Sneaking Up On the Hard Problem of Consciousness*

Benjamin Kuipers
Computer Science Department
University of Texas at Austin
Austin, Texas 78712 USA
kuipers@cs.utexas.edu

Abstract

We discuss several aspects of consciousness — the Easy Problem, the Intentionality Problem, and the Hard Problem — from the pragmatic perspective of artificial intelligence and robotics. Our computational approach is driven by the enormous information content of the sensory stream, and the properties of methods by which an agent may cope with its demands.

Introduction

Consciousness is one of the most intriguing and mysterious aspects of the phenomenon of mind. Artificial Intelligence (AI) is a scientific field built around the creation of computational models of mind (including such methods as neural networks, probabilistic inference, and dynamical systems as well as logic-based knowledge representation and inference). Computational approaches to understanding the phenomena of mind have been controversial, to say the least, but nowhere more than when applied to the problem of consciousness.

In a recent paper (Kuipers 2005), I described how the problem of consciousness looks to a researcher in AI and robotics, sketched out a computational model of consciousness, and evaluated its prospects against a set of eleven criteria "that any philosophical-scientific theory should hope to explain" according to John Searle (2004), a prominent philosopher and critic of AI. This paper extends that argument, fills some gaps, and attempts to make some useful distinctions.

Inspired by the important distinction between the "Easy" and "Hard" problems of consciousness (Chalmers 1996), and by the core issue behind the famous "Chinese room" story (Searle 1980)), we will consider three major aspects of the problem of consciousness: the Easy Problem, the Intentionality Problem, and the Hard Problem.

*This work has taken place in the Intelligent Robotics Lab at the Artificial Intelligence Laboratory, The University of Texas at Austin. Research of the Intelligent Robotics lab is supported in part by grants from the National Science Foundation (IIS-0413257 and IIS-0713150), from the National Institutes of Health (EY016089), and by an IBM Faculty Research Award.

Overview

The key ideas here are the following.

1. The sensory data stream provides information to the agent at an extremely high rate (gigabits/second).
2. This information is managed and compressed by selecting, tracking, and describing spatio-temporal portions of the sensory input stream.
3. A collection of parallel processes operates on information from the sensory input stream to construct a coherent sequential narrative describing the agent's sensorimotor interaction with the world.
4. The agent autonomously learns intentionality by constructing models of hypothetical entities in the external world. These entities explain regularities in the sensorimotor interaction, and serve as referents for the symbolic knowledge representation.
5. The high information content of the sensory stream allows the agent to continually evaluate these hypotheses, refuting the ones that result in poor predictions. The high information content of the sensory input stream explains certain key features of subjective experience.

The Easy Problem

The "Easy Problem" is: *What does consciousness do for us, and how does it work*? Only a philosopher could call this problem "Easy", since solving it will likely require decades at least, and dozens or hundreds of doctoral dissertations. What the name means is that scientists applying the methods of various disciplines have been able to formulate useful technical statements of the problem, and they have tools that apply to those problem statements. Progress may be difficult, but we know what it means. (The "Hard Problem" does not enjoy these benefits.)

Trackers into the Firehose of Experience

To a researcher in AI and robotics, one of the driving forces behind cognitive architecture is the need to cope with the enormous volume of sensory data, arriving asynchronously along many different channels. This is the "firehose of experience" in the title of (Kuipers 2005), where the first version of this theory is presented in more detail.

The primitive elements in this architecture are called *trackers*. Each tracker can be thought of as having two ends. One end consists of a number of pointers into the firehose of experience, designating a spatio-temporal region in the input stream, tracking that region as its natural boundaries evolve in real time. The other end consists of a dynamic symbolic representation of that particular portion of sensory experience, supporting inference about its static properties, its current state, and its history. The symbolic representation is "dynamic" in the sense that the values of certain attributes are automatically updated by processes operating on the tracked elements of the sensory stream. For example, one tracker might describe the changing location and shape of a pedestrian walking through the agent's field of view. Another might dynamically describe the agent's pose within the frame of reference of the enclosing room as the agent moves through it.

