The Journal of Mind and Behavior
Winter and Spring 2005, Volume 26, Numbers 1 and 2
Pages 1–22
ISSN 0271–0137

Brain-Inspired Conscious Computing Architecture

Włodzisław Duch

Nanyang University of Technology
and
Nicolaus Copernicus University

What type of artificial systems will claim to be conscious and will claim to experience qualia? The ability to comment upon physical states of a brain-like dynamical system coupled with its environment seems to be sufficient to make claims. The flow of internal states in such systems, guided and limited by associative memory, is similar to the stream of consciousness. A specific architecture of an artificial system, termed *articon*, is introduced that by its very design has to claim being conscious. Non-verbal discrimination of the working memory states of the articon gives it the ability to experience different qualities of internal states. Analysis of the flow of inner states of such a system during typical behavioral process shows that qualia are inseparable from perception and action. The role of consciousness in learning of skills — when conscious information processing is replaced by subconscious — is elucidated. Arguments confirming that phenomenal experience is a result of cognitive processes are presented. Possible philosophical objections based on the Chinese room and other arguments are discussed, but they are insufficient to refute articon's claims that it is conscious. Conditions for genuine understanding that go beyond the Turing test are presented. Articons may fulfill such conditions and in principle the structure of their experiences may be arbitrarily close to human.

Keywords: consciousness, qualia, brain-like computing

In his famous article "Computing Machinery and Intelligence" Alan Turing (1950) considered the question "Can machines think" to be too ambiguous to answer. Bearing in mind the problems with defining consciousness and the various uses of the word, the question "Can artificial systems be conscious?" can also be too ambiguous to answer. In this paper another question is consid-

This article was written during my sabbatical at the School of Computer Engineering, Nanyang University of Technology, Singapore 639798. Requests for reprints should be sent to Włodzisław Duch, Department of Informatics, Nicolaus Copernicus University, ul. Grudziądzka 5, 87–100 Toruń, Poland. Email: wduch@phys.uni.torun.pl

ered instead: What types of artificial systems may claim to be conscious, and are there any strong arguments against such claims?

A computer that is programmed to repeat "I am conscious" does not make a justified claim. To understand when a claim like that is justified one needs to understand the processes leading to conscious experience in our brains. Trying to elucidate this process I will distinguish three levels of difficulty. First, any useful theory of consciousness should explain the difference between those processes that we are conscious of, and have phenomenal experience of, and similar processes that have no such component. Such contrastive heterophenomenological approaches have been proposed by James (1904) and developed by Baars (1988). Habituation, a process of vanishing phenomenal experience in spite of persisting physical stimuli, is a good example of minimally contrastive situations. What has changed, why has conscious experience vanished? Most theories of consciousness (Chalmers, 1996) fail already at this basic level.

The second, more difficult level is to understand the structure of the qualia. Each sensory modality has a specific structure of its perceptual space. Auditory, visual, tactile and olfactory perceptions elicit conscious experiences, but there is a qualitative difference between the types of qualia that accompany these experiences. If all information processing was simply accompanied by phenomenal experience, based on quantum effects (Chalmers, 1996), if consciousness was all-pervading (De Quincey, 2002), or involved a spiritual soul (Eccles, 1994), then the differences between the structure of different qualia would still remain undiscovered. Moreover, the qualia that we are experiencing change in time. Learning new skills (driving, horse-riding, playing an instrument) initially requires conscious control, but after some time control becomes automatic or subconscious. How can a conscious process become subconscious? I will argue here that only theories based on cognitive brain mechanisms may explain all facts related to qualia.

The third, even more subtle and difficult level, is to explain why there is any feeling at all, why are we not zombies. At first this hard problem of consciousness (Chalmers, 1995) may seem to be hopelessly difficult. Intensive debate on this topic in the last decades has not brought much agreement (Chalmers, 1997). I will argue that artificial systems based on brain-like computing principles *must* claim to experience qualia states. Systems of this type will be called *articons* (from *arti*-ficial *con*-sciousness), and they will be something different than just artificial intellects (or *artilects*).

Brain-like Computing Leads to Systems that Claim to be Conscious

An industrial robot, an animal with the brain stem intact and most of the brain removed, or a human in a coma react to specific stimuli that elicit a spectrum of automatic responses. For the nineteenth-century neurophysiolo-

gists, such as Thomas Laycock who formulated the concept of "unconscious cerebration," the brain was the seat of consciousness, while the brain stem and the spinal cord were responsible for unconscious responses. There was a strong resistance to the idea that the brain may also react in an automatic way.[1] William James (1904) claimed that consciousness is not an independent entity, but is a function of particular brain-based experiences. Consciousness cannot be defined independently of the object we are conscious of; both form the same functional complex.

A close connection between mind functions and brain complexity is evident. Inner life, leading to sophisticated behavior, requires sophisticated brains. On the other hand a series of reflexes and automatic responses triggered by specific stimuli may produce interesting behavior without any inner life. Computers are great deceivers, already capable of creating graphics that are hard to distinguish from reality. External observation may not be sufficient to differentiate between simulacrum and genuine mind. One way out of this dilemma is to propose a specific architecture for brain-like computing and justify why it should be sufficient to produce mind-like behavior with inner life behind it.

Active brains are the only systems that are undoubtedly associated with minds. Understanding brains (and anything else) requires simple models, but oversimplification or wrong metaphors lead to insurmountable problems. Popularity of the grossly simplified models of the brain, such as the left–right hemisphere division of functions, or the triune brain models (Humpden–Turner, 1981), show this need for simplicity. Turing machine information-processing metaphors are not well suited to represent dynamical processes in the brain. Multi-dimensional neurodynamical systems provide much better metaphors and models, but are more difficult to grasp (connection between neurodynamics and psychological spaces is outlined in Duch, 1997).

