7 Recognizing Activity and Intent

Introduction

As shown in chapter 6, concepts derived from a broad appreciation of "embodiment" have made inroads into the discipline of human-computer interaction (HCI). The ways in which the various flavors of "embodiment" have been applied in this field are instructive, and they echo work across other design domains. Broadly, "embodiment" in HCI relates to the manner in which humans engage in physical interaction with artifacts and their physical surroundings. At one level, this concerns the design and development of tangible user interfaces discussed in chapter 6. As I noted, there is less work that has an obvious link to radical embodied cognitive science (RECS), particularly in terms of understanding the dynamics of continual reciprocal engagement within the human-artifact-environment system. RECS offers the potential to provide an account for ontology and epistemology in HCI (and interactions with other artifacts).

Within a human-artifact-environment, the environment can be characterized as a set of features, and a subset of these features will be salient to an individual's task ecology. In the information-processing approach, salience relates to information-as-content (where features are selected on the basis of their meaning). In contrast, perception-action coupling emphasizes information-as-context. We have considered how Newell's notion of organism, task, and environment constraints (discussed in chapter 3) might influence this notion of salience, and RECS can provide the framework within which to consider how salience is defined within the human-artifact-environment system. Moreover, salience relates to the objective that the system is optimizing (or satisficing). From an information-processing

perspective, the "objective" could be a predefined "goal" that the human (as "intentional agent") is seeking to achieve. In terms of computer recognition of activity or intent, we might define the objective in terms of parameters for a Markov decision process, or "reward function" in reinforcement learning, or "priors" in a Bayesian model. In these instances, the computer will be seeking to optimize these values, possibly in terms of encouraging or discouraging particular actions. We discussed, in chapter 6, the notion of "agency" in interaction between human and technology and suggested that either the human is in charge, as the "intentional agent," or the computer is in charge, optimizing specific parameters. However, this contrast relies on an assumption that there is a directed relationship between human and computer which is contrary to the system perspective taken in this book.

Within the human-artifact-environment system, certain interactions will be possible (between human-artifact, human-environment, artifact-environment), and the outcome of these will create "states" in which the system is stable. These stable states represent objectives for the system (in dynamic systems terms, these are "attractor states" in which the system is most likely to rest). In this way, activity involves seeking a stable state (as opposed to seeking a defined goal). Of course, the stable state might correspond to a goal but the point of this description is that (just as we saw in chapter 3, Suchman suggested that plans can be situated, i.e., discovered opportunistically in the ongoing, reciprocal interaction), so dynamic systems show how stable states can be "discovered" through activity. The transition from one state to another will be determined by the constraints that apply to elements in the system and their interactions. RECS provides the framework for the ways in which these constraints affect activity. From the proposal that a designer should understand the ongoing, reciprocal engagement with the environment and that the environment offers a "landscape of affordances," an account that follows RECS should be able to reflect the richness and complexity of such interactivity. To begin the discussion, I consider the simple activity associated with reaching for artifacts to pick them up.

Reaching for Artifacts

People respond to affordances "automatically," with little conscious awareness of the features to which they are attending.[1] So, how is information acquired and used, in relation to affordance? In a series of neat experiments,

Tucker and Ellis presented people with images of handled artifacts (say, a saucepan) and asked them to press a button with their left or right hand, depending on whether the image was inverted or not.[2] Irrespective of decision (inverted or not), response times were much faster when the handle of the artifact pointed toward the hand required for the response. The implication was that the orientation of the handle "primed" (that is, preactivated) movement of the hand that would be used to grasp the handle, and subsequent studies replicated this effect.[3] This suggests that the presence of an artifact that *could* be grasped initiates activity that would support grasping. It is a moot point as to how fine-tuned this relationship might be.

Reach-to-grasp movements are adapted to artifact properties. That is, the manner in which reaching is performed is affected by the artifact's width, weight, and slipperiness and by the subsequent action. The distance between thumb and fingers changes depending on the type of artifact that we will grasp. What is apparent from many of these studies is that there is a commitment to a specific form of grasp (as evidenced by the orientation of the hand and by the width between thumb and fingers as the hand approaches the artifact). A series of experiments have shown that it is possible to adjust this commitment during the performance of the action. This adjustment could involve the person reaching for an artifact and avoiding a set of distractor artifacts,[4] which slows movement time, or cuing a person to reach for an artifact in a specific location, and then changing the cue during their movement toward that artifact so that they need to change movement to a new location.[5] In anticipation of subsequent action, people adopt uncomfortable postures because they are seeking comfort in an end state (wine glasses) or in order to exert maximal torque (faucets).[6] Thus, the type of grasp to make when reaching for an artifact adapts to the relative weight given to artifact size, temperature, handle orientation, distance from the person, and so on. One implication (which reiterates points raised earlier in the book) is that we ought to attend to the most salient (rather than all) the available features, and that selection of features could be influenced by action pattern (in a reciprocal manner to the action pattern being influenced by the attended features). This ability to adapt movement, posture, and grasp according to the properties of the artifact or the demands of the task suggests that there is a process by which the exploration (of available features) occurs in an optimal manner—it doesn't make sense to assume that every detail is extracted and processed prior to performing an action.

In the simple act of reaching for an artifact, "decisions" are made and amended rapidly. For some researchers, this suggests that we have sophisticated "feed forward" control systems, in which a model of the world is used to plan and then guide our movements in the world. Other researchers take a very different perspective, arguing that our actions are model-free and that we respond to opportunities offered by artifacts in the world. It is this latter perspective that is most aligned to the claims made in this book. But, as figure 7.1 indicates, it can be challenging for a computer (using cameras to provide data) to determine that a person is reaching for a jug; it can be even more challenging for the computer to predict *why* the person is reaching for the jug or how this jug will be used.

