Consciousness: Drinking from the Firehose of Experience[*]
(revised and expanded)

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

Abstract

The problem of consciousness has captured the imagination of philosophers, neuroscientists, and the general public, but has received little attention within AI. However, concepts from robotics and computer vision hold great promise to account for the major aspects of the phenomenon of consciousness, including philosophically problematical aspects such as the vividness of qualia, the first-person character of conscious experience, and the property of intentionality. This paper presents and evaluates such an account against eleven features of consciousness "that any philosophical-scientific theory should hope to explain", according to the philosopher and prominent AI critic John Searle.

The Problem of Consciousness

Artificial Intelligence is the use of computational concepts to model the phenomena of mind. Consciousness is one of the most central and conspicuous aspects of mind. In spite of this, AI researchers have mostly avoided the problem of consciousness in favor of modeling cognitive, linguistic, perceptual, and motor control aspects of mind. However, in response to a recent discussion of consciousness by the well-known philosopher and AI critic John Searle (Searle 2004), it seems to me that we are in a position to sketch out a plausible computational account of consciousness.

Consciousness is a phenomenon with many aspects. Searle argues that the difficult aspects of consciousness are those that make up the subjective nature of first-person experience. There is clearly a qualitative difference between thinking about the color red with my eyes closed in a dark room, and my own immediate experiences of seeing a red rose or an apple or a sunset. Philosophers use the term *qualia* (singular *quale*) for these immediate sensory experiences. Furthermore, when I see a rose or an apple, I see it as an object in the world, not as a patch of color on my retina. Philosophers refer to this as *intentionality*.

The position I argue here is that the subjective, first-person nature of consciousness can be explained in terms of the ongoing stream of sensorimotor experience (the "firehose of experience" of the title) and the symbolic pointers into that stream (which we call "trackers") that enable a computational process to cope with its volume.

Consciousness also apparently constructs a plausible, coherent, sequential narrative for the activities of a large, unsynchronized collection of unconscious parallel processes in the mind. How this works, and how it is implemented in the brain, is a fascinating and difficult technical problem, but it does not seem to raise philosophical difficulties.

Other Approaches to Consciousness

There have been a number of recent books on the problem of consciousness, many of them from a neurobiological perspective. The more clinically oriented books (Sacks 1985; Damasio 1999) often appeal to pathological cases, where consciousness is incomplete or distorted in various ways, to illuminate the structure of the phenomenon of human consciousness through its natural breaking points. Another approach, taken by Crick and Koch (Crick & Koch 2003; Koch 2003), examines in detail the brain pathways that contribute to visual attention and visual consciousness in humans and in macaque monkeys. Minsky (1985), Baars (1988), and Dennett (1991) propose architectures whereby consciousness emerges from the interactions among large numbers of simple modules.

John Searle is a distinguished critic of *strong AI*: the claim that a successful computational model of an intelligent mind would actually *be* an intelligent mind. His famous "Chinese room" example (Searle 1980) argues that even a behaviorally successful computational model would fail to have a mind. In some sense, it would just be "faking it."

In Searle's recent book on the philosophy of mind (Searle 2004), he articulates a position he calls *biological naturalism* that describes the mind, and consciousness in particular, as "entirely caused by lower level neurobiological processes in the brain." Although Searle rejects the idea that the mind's relation to the brain is similar to a program's relation to a computer, he explicitly endorses the notion that the body is a biological machine, and therefore that machines (at least biological ones) can have minds, and can even be conscious. In spite of being nothing beyond physical processes, Searle

[*] 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), from the National Institutes of Health (EY016089), and by an IBM Faculty Research Award.

Revised and expanded version of a paper presented at
National Conference on Artificial Intelligence (AAAI-05)
http://www.cs.utexas.edu/˜qr/papers/Kuipers-aaai-05.html

holds that consciousness is not *reducible* to those physical processes because consciousness “has a first-person ontology” while the description of physical processes occurring in the brain “has a third-person ontology.” He lays out eleven central features of consciousness “that any philosophical-scientific theory should hope to explain.” In the following three sections, I describe how a robotics researcher approaches sensorimotor interaction; propose a computational model of consciousness; and evaluate the prospects for using this model to explain Searle’s eleven features of consciousness.

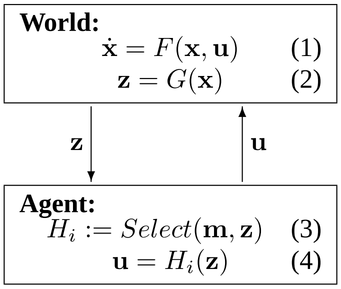
Sensorimotor Interaction in Robotics

When a robot interacts continually with its environment through its sensors and effectors, it is often productive to model that interaction as a continuous dynamical system, moving through a continuous state space toward an attractor. In the situations we will consider, such a dynamical system can be approximated by a discrete but fine-grained computational model, so by taking this view of the robot we are not moving outside the domain of computational modeling.

A Simple Robot in a Static World

Consider a simple robot agent in a static environment. In a static world, the only possible changes are to the state of the robot’s body within the environment, which is represented by a time-varying state vector $\mathbf{x}(t)$. The derivative of $\mathbf{x}$ with respect to time is written $\dot{\mathbf{x}}$. For a simple mobile robot moving on a planar surface, $\mathbf{x}$ would have the form (x, y, θ), representing the *pose* (position (x, y) plus orientation θ) of the robot within its environment. A robot with a more complex body would have a larger state vector $\mathbf{x}$.