Perception has evolved to facilitate action. Sensory information is processed in several stages by relatively independent brain subsystems (modules) in a very complex manner. Partial results of this processing are not perceived consciously, and there is no place in the brain where results of all this processing come together, forming a final percept. How exactly is the final percept formed, if this happens at all, is still not known. Binding of neural activities is one possibility, but the issue is controversial. Perhaps the most important role of perception is to enable the knowledge and exercise of "sensorimotor contingencies," without the need for internal representations, as it was persuasively argued by O'Regan and Noë (2001).

---

[1]The fascinating history of early development of these ideas was presented by J. Miller in "Going Unconscious" (Miller, 1995).

Memory plays a crucial role here, enabling perceptual learning at the basic level, and associative learning at the higher level (Goldstone, 1998). For example, in the olfactory system "Cortical synthetic coding reflects an experience-dependent process that allows synthesis of novel co-occurring features, similar to processes used for visual object coding" (Wilson and Stevenson, 2003, p. 307). This mechanism is used to solve the main olfactory system task; that is discrimination of one odorant from another (or one visual object from another). For our purpose a simplified model is sufficient. Imagine a number of specialized "feature detecting odorant receptive fields" or modules in the olfactory cortex, tuned to specific odorants. A specific signal received from the olfactory bulb will activate one or more of these modules in a resonant manner, and their contribution will be added to the global dynamical state of the brain. More than three modules resonating at the same time send interfering activation patterns, making precise discrimination of odors difficult, as demonstrated by Wilson and Stevenson (2003).

Recognition of objects is memory-based, thus every cognitive system must have associative memory capable of storing new facts and providing content-addressable retrieval. This is achieved by tuning the resonance properties of modules to excitation patterns provided by modules specializing in feature recognition. For example, at the lower level of speech processing elementary phonemes are discovered, and modules coding phonemes that become active resonate, sending their patterns of excitations to a set of modules specializing in word recognition. All these resonant processes contribute to the global dynamical state of the brain. These general facts are rather uncontroversial. Gaffan (1996, p. 69) sees associative learning "as an expansion of the cortical representation of a complex event" and thinks that "the distinction between perceptual and memory systems will need to be abandoned as deeper understanding of cortical plasticity is achieved." Experiments show that long-term memory is referenced in sound processing even in REM sleep (Atienza and Cantero, 2001).

A number of attempts to create models of brain activity based on coupled oscillators has been made (see for example Frank, Daffertshofer, Peper, Beek, and Haken, 2000). Adaptive Resonance Theory (ART), developed by Grossberg (2000) over two decades, proposes that learning may fine-tune the neural modules that respond (resonate) strongly with incoming stimuli. This helps to solve the stability–plasticity dilemma, allowing for rapid learning and preserving the stability of the knowledge already acquired. Adaptive resonant states are formed in the brain from the up-going activity stream (sensory to conceptual areas) and the down-going streams (conceptual to sensory areas), forming reverberating and self-organizing patterns (for vision models based on such processes see Ullman, 1996).

Articons, simplified brain-inspired computing models, will be based on resonant processes and may learn using ART and similar techniques. The articon system is composed of a large number of modules that may resonate in a specific way, some of them being sensory modules, some motor modules, and some associative memory modules. The resulting activity of a module contributes to the global dynamical state that in turn activates all other modules, producing a flow of dynamical states. In Figure 1 a sketch of a minimal articon system is shown, with additional modules that allow the system to comment on its own global dynamical states. How does it help to solve the problem of understanding consciousness?

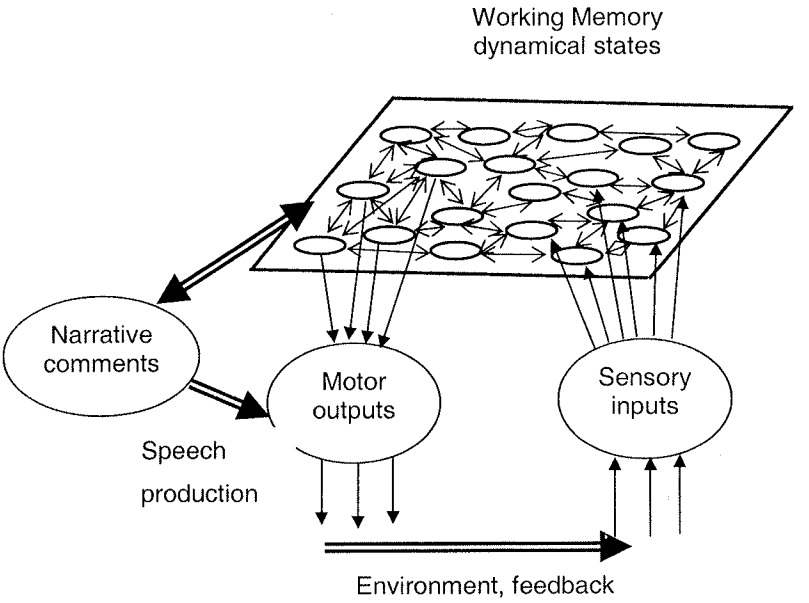

Figure 1: Information flow in the brain-like computing systems used in articon construction.

The long term memory of the human brain is vast, stored by 100 trillion synapses of about 10 billion cortical neurons. The number of modules (cortical columns) is much smaller, around one million, with each using 10–100 thousands of neurons. Since many cortical columns are involved in storing memory traces in a combinatorial fashion, the number of potential memory states is almost infinite. Internal states of cortical columns are the basic building blocks of the inner space of the articon. The ground state, or lowest excitatory state of cortical modules of the living brain, corresponds to a minimal activity as seen in deep sleep. To stay alive neurons need to send about

two spikes per second. The brain dynamics must have a global stable attractor, otherwise all activity will stop and neurons will die. Excitations of this global ground state, due to both the external stimulations and intrinsic internal dynamics, create intricate configurations of neural activity patterns, an internal state reflecting both environmental states and memory of previous states. We can not experience reality directly; only the internal states of the brain are experienced and may be commented upon.