Not only can we respond to the affordances of physical artifacts in anticipation of acting on them, we can also respond to the affordance of moving artifacts. A well-known example of this is the "outfielder problem,"[7] which relates to the challenge of catching a ball hit into the air. In order to catch the ball, you need to position yourself at the most likely point

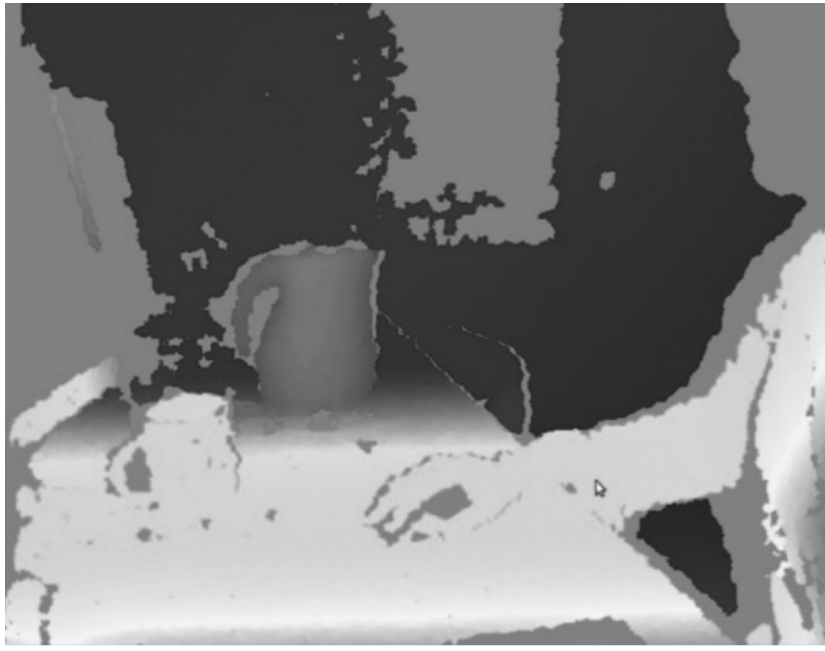


Figure 7.1
A computer's view of a person reaching for a jug.

to intercept it as it drop. The mathematics required to do this—direction, velocity, angle, and so on—are not beyond the ability of the average high school student,[8] but, as with many of the points in this book, the question is not whether are we capable of doing such calculations but whether we do these, as a matter of normal activity, or something else. There is plenty of evidence to suggest that, rather than performing complicated calculations, we employ strategies that rely on the visual relationship of the ball to catcher such that the ball appears to fly in a straight line[9] or to follow a constant velocity.[10] In either case, we are exploiting the optic flow in terms of a specific outcome in order to apply a simple (nonmathematical) adaptation to the visually perceived movement of the ball. These examples highlight the importance of considering the changing relationship between ball and catcher in the environment with the outcome being a "state" in which human (catcher)–artifacts (ball, catcher's mitt)–environment (sports field) be one of equilibrium—that is, ball in catcher's hand on the field. The "other states" of this system (e.g., the ball is dropped, the catcher runs into another player or advertising hoardings, the catcher trips, and the like) indicate the need to coordinate movement in order to achieve one outcome rather than others. The 'fly-ball' examples also illustrate the proposal that the "states" of the "system" can be important in defining the outcome (and provides a way of thinking about "intention" and "anticipation"). This implies that the action will be constrained by what constitutes a catch, which will be governed by the social conventions surrounding a particular sport, which constrain the types of actions that can be performed and the types of artifacts that can be used. One could imagine how different definitions of "catch" could result in different designs for gloves—for example, a "catch" might require the ball to firmly grasped by all fingers of the hand, in which case the glove would need to be fitted to allow each individual finger to move, or a "catch" might require the ball to be held in a container, in which case the glove would need to incorporate a receptacle that can hold the ball and need not separate the fingers. The contemporary baseball mitt is closer to the latter than the former but is still recognizable as a "glove" into which fingers are fitted. How these different glove designs define the concept of "catch" depends on the affordance offered by the context in which the human-artifact-environment system operates. That is, the human has capabilities (from prolonged experience of the sport) and wears the artifact (glove of particular design) in an environment (defined

by both the physical ball park with the moving ball, other players, spectators, and so on, and the social conventions surrounding the sport, such as the rules) with the objective of having the ball end up in the glove.

The idea that we respond to patterns of environmental features is not simply to claim that animals (including humans) are "programmed" to respond to aspects of the environment around them. This would return us to "behaviorism" and its reductionist view of human ability. Rather, RECS recognizes that the salience of information defined by environmental features depends on the capability of the animal. For example, bees are able to see light in the ultraviolet spectrum, and flowers possess ultraviolet patterns. That is to say, flowers "mean" nectar to the bees, not because of some interpretative act but because of the complementarity between the capability of the bee and the property of the flower. While this might seem obvious, it leads to the radical proposal that the "interpretative act" that is the root of semiotics might be less about cognition (and information processing as interpretation) and more about complementarity. It might be, for instance, that seeing the artifact is sufficient to activate regions of the motor cortex associated with actions that *could* be performed with that artifact, and the immediate visual information from the environment is used to tune the subsequent action. From the perspective of enactive or embodied cognition, the "knowledge" that the person would draw on need not be represented in the form semantic knowledge of artifacts, but rather would take the form of a filter that actively selects "salient" information (however that is defined) from the environment.

From a RECS perspective, the relationship between a person's prior experience, the actions that they can perform, and the environment in which they perform these actions (which, of course, includes the artifacts being used) combine to form the perception-action coupling that underpins affordance. In other words, affordance arises out of an interacting web of relations. "Affordances are neither properties of the animal alone nor properties of the environment alone. Instead, they are relations between the abilities of an animal and some feature of a situation."[11] Affordances thus arise through relations in human-artifact-environment systems, rather than existing as discrete properties of any constituent component or as the outcome of some interpretive act. In other words, affordances represent stable states in the human-artifact-environment system, and, in order to

define these stable states, we need to understand how the dynamics of this system are managed. RECS uses dynamic systems methods to do so.