Some trackers integrate information from multiple asynchronous sensor streams. If an explosion occurs, information from the sudden noise and sudden flash of light arrive at common parts of the brain after different delays (about 50 ms for the auditory channel; about 200 ms for the visual channel). They are experienced as simultaneous, and described as aspects of the same event, in spite of significant differences in absolute time.

The idea of trackers is not new. Versions of the sensorimotor tracker concept include Minsky's "vision frames" (1975), Marr and Nishihara's "spatial models" (1978), Ullman's "visual routines" (1984), Agre and Chapman's "indexical references" (1987), Pylyshyn's "FINSTs" (1989), Kahneman and Triesman's "object files" (1992), Ballard, et al, "deictic codes" (1997), and Coradeschi and Saffiotti's "perceptual anchoring" (2003).

Trackers are created and destroyed quite frequently, with perhaps dozens or even hundreds active at any given time. They make it possible to "use the world as its own model", directing attention to a particular aspect of the world to answer a query, rather than attempting to retrieve or infer an answer from stored knowledge. Trackers may have a hierarchical structure, allowing the sensory image of a person, for example, to be tracked at varying levels of detail: entire body; head-torso-arms-legs; upper-arm-forearm-hand; palm-fingers; etc (Marr & Nishihara 1978).

Creating a Coherent Sequential Narrative

The cognitive architecture must be organized to make use of the information provided by the trackers. There appears to be a growing consensus that the mind includes a collection of processes that interact to create a *coherent sequential narrative* from its multiple parallel asynchronous sensory input streams. This narrative is the agent's explanation to itself of what is going on around it. Just as an individual tracker provides an index into the sensory stream corresponding to a symbolic description, this coherent sequential narrative provides organizing structure on the agent's overall experience.

There is a great deal of work to be done to determine the precise structure of this architecture, but the consensus appears to be converging on some sort of Global Workspace Theory (Baars 1988), drawing on Minsky's "Society of Mind" (1985), Dennett's "multiple drafts" (1991), and others. Neuroscientists in search of the "neural correlates of consciousness" (Koch 2003; Edelman 1987) are identifying neural processes in the brain that appear to be participating in a similar architecture.

Natural selection ensures that, on average, these processes do a pretty good job of describing the relevant objects and events in the external world. Some of the processes are innate to the individual, wired into the brain, but were "learned" by the species over evolutionary time. Other processes are learned by the individual from its own experience.

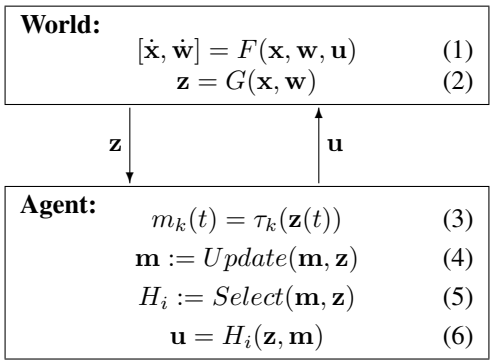


Figure 1: **Dynamical model of a cognitive agent**
The cognitive agent can be modeled at a high level as a dynamical system interacting with its environment. For the cognitive agent, its physical body is part of its environment. At any time t, the agent receives a sense vector $\mathbf{z}(t)$ and sends a motor vector $\mathbf{u}(t)$ to the world. The robot's body has state vector $\mathbf{x}(t)$ whose derivative is denoted $\dot{\mathbf{x}}(t)$, and the state vector of the rest of the world is described by $\mathbf{w}$ and $\dot{\mathbf{w}}$. The functions F and G represent the physics of the world and the sensor model, and neither is known to the agent. The agent acts by selecting a control law H_i based on the current sensor input $\mathbf{z}$ and the symbolic state $\mathbf{m}$ of its internal computational processes. Given this control law, equations (1,2,6) define a dynamical system, describing how the robot-environment system evolves until a new control law is selected. Meanwhile, trackers τ_k (equation 3) contribute dynamic descriptions to the computational state $\mathbf{m}$.