Articons learn to recognize external stimulations creating associative memory patterns in the form of excitations of a few modules that code particular combination of features extracted from these stimulations. External objects activate sensors that extract specific features and send signals to the long term memory, activating memory traces of similar experiences remembered from the past. This activity results in the working memory state that may be identified as percept, in the form of persistent cooperative activity of several columns (modules). In particular auditory signals may be recognized as words (which are initially also percepts, patterns of neural activations). The associative character of memory in such a system will lead to a chain of activation patterns, each corresponding to some inner object, for example a train of words or percepts. Because of the associative, attractor nature of brain dynamics that largely restricts the admissible inner states of the articon to combinations of those states that have been learned, a negligibly small portion of all potential states may be actualized.

From the outer (third person) perspective the inner states of the articon are a stream of dynamical patterns, activation of modules flowing from one quasi-stable pattern to another. Some of these states are images, some are words, and some are sensorimotor actions, forming "mind objects" that interact, defining events in the inner space (Duch, 1997). Combined activity of a few modules is sufficient to recall patterns of excitations stored in the long term memory, reinstating similar patterns of global activity in the articon as those that were present in the moment of actual observation, when the system had learned them. Working memory contains those activity patterns that won the competition at a given moment, a subset of the global dynamical state of the system. Usually a few long term memory modules are strongly active (resonating) at a given moment, contributing to the working memory. The global dynamical state includes the state of all subsystems (i.e., the whole organism, not only the brain).

Note that resonant states in the working memory are spatio-temporal, defined in four-dimensional space-time. A fragment of music recalled unfolds as a series of states in working memory, each state "dressed" in associations, memories, with motor or action components, in one dynamical flow. Internal state of the articon is constantly evolving, therefore every time the same fragment of music is recalled these associations may be slightly different.

Compare the flow of spatio-temporal pattern excitations in articon with information processing in the Turing machine. Instead of a dynamic flow based on associations, in the Turing machine the internal state is changed by a program according to some instructions. Registers in Turing machines are "dead," abstract binary entities without spatio-temporal structure. Articons are not only data flow machines, but have specific architecture that due to the non-linear character of interactions among modules creates working memory states that cannot be decomposed into separate, independent components.

The "narrative interpreter module" has access to the articon working memory states and the ability to recognize and symbolically label them. In the brain this module (or a number of modules, composing the language areas) is made from cortical columns, but in the model presented here it is irrelevant. Working memory states are composed from physical patterns, specific resonance activities of subsets of modules, yet they are intentional, corresponding to "something out there," because they have been learned from sensory data. Relations between working memory states have intrinsic semantics, reflecting relations between the states of the articon's environment and leading to specific actions (external motor actions or new internal states). The interpreter tells the narrative story, symbolically labeling the working memory states, objects and their relations, ready to report the inner states of the articon to the external world. Of course such reports capture the internal states only in a rough way, because it is impossible to capture the richness of the flow of continuous non-decomposable internal states in discrete symbols, but such is the nature of language.

The Rat Example

Imagine a rat that has found a new food. First the rat will smell and taste a bit of the food. Then in a fraction of a second the rat has to decide whether to swallow it or to spit it out. A query is sent to the long term memory from the gustatory cortex (not unlike "request for comments" in Internet newsgroups) and the most relevant previous experience of the rat should be recalled. Long term memory is distributed; the whole brain has to be searched for associations with the content of the particular pattern created by the input information. Some cortical columns of the gustatory cortex resonate contributing to the working memory states that are available to all modules in the rat's brain. Working memory is small, it can hold just a few patterns (about 7 ± 2 in humans). The small size of the working memory in the rat allows it to focus on a single task. Resonant states are formed activating relevant memory traces and the answer appears: bad associations! probably poison! spit! Perception serves action: to remember the episode in the future (type of food, its smell, the place) brain centers that control emotional reactions of

the rat release specific neurotransmitters, increasing plasticity of the neocortex. In addition, a strong physiological reaction starts, cleaning the organism of poison. All this leads to a "system reaction" of the whole organism, creating various sensory inputs (including internal, proprioceptive and hypothalamic inputs from blood chemistry sensors), that at the working memory level form rather unique patterns.

If the rat could comment on this episode, what would it say? And if the rat's brain was replaced by an articon controlling the rat's body, and the articon was trained using rat's experiences, with articon's ability to comment upon its working memory state, what would the "comment" be? Would the rat (or articon) taste again the same food? Seeing it, the repulsive episode would be partially re-instated in its brain. Obviously the rat has different "feelings" for different tastes. These feelings are real, physical states of the brain, due to specific brain activation patterns that — through associations — lead to different actions and states of the organism. Articons, rats and humans have something in common that distinguishes them from computers: qualia, or the difference between real experienced inner states and mere information processing. The stream of their inner states and associations they bring may be commented upon in many ways. "This red is awful, it reminds me of blood . . . and that is a nice red bag, it fits so well to my skirt . . . and this particular red I have never seen and it has a special feel to it." Without reference to memory it will be impossible to distinguish between feelings (internal states in general) associated with seeing particular colors and those coming from different senses. Qualia or phenomenal experiences, independent of cognitive mechanisms, are philosophical fiction.

Comments that articons will make are not just a result of information processing, something that a computer mindlessly repeats instructed by the program, but are comments on physically existing states. Articons that build on the brain-like computing principles will have to claim qualia, inner life and consciousness. Creating articon systems of increasing complexity in the limit all subtle responses and features known from human psychology may be produced. Interacting with such complex articons and observing their inner life we should be able to distinguish "the real thing" from a computer simulation programmed to respond in the same way.

Learning New Skills

Learning new skills over a period of time involves a shift from initially conscious activity, engaging large brain areas, to the final subconscious, intuitive, automatic actions that engage only a few well-localized, specialized brain centers. Skill learning expressed in terms of diminishing conscious control seems to be a rather mysterious process. How does it proceed in the arti-

con that controls a robot body or some other device? The task to be learned must recruit a number of modules that can control the effectors, correlate their activity with activation of the sensory modules, and compare the result with the desired (or imagined) one. This may require retraining of existing sensorimotor maps, or formation of new modules for prototype actions, tuned during further learning. As with all complex tasks the results have to be present in the working memory.