In broad terms, the dynamic systems approach provides a mathematical basis for describing human activity. Much of the fundamental work in the application of dynamic systems to human activity is based on simple patterns of movement. This is for two reasons. First, the mathematical descriptions and analysis become increasingly challenging when we move to more complicated patterns of movement. Second, many of the underlying models can be characterized as coupled oscillators. The reason for the latter is both empirical (certain types of movement and certain types of neural activity can be described in terms of time-varying activity, e.g., rising and falling in signal strength) and theoretical (oscillators tend to have nonlinear relations to each other such that their dependencies are definable but not predictable). As an example, a classic demonstration of dynamic systems involves the simple action of tapping your index fingers on a table.[12] Begin by slowly alternating between right and left fingers to tap in sequence, and then speed up to as fast as you can. For most people, at some point the two fingers more together. That is, you begin by deliberately tapping out of phase but as your speed up, tapping becomes in phase. In terms of dynamics, this "system" has an order parameter that is defined by phase. Thus, the "system" has a tendency toward order in two states: fingers moving out of phase or fingers moving in phase. An interesting finding (typical of nonlinear dynamics) is that the transition from in-phase to out-of-phase is abrupt. That is, the transition between the two states of the order parameter is not gradual but happens suddenly. To appreciate these order and control parameters, we need to look at the basics of control theory.

Control Theory and the Human-Artifact-Environment System

If we begin with the view that activity occurs with the human-artifact-environment system, the question is how do the various types of interaction produce outcomes and how are the interactions performed as efficiently as possible? One way of thinking about this, as we discussed in chapter 1, is in terms of coordinating the interactions so as to minimize the degrees of freedom (DoF) of the system. Accordingly, coordination can be thought of as the process of selecting as few parameters as necessary to manage. From

the theory of control systems (derived from cybernetics), we can say that the system defines an order parameter (which defines how well ordered, or stable, the system is in any given state) and a control parameter (which indicates the actions that can be performed to alter the order parameter). In this case, an order parameter could be a single variable or it could be the product of two or more variables, depending on how the system operates. In other words, the order parameter can be thought of in terms of the objective that the system is seeking to optimize (in line with our discussion in chapter 1). In the simplest version of this, a feedback loop compares the current state of the system with a defined level of the order parameter and calculates an error, which is corrected by performing an action. In this case, the aim of the system is to maintain stability (or homeostasis) in the face of disturbances arising from sources in the environment external to it. Systems of this type align nicely with the homeostatic models that Craik[13] proposed and that we considered in chapter 1. Rather than sampling the environment to create a model of it, the system defines a model based on a setting of the order parameter and its actions, are directed toward keeping the state of the system within the limits set by this model. Recall that the term "cybernetics" refers to a person who steers a ship, and you can readily see how such a model can describe a simple process in which deviation from a defined course is minimized.

From dynamic systems research, control-theoretic models can describe how DoFs can be managed. One implication for RECS is that the solution to the DoF problem requires the definition of an appropriate order parameter. A point to note is that the state of the system (perhaps obtained through sense data) is *not* a mental model of the environment to be constructed but an error to be corrected either to maintain the current state or to ensure transition to another state. In biomechanics, this concept can be considered from two perspectives. In one, movement is considered in terms of the "product of force (kinetic) fields and flow (informational) fields" such that coordination emerges from dynamic environments.[14] From this perspective, systems self-organize as they adapt to changes in the balance between internal (individual) and external (environmental) constraints. In the other, synergetics[15] can be described in terms of variability and consistency in movement, as reflected by "dynamical equations."

Given the range of actions that people could perform to achieve a specific goal, there is a Degrees of Freedom (DoF) problem (outlined in chapter 1).

An elegant solution to DoF is proposed by Bril,[16] who follows Bernstein in proposing a hierarchical control model. In Bril's approach, functional (order) parameters can be achieved through regulatory (control) parameters through which the person controls specific movement parameters. For instance, experts (flint knappers and stone bead makers) seek to hold the functional parameter (kinetic energy) constant when they use different types of hammer or material, while novices vary kinetic energy with different types of hammer. Recently, we applied this finding to the comparison of jewelers performing simple sawing tasks, showing how experience relates to the grip force applied to the handle and to the velocity of the saw blade during cutting.[17] As Bernstein noted, it is important to incorporate feedback into the closed loop control of motion, in terms of the interaction between person and environment. This feedback can be seen as a means of managing the dynamics of the human-tool-environment system. Rather than considering movement as the enactment of a program or schema, an alternative view is to consider the control parameters that need to be optimized. Thus, an optimal control model would seek to determine the "cost function" that is being minimized while allowing the goal of the movement to be achieved. Bernstein spoke of coordinative structures in which combinations of muscle activation become associated with specific movements in levels of synergy.

From these basic control-theoretic or comparator models, we can draw several conclusions that inform RECS. The first is that there is no need for a central "controller" to manage interactions because feedback loops between the system components will allow the output of one component to affect another. At some point, discrepancies between elements in the system will decease, and the system will be in a state of equilibrium. That is, through these feedback loops, the system self-organizes. The feedback loops create a circular causality in which prior states of components lead to hysteresis (literally "history matters"), but once the system achieves equilibrium, this constrains the value of the order parameter and brings the other elements into defined interactions. As the feedback loops can change the state of the elements in the system, the initial state of the system is important in defining how interactions might develop. What is critical in this approach is that we are less concerned with discrete interactions between elements and more concerned with the overall activity in the system. In this respect, the objective can be defined by the order parameter. The set of possible

states in which the system is stable defines a state space. Moving from one state to another involves a phase transition, typically in response to an external change and typically in a way that is abrupt. As control parameters change, so the system shifts from one state to another. As the order parameter changes, the probability of moving to (or away from) a specific state increases. This means that one can consider these states in terms of attractors and repellents that pull the system toward (or push the system away from) parts of the state space. If we think of speed-accuracy trade-offs in a reaction time experiment (where you emphasize speed of response or correctness of response), the order parameter (time to respond to a signal) depends on two control parameters. We might claim that a strategy (favoring speed or accuracy) involves a phase transition that emphasizes one control parameter over the over.