Defining Consciousness

According to the model proposed in (Kuipers 2005), what it means for an agent to be conscious, to have a subjective, first-person view of the world, is for the agent to have the interaction with its world described in Figure 1. This includes:

1. a high-volume sensor stream $\mathbf{z}(t)$ and a motor stream $\mathbf{u}(t)$ that are coupled, through the world and the robot's body, as described by equations (1-2);
2. a non-trivial collection of trackers $m_k(t) = \tau_k(\mathbf{z}(t))$ grounded in the sensor stream (equation 3) capable of providing dynamically updated symbolic descriptions for the agent's knowledge representation system, with top-down and bottom-up activation methods;

3. a non-trivial collection of control laws $\mathbf{u}(t) = H_i(\mathbf{z}(t), \mathbf{m}(t))$ (equations 5-6) that can be used to implement reasonably reliable actions in the world;
4. a sufficiently good correspondence between the actual properties of action and perception in the physical world (1-2), and the agent's symbolic theory of the world ($\mathbf{m}(t)$ (equation 4)) including symbols grounded via trackers (3) and actions implemented by control laws (5-6), so that the agent can interact effectively with its world. (This correspondence is the subject of the Intentionality Problem.)

This definition omits an important factor. High-volume closed-loop interaction is necessary but not sufficient. In the metaphorical language that Bernard Baars uses to describe his Global Workspace Theory, this definition covers not only the processes in the spotlight of attention or waiting on the stage as candidates for the spotlight, but also all the active processes out in the "audience", participating in cognition but far from the spotlight of consciousness.

The coherent sequential narrative constructed to explain experience seems to be a necessary additional factor. That narrative describes the events and activities that occupy the focus of consciousness. Sometimes the shifts in the focus of consciousness from one thing to another have explanations in terms of events represented in the narrative ("Suddenly there was an explosion!"), but other times, the cause of the shift may not be a represented event, so it appears to "just happen", like flipping the Necker cube.

McDermott (2001) and others argue that vivid conscious experience — the experience of qualia — is a retrospective phenomenon, using the coherent sequential narrative to access portions of the sensory input stream from the recent past. "Recent" in this case means within the last 50 to 500 milliseconds, so a great deal of sensory information remains available in short-term sensory memory. Processes operating on the coherent sequential narrative determine which trackers fall within the "spotlight of consciousness", making them globally visible to the "audience" of active processes, and which may simply participate in a particular low-level control loop without conscious experience (Baars 1988).

The Intentionality Problem

The "Intentionality Problem" is: *How can symbols in an internal cognitive knowledge representation refer to objects and events in the external world*? Or equivalently, *Where does meaning come from*? The core of Searle's "Chinese room" argument (Searle 1980) is that the mind necessarily *has* intentionality (the ability to refer to objects in the world), while computation (the manipulation of formal symbols according to syntactic rules) necessarily *lacks* intentionality. Therefore (claims Searle), the mind must be more than a computation.

The same problem comes up in a much more pragmatic form in point 4 of the definition of consciousness presented previously, which specified that the internal knowledge representation, grounded by trackers in the sensory input stream, must correspond sufficiently well with the properties of the external world for useful predictions and actions to be possible. Since the agent has no direct access to the state of the world, and its only indirect access is through its own sensory and motor streams, the problem of establishing and maintaining such a correspondence is critical.

Note that Searle, and I, and everyone else, are born locked inside our own skulls, receiving coded information along nerve fibers, and sending coded responses along other nerve fibers. We humans have the same intentionality problem that the Chinese Room illustrates. The source of the meaning by which knowledge representations refer to the external world is as much a mystery for us biological humans as it is for computational systems.

William James describes the baby as perceiving the world as "one great blooming, buzzing confusion". Any roboticist recognizes in this phrase the difficulty of interpreting the raw elements of the sensory stream, and the difficulty of accomplishing anything useful with incremental motor signals. We will refer to these together as the "pixel level" of understanding of the sensorimotor system.