A feedback loop — commonly called "the conscious monitoring" loop — is necessary for learning of a skill. Learning using reinforcement techniques (Sutton and Burto, 1998) requires observing and evaluating how successful are the actions that the articon has planned and is executing. Sensory observations carry qualia, failures are interpreted as painful, and should be remembered as an episode. This allows the resonant processes to use the experience from failures in the next trial. The moments of failure are especially rich in qualia, and the system is aware of such moments, commenting upon them and remembering them. Real brains learn changing the strength of synaptic connections (cortical plasticity). Articon modules learn regulating the amount of new versus old knowledge, changing the level of activity of specific modules recruited for the task. Relating current performance to memorized episodes of recent performance requires bringing all relevant observations together, as well as evaluation and comparison of the information collected in the working memory. Evaluation has both cognitive and affective components, emotional reactions that provide reinforcement for learning (in the brain frequently via dopamine release), facilitating rapid learning of specialized neural modules.

There is no transfer from conscious to subconscious processing; this division is simply a bad conceptualization. Expectations that match the observations slowly begin to lose the competition and do not make it to the working memory, becoming "subconscious." Only a few specialized modules that win the adaptive resonance competitions are left to control the task, and further fine-tuning (learning) proceeds at a much slower pace.

Qualia in Articons

Are the qualia in articons real? As real as it gets, to use a popular expression; qualia are quasi-stable patterns of physical excitations of brain neural tissue, created by specific cognitive processes, capturing the brain/body reactions to stimuli or to the internal states preceding them. This is in fact an old idea and among all theories of qualia only those based on such understanding of qualia have made steady progress in explaining the details of human experience. Emotions are connected with strong qualia. In 1911 the surgeon George Crile, speaking about emotions to the American Philosophical Society made

the following claims: ". . . it is possible to elicit the emotion of fear only in those animals that utilize a motor mechanism in defense against danger or in escape from it." In fear "the functions of the brain are wholly suspended except those which relate to the self-protective response against the feared object"; and "we fear not in our hearts alone, not in our brains alone, not in our viscera alone — fear influences every organ and tissue." Crile (1911/1998, chapter 3) gives an interesting analogy here: "An animal under the stimulus of fear may be likened to an automobile with the clutch thrown out but whose engine is racing at full speed." Of course such a state of the organism is reflected in a specific brain activation pattern, and has a unique "feel" to it that is expressed in interpretive commentaries of the system.

Crile (1911/1998, chapter 4) also stated "We postulate that pain is one of the phenomena which result from a stimulation to motor action." Our internal organs cannot send us symbols informing, for example, "your left toe has just been cut by a sharp stone." The only way to pass information requiring attention is to bring it to the working memory level, where it appears as a specific excitation pattern disrupting other activities and demanding immediate action. Since the required action usually goes beyond simple reflex what is called for is the highest-level control with access to all brain resources. What would pain look like without cognitive interpretation? We would not know where it is, what it is, and how to react to it, so the qualia associated with it would be very different. This is what happens in pain asymbolia (Ramachandran, 1999), resulting from lesions of the secondary somatosensory cortex that specialize in giving meaning to tactile sensory signals. Other examples of wrong cognitive interpretations of the brain data are found in unilateral neglect and phantom limbs (Ramachandran, 1999).

Phenomenology of pain shows clearly how cognitive mechanisms determine the qualia. Placebo can be as effective as powerful anesthetics. Most people experience pain as something unpleasant but masochists find pleasure in it. Full intensive concentration on actual experience of pain may change qualia completely. Causalgia is a post-injury blazing pain condition that may be initiated by touch, noise or anything else. Analysis of this and other pain conditions shows that without cognitive interpretation there are no pain qualia (as noticed by Beecher, 1946, investigating wounded soldiers). Without proprioceptive information people are not even conscious of their own motor actions (Fourneret, Paillard, Lamarre, Cole, and Jeannerod, 2002).

Perceptual learning (Gaffan, 1966) enhances the ability to experience qualia through training. Better memory for sensory stimulation allows for more subtle discrimination, changing our phenomenal experience and bringing new qualia. This is true for vision, hearing, taste and all other senses. Try to put a puzzle of a few thousand pieces together and you will notice how this enriches perception of shapes and colors — that is, it enriches qualia. New

qualia are also experienced in dreams. Interpretation of brain states is clearly based on memory traces.

If for some reason cognitive mechanisms used for interpretation of the brain states stop working, experience will vanish. In particular, habituation, intensive concentration on some stimuli, or shifting of attention may remove qualia associated with the experience. A good example is the segmentation of visual stimuli from a background — qualia may arise only if a correct interpretation of the stimuli is made, otherwise one may look without noticing quite large objects or significant changes in one's environment (cf. O'Regan and Noë, 2001). Because memory references are involved in cognitive interpretation qualia are influenced by drugs acting on memory. In case of covert perception, for example in blindsight (Weiskrantz, 1997), cortical structures that provide appropriate representations for discrimination are damaged and thus visual qualia also vanish. Such damage leads to a serious impairment of behavioral competence. Information available in the brain is still sufficient to make partially correct decisions but this information enters global brain neurodynamics in a quite different way than information from the visual cortex.

The feeling of laughter may also be understood as interpretation of a system's reaction. Electrical stimulation in the anterior part of the human supplementary motor area (SMA) can elicit the physical reaction of laughter, and this is followed by cognitive interpretation leading to a feeling of laughter, and even confabulations to justify it (Fried, Wilson, MacDonald, and Behnke, 1998).

Qualia have different structural properties, matching their specific roles. For example, they have spatial structure in case of visual, tactile, temperature or pain stimuli, and non-spatial structure in case of taste, olfaction, thoughts or imagery. Words and thoughts " . . . are symbolic of motor acts" (Crile, 1911/1998, chapter 4), therefore they also may have certain quality to them, although it is felt stronger by those who experience synesthesia, where not only motor, but also sensory areas are stimulated by symbolic thought patterns.

