioning. Continuous-time recurrent neural networks are arguably the simplest nonlinear, continuous-time neural network model. They are relatively computationally and analytically tractable, have a straightforward neurobiological interpretation that captures some of the key features of nerve cells, and are universal approximators of smooth dynamics (Funahashi & Nakamura, 1993; Kimura & Nakano, 1998). Thus, they provide an excellent substrate for exploring the dynamics of neural circuits. Evolutionary algorithms capture the two key features of Darwinian evolution: (1) heritable variation; (2) differential reproduction under resource limitations (or natural selection). A significant advantage of an evolutionary approach is that, by selecting only for the overall behavioral efficacy of the entire agent rather than a particular neuronal solution, we can minimize our a priori commitments to how a given behavior must be instantiated and thus better explore the design space of possibilities that exists in a relatively unbiased way.

The study of simpler idealized model agents is not without its drawbacks. First, such models are never likely to make specific quantitative predictions that are empirically testable in the way that more realistic models often can. However, the intent of idealized models is not empirical prediction, but conceptual clarification. Detailed, realistic models are often as difficult to understand as the original system. Clearly the study of idealized models and empirically driven analysis of particular cognitive processes are complementary approaches, and neither should proceed in isolation from the other. Second, there are many subjective decisions that must be made in order to formulate an idealized model, such as choosing a task to study, choosing a body model, choosing an evolutionary algorithm, etc. These decisions introduce their own biases. But far from being a shortcoming, I view this as an opportunity. Each set of decisions constitutes hypotheses to be explored. Of course some of the choices will be wrong, but how will we know until we try them and work through the consequences? Third, many of the details of an idealized model's analysis will be tied to the particular decisions made in its formulation, and thus be of questionable direct relevance to any real cognitive system. But it is only by working through the details of tractable concrete examples of the cognitive phenomena of interest that we can begin to glimpse whatever general theoretical principles may exist and whatever mathematical and computational tools will be necessary to formulate them.

10 Conclusion

In response to a question about the relevance of thinking of the brain as a complex dynamical system, the philosopher of mind Patricia Churchland once replied "It's obviously true, but so what? Then what is your research program?" (Lewin, 1992, p. 164). The present paper has described one possible way to begin to answer this question: (1) develop simpler idealized models of minimally cognitive behavior in embodied, situated agents with as few a priori theoretical commitments as possible; (2) analyze the dynamics of the resulting brain/body/environment system with an open mind about the mechanisms by which they might operate; (3) carefully consider the implications of these analyses for more realistic systems. The extent to which dynamical language will eventually become a key element of the explanatory structure of cognitive science remains to be seen. What does seem certain, however, is that dynamical analyses of cognitive behavior, and the critical questions they raise, can only broaden and clarify our understanding of the mechanisms of cognition.

Acknowledgments

Portions of this analysis have been previously presented at the Workshop on Dynamics, Computation and Cognition (held at the Santa Fe Institute on 12–14 May 1996) and at the Fourth International Conference on Simulation of Adaptive Behavior (held in Cape Cod, MA, USA on 9–13 September 1996). I would like to thank Dr Hillel Chiel and Doug Downey for their comments on an earlier draft of this paper, as well as the reviewers. Part of this work was performed while I was on a sabbatical leave at the Santa Fe Institute, and I gratefully acknowledge their support throughout this time. This work was also supported in part by grant RG0084/1997-B from the Human Frontier Science Program and by grant EIA-0130773 from the NSF.

References

Abraham, R. H., & Shaw, C. D. (1992). *Dynamics: The geometry of behavior*. Addison–Wesley.

Abarbanel, H. D. I., & Rabinovich, M. I. (2001). Neurodynamics: Nonlinear dynamics and neurobiology. *Current Opinion in Neurobiology*, *11*, 423–430.

Agre, P. E. (1997). *Computation and human experience*. Cambridge, UK: Cambridge University Press.

Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science*, *273*, 1868–1871.

Arkin, R. C. (1998). *Behavior-based robotics*. Cambridge, MA: MIT Press.

Ashby, W. R. (1952). *Design for a brain*. Wiley.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice*. Oxford: Oxford University Press.

Ballard, D. (1991). Animate vision. *Artificial Intelligence*, *48*, 57–86.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science*, *22*(3), 295–318.

Beer, R. D. (1990). *Intelligence as adaptive behavior: an experiment in computational neuroethology*. Academic Press.

Beer, R. D. (1995a). A dynamical systems perspective on agent–environment interaction. *Artificial Intelligence*, *72*, 173–215.