We distinguish the environment and the robot’s body from the computational process (which we will call the “agent”), that receives the sense vector $\mathbf{z}(t)$ from the environment and determines a motor vector $\mathbf{u}(t)$ to send out to its body in the environment. Let $\mathbf{m}$ be the symbolic state of the agent’s internal computational process. Note that the agent has access to its sense vector $\mathbf{z}$, and can set its own motor vector $\mathbf{u}$, but it only has indirect access to its own state vector $\mathbf{x}$. The coupled system consisting of the robot agent and its environment can be described as a dynamical system. (Here we superimpose two useful standard representations: the block diagram and the differential equation.)



$$\dot{\mathbf{x}} = F(\mathbf{x}, \mathbf{u}) \qquad (1)$$

$$\mathbf{z} = G(\mathbf{x}) \qquad (2)$$

$$H_i := Select(\mathbf{m}, \mathbf{z}) \qquad (3)$$

$$\mathbf{u} = H_i(\mathbf{z}) \qquad (4)$$

Equation (1) describes how the robot’s state changes as a function of its current state and the motor vector. The function F represents the physics of the world and the robot’s body, including all the complexities of motor performance, wheel friction, barriers to motion, and so on. F is not known to the agent (or to the human researcher, who typically uses simplified approximations). Equation (2) describes the dependence of the robot’s sensor input on its current state. The function G is also extremely complex and not known to the agent. From time to time, based on its internal symbolic state $\mathbf{m}$ and its current observations $\mathbf{z}$, the agent selects (equation 3) a reactive control law H_i which determines the current value $\mathbf{u}(t)$ of the motor vector as a function of the current value $\mathbf{z}(t)$ of the sensor input (equation 4).

For a particular choice of the control law H_i, equations (1,2,4) together define the coupled robot-environment system as a dynamical system, which specifies trajectories of $(\mathbf{x}(t), \dot{\mathbf{x}}(t))$ that the robot must follow. The robot’s behavior alternates between (a) following a trajectory determined by a particular dynamical system until reaching a termination condition, and then (b) selecting a new control law H_j that transforms the coupled robot-environment system into a different dynamical system, with different trajectories to follow (Kuipers 2000).

The Firehose of Experience

When engineering a robot controller, a human designer typically works hard to keep the model tractable by reducing the dimensionality of the state, motor, and sensor vectors. However, as robots become more complex, or as we desire to apply this model to humans, these vectors become very high-dimensional.

The sensor stream $\mathbf{z}(t)$ is what I call “the firehose of experience” — the extremely high bandwidth stream of sensor data that the agent must cope with, continually. For a biological agent such as a human, the sense vector $\mathbf{z}$ contains millions of components representing the individual receptors in the two retinas, the cochleal cells in the two ears, and the many touch and pain receptors over the entire skin, not to mention taste, smell, balance, proprioception, and other senses. Robot senses are much simpler, but they still provide information at an overwhelming rate. (A stereo pair of color cameras alone generates data at over 440 megabits per second.) With such a high data rate, any processing applied to the entire sensor stream must be simple, local, and parallel. In the human brain, arriving sensory information is stored in some form of short-term memory, remains available for a short time, and then is replaced by newly arriving information.

In a biological agent, the motor vector $\mathbf{u}$ includes control signals for hundreds or thousands of individual muscles. An artificial robot could have dozens to hundreds of motors (though a simple mobile robot will have just two).

Modifying Searle’s Chinese Room metaphor (Searle 1980), in addition to comparatively infrequent slips of paper providing symbolic input and output, the room receives a huge torrent of sensory information that rushes in through one wall, flows rapidly through the room, and out the other wall, never to be recovered. Inside the room, John can examine the stream as it flows past, and can perhaps record fragments of the stream or make new symbolic notations based on his examination, in accordance with the rules specified in the room.

The “firehose of experience” provides information at a rate much greater than the available symbolic inference and storage mechanisms can handle. The best we can hope for is to provide pointers into the ongoing stream, so that relevant portions can be retrieved when needed.

Trackers in the Sensor Stream

The key concept for making sense of the “firehose of experience” is the *tracker*, a set of symbolic pointers into the sensor stream that maintains the correspondence between a higher-level, symbolically represented concept and its ever-changing image in the sensor stream. (Think of tracking a person walking across a scene while you are attending to something else in the scene.)

We will describe trackers in the framework of equations (1-4) by equations of the form

$$m_k(t) = \tau_k(\mathbf{z}(t)) \tag{5}$$

meaning that an individual tracker τ_k takes as input the sensor stream $\mathbf{z}(t)$ and produces as output the symbolic description $m_k(t)$, which is part of the symbolic computational state $\mathbf{m}(t)$ of the agent. The subscript k indicates that multiple trackers τ_k may be active at any given time.

An individual tracker τ_k may be created “top-down” by a symbolic inference process, or “bottom-up” triggered by detection of a feature in the sensory stream $\mathbf{z}(t)$. The human visual system includes parallel feature-detection mechanisms that trigger on certain “pop-out” colors and textures.

This is not a new idea. 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).

A tracker has its own control laws, updating its parameters from information within the sensor stream. These control laws embody its expections about the dynamics of the tracked concept, and how its perceptual image will appear (Blake & Yuille 1992; Hutchinson, Hager, & Corke 1996). The high volume and high temporal granularity of the sensor stream helps the trackers track more successfully. The feedback-based technology for tracking objects from changing sensor input has its roots in radar signal interpretation from the 1940s (Wiener 1948; Gelb 1974).