It seems very doubtful that any other understanding of qualia may explain all that cognitive science and neuroscience have uncovered about the conditions (why and when) and the way qualia are experienced and structured. Articons have no choice but to claim that they experience qualia. Are there any arguments to refute their claims?

Chinese Room

The Turing test based on a dialog with a computer program, or in general any test based on external observations of seemingly intelligent behavior, may fool us to believe that "there is someone in there." An expert system, that is, an AI program that receives an input, analyzes it, finds matching rules and produces outputs, obviously has no inner life, no thoughts buzzing in its mind,

it is just a series of program steps executed on computer hardware. Results obtained with the expert system technology may impress some people, but no matter how smart it may look from outside, there is nothing inside.

The Chinese room argument of John Searle (1980, 1984) is the most famous critique of the Turing test. Searle assumes that certain brain processes are sufficient for intentionality, but that "instantiating a computer program is never by itself a sufficient condition of intentionality" (Searle, 1980, p. 417). To prove that this is the case he shows that "a human agent could instantiate the program and still not have the relevant intentionality" (p. 417). In his Gedankenexperiment he imagines himself being locked in the room and given a large batch of Chinese writing. He does not know Chinese, therefore these writings are to him only meaningless squiggles. In addition to the Chinese writings he is given a set of rules in English, enabling him to identify the Chinese characters received through the input window by their shapes and to correlate them with other Chinese characters using the batch of writing at his disposal. The results are then passed to the Chinese-speaking people outside the room, who consider them to be the answers to their questions.

Searle being inside the room obviously does not understand the meaning of questions given to him in Chinese, nor does he understand the meaning of the answers that he gives back using the rule book written in English to correlate characters. He acts as a computer program, blindly following instructions. The people outside may be perfectly convinced that they are dealing with someone who understands them although there is no understanding inside the room. Programs for natural language analysis instantiated on computers simply follow instructions without any intentionality or understanding. A Turing test based on external observations of behavior is not sufficient to grant understanding to an artificial system.

Can this argument be used to refute claims that an articon has real understanding and is conscious of its states? There is a vast literature criticizing or supporting Searle's claims, but somehow the most important issues have been missed. First, the Chinese room argument is not a test — the outcome is always negative! Second, a feeling "I understand" is confused here with real operational understanding. Third, the conditions under which a human observer could recognize that an artificial system (or an alien brain) understands have not been discussed.

Consider the first issue: Under which condition the Chinese room experiment could grant any system real understanding? For example, could a person placed in a human brain to watch neural processes there find any signs of understanding? Already Leibnitz in his *Monadology* (1714/1982, par. 17) stated: "Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior, as he would into

a mill. Now, on going into it he would find only pieces working upon one another, but never would he find anything to explain perception." Searle concludes (1980, 1984) that "brain processes *cause* conscious processes," but what is it about the brain that gives us genuine understanding? His solution is that neurons must have some mysterious "intentional powers" that computer elements do not have. From this point of view even a detailed simulation of all neurons will not be sufficient. Imagine that in a distant future a robot with a brain identical to the human brain will be created; according to Searle no matter how human-like the robot will be there will be no understanding in the robot's brain due to the lack of "intentional powers" of its electronic elements. It is quite likely that brain prosthesis will be mounted in human brains in the not so distant future: Will these people lose their intentional powers?

These conclusions seem absurd. The articon example shows that it is the organization of the system, rather than the elementary unit, which is important. Biological properties and functions of complex organisms emerge from interactions of their cells, not from some special properties of elementary units. The Turing test is an important step, a necessary, although insufficient, condition to grant a system genuine understanding. The Chinese room experiment fails to tell us anything about the inner world of the system under observation. This experiment will never find understanding in any system, artificial or biological.

The second problem with the Chinese room experiment is the lack of discussion concerning what does it mean "to understand"? This issue is more complex than it seems. Searle contends that he understands English but does not understand a word of Chinese. We usually intuitively know when we understand, although sometimes we may be wrong about it. What exactly is this feeling of understanding? Gopnik (1998) argues that explanation should be thought of as the phenomenological mark of the operation of the theory-formation system in the brain, and that finding an explanation leads to a reward, felt as satisfaction. Language is the most complex function of the brain and it takes time to understand a sentence, especially if it is long and has compound structure. The brain needs time to parse sentences and has to signal when it is ready to proceed further. This signal is recognized as the feeling "I understand" meaning, among other things, that I am ready to receive more information. I understand if I am able to relate new information to the contextual knowledge that I already have, knowledge that is finally grounded in my perceptions and actions.

On the other hand, understanding implies the ability to answer questions requiring simple inferences. The brain is not always correct in generating the "got it," or understanding signal, as everyone knows from introspection. Some drugs or mental practices induce the illusion of understanding everything, so that the feeling is there but the ability to give correct answers is not (Terrill,

1962). Sometimes the feeling is hardly there, understanding gets more and more fuzzy, and additional questions are asked to clarify the meaning. There may be even no feeling of understanding, but correct answers may be given, indicating that the person in fact understands. Identification of understanding with some inner feelings may be as misleading as granting it to the AI computer program. The Turing test checks for understanding by asking questions, not by looking for the signs of feeling of understanding. A person inside the Chinese room may finally start to understand some questions and answers as we might understand a foreigner gesturing with his hands. The feeling of understanding is an additional brain signal that is not necessary for genuine understanding.