Beer, R. D. (1995b). Computational and dynamical languages for autonomous agents. In R. Port, & T. van Gelder (Eds.), *Mind as motion, explorations in the dynamics of cognition* (pp. 121–147). Cambridge, MA: MIT Press.

Beer, R. D. (1995c). On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, *3*(4), 471–511.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, M. Mataric, J. A. Meyer, J. Pollack, & S. Wilson (Eds.), *From animals to animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior* (pp. 421–429). Cambridge, MA: MIT Press.

Beer, R. D. (1997). The dynamics of adaptive behavior: A research program. *Robotics and Autonomous Systems*, *20*, 257–289.

Beer, R. D. (1998). Framing the debate between computational and dynamical approaches to cognitive science. *Behavioral and Brain Sciences*, *21*(5), 630.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4*(3), 91–99.

Beer, R. D., & Gallagher, J. C. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior*, *1*, 91–122.

Beer, R. D., Chiel, H. J., & Gallagher, J. C. (1999). Evolution and analysis of model CPGs for walking II. General principles and individual variability. *J. Computational Neuroscience*, *7*(2), 119–147.

Bickhard, M. H., & Terveen, L. (1995). *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Elsevier.

Boyton, R. M. (1975). Color, hue and wavelength. In E. C. Carterette, & M. P. Freidman (Eds.), *Handbook of perception, volume 5* (pp. 301–347). Academic Press.

Brooks, R. A. (1991a). New approaches to robotics. *Science*, *253*, 1227–1232.

Brooks, R. A. (1991b). Intelligence without representation. *Artificial Intelligence*, *47*, 139–159.

Carello, C., Turvey, M. T., Kugler, P. N., & Shaw, R. (1984). Inadequacies of the computer metaphor. In M. S. Gassaniga (Ed.), *Handbook of cognitive neuroscience* (pp. 229–248). Plenum Press.

Cangelosi, A. , Greco, A., & Harnad, S. (2000). From robotic toil to symbolic theft: Grounding transfer from entry-level to higher-level categories. *Connection Science*, *12*(2), 143–162.

Chiel, H. J., Beer, R. D., & Gallagher, J. C. (1999). Evolution and analysis of model CPGs for walking I. Dynamical modules. *Journal of Computational Neuroscience*, *7*(2), 99–118.

Churchland, P. S., Ramachandran, V. S., & Sejnsowski, T. J. (1994). A critique of pure vision. In C. Koch and J. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23–60). Cambridge, MA: MIT Press.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge, MA: Cambridge University Press.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge, MA: MIT Press.

Clark, A., & Grush, R. (1999). Towards a cognitive robotics. *Adaptive Behavior*, *7*(1), 5–16.

Clark, A., & Toribio, J. (1995). Doing without representing? *Synthese*, *101*, 401–431.

Cliff, D. (1991). Computational neuroethology: A provisional manifesto. In J. A. Meyer, & S. W. Wilson (Eds.), *From animals to animats*, *Proceedings of the First International Conference on Simulation of Adaptive Behavior* (pp. 29–39). Cambridge, MA: MIT Press.

Cliff, D., Harvey, I., & Husband P. (1993). Explorations in evolutionary robotics. *Adaptive Behavior*, *2*, 73–110.

Cummins, R. (1989). *Meaning and Mental Representation*. Cambridge, MA: MIT Press.

Damper, R. I., & Harnad, S. R. (2000). Neural networks models of categorical perception. *Perception*, *and Psychophysics 62*(4), 843–867.

Dean, J. (1998). Animats, & what they can tell us. *Trends in Cognitive Sciences*, *2*(2), 60–67.

Dennett, D. C. (1980). The milk of human intentionality. *Behavioral and Brain Sciences*, *3*, 4128–430.

Ehret, G. (1987). Categorical perception of sound signals: Facts and hypotheses from animal studies. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 301–331). Cambridge, UK: Cambridge University Press.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port, & T. van Gelder (Eds.), *Mind as motion* (pp. 195–225). Cambridge, MA: MIT Press.

Funahashi, K., & Nakamura, Y. (1993). Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, *6*, 801–806.

Gardner, H. (1985). *The mind's new science*. Basic Books.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Lawrence Erlbaum.

Glenberg, A. M. (1997). What is memory for? *Behavioral and Brain Sciences*, *20*, 1–55.

Harnad, S. (Ed.) (1987). *Categorical perception: The groundwork of cognition.* Cambridge, UK: Cambridge University Press.