Quine (1961) describes human knowledge as a symbolic “web of belief” anchored at the periphery in sensorimotor experience. Trackers are the anchoring devices for symbols. We say that a tracker is *bound to* a spatio-temporal segment of the sensor stream when that portion of ongoing experience satisfies the criteria of the tracker’s defining concept, and when tracking is successful in real time. The tracker mediates between signal processing and symbol manipulation. At the signal processing end, the tracker implements a dynamical system keeping its pointers corresponding as closely as possible with the relevant portion of the sensor stream. At the symbol manipulation end, the tracker serves as a logical constant, with time-dependent predicates representing the attributes of the tracked object.

The location of the tracker within the sensor stream is regulated and updated by control laws, responding to the image properties expected for the tracked concept and their contrast with the background. Image processing strategies such as dynamical “snakes” (Blake & Yuille 1992) represent changing boundaries between figure and ground. With adequate contrast, the high temporal granularity of the sensor stream means that updating the state of the tracker is not difficult. With increasing knowledge about the sensor image of the tracked concept, the tracker can maintain a good expectation about the relevant portion of the sensor stream even in the presence of occlusion, poor contrast, and other perceptual problems.

Trackers implement the principle that “the world is its own best model.” When a tracker is bound to a portion of ongoing experience, current sensor data is easily available to whatever symbolic cognitive processes might be active, because the tracker provides efficient access into the correct portion of the sensor stream.

Properties of an actively tracked object can be retrieved more accurately and efficiently directly from the sensor stream, rather than by attempting to retrieve facts previously stored in memory. Human confidence in the completeness and quality of perception come from this ability to retrieve high-quality data from the sensor stream on demand, not from complete processing of the image (Ballard, Hayhoe, & Pelz 1995; O’Regan & Noë 2001).

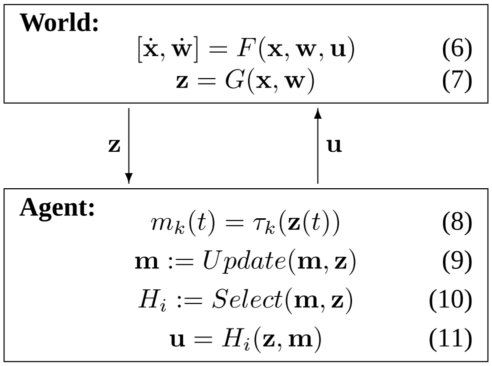
The phenomenon of “change blindness” (O’Regan & Noë 2001) illustrates some of the properties of trackers. People can fail to notice startlingly large changes in a visual scene due to intense focus of attention, or due to momentary breaks in the sensor stream that require the trackers to be rebound.

Trackers can be hierarchically structured. For example, a person tracker would have subordinate trackers for torso, head, arms, legs, and so on (Marr & Nishihara 1978). If an object moves to the periphery of the visual field, a hierarchical tracker may lose its more detailed components, becoming little more than a “blob” tracker representing only location and extent. But if more detailed information is needed, a quick saccade can bring the target back to the fovea, the hierarchical structure of the tracker is recreated, and questions are answered from vivid visual input as if it had been continually available (Ballard *et al.* 1997).

A Computational Account of Consciousness

We extend our previous simple model of the robot agent to incorporate trackers. We also extend it to non-static worlds by distinguishing between the state $\mathbf{x}$ of the robot’s body and the state $\mathbf{w}$ of the external world.

The trackers (8) provide active, time-varying assertions directly grounded in the sensor stream. Equation (9) refers to the update of the agent’s knowledge representation, based on its prior state and sensor input. The $Update$ function (9) encapsulates several important issues, including how the coherent sequential narrative of subjective consciousness is constructed (cf. *Unity*, below).



World:

$$[\dot{\mathbf{x}}, \dot{\mathbf{w}}] = F(\mathbf{x}, \mathbf{w}, \mathbf{u}) \qquad (6)$$

$$\mathbf{z} = G(\mathbf{x}, \mathbf{w}) \qquad (7)$$

Agent:

$$m_k(t) = \tau_k(\mathbf{z}(t)) \qquad (8)$$

$$\mathbf{m} := Update(\mathbf{m}, \mathbf{z}) \qquad (9)$$

$$H_i := Select(\mathbf{m}, \mathbf{z}) \qquad (10)$$

$$\mathbf{u} = H_i(\mathbf{z}, \mathbf{m}) \qquad (11)$$

Defining Consciousness

According to the model proposed here, what it means to be conscious, what it means for an agent to have a subjective, first-person view of the world, is for the agent to have:

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 (6-7);
2. a non-trivial collection of trackers $m_k(t) = \tau_k(\mathbf{z}(t))$ grounded in the sensor stream (equation 8) capable of providing symbols 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 10-11) that can be used to implement reasonably reliable actions in the world;
4. a sufficiently good correspondence between the agent's symbolic theory of the world ($\mathbf{m}(t)$ (equation 9), with symbols grounded via trackers (8) and actions implemented by control laws (10-11)) and the properties of action and perception in the physical world (6-7), so the agent can interact effectively with its world.