What are then the sufficient conditions to recognize understanding in other minds? Learning of behavior in monkeys and humans is largely based on imitation. This is possible because in our prefrontal cortex we have "mirror neurons" that respond to specific actions performed both by oneself and to observed actions performed by others (Carey, 1996). This is the bridge between two minds, allowing for intuitive and direct communication based on observation and common brain structures. We can understand only the systems that have minds of similar structure to ours, by "resonating" with such minds, trying to assume similar dynamical states. A way to create such resonance between minds is through language-based communication and observation of other people's behavior. The Chinese room experiment does not try to discover if there is an understanding mind inside the room and thus it does not teach us anything about the mind of artificial systems, refusing to attribute "genuine understanding" to machines as well as to humans. How can we know that Searle's neurons still have their "intentional powers"? How can we tell whether an alien from Andromeda is a robot or is a real, intentional being? If we could get into resonance with the alien's brain perhaps we could recognize if there are genuine images, thoughts and emotions arising there, or just blank "wait states," and "run the instruction" in response to questions.

Cognitive Resonance

Although the Chinese room argument is flawed it may be twisted a bit to show that an articon really understands. Learning in natural environments articons will behave in an individual, unique way, with exponentially large numbers of potentially accessible internal states. There is no set of rules that can reproduce the dynamics of such a system. Suppose that some articon has learned to answer questions in Chinese and that you can observe its working memory, receiving all information that appears there in the form of iconic images (perhaps less abstract than Chinese characters). Your instruction book contains explanations of all incoming patterns, referring it back to the observa-

tions that contributed to creation of such memory states. You see the whirling of thoughts and images in the working memory, the flow of associative processes leading to answers. This is what you would see also in other people's brains if we knew how to convert EEG or MEG patterns into a combination of internal representations that gave rise to these specific activations, and to refer those representations to external observations.

Watching the inner states of articon's working memory your brain may start to "resonate" with the flow of some of the observed patterns and you will develop the feeling that you understand what the conversation is about. You may have a glimpse of the first-person view of the articon's internal world projected into your world. It is impossible to enter fully someone else's internal world, to have identical experience of "what it is like to be someone or something else." We always see the world through the filter of our own brain, states that it supports, memories and associations that the incoming information elicits. Although the difference between real brains and articon systems may be large, some mutual understanding should be possible. If the articon passes the Turing test, and if it also passes the "resonance test," there will be no reason to reject its claims of genuine understanding, experiencing qualia and being conscious of the flow of its processes.

The Adaptive Resonance Theory (Grossberg, 2000) describes some brain states and mental phenomena in terms of resonant states between the interacting brain modules. The ability to understand emotions of other people leads to emotional resonance (Ekman, Campos, Davidson, and De Waals, 2003), that is, sharing or directly understanding of other people's emotional states. Two or more separate brains may also be in resonance although there is no direct physical coupling between their neurons; interactions via speech and visual channels may be sufficiently strong to create such states. Although the concept of cognitive resonance has not yet been widely adopted and investigated the brain mechanism behind it should be similar as in the adaptive resonance states within the brain. On the psychological level cognitive resonance may manifest itself as a feeling of perfect understanding of each other's minds, a rare situation that happens when close friends or collaborators understand each other perfectly during a discussion.

To be a bit more technical, the necessary condition for such resonance is defined as follows: dynamics of both brains B_1, B_2 should admit attractors A_1 and A_2, with similar relational structure $A_1 \sim A_{1i}$ in respect to other attractors in B_1, as $A_2 \sim A_{2i}$ has in the brain B_2. An approximate correspondence between these states in both brains should be established. The two brain states A_1, A_2 do not have to be similar, but the structure of the network A_{1i} A_{2i} and transitions between states within each network should be roughly similar. If, as a result of their interaction, both brains are in dynamical states $A_1 \sim A_2$ a cognitive resonance of minds may occur and experiential understanding between them is

established. We are able to share such mind states to a high degree with our family members, with other members of the same culture, and to a somewhat lesser degree with members of different cultures and to a lesser degree with animals, because not only are their minds formed by very different environments, but also their brains and their senses are physically different. Computers are incapable of any experiential understanding of humans, but articons may be able to achieve some level of experiential understanding.

Can we understand what is it like to be someone else? There are at least two kinds of understanding: intellectual and experiential. Intellectual understanding, involving mostly frontal and temporal lobes, is based on models of the world and communication with others on that basis. This is what can be captured in artilects, or even expert systems to a large (although not perfect) degree. Experiential understanding, engaging mostly the limbic system and sensory cortices, is based on sharing the feelings of our family, friends and other people, and it requires certain emotional and cognitive resonance among brain states responsible for the contents of these minds.

There is something it is like to be a bat and something it is like to be a man, since "to be" means to be a flow of mind states produced by the brain of a bat or of a man, implying a subjective view. Intellectual understanding requires an objective, external description and one is not reducible to the other. To know what it is like to be a bat for a bat requires a bat's brain; a human brain is not sufficient. Nevertheless, a fairly detailed description of a bat's internal states may be formed, and some intellectual understanding achieved through modeling of a bat's behavior. When we find a particular state of the brain we may infer that a particular experience, whatever that might be for a bat, is correlated with it. Since humans share several needs with bats, such as the need for food and sleep, drawing on our own experiences we may assign reasonable interpretations to some behaviors of bats. Nagel (1974) himself admits that perhaps all robots complex enough to behave like a person would have phenomenal experience. His main objection is not to physicalism itself, but rather to the lack of the beginnings of a conception of how the above statement might be true. This is precisely what I have tried to show here, although it is not just the complexity, but the specific organization of the robot's brain that is important.

Another well-known thought experiment concerns Mary, the colorblind neuroscientist, who gains color vision and learns first-hand about red color (Jackson, 1982). There are inner facts that are beyond the physical facts, but the conclusion that physicalism is false because knowing everything about neuroscience does not imply knowledge about qualia, is premature. Dennett's (1996) solution is to deny the problem by claiming that to know everything means to be able to correlate the qualia with brain states. In his version of the experiment Mary is presented with a bright blue banana and immediately rec-

ognizes that fact (perhaps with access to the maps of activity of the V4 visual area it could be done even today). Dennett concludes that the story does not prove that Mary has learned anything new. She has not learned anything new only in the sense of verbal, intellectual learning, but certainly her brain, stimulated for the first time by color light, assumes a new dynamical state, so this state must be interpreted internally as a new phenomenal experience. Her previous knowledge was abstract, symbolic, engaging temporal and frontal lobes only, not occipital cortex. There is no great mystery in the fact that new brain states are experienced as mind events having new qualities. People that are born blind and gain their sight after adulthood certainly learn quite a lot, and it helps them little if they have great intellectual knowledge of geometry. Inner life is real, although it is in a way "a shadow" of neurodynamics (Duch, 1997). Articons support a similar flow of inner states as real brains and seem to be immune to philosophical critique that applies to computers.