We have taken significant steps toward learning intentionality. The Spatial Semantic Hierarchy (Kuipers & Byun 1991; Kuipers 2000) maps an unknown environment by identifying *locally distinctive states* and linking them into a topological map. The ability of a symbol to refer to a distinctive state in the physical environment depends on the behaviors of the dynamical systems defined by the control laws, not on any pre-existing intentionality in the set of symbols. Pierce and Kuipers (1997) showed that these control laws could be learned from the dynamical regularities in the robot's own experience with its uninterpreted sensors and effectors, constrained by their causal connections with the environment. (An appendix provides more detail.) Modayil and Kuipers (2007) have developed these into methods for learning to individuate, track, and describe coherent objects from the "blooming, buzzing confusion" of sensory input, and then to learn meaningful actions to perform on them.

A tracker follows a spatio-temporal region in the sensory input stream, and creates a dynamic symbolic description of it. We observers may know that such a region is the sensory projection of an object in the external world, but the agent doesn't have that knowledge.

The learning agent constructs a model of the world. Each tracker posits the existence of external entities whose projections onto the sensors account for its portion of the sensory stream. The properties of those entities are inferred from, and account for, the information observed in the sensory stream. This is a version of Quine's "web of belief" (Quine 1961).

The meanings to which the symbols in the agent's knowledge representation refer are the entities in that constructed model. Meanings therefore do not reside in the external world, but in the constructed model we build of it. If the agent's constructed model corresponds sufficiently well with the external world, then it can function effectively, and it has created its own intentionality. When the internal model and external world diverge, plans and predictions fail, and the sensory system provides relevant information that can be used to correct the model. If the divergence is sufficiently serious, the agent becomes non-viable, so natural selection ensures the quality of the constructed model.

The Hard Problem

The "Hard Problem" is: "*Why does consciousness feel like anything at all*?" Suppose that the mind is a computation, running on the physical substrate of the brain. Why should a computational process — even one that constructs its own intentionality and builds a coherent sequential narrative from its experience — feel like anything at all to the agent?

It is undeniable that many experiences "feel like" something. Pain hurts, sugar tastes sweet, the sight of a loved one after an absence raises feelings that are strong and real, even and especially though they can't be fully articulated. In the words of Francisco Varela, "*. . . why is it that consciousness feels so personal, so intimate, so central to who we are*, . . . " (Blackmore 2006, p. 226).

We are not zombies (or at least I am not). Why not?

This problem is Hard, even to a philosopher. So far it has resisted all attempts even to state what it would mean to provide a solution. It's not just that we can't find a solution. We can't even figure out what a solution would look like.

However, perhaps we can sneak up on the Hard Problem and get a closer peek at what makes it tick. Rather than directly approach the question of why anything feels like anything at all, we will ask why some experiences are more vivid than others.

Why are some experiences more vivid than others?

Looking *now* at the apple sitting in front of me, I experience a vivid perceptual image. This apple is very round, and a light greenish-yellow with a few brownish-red streaks. (Someone else might describe it as yellowish-green or with reddish-brown streaks.) This is a classic *quale*, a primary sensory experience. I claim that it is vivid, in part, because my internal symbolic concept of this apple is directly bound to the corresponding region in my sensory input stream, which provides a huge flow of information in real time.

Writing *now* some days later, my recalled image of that apple is still vivid, but not nearly as vivid as the experience itself. I can recall fragments ("snapshots?") of the sensory experience of perceiving the apple, but they are clearly incomplete. Even with effort, there are questions about the apple that I cannot answer now, with the amount of stored information available to me, that could have been effortlessly answered during the experience itself, simply by shifting my focus of attention.

You, the reader, have just read descriptions of my experience with a particular apple. But it is not your apple, or your experience, so your concept of this apple, like mine of some apple I read about, is much less vivid than either direct experience or personal memory.