This is an unabashedly "Strong AI" claim. It does not appeal to the Turing Test (Turing 1950), to claim that such a robot behaves *as if* it were conscious. If a robot satisfies the structural criteria above, we assert that it genuinely *is* conscious.

How can we justify this? In the next section, we will consider a catalog of properties of human consciousness. We argue that this computational model explains some of these properties, and is compatible with a number of different explanations of the rest (though it may not select among them). Therefore, we propose that a computational model of this form (once it is filled out technically) explains consciousness in humans.

If all of the known properties of consciousness in humans can be explained by a computational model, where the explanations depend on a certain set of structural conditions, then there seems to be no basis for denying that other systems that satisfy the same structural conditions are also conscious.

Admittedly, the structural model doesn't help us know what it "feels like" to be such a robot, any more than we can know what it feels like to be a bat, or a dolphin, or John Searle. For its consciousness to be "human-like", a robot would have to be sufficiently similar to humans in sensorimotor system, symbolic processing capabilities, knowledge base, and cultural background. What "sufficiently" means here remains to be seen. Consider the difficulty in comparing human and dolphin consciousness, due to differences in sensorimotor system, environment, knowledge, and culture.

Evaluating a Theory of Consciousness

It is not yet possible to build a robot with sufficiently rich sensorimotor interaction with the physical environment, and a sufficiently rich capability for tracking and reasoning about its sensor and motor streams, to be comparable with human consciousness. The remaining barriers, however, appear to be technical rather than philosophical.

We begin evaluating this theory of consciousness by discussing how well such a computational model might account for eleven central features of consciousness "that any philosophical-scientific theory should hope to explain" (Searle 2004, pp. 134–145). Each of the following subsections is titled with Searle's name for a feature, followed by a quote from his description of it.

For some of these features — Qualitativeness, Subjectivity, Intentionality, Distinction between Center and Periphery, and Active and Passive — the dynamical tracker model of consciousness provides a specific *explanation*. For other features — Sense of Self, Unity, Situatedness, Gestalt Structure, Mood, and Pleasure/Unpleasure — there may be several possible explanations, all of which are *expressible* within the dynamical tracker model.

Qualitativeness

> *Every conscious state has a qualitative feel to it.* . . . [This includes] *conscious states such as feeling a pain or tasting ice cream* . . . [and also] *thinking two plus two equals four* (Searle 2004, p. 134).

The vividness, intensity, and immediacy of subjective experience are due to the enormous information content of the sensor stream $\mathbf{z}(t)$. There's a difference between thinking about the color red with my eyes closed in a dark room, and the immediate experiences (*qualia*) of seeing a red rose or apple or sunset. The intensity of subjective experience increases with the information content of the input: from text or verbal descriptions, to viewing a color photograph, to memories or dreams of experiences, to live multisensory experience.

Trackers do not capture the experience itself, but they provide structure to William James' "one great blooming, buzzing confusion". By providing rapid access to specified parts of the sensory stream, trackers (in vision at least) maintain the illusion that the entire visual field is perceived with the same high fidelity as the point of foveal attention (Ballard *et al.* 1997; O'Regan & Noë 2001).

If an attribute value such as the color red is stored as a symbol "red" in memory, its information content is determined by the number of other possible color symbols that could have been stored as a value of that attribute: at most a dozen bits or so. On the other hand, if a tracker is bound to a region in the sensor stream, the number of bits of color information streaming past, even in a small region, is orders of magnitude larger.

The higher information content of the sensor stream means that attribute values drawn from sensory experience necessarily include more distinctions than are available in, for example, common vocabulary. The reds of roses, apples, and sunsets are different, though their distributions may overlap. The agent who has experienced these qualia possesses more distinctions than one who hasn't, and can recognize the rarity of a sunset-colored rose.

Although it is implausible for the entire sensory stream to be stored in long-term memory, at least some qualia (e.g., the pain of a kidney stone or the smell of a freshly-baked madeleine) can be stored in memory and can serve as associative links to memories of previous experiences. The high information content of a quale makes it possible to select out a single distinctive association from the huge contents of long-term memory.

The practical value of qualia is that they help keep the hallucinations down. Symbolic (logical) theories are subject to multiple interpretations. Larger theories have fewer interpretations. Sensory grounding through trackers provides a huge number of additional axioms to such a theory, and thereby constrains its possible interpretations. Active trackers provide strong evidence, solidly grounded in the sensor stream, to eliminate incorrect perceptual hypotheses.

There is a compelling argument that perception requires abduction (Shanahan 2005). There must be a process that proposes hypotheses to account for ongoing sensor data. If this process were purely bottom-up (i.e., driven by the sensor data), then in the absence of the sensor stream, no hypotheses would be generated and no perception would take place. However, experience suggests that there are significant top-down and perhaps random processes for generating hypotheses. Under conditions of sensory deprivation, people tend to hallucinate, that is, to generate perceptual hypotheses poorly grounded in sensor input. The high information content of the sensor stream helps to keep the generation and refutation of perceptual hypotheses in balance.

Subjectivity

> *Because of the qualitative character of consciousness, conscious states exist only when they are experienced by a human or animal subject. . . . Another way to make this same point is to say that consciousness has a first-person ontology* (Searle 2004, p. 135).

Consciousness is experienced exclusively from a first-person point of view. (I reject Searle's explicit restriction of conscious experience to "a human or animal subject".)