Kalman Filters

Control-theoretic models of human activity were the direct descendants of cybernetics, in that they were feedback loops in which a servomechanism corrected movement in response to deviations from a defined path. As an example, imagine steering a car on a winding country road. Assuming that there is no other traffic or other obstacles, you could perform this activity by looking at the road ahead and making small corrective adjustments to the steering wheel to keep the car in the center of the lane in which you are driving. If this was all that driving involved, a basic servomechanism would be sufficient. Of course, we do not believe that this model describes driving because we typically have to attend to more things than the empty road ahead of us. However, as an initial example, this gives a flavor of basic cybernetic, closed-loop control.

Among the problems with the simple closed-loop servomechanism presented here, one of the most pressing is the way in which it handles uncertainty in the input signal. What you would not want was to drive the car by swinging the steering wheel in response to any perceived change in the environment. Not only would such control by ineffective, it would also be really uncomfortable for you and your passengers. So, this requires a way of deciding whether or not to react to changes in the input. A common approach to modeling manual control (at least for this sort of "tracking" task) involves the use of Kalman filter. The purpose of the Kalman filter is to reduce uncertainty in the input signal. In this case, the "controller"

samples the environment and issues a control signal to maintain the state of the system within acceptable limits. For driving, "acceptable limits" will be defined by the position of the vehicle in its lane; if the road curves, then the vehicle needs to turn to keep in the middle of its lane. Samples from the environment could be affected by uncertainty (perhaps it is twilight or foggy or raining, so the road ahead is not so easy to see clearly). The "controller" needs to decide the degree of confidence to give to each sample before it issues a control signal. If the input signal has high levels of uncertainty, and the controller responds to all samples with equal confidence, this could result in very jittery control. Consequently, the Kalman filter compares each sample with an expected signal. The expected signal reflects the average of prior samples (as the input signal) and the current control signal. The decision to change the control signal depends on the confidence given to the input or output signals, together with the rate at which the samples were obtained in order to define and correct "error" (between input and output).

What is particularly important for a Kalman filter control system is to have a continual stream of information on which to base its analysis. Indeed, if there is no new information (either because the input signal has stopped completely or because there is no change in the input signal), then the controller becomes very sluggish in its response (because it cannot detect the error signal that it requires or because any new information might require a large adjustment). Thus, one might expect that the brain (if it behaved like a Kalman filter) would be continually sampling the environment, adjust body posture or move sensory organs to provide an ongoing stream against which it could update and maintain its model. Without committing to any claim that the brain is a Kalman filter, it is worth noting that the saccadic movement of the eyes[18] or a phenomenon such as postural sway[19] indicate a continuously varying input.

For me, the Kalman filter, as an error-correcting servomechanism, provides a simple analogy for how the brain might be for "coping not copying" in its interactions with our environment. I use an analogous argument to explain how the concept of recognition-primed decision-making (central to naturalistic decision-making) could be described as a closed-loop control system, so that it did not need recourse to schema or mental models.[20] What these mechanisms suggest is that (certain) activity can be described in ways that allow accurate prediction and that have no need of a mental model of the environment. Rather, they rely on models that reflect the

"relation-structure" that Craik described (see chapter 1). The idea that the environment provides an "input signal" to a controller is not so far removed from the perception-action coupling of Gibson. However, the analogy of a servo-mechanism might feel dangerously close to a totally mechanistic (or worse, behaviorist) account of human activity; it might be acceptable to think of machines or robots as behaving in this manner, but how well does this fit with human behavior (especially if we want to capture "cognition" and "creativity")? Equally, is there a risk of replacing one form of internal representation (mental models and the like) with another (probabilistic or other weighting of salience)? Before answering this, I want to pose a counter-question: If one accepted that the information-processing metaphor (with all of the attendant baggage that I challenged in chapter 1) could describe how the brain functions as a "copying" machine, why balk at the suggestion that servomechanisms can be provide a metaphor for the brain as a "coping" machine? Both approaches (information-processing and servomechanisms) are reductionistic, both are based on machines, and both are intended to guide thinking through metaphor, and yet, only (I suggest) the servomechanism and its related concepts provide opportunity to rigorously describe how activity might be coordinated. I say this because the information-processing approach has a tendency to reduce itself to a set of interconnected "boxes" (describing particular functions), in a "production line," with an over-reliance on assumptions about what "information" is being "processed." A Kalman filter replaces the "production line" with a neater system that adapts to changes in the environment without the need for multiple stages of translation of "information." What a Kalman filter does not tell us is how features in the environment have salience. Kalman filters assume that the input is in a defined format (which is why a "tracking" or "steering" task is a useful way to conceptualize it). Nor do they tell us how the output (the actions we perform) adapt to environmental, task, or human constraints. To consider these issues we need to be able to define salience of cues and to define how the system manages its objectives.

The Bayesian Brain

One could characterize information-processing approaches to cognition as saying that cognition *causes* action. Indeed, there is so little consideration of action in a conventional information-processing experiment that the

response a participant makes is often reduced to pressing a button. However, as we saw earlier in this chapter, even the act of pressing a button can be primed by the action context in which it occurs (i.e., pressing a button that is on the same side as a handle on an image results in faster response). This suggests that cognition and action must be more closely intertwined than most theories of cognition assume. Early accounts of action assumed a close-loop relationship between the brain and the environment. William James, for instance, used ideomotor theory to account for the ways in which humans learn to control activity.[21] Babies kick and wriggle and through these seemingly random movements begin to sense differences in afferent information, which, in turn, become available to perception and result in association between a specific movement and specific neural patterns. This is the basis of Hebbian learning, in which neural pathways become entrained and reinforced through the practice of specific movements. When the specific movement is required at a later time, these pathways become reactivated (assuming that they have sufficient resting potential and that they continue to be primed). While this process might account for our ability to perform specific movements, it does not seem to say anything about cognition. However, this misses the point that such learning creates the ability to intentionally perform action—although, of course, there is still the feeling that the input to this intentional control comes from some "cognitive" activity.