These stories illustrate three widely separated points on a spectrum of the vividness of subjective experience. I claim that the differences in vividness are well-explained by the differences in information content. Ongoing visual experience of the apple means that the agent has trackers directly connected to the "firehose of experience", with its vast information content and its potential for answering new questions with a quick change of focus of attention. Memory of personal experience draws on whatever can be captured from the sensory "firehose" and stored in long-term memory. (Vivid dreams and evoked memory from direct neural stimulation probably fall between these first two points on the spectrum.) A word like "apple" in a written story transmits perhaps a few dozen bits. What vividness it has comes from evoking personal memories in the reader.

Why is subjective experience so personal?

Long-term memory is enormous. Like snowflakes, no two separately-created multi-mega-pixel digital images are ever identical, simply because of the huge number of bits they encode, and the number of unpredictable processes that determine those bits. Likewise with sensory experiences. Each snapshot from the sensory stream, and each fragment stored in long-term memory has huge information content.

The sheer number of bits in a particular sensory experience makes it astronomically unlikely that any other individual could have precisely the same experience. Even more so the entire contents of long-term memory. The sheer number of bits, created through individual sensory experience and stored in an agent's long-term memory, ensures that long-term memory must be unique and personal to its self.

Note that this argument depends on the number and unpredictability of the low-level processes that create sensory experiences. If memory is created purely through symbolic input, or if bulk memory can be backed up and restored as with a disk drive, then it might be possible to create identical individuals.

Specific pieces of information are easily retrieved, even from the enormous store of long-term memory, by following paths of associations. Each step in the path might require only a few bits of information, for example from a word in a sentence. However, it is clear from personal experience that, on occasion, a vivid sensory experience can leapfrog over the associative paths to evoke something buried deep within long-term memory.

The information content of such a link must be enormous, because it is activated by a sensory experience (consisting of many bits), and it retrieves a target from within the large space of long-term memory (requiring an address of many bits). Like sensory experience, and like the pattern of the simpler symbolic links, the sheer information content of such a direct link makes it, with very high probability, distinctive to the individual agent and therefore unique and personal.

Can qualia be learned and taught?

Certainly! Anyone who has deliberately learned to discriminate among experiences with chocolate or wine (or any other domain of sensory experience) knows that the categories for subjective experience can and do change with experience, even explicitly guided pedagogical experience.

The young child might distinguish sweet from non-sweet, and then among vanilla, chocolate, and strawberry. The older child distinguishes milk chocolate from dark. The adult learns to recognize and appreciate (or not) the distinctive flavors of Hershey, Ghirardelli, Lindt, and many others.

Qualia, therefore, do not correspond to biologically determined categories of sensory input. They correspond to

learned categories of represented experience, bound to portions of the sensory input stream. In the framework we are describing, sensory trackers are learned from experience, and qualia correspond to active trackers.

Different emotions feel different. But why?

Another way to sneak up on the Hard Problem is to consider why different emotions feel different. After all, emotions are a particular class of subjective experience, and there is a long literature on the nature and determinants of emotion.

William James (1884) attributed emotion to a combination of bodily arousal and the perceived situation. However, bodily arousal by itself is not sufficiently discriminating to account for the range and variety of experienced emotions. While arousal is important, other factors contribute to determining the emotional content of experience. A classic series of experiments showed that subjects who were aroused by injections of adrenaline, but then presented with different perceptual or cognitive influences, had dramatically different perceptions of their own emotional state (Maranon 1924; Cantril & Hunt 1932; Schachter & Singer 1962).

This suggests that what makes emotion feel like anything at all is bodily arousal, but what it actually feels like (the emotional *content*) is determined by cognitive and situational factors.

Why should information feel like anything?

Generalizing from emotion, we conjecture that what subjective experience "feels like" is determined in part by bodily arousal due to information transfer (the firehose of experience), combined with factors from the content of the information and the cognitive and behavioral context of the agent.

High-bandwidth information transfer is necessarily realized as a physical process involving rapid state-changes in a physical device. Such a process consumes and dissipates energy. An agent with suitable sensors to monitor its own physical state can sense the rate at which information is being processed.

This is a type of physiological arousal, which can then be interpreted in the context of the content of the information being provided, as well as the other goals and activities of the agent. Thus, qualia "feel like something" because the body senses the rush of information transfer, and associates a content-dependent interpretation with that feeling.