What it means for an agent to have a first-person point of view is for it to have access to the sensor and motor streams from its own body. That is, its body is physically embedded in the world, and equations (6-7) describe the causal path from its actions $\mathbf{u}$ to its perceptions $\mathbf{z}$. By selecting a control law H_i, the agent creates a causal path from its sensory input $\mathbf{z}$ to its motor output $\mathbf{u}$, closing the loop and giving it some degree of control over its body. Only the agent itself has access to the sensor stream $\mathbf{z}(t)$ from its own body, or to the motor output stream $\mathbf{u}(t)$, and only the agent is in a position to select and impose a control law H_i relating them. (This individuation reflects biology. Robots may not have the same constraints. Also see the movie "Being John Malkovich.")

The agent can learn from experience which aspects of its perceptions are under its direct control, and which are not, therefore learning to separate its concept of itself ($\mathbf{x}$) from its concept of its environment ($\mathbf{w}$) (Philipona, O'Regan, & Nadal 2003). This distinction comes not from anatomy, but from the existence of tight control loops. Virtual reality and telepresence are subjectively compelling exactly because humans are quickly able to learn novel models of senses, actions, body, and world from interactive experience.

Exceptions to the agent's privileged access to its own sensorimotor system confirm this description of the first-person point of view. Searle (Searle 2004, p.142) cites an experiment by the neurosurgeon Wilder Penfield, who was able to stimulate a patient's motor neurons directly, to raise the patient's arm, prompting a response from the patient who said, "I didn't do that, you did." This corresponds to the surgeon being able to set $\mathbf{u}(t)$ directly, without the patient selecting a control law.

Unity

> *At present, I do not just experience the feelings in my fingertips, the pressure of the shirt against my neck, and the sight of the falling autumn leaves outside, but I experience all of these as part of a single, unified, conscious field* (Searle 2004, p. 136).

We experience the audio-visual surround as a single unified field, continuous in space and time, in spite of a variety of disturbing facts about our actual sensory input (Koch 2003). The fovea has vastly higher resolution than the periphery of the retina, and the blind spot has no resolution at all. The density of color-receptive cones is even more strongly biased toward the foveal area and away from the periphery. Auditory and visual evidence from the same event reaches the brain at different times. For example, it is well-known that reaction time to an auditory stimulus is about 50 ms faster than to a visual stimulus. Furthermore, the auditory stimulus from an event can be delayed up to about 80 ms from the visual stimulus without interfering with the perception of simultaneity.

The apparent unity of perception is a plausible coherent narrative, constructed 50-500 ms after the fact from evidence from parallel and irregular sources (Ballard *et al.* 1997; O'Regan & Noë 2001). Several mechanisms and cognitive architectures have been proposed to explain how this narrative is constructed. For example, Minsky's "Society of Mind" (1985), Baars' "Global Workspace Theory" (1988), Dennett's "Multiple Drafts Model" (1991), and others, propose that consciousness arises from the interaction of many simple cognitive modules that observe and control each other. The generally-linear stream of conscious thought is constructed, typically in fragments, by these modules from each others' outputs. Within this kind of architecture, trackers are the modules that interface between the sensor stream and the symbolic cognitive modules.

A number of technical and scientific questions remain to be answered about how the coherent conscious narrative is actually constructed from parallel and irregular sources of input. Global Workspace Theory (Baars 1988; 2002) appears to be the current best detailed computational model of this process.

In robotics, the Kalman Filter (Gelb 1974) is often used to predict the most likely trajectory of a continuous dynamical system (along with its uncertainty), given a model and an irregular collection of sensor observations (along with their uncertainties). The technical methods are different, but philosophically, the slightly retrospective construction of a coherent sequential narrative from irregular observations is no more problematical than a Kalman Filter.

Intentionality

> *My present visual perception, for example, could not be the visual experience it is if it did not seem to me that I was seeing chairs and tables in my immediate vicinity. This feature, whereby many of my experiences seem to refer to things beyond themselves, is the feature that philosophers have come to label "intentionality"* (Searle 2004, p. 138).

The core of Searle's "Chinese room" argument (Searle 1980) is that strong AI commits a category error with regard to intentionality. 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, the mind cannot be a computation.

However, intentionality is exactly what the tracker for a high-level concept delivers: it binds a portion of the current sensor stream to the symbolic description of an object (believed to be) in the external world. The relationship of intentionality follows from the causal connection from the external, physical world to the contents of the sensor stream, and thence to the internal symbols created by the trackers.

Searle's response to the "Robot Reply" (Searle 1980) acknowledges the importance of the causal connection between a robot's sensorimotor system and the world, but he claims that uninterpreted sensor and motor signals are just as free of intentionality as any formal symbols.

Presumably, Searle would argue that the intentionality provided by a tracker is merely "derived intentionality," coming from the mind of the human who programmed the algorithms and control laws that make the tracker work. This argument is vulnerable to a demonstration that effective trackers can be learned automatically from experience with uninterpreted sensors and effectors. In a preliminary form, Pierce and Kuipers (1997) have made just such a demonstration.

We have taken significant steps toward learning intentionality. The Spatial Semantic Hierarchy (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 intentionality in the preexisting 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.[1] Modayil and Kuipers (2004) have used related methods to learn to individuate and describe coherent objects from the "blooming, buzzing confusion" of sensory input.

We believe that learning methods like these can be extended to learn trackers for many kinds of distinctive configurations in the sensory stream. New symbols are defined, and their properties are learned, to refer to the objects of the trackers in the external world. The agent thus acquires intentionality *of its own*.