One way of conceptualizing a closed-loop control for human action is to use the Bayesian brain approach,[22] which assumes that the brain does not act as a passive filter; rather it (1) has a set of probabilistic models, "Bayesian beliefs," of the sources of information available to the senses, and (2) uses these "beliefs" to make predictions about how the information will change. As soon as there is a discrepancy between the prediction and the information, there needs to be either an effort to collect more information or to change the beliefs. There are some similarities between the manner in which the beliefs are used to define sources of information and a Kalman filter discussed previously and the manner in which these beliefs are updated and the Brunswik lens model discussed in chapter 3. To a great extent, the problems associated with the Brunswikian model apply to the Bayesian brain concept (i.e., an assumption that the world is sampled in terms of internal states and that these internal states are used to determine action). On the other hand, one could interpret the Bayesian brain in cybernetic

terms as an example of Ashby's law of requisite variety (in that the Bayesian beliefs should be sufficiently complex to create expectations of the state of the environment relative to a person's actions). Indeed, in the Bayesian brain literature, there is assumed to be a "Markov blanket" in which a given state can be predicted because the model contains sufficient states to make such a prediction. A potential problem here lies in the scope of the blanket; as with the law of requisite variety, there is an implication that the "model" can contain all possible states that the system will encounter. In a cybernetic system, say, geared to managing temperature or water pressure, one can imagine that a finite set of states can be defined, which are sufficient to explain activity (and even here, one probably needs to have a couple of "wild-card" states to reflect unusual causes of puncture or damage to the pipes or damage to heating elements). But would one commit to the idea of a sufficient "model" for the brain in its interactions with the environment?

In a Bayesian description of brain activity, beliefs are specified in a hierarchy of layers in which high-level goals are defined as the prior probabilities, which then influence lower layers (Friston claims that this hierarchical structure can be found in the cortical structure and that it involves the activity of pyramidal cells).[23] Sensory information can be broadcast across the brain, and this requires adjustment (of the gain of channels over which the information flows) so that specific prediction errors can be managed. Consequently, optimizing the operation over the different layers can involve seeking and reconciling error between what is expected and what is observed. Such adjustment involves a process in which prediction errors are minimized by either updating the priors or by seeking additional information—that is, "active inference" is performed to guide sensory activity to reduce such errors. Consequently, this approach has also been termed "predictive processing."[24] The overarching goal of this activity is to maintain the brain in a state in which entropy is as low as possible (in other words to avoid increased entropy or "surprises" arising from uncertainty).[25] As soon as entropy increases, actions are performed to collect more sense data (or, by analogy with control theory, to modify the order parameter), which is used for comparison. Given that collecting data (or modifying the order parameter) can have associated costs, a further goal is to ensure that such costs are minimized.

This Bayesian brain approach describes some aspects of the ongoing reciprocal engagement between human and environment in terms of

continuous perception-action cycles in which the order parameter of the system in defined in terms of sense data. As with our previous discussion of multiple objective optimization (in chapter 2), we can assume that there is a large number of states in which the brain can operate, but that its state at any given moment will be defined by a much smaller subset of states (with the aim of maintaining equilibrium or homeostasis as far as is practicable). If there are prediction errors (due to a mismatch in the current level of the order parameter or to sudden changes in the environment), these will increase entropy and cause homeostasis to be disrupted.

The Bayesian brain system seeks to manage "free energy," which, from information theory, means that the brain seeks to minimize any discrepancy between belief and sensory information in order to keep the long-run average surprise (unexpected or out-of-model) events as low as possible.[26] Free energy depends on incoming sensory signals, conditional expectations, and a model that relates conditional expectations to states of the world. This results in a scheme in which conditional expectations are replaced by sensory signals and the model is updated. The error (between prediction and model state) then informs the resulting response. From this perspective, perception is not a process of creating mental model that contributes to cognitive processing, but a means of managing sensory information within constraints set by the prior probabilities of information in the world that the brain is configured to respond to. Action, from this point of view, becomes a way of either updating these prior probabilities or seeking further sensory information. Accordingly, the purpose of cognition is to maintain homeostasis of activity in response to salient information.

The Bayesian Body

The Bayesian brain hypothesis provides a way of theorizing how affordance operates and an elegant set of testable hypotheses about how the choice between seeking further information or performing an action is made. At present, the Bayesian brain (and predictive processing) seems to situate all of the activity in the neural architecture of the brain and to rely on data from brain imaging to provide support to the argument. From an enactive and embodied perspective, this is troublesome because it offers little opportunity to include the body in the theory.[27] One approach would be to align a Bayesian brain approach with sensorimotor contingency theory.[28]

In this approach, the environment acts as the "external memory" from which to derive action. Sensory signals correspond to actions in this environment such that "rules or regularities relating sensory inputs to movement, changes and action."[29]

If one accepts that prior probabilities of perception-action pairings are adapted as features are attended to (and that some of these priors persist, perhaps as resting activations, between situations), then one could also accept that, by analogy, resting activations exist for the body. In a sense, this is what Bernstein meant by coordinative structure (see chapter 1). As an athlete or craftworker repeats a particular action, so the musculoskeletal system becomes tuned to that movement. In other words, "goals make perception enactive."[30]