Conclusions

We approach the problem of consciousness from the pragmatic design perspective of AI and robotics. One of the major requirements on an embodied agent is the ability to cope with the overwhelming information content of its own sensory input (the "firehose of experience"). A plausible cognitive architecture that meets this requirement includes *trackers* that ground dynamic symbolic descriptions in spatio-temporal regions of the sensory stream, and a *coherent sequential narrative* that explains the objects and events from the external world that are observed in the sensory stream. Researchers from a variety of perspectives appear to be converging on such an architecture, which would be a solution to the Easy Problem of consciousness.

The Intentionality Problem applies to any embodied agent, human or robot, that interacts with the world only through coded sensor and motor signals. We argue that there is no magic, for humans or robots, whereby symbols inside the mind can refer, directly and correctly, to corresponding objects in the outside world. On the other hand, we can exhibit early versions of learning algorithms that can construct explanations for the regularities of pixel-level sensorimotor interaction in terms of higher-level entities such as places, paths, objects and actions. The "meaning" of a symbol in the internal knowledge representation is an entity hypothesized by such a learning algorithm, that is, another internal construct. If these internal entities correspond usefully with the external world, the agent will be able to plan and act effectively. If not, not.

We have attempted to "sneak up" on the Hard Problem by offering relative information content as an explanation for why different experiences have different levels of vividness. This leaves open the Hard Problem itself: *Why should any amount of information transfer feel like anything at all*? However, we do know that information transfer in an embodied agent necessarily corresponds to some sort of physical state-changes, which must have physical correlates that can be sensed. Drawing on an analogy with classic models of emotion, we may speculate that it is the physical correlates of raw information transfer that "feels like anything at all", and that "what it feels like" depends on the content of the information and the cognitive and behavioral context of the agent.

References

Agre, P. E., and Chapman, D. 1987. Pengi: An implementation of a theory of activity. In *Proc. 6th National Conf. on Artificial Intelligence (AAAI-87)*. Morgan Kaufmann.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge University Press.

Ballard, D. H.; Hayhoe, M. M.; Pook, P. K.; and Rao, R. P. N. 1997. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences* 20(4):723–767.

Blackmore, S. 2006. *Conversations on Consciousness*. Oxford University Press.

Cantril, H., and Hunt, W. A. 1932. Emotional effects produced by the injection of adrenaline. *American Journal of Psychology* 44(2):300–307.

Chalmers, D. J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

Coradeschi, S., and Saffiotti, A. 2003. An introduction to the anchoring problem. *Robotics and Autonomous Systems* 43(2-3):85–96.

Dennett, D. 1991. *Consciousness Explained*. Little, Brown & Co.

Edelman, G. 1987. *Neural Darwinism: The Theory of Neuronal Group Selection*. NY: Basic Books.

James, W. 1884. What is an emotion? *Mind* 9:188–205.

Kahneman, D., and Treisman, A. 1992. The reviewing of object files: object-specific integration of information. *Cognitive Psychology* 24:175–219.

Koch, C. 2003. *The Quest for Consciousness: A Neurobiological Approach*. Englewood CO: Roberts & Company Publisher.

Kuipers, B. J., and Byun, Y.-T. 1991. A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Journal of Robotics and Autonomous Systems* 8:47–63.

Kuipers, B. 2000. The Spatial Semantic Hierarchy. *Artificial Intelligence* 119:191–233.

Kuipers, B. 2005. Consciousness: drinking from the firehose of experience. In *Proc. 20th National Conf. on Artificial Intelligence (AAAI-05)*, 1298–1305. AAAI.

Maranon, G. 1924. Contribution à l'etude de l'action émotive de l'adrénaline. *Française d'Endocrinologie* 21:301–325.

Marr, D., and Nishihara, H. K. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society B* 200:269–294.

McDermott, D. V. 2001. *Mind and Mechanism*. Cambridge MA: MIT Press.

Minsky, M. 1975. A framework for representing knowledge. In Winston, P. H., ed., *The Psychology of Computer Vision*. NY: McGraw-Hill.