Conclusion

Instead of trying to define consciousness I have tried to show that articon systems based on brain-like computing principles will not only have inner life, but will also claim to have qualia and will claim to be conscious of them. I have proposed minimal architecture of a system called articon that will have to make such claims. Claims of qualia are based on interpretation of real, physical states supporting working memory of such systems. These continuous dynamical states differ in a fundamental way from states of a Turing machine by including potential associations through peripheral "dressing" components that lead to subtle variation of the interpretation of the meaning of the state. The articon will recognize these differences as different feelings, or qualia associated with the perceived object.

Such understanding of qualia is in agreement with a large body of data from cognitive science. The inner states in articon systems may have properties that come arbitrarily close to the properties of phenomenological states. The flow of the inner states is controlled by associative properties of memory, and only in unusual circumstances (corresponding to mental illness or intoxication) will inner experiences significantly deviate from those in normal, awake states. Associative memory models capable of hallucinations resulting from formation of spurious memory states are useful in computational psychiatry (Reggia, Ruppin, and Berndt, 1996).

Taylor (1998) has described in some detail the possible neural underpinning of phenomenal experience characterized in terms of transparency, presence, unity, intentionality and perspective. Qualia in articons can have the same properties as human qualia, provided that organizational principles of information processing in the artificial system are sufficiently similar to that

of human brains. Claims of qualia are a necessary consequence of brain-like organization of computations, in particular the ability to comment upon physical states of the architecture carrying out these computations. These qualia may have a wide range of structural properties, depending on the complexity of the artificial system, its sensors, modalities, the groundings of its concepts via perceptual learning, and the ability to discriminate and comment on different states of its working memory.

Because artificial systems will never be identical with biological systems, providing only a rough functional approximation to brain-like organization, their qualia will be different than ours. The same is true for people with abnormal or damaged brains, or for animals. The word "pain" describes rather different reactions of organisms across different species. Pain will obviously be rather different for artificial systems capable of sustaining an internal state with pain-like characteristics that result, for example, from temperature sensor overheating. Burned sensors may send signals disturbing normal flow of inner states, demanding attention at the highest "conscious" control level. An articon will report the disruption as pain and complain about it until the damage is repaired.

There is a growing consensus that the real grounding of the meaning of words is in sensorimotor activities (Harnad, 2003). Perhaps similar consensus will slowly grow for the idea that qualia are physical states of the brain. Weiskrantz (1997) analyzed blindsight and amnesia patients and came to the conclusion that the ability to render a parallel acknowledged commentary is indispensable for consciousness. Similar conclusions are drawn from the work with deafferented patients (Fourneret, Paillard, Lamarre, Cole, and Jeannerod, 2002).

Are articons kinds of computers? Yes, in the same sense that physical processes are kinds of computations, for example, if one agrees that gravitational forces in the solar system solve the N-body problem. Rules and computations are not good replacements for real physical states of brain/body. Classical logic and discrete symbols are not a good way to approximate continuous brain states. There is no way to represent accurately the states of a dynamical system by logical rules — any approximation to experiential understanding based on expert systems shuffling symbols is not likely to converge to similar behavior and to reach a high level of competence.

Can articons be implemented using today's hardware? No, if the von Neumann architecture of ordinary computers is used. They are much closer to the data flow computer architectures that have proved to be very difficult to create. Construction of the articon system is much more difficult than construction of a rule-based expert system. The data flow in the SOAR architecture (Newell, 1990) is somewhat similar to the data flow in the articon, and the development of SOAR in the form of a rule-based expert system

shows what can be achieved in artificial intelligence at the symbolic level. Yet processing of expert system rules does not lead to processor states with qualia characteristics. Sustained, dynamical internal states of sufficiently rich relational structure are the first step towards physical realization of articons. An open question is to what extent digital technology can imitate such processes. Steps in right directions have already been made by Haikonen (2003), who looked for consciousness in winner-takes-all associative memory circuits, and Holland and Goodman (2003), who concentrated on robots with internal models. Silicon models of analog neurons already exist, capable of sustaining dynamical states, and they may be used as the building blocks of articons.

It remains to be seen whether there is something more about the phenomenal experience that is left to be explained. Scientific discussions on consciousness should be based on careful observations and critical evaluation of our inner experience. This is usually not the case, since almost everyone makes casual observations of his/her own state of mind. A few recent exceptions include the neurophenomenology of Varela (1996; see also Shanon, 1998; Shear, 1997), and the ancient Indian philosophy, especially Buddhist philosophy, based on introspection and critical reflection (Novak, 1996). In Theravada Buddhist philosophy mind and body are on equal footing. When the mind learns how to focus attention it sees that "all skandhas are empty," as one reads in the *Heart Sutra* (Conze, 1978), a text written more than sixteen centuries ago. Five skandhas, or mutually conditioning factors, include physical body, sensations, perceptions, impulses (dispositional tendencies) and consciousness. "Feeling, perception, volition, even consciousness itself" all are called empty because they do not have permanent, independent existence; everything arises as activations of brain modules. If we really look deeply everything in our mind and in the material world is constantly changing (impermanent) and is mutually dependent, everything is a flow of dynamical states sustained by activations of memory.

Associative memory and various neural structures shape the potentially accessible states of inner world, forming minds. Brain processes should be understood as the substrate of the inner world, in which mind contents and mind events are shadows of neurodynamics (Duch, 1997). Relations between mind events are not caused by the brain, but by the history of the individual, by environmental factors and social context reflected in memory. This agrees with contextual coemergence in the biocognitive epistemology of Martinez (2001). Psychological processes admit more fruitful analysis if minds are considered on their own footing. Minds of articons, systems based on the brain-like computing principles, will emerge through developmental processes with all mind characteristics, including consciousness.

