BRAIN-INSPIRED CONSCIOUS COMPUTING ARCHITECTURE.

Włodzisław Duch,

School of Computer Engineering, Nanyang University of Technology

Nanyang Avenue, Singapore 639798

&

Department of Informatics, Nicholas Copernicus University,

ul. Grudziądzka 5, 87-100 Toruń, Poland

http://www.phys.uni.torun.pl/~duch

Abstract

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 system, guided and limited by associative memory, is similar to the stream of consciousness. Minimal requirements for an artificial system that will claim to be conscious were given in form of specific architecture named *articon*. 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 inner state flows 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 claims articon's claims. 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.

1. Introduction.
2. Brain-like computing leads to conscious systems.
3. Consequences and the resonance test.
4. Philosophical critique.
5. Conclusions.

INTRODUCTION

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 various uses of the word "consciousness" the question "can artificial systems be conscious" can also be too ambiguous to answer. In this paper another question is considered instead: what type 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. I will distinguish here three levels of difficulty. First, any useful theory of consciousness should explain the difference between those proc-

esses that we are conscious of, and have phenomenal experience, and similar processes that have no such component. Such contrastive heterophenomenology approach has been proposed by James 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 fail already at this basic level.

The second, more difficult level is to understand the structure of the qualia. Each sensory modality has 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 accompanies 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 1985), the differences between the structure of different qualia will still remain to be answered. Moreover, the qualia that we are experiencing change in time. Learning new skills (driving, horse-riding, playing) requires initially 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, most 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. After all intensive debate on this topic in the past 8 years 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. A minimal architecture of an artificial system that has this ability will be presented and analyzed in the next sec-

tion. System of this type will be called an *articon* (from *arti*-ficial *con*-sciousness), and it will be something different than just artificial intellect (or an *artilect*).

Our brains are not just of artilect type, but very complex articon types, therefore our claims of being conscious are fully justified. Phenomenal consciousness – the way a particular state feels – may be understood in ontological, not in a functional sense, since it amounts to existence of specific physical states of neural tissue of our brains. To show that this understanding of consciousness is correct in the third section of this paper several questions at different difficulty levels are answered. Only theories of consciousness based on cognitive processes allow for such predictions and explanations. In the fourth section philosophical objections that might refute the claims of an artilect to be conscious are recalled and found irrelevant. The last section contains conclusions.

BRAIN-LIKE COMPUTING LEADS TO CONSCIOUS SYSTEMS.

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 eliciting a spectrum of automatic responses. For the XIX century neurophysiologists, 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 (the fascinating history of early development of these ideas was presented by J. Miller in “Going unconsciouss”, in: Silvers 1995). 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.

Mind functions and brain complexity are closely connected. 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. Perhaps such programmed responses may even lead to advanced systems that will pass the Turing test. Computers are great deceivers, already capable of creating graphics that is 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 leads to insurmountable problems. 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, but are easy to understand. Multidimensional dynamical systems provide much better metaphors and models, but are more difficult to grasp (for an attempt to go from neurodynamics to psychological spaces see Duch, 1997). A good model of the flow of mind events is needed to elucidate the mysteries surrounding conscious behavior. An attempt to provide such model is done below.

Perception has evolved to facilitate action (O'Regan & Noë, 2001). 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 all this comes 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 and perhaps the most important role

of perception is indeed to “enable the knowledge and exercise of sensorimotor contingencies” (O’Regan & Noë, 2001). Certainly each module of the brain that processes sensory information contributes to the dynamical state of the whole brain.

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). This mechanism is used to solve the main task, 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 indeed the experiments show (Wilson and Stevenson, 2003).