The Distinction between the Center and the Periphery

> *Some things are at the center of my conscious field, others are at the periphery. A good mark of this is that one can shift one's attention at will. I can focus my attention on the glass of water in front of me, or on the trees outside the window, without even altering my position, and indeed without even moving my eyes. In some sense, the conscious field remains the same, but I focus on different features of it* (Searle 2004, p. 140).

An individual tracker maintains a set of pointers into the sensor input stream that defines the features it attends to. The rest of the sensor stream is examined only enough to continue to track successfully, and to allow "pop-out" detection.

Some trackers are within the agent's focus of attention, in which case they constitute the "figure" part of the "figure-ground" distinction in the visual field. Other "ground" trackers outside the current focus of attention may track objects that could be attended to later, or they may contribute to maintaining *Situatedness* (below).

Thermostats and robot vacuum cleaners are coupled with the world to form simple dynamical systems. However, they fail to be conscious because they have a single fixed "figure", no "ground" at all, and no ability to shift focus of attention.

Situatedness

> *All of our conscious experiences come to us with a sense of what one might call the background situation in which one experiences the conscious field. The sense of one's situation need not be, and generally is not, a part of the conscious field. But, normally I am in some sense cognizant of where I am on the surface of the earth, what time of day it is, what time of year it is, whether or not I have had lunch, what country I am a citizen of, and so on with a range of features that I take for granted as the situation in which my conscious field finds itself* (Searle 2004, p. 141).

While the concept of the tracker is particularly clear when applied to images of objects that move within the visual

[1] See a newly-added appendix for more detail.

field, it applies equally well to tracking the location of the robot within a given frame of reference, for example, the current enclosing room. This concept of tracker can, in turn, be generalized to track motion through an abstract space such as time or a goal hierarchy. Such background situation trackers could potentially continue tracking with little or no attention.

Active and Passive Consciousness

> *The basic distinction is this: in the case of perception (seeing the glass in front of me, feeling the shirt against my neck) one has the feeling, I am perceiving this, and in that sense, this is happening to me. In the case of action (raising my arm, walking across the room) one has the feeling, I am doing this, and in that sense, I am making this happen* (Searle 2004, p. 142).

The agent's sensorimotor interface (equations 7 and 11) clearly divides into the sensor stream $\mathbf{z}(t)$, which is happening to the agent, and the motor stream $\mathbf{u}(t)$, by which the agent makes things happen.

Active control of perception by moving the eyes to bring a target into the fovea is accomodated by the current model, since the state of the eyes would be part of the robot's state vector $\mathbf{x}(t)$, and would be controlled by the motor vector $\mathbf{u}(t)$. Attentional processes such as giving a particular tracker more resources and allowing it to fill out its hierarchical structure more fully, could also be modeled as control laws whose effect is on the internal state $\mathbf{m}(t)$ of the agent.

The Gestalt Structure

> *We do not, for example, in normal vision see undifferentiated blurs and fragments; rather, we see tables, chairs, people, cars, etc., even though only fragments of those objects are reflecting photons at the retina, and the retinal image is in various ways distorted. The Gestalt psychologists investigated these structures and found certain interesting facts. One is, the brain has a capacity to take degenerate stimuli and organize them into coherent wholes. Furthermore, it is able to take a constant stimulus and treat it now as one perception, now as another* (Searle 2004, p. 143).

Each tracker looks for a certain structure in the sensor stream. When it finds it, that structure is foreground for that tracker, and the rest of the sensor stream is background. The findings of the Gestalt psychologists provide clues about the properties of individual trackers, of the process by which potential trackers are instantiated and become active, of the ensemble of active trackers, and perhaps even of the learning process by which trackers for new types of objects are learned.

For example, interpretation-flipping figures such as the Necker cube or the duck/rabbit figure suggest properties of the ensemble of active trackers, such as mutual exclusion and continued competition among the higher level of hierarchical trackers, while lower levels preserve their bindings and can be used by either competing interpretation.

Mood

> *All of my conscious states come to me in some sort of mood or other. ... there is what one might call a certain flavor to consciousness, a certain tone to one's conscious experiences* (Searle 2004, p. 139).

The relation between a human agent's psychochemical state (a component of $\mathbf{x}(t)$), mood (a component of $\mathbf{z}(t)$), and the rest of the agent's perception, is presumably embedded in the complex and unknown functions F and G. How mood affects behavior is embedded (in part) in the mechanism for selecting the next control law H_i.

Pleasure/Unpleasure

> *Related to, but not identical with, mood is the phenomenon that for any conscious state there is some degree of pleasure or unpleasure. Or rather, one might say, there is some position on a scale that includes the ordinary notions of pleasure and unpleasure* (Searle 2004, p. 141).

The pleasure/unpleasure scale has a natural role as a reward signal during reinforcement learning. The links between particular qualia and their positions on the pleasure/unpleasure scale are very likely determined by evolutionary learning (Dennett 1991). For example, pain is unpleasant and sex is pleasant because of their functional roles in the survival of the individual and the species.

The Sense of Self

> *It is typical of normal conscious experiences that I have a certain sense of who I am, a sense of myself as a self* (Searle 2004, p. 144).

In many ways, the most pragmatically useful aspect of consciousness is the ability to observe, describe, store, recall, reflect on, and in some ways control one's own thoughts, experiences, goals, plans, and beliefs.