As activity is performed, the interactions between elements in the coordinative structure vary, depending on the way in which the structure is being controlled and the way in which it is affected by the environment around it. This notion frames the point made by Ingold that there is a moment-by-moment, stroke-by-stroke variation in the tool-wielding movements of the skilled craft-worker.[31] In this way, one can consider activity in terms of softly assembled systems in which activity is contextually constrained and embodied and in which repetitive actions share a "family resemblance" but exhibit variability. Local interactions among embodied processes on different timescales weave the intrinsic fluctuations of the component processes into a coherent fabric of flux, despite inherent tendencies of the different processes to vary at their own different rates (on their own timescales). In other words, the challenge for understanding activity is less one of understanding discrete actions and more one of understanding the ways in which activity balances between consistency and variability, which is what Bernstein defined as "dexterity." In other words, skillful coping is not simply a matter of performing an action but rather is about acting in order that the human-artifact-environment system reaches a state that matches an objective, or an order parameter. As each element in the system might be subject to change, there is a need to adapt to change. Such adaptations, in dynamic systems terms, can be considered in terms of changes to the human-artifact-environment system, which can be measured in terms of stability or instability of the system. Measures of stability, over time, are derived from various definitions of entropy. When a system is stable, it will be low entropy. "Competitions among local rates of change strike a precise

balance with globally emerging cooperative activity. In the precise balance of (or near) the critical state, they produce a long-range correlated, aperiodic pattern of change or flux in behaviour."[32]

With entropy analysis, we are in a better position to understand the underlying dynamics of activity. One approach is to use 1/frequency (1/f), which describes the fluctuations in time-varying data between highly predictable and totally random. In other words, it provides a measure of the underlying stability of the system that generated the signal. What makes this measure interesting is that many phenomena produce time-varying data that at local levels appear random, but that over longer timescales show repeatability. From this, 1/f scaling can also be considered in terms of long-term memory in signals. The reason why this is of interest to dynamic systems models, particularly in terms of human activity, is that it allows us to make sense of activity that might look unstructured or random on a moment-to-moment basis but that demonstrates a repeating pattern over many instances. To take a simple example, recall the reaction-time experiment in which you have to press a button each time a light turns on. Your time to respond ("reaction time") is a standard metric for a host of cognitive studies. Usually, the results of thousands of trials will be collected, and the average (mean) and variability (standard deviation) reported. What these statistics do not reflect is the way in which your attention (and enthusiasm) for the task might wax and wane, particularly over thousands of repetitions. If, instead of averaging reaction times, we treated these actions as a series of events over a time period, we can explore the strategy that is being applied.

Such 1/f scaling can be applied across different cognitive tasks to indicate a "softly assembled" system by focusing on interaction-dominant dynamics (in which component dynamics alter interactions) rather than component-dominant dynamics (in which behavior arises from components, demarcated and assigned specific functions).[33] In part, 1/f scaling reflects the motor component of the activity being studied and the ability of people to adapt to situational demands as embodied systems. For example, hand-mouse coordination in a simple video game exhibits 1/f scaling during normal operation but not when the task is disrupted.[34] This result indicates that during normal operation hand-mouse control can be described as an interaction-dominant system. Applying this concept to jewelers, 1/f scaling can distinguish skill levels in the use of jewelry saws.[35] In addition to dynamics being detected in physical performance, these are

also apparent in cognitive and perceptual tasks. 1/f scaling has been shown in cognitive tasks,[36] and dynamic systems measures can be applied to reaction time experiments[37] and problem solving.[38]

Computer Recognition of Human Activity

The ability of computers to recognize and respond to human activity has grown dramatically since the 1990s. In this section, by "human activity" I mean speaking and moving. The proficiency of speech-recognition technology and wearable fitness monitors is such that these have now slipped over from being technology (with all its implications of the magical and beguiling) to the status of a commodity (so quotidian that we barely notice or question it operation—until it goes wrong). When I was doing my PhD on speech technology in the 1980s, the majority of speech-recognition algorithms (particularly for commercial applications) would use a limited number of words and a highly restricted syntax for combining these words, often requiring a period of "training" so that the device could modify its models to your manner of speaking. Most of these devices seemed to favor a sort of transatlantic English and struggled with pronunciation or accents that deviated too far from this.

Even with the major leaps in algorithmic complexity, both speech technology and wearable devices are essentially signal-processing devices. That is, the basic challenges in their operations arise from the collection, cleaning, and analysis of data from their sensors (microphones, inertial measurement units, and so on) so that these data can be used to create models against which new signals can be compared and classified

For the most part, speech- and activity-recognition technologies are concerned with isolating discrete "units" (e.g., words, actions) from the continuous stream of data coming from the sensors. One might assume that, for speech, the unit could be human-scaled—for example, a word. Unfortunately, defining such units at the "word" level tends to produce quite poor performance. This is partly because isolating words as discrete units can be challenging; in speech, words overlap and run into each other. This leads to problems of "end-point" detection (where each word begins or ends). The acoustic parameters of the signal can be affected by "linguistic" context (the words before or after it) or the "extra-linguistic" context (the emotional state or age of the speaker, the background noise). Noting that

few speech-recognition systems make use of detailed semantic knowledge (they are, as we noted above, sophisticated signal-processing systems), these problems are not dealt with through understanding the meaning of the words. Rather, speech technology (since the 1990s) has focused on "units" that can be assumed to be fairly stable, or at least to have variability that can be predicted. To this end, the "units" are phonemes (or the acoustic equivalent: sounds that can be labeled as phonemes), with statistical models defining the probability of phonemes being combined in sequences. This is the basis of Markov models, which heralded a step change in speech recognition in the 1990s and was the basis of many commercial systems. Recognition performance (particularly in a benign environment of the laboratory) could reach above 90 percent in terms of accuracy—so you would have to repeat one or two words out of every ten. The advances in this technology over the intervening years have been remarkable, particularly with the widespread use of deep neural networks. In deep neural networks, the statistical patterns are discovered by computers through the correlations between phonemes in massive corpora of speech. For the purpose of this discussion, it is sufficient to accept that speech recognition involves the definition of discrete "units" (phonemes), that these units are probabilistically related to each other, and that all the information required can be obtained from the speech signal. Consequently, while the signal that this technology processes contains human speech, it is a moot point as to whether it "knows" that is dealing with "speech." My point is that few, if any, of these technologies begin their analysis from an understanding of how a person produces speech.