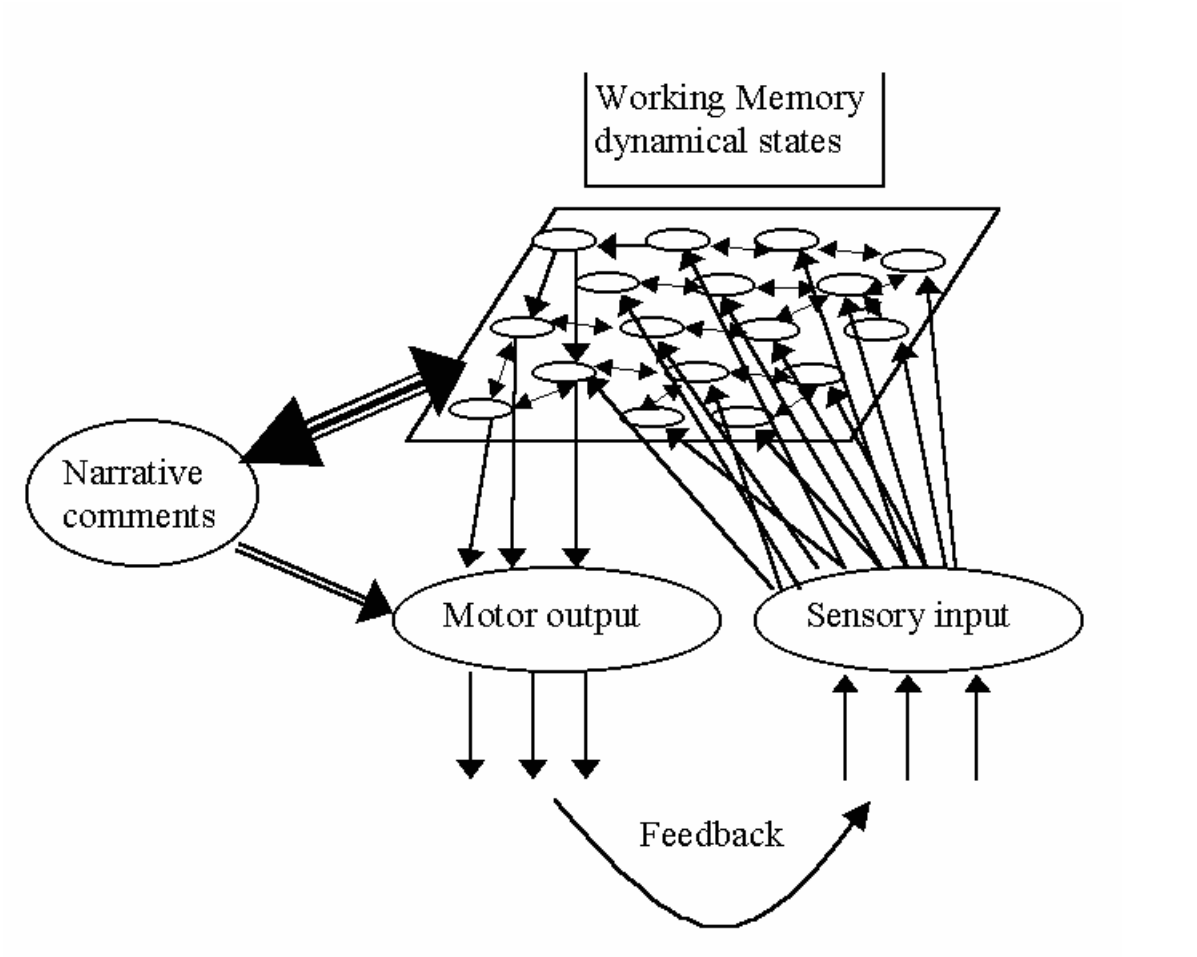
Recognition of objects is memory-based, thus every cognitive system must have associative memory capable of storing new facts and providing content-addressable retrieval. This is achieved by tuning the resonance properties of modules to excitation patterns provided by modules specializing in feature 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) 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 & Cantero, 2001).

A number of attempts to create models of brain activity based on coupled oscillators have been made (see for example Frank et al, 2000). Adaptive Resonance Theory (ART) developed by Grossberg (2000) in 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, should be based on resonant processes and may learn using ART or 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 module, producing a flow of dynamical states. In Fig.1 a sketch of such a system is shown, with additional modules that allow the articon system to comment on its own global dynamical states. How does it help to solve the consciousness problem?

The long term memory (LTM) of the human brain is huge, 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 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 the 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 grounds state due to the external stimulations or intrinsic internal dynamics create configurations of excitation patterns, an internal state that may be commented upon.

Articons are able to learn to recognize external signals repeatedly presented using associative memory patterns in form of excitations of a few modules that code particular combination of features. Thus an external signal coming from some object elicits an inner state that may be identified as percept, in 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 activations). Associative character of memory in such system will lead to a chain of activation patterns, each corresponding to some inner object, for example a train of

words or percepts. Negligibly small portion of all potential states of such system may be actualized in a non-transitory way, thanks to the associative, attractor nature of its dynamics that largely restricts the admissible inner states of the articon to combinations of those states that have been learned.

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 sensomotoric actions, forming "mind objects" that interact, defining events in the inner space (Duch, 1997). It is worthwhile to note the similarity of this picture to that of the objects and events in the material world. In quantum theories (Messiah, 1976) elementary particles are created by excitations of the ground state which does not correspond to empty space, but zero-point vacuum vibrations. Elementary particles form atoms and molecules, while material objects are composed of configurations of many atoms and molecules, the building blocks of physical objects. Excitations of the brain ground state, with its own minimal vibrations, contribute to the global dynamical state of the brain that contains several mind objects. Of course the number of mind objects is rather limited and their lifetime is short, but microworld objects have even shorter lifetimes.