Minsky, M. 1985. *The Society of Mind*. NY: Simon and Schuster.

Modayil, J., and Kuipers, B. 2007. Autonomous development of a grounded object ontology by a learning robot. In *National Conference on Artificial Intelligence (AAAI-07)*.

Pierce, D. M., and Kuipers, B. J. 1997. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence* 92:169–227.

Provost, J.; Kuipers, B. J.; and Miikkulainen, R. 2006. Developing navigation behavior through self-organizing distinctive-state abstraction. *Connection Science* 18(2):159–172.

Provost, J. 2007. *Reinforcement Learning in High-Diameter, Continuous Environments*. Ph.D. Dissertation, Computer Science Dept., University of Texas at Austin.

Pylyshyn, Z. W. 1989. The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition* 32:65–97.

Quine, W. V. O. 1961. Two dogmas of empiricism. In Quine, W. V. O., ed., *From a Logical Point of View*. Harvard University Press, second, revised edition.

Schachter, S., and Singer, J. E. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological Review* 69:379–399.

Searle, J. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–424.

Searle, J. R. 2004. *Mind: A Brief Introduction*. Oxford University Press.

Ullman, S. 1984. Visual routines. *Cognition* 18:97–157.

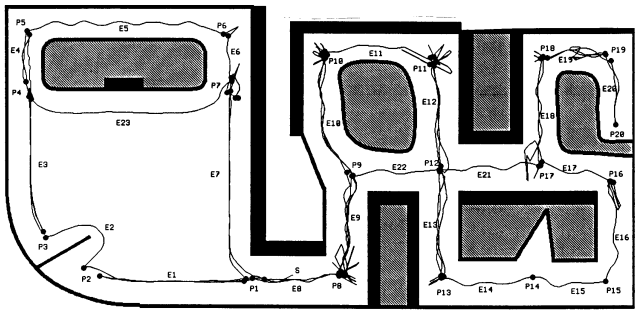


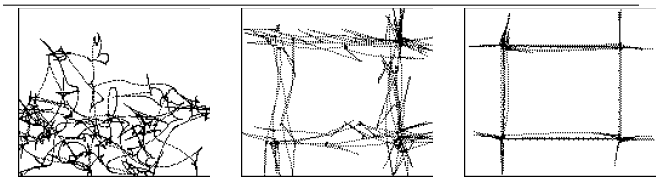
Figure 2: Hill-climbing control laws (seeking states equidistant from obstacles) define locally distinctive places, and trajectory-following control laws (midline- or wall-following) define path segments joining them. By providing reliable motion among distinctive states, these control laws enable a principled abstraction from the continuous world to a discrete topological map.

Appendix: Learning from Uninterpreted Sensors and Effectors

In the Spatial Semantic Hierarchy (SSH) (Kuipers & Byun 1991; Kuipers 2000), a robot abstracts its continuous environment to a discrete graph — the topological map — which includes symbols for places, paths, and the actions linking them. These symbols are grounded in the behavior of control laws in the environment (Figure 2).

Pierce and Kuipers (1997) showed how a robot learning agent, starting with an uninterpreted set of sensors and effectors, could learn these these symbols *for itself*, including its own collection of hill-climbing and trajectory-following control laws (Figure 3).

More recent work (Provost, Kuipers, & Miikkulainen 2006; Provost 2007) has used self-organizing maps and hierarchical reinforcement learning to improve these methods. We have also taken steps toward learning objects, actions, and their affordances and effects, to the point of being able to devise and carry out simple plans (Modayil & Kuipers 2007).



(a) (b) (c)

Figure 3: Exploring a simple world at three levels of competence. (a) The robot wanders randomly while learning a model of its sensorimotor apparatus. (b) The robot explores by randomly choosing applicable homing and open-loop path-following behaviors based on the static action model while learning the dynamic action model (see text). (c) The robot explores by randomly choosing applicable homing and closed-loop path-following behaviors based on the dynamic action model. (Pierce & Kuipers 1997)