As we have seen, the apparently sequential and continuous nature of conscious experience is the post-hoc construction of a plausible coherent narrative to explain a somewhat irregular collection of sensory inputs.

Once such a narrative exists, it can be stored in long-term memory, recalled, and reasoned about like any other piece of symbolic knowledge. The construction, storage, recall, and manipulation of this kind of knowledge poses no fundamental difficulties for computational modeling (Minsky 1968).

It is important to acknowledge that memory can store more than abstracted symbolic descriptions of experience. Memory can include qualia such as snapshots or fragments of the sensory stream with high information content. The *content* of the conscious narrative, as well as the content of these qualia, and the associations they provide into the agent's long-term memory, are highly specific to the agent whose experience they reflect, so they contribute to a unique "sense of self."[2]

[2] "*What do you see when you turn out the light? I can't tell you, but I know it's mine.*" – John Lennon.

Discussion

The claim presented here (a "strong AI" claim) is that the conditions for consciousness are expressible as a computational model, including dynamical trackers maintaining symbolic references to perceptual images in the sensor stream. The phenomenal character of consciousness ("what it is like") comes from the enormous flow of information in the sensory stream, and from the turbulent "churn" of process activation, on the way to being serialized as conscious thought (Baars 1988).

Qualia reflect the information density of the sensor stream. Trackers ground a symbolic knowledge representation in the "firehose of experience" and constrain its interpretations. Intentionality is intrinsic if useful trackers can be learned from uninterpreted experience. And the sequential stream of subjective consciousness is a plausible, coherent narrative, constructed retrospectively (by 500 ms or so) from data provided by parallel unconscious processes.

The empirical and philosophical study of consciousness in humans helps clarify the nature of the phenomenon. The study of the brain helps us understand the one implementation of a conscious system in whose existence we have confidence. But according to our claim, consciousness is not restricted to biological implementation. The essential features of consciousness can, in principle, be implemented on a robot with sufficient computational power and a sufficiently rich sensorimotor system, embodied and embedded in its environment.

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.

Baars, B. J. 2002. The conscious access hypothesis: origins and recent evidence. *Trends in Cognitive Sciences* 6(1):47–52.

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.

Ballard, D.; Hayhoe, M.; and Pelz, J. 1995. Memory representations in natural tasks. *Cognitive Neuroscience* 7:66–80.

Blake, A., and Yuille, A. 1992. *Active Vision*. Cambridge, MA: MIT Press.

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

Crick, F., and Koch, C. 2003. A framework for consciousness. *Nature Neuroscience* 6(2):119–126.

Damasio, A. 1999. *The Feeling of What Happens*. Harcourt, Inc.

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

Gelb, A. 1974. *Applied Optimal Estimation*. Cambridge, MA: MIT Press.

Hutchinson, S.; Hager, G. D.; and Corke, P. I. 1996. A tutorial on visual servo control. *IEEE Trans. on Robotics and Automation* 12(5):651–670.

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

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.

Minsky, M. L. 1968. Matter, mind, and models. In Minsky, M., ed., *Semantic Information Processing*. MIT Press. 425–432.

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. 2004. Bootstrap learning for object discovery. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*.

O'Regan, J. K., and Noë, A. 2001. A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences* 24(5):939–1011.

Philipona, D.; O'Regan, J. K.; and Nadal, J.-P. 2003. Is there something out there? Inferring space from sensorimotor dependencies. *Neural Computation* 15:2029–2049.

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

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.

Sacks, O. 1985. *The Man Who Mistook His Wife for a Hat and Other Clinical Tales*. Simon & Schuster.

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.

Shanahan, M. 2005. Perception as abduction: turning sensor data into meaningful representation. *Cognitive Science* 29:103–134.

Turing, A. M. 1950. Computing machinery and intelligence. *Mind* 59:433–460.

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

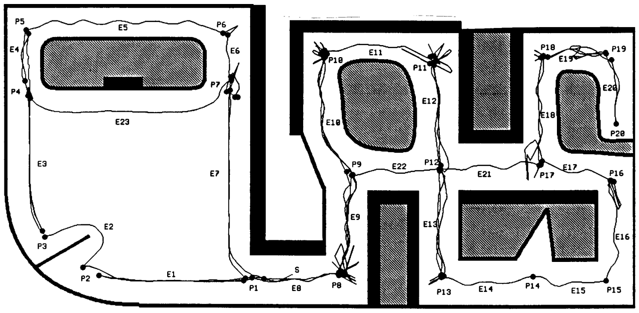


Figure 1: A simulated robot applies the SSH exploration and mapping strategy. It identifies a topological graph of distinctive places and connecting path segments according to the behavior of control laws in the environment.

Wiener, N. 1948. *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge MA: MIT Press.

Learning from Uninterpreted Sensors and Effectors

[3]

In the Spatial Semantic Hierarchy (SSH) (Kuipers 2000), a robot explores and maps an unknown environment by identifying neighborhoods within which hill-climbing control laws can bring the robot reliably to isolated *locally distinctive states*. A trajectory-following control law carries the robot reliably from one distinctive state to the neighborhood of another, where hill-climbing brings it to the next distinctive state and prevents the accumulation of position error.

The robot can thus abstract the continuous environment to a discrete topological map, with symbols representing places and paths, as well as distinctive states and the actions linking them. Figure 1 shows the behavior trace of a robot exploring a simulated environment, hill-climbing to distinctive states equidistant from multiple obstacles, and following trajectories defined by midline- or wall-following control laws.