To some degree mathematical structure of quantum mechanics may be applied to describe excitations of the neural modules. The activity of neural modules is quite accurately modeled using classical physics, and so far no effects specific to quantum mechanics have been discovered in neurons. Some quantum theory ideas may nevertheless be useful to describe dynamics of articon systems. For example, an analog of Heisenberg uncertainty relations may be observed. Starting from the articon system prepared in some internal state P, and asking a question Q_1 followed by Q_2 will usually create a different final state $P \rightarrow F(Q_1,Q_2)$ than if question Q_2 is asked first, followed by Q_1, leading to $P \rightarrow F(Q_2,Q_1)$. In psychology this is known as the priming effect (asking the first question changes or prepares the system). In

quantum mechanics it is expressed by the Heisenberg uncertainty relations for the two questions (observables) that cannot be answered (measured) simultaneously, like a question about the position and the momentum of a particle (measuring one changes the value of the other). Although more formal analogies of this type may be found it should be remembered that quantum mechanics is a linear theory while the dynamics of articon systems is based on non-linear processes.

Combined activity of a few modules recalls patterns of excitations stored in the LTM, reinstating similar patterns of global activity of the articon as those that were present in the moment of actual observation, when the system has learned to remember them. Working memory (WM) 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 LTM modules are strongly active (resonating) at a given moment, contributing to WM. The global dynamical state includes the state of all element in the system (the whole organism, not only the brain).

Note that resonant states in the working memory are spatio-temporal, defined in 4-dimensional space. A fragment of music remembered unfolds as a series of states in WM, each state “dressed” in associations, memories, 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 spatiotemporal pattern excitations in articon with information processing in the Turing machine. In this case the internal state is changed by a program according to some instruction, instead of dynamic flow based on associations, registers in Turing machines are “dead”, abstract entities jumping from 0 to 1 values. Articons are not only data flow machines, but have specific architecture that due to the non-linear character of module interactions create working memory states that cannot be decomposed into independent components.

"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 here it is irrelevant. Working memory states are composed from physical patterns, specific resonance activities of subsets of modules, yet they correspond to 'something out there', since they have learned from sensory data. Relations between WM states have sense and reflect the relations between the states of articon environment. The interpreter tells the narrative story symbolically labeling the working memory states, objects and their relations, reporting 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.

CONSEQUENCES AND THE RESONANCE TEST

Imagine a rat that has found some food. In a fraction of a second the rat has to decide: eat or spit? First it will smell and taste a bit of the food. Now "request for comments" is send to the long term memory from the gustatory cortex and the most relevant previous experience of the rat should recalled. LTM memory is distributed, all 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 (WM) states that are available to all modules in rat's brain. WM is small; it can hold just a few patterns (about 7±2 in humans). The small size of the working memory in 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 future (the place, type of food, its smell) centers controlling emotional reactions in rat's brain release neurotransmiters to increase the plasticity of the cortex. In addition strong physiological reaction starts to clean 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 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 come to its mind partially re-instating original response. Obviously the rat has different 'feelings' for different tastes. These feelings are real, physical states of his brain, with specific working memory 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 inner states and the ability to comment on these states allows them to express their feelings 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 any feelings (internal states in general): those associated with seeing particular colors, as well as those coming from different senses. Qualia or phenomenal experiences, independent of cognitive mechanisms, are philosophical fiction.

All such comments are not just a result of information processing, something that computer mindlessly repeats instructed by the program, but are comments on something that really exist. Articons build on the brain-like computing principles will claim to have inner

life, qualia and will claim to be conscious. In the limit of creating more and more complex models articons may acquire all subtle responses and features known from human psychology. Observing their inner life we should be able to distinguish the real thing from computer simulation programmed to respond in the same way.

Learning new skill 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 articon that controls a robot body or some other device? Conscious control is an illusion, there is only a flow of inner states. The task to be learned has to 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” – is necessary for the reinforcement learning of the skill. Learning 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, remembering and commenting upon them. Articon modules learn (real brains have cortical plasticity) regulating the amount of new versus old knowledge, or levels of activity of specific modules recruited for the task. Relating current performance to memorized episodes of recent performance requires bringing all relevant observations together (to WM), evaluation and compari-

son. Evaluation is followed by emotional reactions that provide reinforcement for learning (in the brain frequently via dopamine release), facilitating rapid learning of specialized neural modules.

Thus the role of conscious experience – interpreter continuously commenting on the state of WM – is to provide reinforcement. There is no transfer from conscious to subconscious, this idea comes from bad conceptualization started with Freud. Expectations that match the observations slowly begin to loose the competition and do not make it to the working memory, becoming "subconscious". Only a few specialized modules that won the adaptive resonance competition are left to control the task, and further fine-tuning (learning) proceeds at much slower pace.