The recognition of human activity (i.e., movement) can be performed either from sensors on the person or with cameras. For example, in the Microsoft Kinect a depth-camera captures the image of a person and this is translated to a point cloud that is matched to a skeleton model. As long as there is good alignment between point cloud and model, the person's movement is recognized (so the avatar on the screen follows the movements of the player). However, the alignment might not be perfect, and players often need to subtly change the way that they move in order to maintain alignment. A similar adjustment happens with speech technology, such that speakers might alter their pronunciation or choice of words (particularly when the device has made a mistake). To date, much of the work using sensor data makes assumptions similar to those used in speech

recognition (not least because so much of the analysis of sensor data either uses statistical models, such as hidden Markov models, or uses deep neural networks). While the issue of whether speech technology knows that it is processing "speech" (noted above) does not affect its overall performance, for activity recognition I think that are many unresolved issues. For instance, assuming that "actions" can be defined in terms of discrete units that are separable from the flow of activity is quite odd when applied to everyday settings. In some cases, say, counting steps on a digital pedometer, the model could be quite simple, as in defining a threshold for the signal to pass in order to count as a "step." Having said that, step-counting based solely on sensor data is not as trivial as this implies. In particular, deciding when a step has been completed could involve reconciling more than one impact (heel striking, knee locking, weight transfer on to front of foot, and so on), depending on the way that a person was walking (particularly if this person was relearning how to walk following an injury or was wearing braces or calipers), and on the location of the sensor (in the shoe, on the waist, in the pocket, and so on). Furthermore, counting steps is only part of the analysis that one might wish to make—for example, analysis of gait might be more important, particularly in rehabilitation.

For basic activity recognition, action can be defined in simple terms of a threshold beyond which the incoming signal needs to pass (as in the example of step-counting). A more complicated approach might combine parameters from several sensors to cope with contextual factors that could influence the signal. In this "context-aware computing" the challenge is to ensure that data from all the sensors can be combined into reliable models. In almost all cases, however, the models have little need to know about how the signal was produced. That is, these technologies rely on the assumption that all the necessary information can be extracted from the sensor data. The sensor data are then used to create a model. The model is used to evaluate any future sensor data, labeled using the "units" to which the model has been trained. Of course, this is the same process that the information-processing approach to human cognition adopts. My claim is that neither speech or activity recognition nor the information-processing approach to cognition begins its analysis from an understanding of how people produce speech or action. That is, rather than engaging with the embodied nature of human behavior, these approaches assume that this behavior can be reduced to a discretized model.

Some people who develop wearable technology or activity-recognition algorithms might be affronted by my claim that they ignore embodiment. I can imagine them saying something like, our devices attach to the body so we must be doing embodiment, but, perhaps, the majority will shrug and say, so what? Why should technologies that respond to human activity need a concept of embodiment? For wearable technologies, the concept of "activities of daily living" is commonly invoked to define classes of activity into which patterns of sensor activation can be grouped. One reason why these technologies might benefit from a concept of embodiment is to enable them to achieve their aims of adaptation and personalization. Rather than detecting *what* action has been performed, they could ask *how* it has been performed.

Many algorithms used for activity recognition are based on normalization of the data. That is, the models might identify statistical points of consistency, say, a central value in a cluster of similar data, and then create a boundary around this point to define an inclusion zone; any value within this zone would be treated as equivalent to the central point. So, if we give the central point a label, such as a particular phoneme or action, then any subsequent data that fall within the inclusion zone would be given the same label. By definition, this approach seeks to eliminate variability. In signal-processing terms, this makes sense because sources of variability might include noise or other interference to the sensor data and this needs to be minimized to reduce recognition error. But recall that Bernstein's definition of dexterity was based on adaptive variability in human actions. From this, it is unlikely that activity recognition, as it is currently performed, could adequately reflect the ways in which skills are learned (or lost). A common approach to "skill" (in activity recognition) is to define discrete levels, with a model defined from each level. While this aligns to signal-processing approaches, it does not align with theories of human performance or skill acquisition.

A little reflection on the suggestion that "skill" can be discretely compartmentalized shows its failings. Skilled practitioners do not always do different actions than novices do, nor do they always perform tasks more quickly. Rather, a characteristic of skill is the seamless merging of tasks into sequences and the ability to rapidly adapt performance of a task to suit context or the ability to anticipate the needs to a subsequent action and adjust a current action accordingly. Treating actions as discrete units misses that seamless merging (unless, of course, one creates models that reflect

all possible combinations of sequence). If we do not use discrete units to define action, then we shouldwe treat actions as sequences, in time-series, which returns us to our earlier consideration of dynamic systems.

Recognizing Actions and Inferring Thoughts

We are all familiar with "recommenders" on websites, which suggest that, as we have purchased product X, we might want to consider product Y. Early instance of these recommender systems were based on crude matching of purchases (which could often lead to peculiar recommendations). Contemporary versions incorporate more nuanced reasoning and more information (often obtained through "scraping" the records of your interactions on a variety of webpages or with credit cards or store loyalty cards). In this case, the recommendations are developed from a detailed "model" of you as a consumer. While the idea of such models might be worrying (not least because we have little control over who is using our data and for what purposes), the point at issue here is how we are meant to respond to recommenders. For consumer decisions, these might be relatively benign (irritating but easy to ignore). However, there has been a growing class of recommender systems (often running on devices that we wear or carry) that are expressly designed to modify our behavior or change our habits. At present, these apps tend to be focused on health, particularly diet, exercise, smoking cessation, or medication reminders. These apps take data from sensors on the person (such as accelerometers or step-counters) and use these to provide motivational messages, or they have reminders programmed to occur at specific times, such as when to take medication. From the early 2000s, the input to the reminders comes from a broader range of sources; we have already considered location-based adaptation, for one, and personal information assistants can adapt to our previous actions and preferences.