Harnad, S., Hanson, S. J., & Lubin, J. (1995). Learned categorical perception in neural nets: Implications for symbol grounding. In V. Honovar, & L. Uhr (Eds.), *Symbol processors and connectionist network models in artificial intelligence and cognitive modeling: Steps toward principled integration* (pp. 191–206). Academic Press.

Harvey, I. (1996). Untimed and misrepresented. *AISB Quarterly*, Winter 1996, No. *96*, 20–27.

Haugel, J. (1991). Representational genera. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Lawrence Erlbaum.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act*. Cambridge, MA: MIT Press.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Johnson, S. D. (1994). Simple hybrid systems. *International Journal of Bifurcation and Chaos*, *4*(6), 1655–1665.

Keijzer, F. (2001). *Representation and behavior*. Cambridge, MA: MIT Press.

Kelso, J. A. S. (1995). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Kimura, M., & Nakano, R. (1998). Learning dynamical systems by recurrent neural networks from orbits. *Neural Networks*, *11*, 1589–1599.

Lakoff, G., & Johnson, M. (1998). *Philosophy in the flesh*. Basic Books.

Lewin, R. (1992). *Complexity: Life at the edge of chaos*. Macmillan.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Markman, A. B., & Dietrich, E. (2000). In defense of representation. *Cognitive Psychology*, *40*, 138–171.

Maturana, H. R., & Varela, F. J. (1980). *Autopoiesis and cognition*. D. Reidel.

Meyer, J. A. (1996). Artificial life and the animat approach to artificial intelligence. In M. Boden (Ed.), *Artificial intelligence* (pp. 325–354). Academic Press.

Mitchell, M. (1996). *An introduction to genetic algorithms*. Cambridge, MA: MIT Press.

Nolfi, S., & Floreano, D. (2000). *Evolutionary Robotics*. Cambridge, MA: MIT Press.

Parisi, D. (1997). Artificial life and higher level cognition. *Brain and Cognition*, *34*, 160–184.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge, MA: MIT Press.

Platt, M. L. (2002). Neural correlates of decisions. *Current Opinion in Neurobiology*, *12*, 141–148.

Port, R. F., & van Gelder, T. (Eds.) (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Ramsey, W. (1997). Do connectionist representations earn their explanatory keep? *Mind and Language*, *12*(1), 34–66.

Skarda, C., & Freeman, W. J. (1987). How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences*, *10*, 161–195.

Slocum, A. C., Downey, D. C., & Beer, R. D. (2000). Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In J. Meyer, A. Berthoz, D. Floreano, H. Roitblat, & S. Wilson (Eds.), *From animals to animats: Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior* (pp. 430–439). Cambridge, MA: MIT Press.

Strogatz, S. H. (1994). *Nonlinear dynamics and chaos*. Addison-Wesley.

Studdert-Kennedy, M., Liberman, A. M., Harris, K. S., & Cooper, F. S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, *77*, 234–249.

Suchman, L. A. (1987). *Plans and situated actions*. Cambridge, UK: Cambridge University Press.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thelen, E., Schöner, G., Scheier, C., & Smith, L. B. (2001). The dynamics of embodiment: A field theory of infant perseverative reaching. *Behavioral and Brain Sciences*, *24*, 1–86.

Tijsseling, A., & Harnad, S. (1997). Warping similarity space in category learning by brackprop nets. In M. Ramscar, U. Hahn, E. Cambouropolos, & H. Pain (Eds.) *Proceedings of SimCat 1997: Interdisciplinary Workshop on Similarity and Categorization* (pp. 263–269). Edinburgh, UK: Department of Artificial Intelligence, Edinburgh University.

Townsend, J. T., & Busemeyer, J. (1995). Dynamic representation of decision-making. In R. Port, & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 101–120). Cambridge, MA: MIT Press.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21*(5), 615–665.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind*. Cambridge, MA: MIT Press.

Webb, B. (2000). What does robotics offer animal behaviour? *Animal Behaviour*, *60*, 545–558.

Wheeler, M., & Clark, A. (1999). Genic representation: Reconciling content and causal complexity. *British Journal for the Philosophy of Science*, *50*, 103–135.

Wyttenbach, R. A., May, M. L., & Hoy, R. R. (1996). Categorical perception of sound frequency by crickets. *Science*, *273*, 1542–1544.

About the Author

Randall D. Beer is a professor of Electrical Engineering and Computer Science at Case Western Reserve University, with a joint appointment in the Department of Biology. His research is focused on the modeling and analysis of brain/body/environment systems, and its implications for neuroscience, cognitive science and robotics.