For the robot to learn these symbols *for itself*, it must learn its own collection of hill-climbing and trajectory-following control laws, starting with an uninterpreted set of sensors and effectors. Pierce and Kuipers (1997) accomplished this task for a simulated robot with unknown sensors and effectors, which required learning the structure of a ring of sonar-like range sensors and learning an abstract model of motor commands that set the velocities of the right and the left wheel.

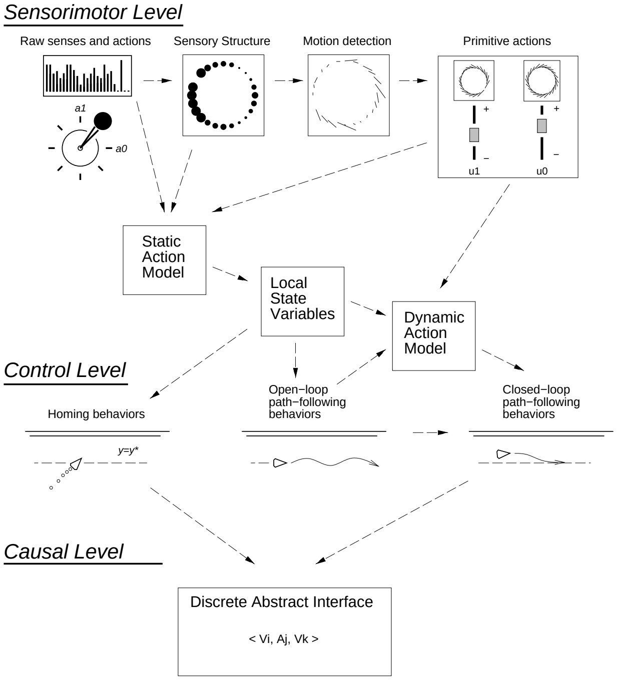


Figure 2: The lattice of learning methods and their results, from Pierce & Kuipers (1997).

Figure 2 shows a lattice of learning methods that analyze data from several different experiments, building a progressively richer description of the sensory and motor systems, and eventually supporting the creation of hill-climbing and trajectory-following control laws. Figure 3 shows exploration traces corresponding to three different levels of competence during learning.

The steps of the learning process are (Pierce & Kuipers 1997):

(1) Gather observations during random sequences of actions. First, coarsely cluster the sensors according to the qualitative properties of a histogram of values returned by each sensor. Then, within appropriate clusters, compute pairwise correlations among sensor values and interpret them as similarity measures.

(2) Assign the sensors in a cluster to positions in a high-dimensional space reflecting their pairwise similarities. Project to a low-dimensional subspace (2D in our examples) that best accounts for most of the variance in the cluster. Once sensor values have a spatial as well as temporal dependence, we can calculate spatial as well as temporal derivatives, and thus define motion fields.

(3) The motion fields corresponding to different motor signals are analyzed using principal component analysis to determine the most significant motion effects and the motor signals that correspond to them. These signals are used as the natural primitives for the motor space.

(4) Higher-level sensory features are proposed, based on the spatial and temporal attributes of the field of primitive sensory values. These include features such as discontinuities, local minimum and local maximum, with magnitude, position, and scope. Proposed features are evaluated according to stability, predictive power, and extensibility.

(5) Evidence is collected of the effects of primitive motor commands on higher-level features, searching for motor commands that change features in predictable ways. “Lo-

[3] The appendix is an addition to the original AAAI-05 version of this paper, describing in more detail some specific robot learning results.

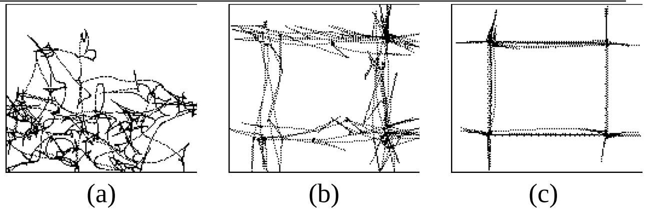


Figure 3: Exploring a simple world at three levels of competence. (a) The robot wanders randomly while learning a model of its sensorimotor apparatus. (b) The robot explores by randomly choosing applicable homing and open-loop path-following behaviors based on the static action model while learning the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

cal state variables" are defined for particular neighborhoods in the environment. Trajectory-following and hill-climbing control laws are defined according to which local state variables correspond to features that are both observable and controllable.

(6) Open-loop control laws are defined by identifying commands that reliably change one feature while keeping another one relatively constant. Closed-loop control laws are defined by searching for and identifying commands that can reduce deviations in the relatively constant feature, actively keeping it close to a desired setpoint. (Think of moving along a wall, turning slightly to maintain a desired distance from it. Compare figures 3(b,c).)

Higher-level sensory and motor features are learned without drawing on prior knowledge of the robot's environment. They are learned by identifying statistical and dynamical regularities in the experiences the robot receives after sending motor commands. In these experiments, the concepts whose intentionality is learned by the robot itself are a set of specific distinctive states (position and orientation), and the places and paths that make up a topological map of the environment.

Extending this learning to include objects, actions, affordances, other agents, and the other aspects of commonsense knowledge occupies several years of learning for a human child, and very likely several more decades of research for artificial intelligene and the cognitive sciences.