Do I wish to claim that these devices somehow "know" what you are thinking? This sounds pretty far-fetched, particularly when you consider the type of data that such a device might be collecting. But is the idea of a wearable device that *is* able to know what you are thinking (or "read your mind") simply a matter of the type of data that it collects? In a sense, this is only a matter of refining the ways in which "recommender" systems currently work. After all, if you plug enormous quantities of well-curated data into deep neural networks, then some consistent and intriguing results are

inevitable. This is not a matter of opinion; it's just math—but, of course, it implies a particular definition of "what you are thinking."

Across much of cognitive science, "thinking" refers to purposeful, goal-oriented activity (such as the problem-solving we discussed in chapter 2), rather than the tumbling chaos of chatter that might intrude on our quieter, less distracted moments. In other words, "thinking" is typically defined in terms of a goal or intention toward which action is directed, rather than the mudging of thoughts about relationships, finances, or what to have for dinner. There is good reason for this focus, in the cognitive sciences at least. If you are going to study "thinking," then you need to know when it is happening, and, if you are running an experiment, you need to make sure that what is happening, happens in a similar manner to all the participants (otherwise you run the risk of the experiment being confounded by individual differences). In other words, thinking, for these experiments, involves the manipulation of information-as-content. Even conceding a narrow definition of "thinking," there remains a challenge of associating an action (or sequence of actions) with an intention and whether such an association necessitates a "theory of mind." For some sequences, this could be a trivial challenge. For example, you fill a kettle with water and put it on to boil, then you open a cupboard door and take out a cup. From knowing the time of day and detecting these actions (e.g., using data from sensors on the handle of the kettle and the cup, the door of the cupboard etc.), it would be probable that your actions will result in making yourself a cup of coffee—and your intention would be "make coffee." At this level, talk of "thinking" might feel redundant. More significantly, does the identification of a sequence of actions that can be associated with a known outcome actually signify intention? Before answering this, let's add a further element to the activity. Suppose that, as a New Year resolution or on medical advice, you have decided to reduce your caffeine intake. As long as one of the next actions in the sequence does *not* involve taking the coffee jar from the cupboard, then we can simply switch the notion of intention to "make a hot drink" (at a higher level of definition) or "make a herbal tea" (at a lower level). The suggestion of higher- and lower-level definitions implies a hierarchy of intentions (which can inform activity recognition by digital technology). But, let's assume that you have picked up the coffee jar. The sequence of actions now points to an intention of breaking your resolution or ignoring medical advice. In this case, the device that is monitoring your actions could intervene, perhaps by activating a buzzer on

your wrist, perhaps by sending you a text message, perhaps by logging this intention, and sending a message to your physician. With this trivial change of context, this example has shifted to something that the reader might find more sinister. The shift has not come from a change in technology or algorithm (in each case, sensors generate data that are interpreted by algorithms tuned to detect and respond to specific features); rather it has come from the change of emphasis from recognizing activity to predicting intention. In the first example, the algorithm defines a "goal" (i.e., a class of activity that is specified by a collection of actions). Pursuit of the goal could be supported by, for example, having the cupboard door handle light up to cue the person to find the cup. In the second example, the algorithm is evaluating the activity in terms of a value structure, in which the "values" represent social or other forms of interpretation of acceptability of an action. This returns us to the discussion, in chapter 4, about the politics of affordance. In this case, the "affordances" relate to the opportunities for action that the device is defining for a given context and raises questions of how we, the users of the device, can accept or dispute such a definition and what options are available to us if we do not agree with the device.

If we are not directly interacting (or even not interacting at all) with smart technology, how should we consider our relationship with it? It feels as if some of the traditional views that HCI offers become redundant, as do the options that an information-processing approach might suggest. For example, if the behavior of the smart technology is opaque, should we simply seek to make it "transparent"? There is a lot of interest in the question of "explanation" of the decisions made by complex artificial intelligence. In this respect, the problem (of transparency or explanation) becomes a matter of information-as-content. My problem with this is that we are probably no longer interacting with smart technology in ways that make explanation possible or plausible. Any "feedback" that the technology presents to us will, at best, create further demands on our attention and decision-making and, at worse, become confusing, misleading, and pointless. However, if we think about how people provide explanations, we might realize that they are every bit a matter of information-as-context: not only do we adapt the content that we provide to our audience, but this adaptation often unfolds and develops in our conversation with them.

Embodying Design

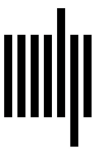
An Applied Science of Radical Embodied Cognition

By: Christopher Baber

Citation:

Embodying Design: An Applied Science of Radical Embodied Cognition
By: Christopher Baber
DOI: 10.7551/mitpress/12419.001.0001
ISBN (electronic): 9780262369886
Publisher: The MIT Press
Published: 2022

The open access edition of this book was made possible by generous funding and support from MIT Press Direct to Open



The MIT Press

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Stone Serif and Stone Sans by Westchester Publishing Services.

Library of Congress Cataloging-in-Publication Data

Names: Baber, Christopher, 1964– author.
Title: Embodying design : an applied science of radical embodied cognition / Christopher Baber.
Description: Cambridge, Massachusetts : The MIT Press, [2021] | Includes bibliographical references and index.
Identifiers: LCCN 2021033926 | ISBN 9780262543781 (paperback)
Subjects: LCSH: Expert systems (Computer science) | Human-machine systems. | Thought and thinking. | Artificial intelligence.
Classification: LCC QA76.76.E95 B22 2021 | DDC 006.3/3—dc23
LC record available at https://lccn.loc.gov/2021033926