References

Atienza, M., and Cantero, J.L. (2001). Complex sound processing during human REM sleep by recovering information from long-term memory as revealed by the mismatch negativity (MMN). *Brain Research, 901*, 151–160.

Baars, B.J. (1988). *A cognitive theory of consciousness*. Cambridge, Massachusetts: Cambridge University Press.

Beecher, H.K. (1946). Pain in men wounded in battle. *Annals of Surgery, 123*, 96–105.

Carey, D.P. (1996). "Monkey see, monkey do" cells. *Current Biology, 6*, 1087–1088.

Chalmers, D.J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies, 2*, 200–219.

Chalmers, D.J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Chalmers, D.J. (1997). Moving forward on the problem of consciousness. *Journal of Consciousness Studies, 4*, 3–46.

Conze, E. (1978). *Selected sayings from the Perfection of Wisdom*. Boulder, Colorado: Prajna Press.

Crile, G.W. (1998). *Origin and nature of emotions: Miscellaneous papers by G.W. Crile* [A.F. Roland, Ed.]. Seattle, Washington: The World Wide School. Electronic version is available from: http://www.worldwideschool.org/library/ (Originally published 1911)

Dennett, D.C. (1991). *Consciousness explained*. Boston: Little–Brown.

Dennett, D.C. (1996). Facing backwards on the problem of consciousness. *Journal of Consciousness Studies, 3*, 4–6.

De Quincey, C. (2002). *Radical nature: Rediscovering the soul of matter*. Montpelier, Vermont: Invisible Cities Press.

Duch, W. (1997). Platonic model of mind as an approximation to neurodynamics. In S. Amari and N. Kasabov (Eds.), *Brain-like computing and intelligent information systems* (pp. 491–512). Singapore: Springer.

Eccles, J.C. (1994). *How the self controls its brain*. Berlin: Springer.

Ekman, P., Campos, J., Davidson, R.J., and De Waals, F. (2003). *Emotions inside out*. Volume 1000. New York: Annals of the New York Academy of Sciences.

Fourneret, P., Paillard, J., Lamarre, Y., Cole, J., and Jeannerod, M. (2002). Lack of conscious knowledge about one's own actions in a haptically deafferented patient. *Neuroreport, 13*, 541–547.

Frank, T.D., Daffertshofer, A., Peper, C.E., Beek, P.J., and Haken, H. (2000). Towards a comprehensive theory of brain activity: Coupled oscillator systems under external forces. *Physica D, 144*, 62–86.

Fried, I., Wilson, C.L., MacDonald, K.A., and Behnke, E.J. (1998). Electric currents stimulate laughter. *Nature, 391*, 650.

Gaffan, D. (1996). Associative and perceptual learning and the concept of memory systems. *Cognitive Brain Research, 5*, 69–80.

Goldstone, R.L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585–612.

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*, 101–118.

Grossberg, S. (2000). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences, 4*, 233–246.

Haikonen, P. (2003). *The cognitive approach to conscious machines*. Exeter, United Kingdom: Imprint Academic.

Harnad, S. (2003). The symbol grounding problem. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. London: Nature Publishing Group/Macmillan.

Holland, O., and Goodman, R. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies, 10*, 77–109.

Humpden–Turner, C. (1981). *Maps of the mind*. New York: Macmillan.

Jackson, F. (1982). Epiphenomenal qualia. *Philosophical Quarterly, 32*, 127–136.

James, W. (1904). Does "consciousness" exist? *Journal of Philosophy, Psychology and Scientific Methods, 1*, 477–491.

Leibniz, G.W. (1982). *Vernunftprinzipien der Natur und der Gnade (Monadologie)*. Hamburg: Meiner. (Originally published 1714)

Martinez, M.E. (2001). The process of knowing: A biocognitive epistemology. *The Journal of Mind and Behavior, 22*, 407–426.

Miller, J. (1995). Going unconscious. In R.B. Silvers (Ed.), *Hidden histories of science*. New York: New York Review.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 4*, 435–450.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, Massachusetts: Harvard University Press.

Novak, P. (1996). Buddhist meditation and the consciousness of time. *Journal of Consciousness Studies, 3*, 267–277.

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

Ramachandran, V.S. (1999). Consciousness and body image: Lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Philosophical Transactions of the Royal Society London B, 353*, 1851–1859.

Reggia, J.A., Ruppin, E., and Berndt, R.S. (Eds.). (1996). *Neural modeling of brain and cognitive disorders*. Singapore: World Scientific Press.

Searle, J.R. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3*, 417–458.

Searle, J.R. (1984). *Minds, brains, and science*. Cambridge, Massachusetts: Harvard University Press.

Shanon, B. (1998). What is the function of consciousness? *Journal of Consciousness Studies, 5*, 295–308.

Shear, J. (Ed.). (1997). *Explaining consciousness: The hard problem*. Cambridge, Massachusetts: MIT Press.

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning: An introduction*. Cambridge, Massachusetts: MIT Press.

Taylor, J.G. (1998). Cortical activity and the explanatory gap. *Consciousness and Cognition, 7*, 109–148.

Terrill, J. (1962). The nature of the LSD experience. *Journal of Nervous and Mental Disease, 135*, 425–429.

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

Ullman, S. (1996). *High level vision: Object recognition and visual cognition*. Cambridge, Massachusetts: MIT Press.

Varela, F. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*, 330–349.

Weiskrantz, L. (1997). *Consciousness lost and found: A neurophsiological exploration*. Oxford, United Kingdom: Oxford University Press.

Wilson, D.A., and Stevenson, R.J. (2003). Olfactory perceptual learning: The critical role of memory in odor discrimination. *Neuroscience and Biobehavioral Reviews, 27*, 307–328.