Are the qualia in articons real? As real as it gets; 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 (Chalmers, 1996) the only one that has made steady progress in explaining the details of human experience. Emotions are connected with strong qualia. In 1911 surgeon George Crile, talking about emotions to the American Philosophical Society (Crile, 2002) 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 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 appears as quite specific working memory pattern, and has a unique "feel' to it that is expressed in interpretive commentaries of the system.

Crile (2002) also wrote “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: 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 the highest-level control with access to all brain resources is called for. How 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 specializes in giving meaning to the tactile sensory signals. Other examples of wrong cognitive interpretations of the brain data are found in the 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 the 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).

Perceptual learning (Gaffan, 1966) leads to the enhanced ability to experience qualia through training. Better memory for sensory stimulations 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 qualia. New qualia

are also experienced in dreams, because interpretation of brain states is 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 the 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 & Nöe, 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 serious impairment of behavioral competence. Information available in the brain is sufficient to make some rough discriminations but there is no reason why the qualia resulting from these discriminations should have visual qualities.

The feeling of laughter may also be understood as interpretation of 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 et. al, 1998).

Qualia have different structural properties, matching their specific roles. For example spatial structure in case of visual, tactile, temperature or pain stimuli, and non-spatial in case of taste, olfaction, thoughts or imagery. Words and thoughts “…are symbolic of motor acts” (Crile, 2002), therefore they also may have certain quality to them, although it is felt stronger by those that 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 has 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?

PHILOSOPHICAL CRITIQUE.

Turing test, or in general any test based on external observations, may fool us to believe that “there is someone in there”. 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, just a series of program steps. Results obtained with the expert system technology may impress some people – especially if they interact with the system for a short time only – 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. It has been designed to show that mere computations are not sufficient to bring real understanding. A person locked in the room and shuffling symbols to correlate incoming Chinese signs with the outgoing Chinese signs according to some rulebook does not understand questions and answers. Therefore formal systems based on rules and symbols are incapable of real understanding and passing the Turing test is not sufficient to believe that an artificial system can really understand. Can this argument be used to refute claims that articon has real understanding and is conscious of its states?

Before addressing this question it is necessary to note that the Chinese room argument has critical flaws that make it worthless. There is a vast literature criticizing or supporting Searle claims, but somehow the most important issues have been missed. First, 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 human observer could recognize that an artificial system (or an alien brain) understands have not been discussed.

Could a person placed in our brain, turning into a demon observing neural processes, find any understanding? Already Leibnitz in his *Monadology* (1714, reprinted 1982) stated that it is impossible. He asks us to enter the thinking, feeling and perceiving machine just to find there mechanisms rather than floating thoughts. Searle concludes (1980, 1984) that "brain processes *cause* conscious processes", but what it is about the brain processes that gives us genuine understanding? His solution is that neurons must have some mysterious "intentional powers" that computer elements do not have. The articon example shows that it is the organization of the system, rather than the elementary units, which is important. Turing test is an important step, a necessary although insufficient condition to grant a system genuine understanding. Chinese room fails to tell us anything about the inner world of the system under observation, focusing at the low level details. Since the outcome of this experiment is always "no" it is useless as a test.

Second, what does it mean "to understand"? Searle writes that he understands English but does not understand a word of Chinese. I intuitively know when I understand. What exactly is this feeling of understanding? 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", go on (Gopnik, 1998, compares the joy of finding an explanation in child's brain to an orgasm). I understand if I am able to relate new information to the contextual knowledge that I already have, knowledge that is finally grounded in perceptions and actions. Quantum mechanics is too remote from concepts that are

well grounded and as a result discussions on its meaning are still vigorous – perhaps we will never have a strong feeling of understanding such abstract theories.

On the other hand understanding implies the ability to answer questions requiring simple inferences. The brain is not always correct in generating the understanding signals. Some drugs or mental practices induce the illusion of understanding everything, so that the feeling is there but the ability to answer is not. Sometimes the feeling is hardly there, understanding gets more and more fuzzy, and additional questions are asked to clarify the meaning. Sometimes there is no feeling of understanding, but correct answers are given, indicating that the person in fact understands. Turing tests checks for understanding by asking questions, not by looking for the signs of feelings of understanding. A person inside the Chinese Room may finally start to understand some questions and answers as we do understand a bit foreigner waving his hands. The feeling of understanding is an extra signal that is not necessary for genuine understanding.

What are then the sufficient conditions to recognize understanding in other minds? Learning in monkeys and humans has a lot to do with imitation of behavior. This is possible because in our prefrontal cortex we have neurons that respond to specific actions performed both by oneself, and to observed actions of the 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 behavior. The Chinese Room experiment does not try to discover the mind of a system by bringing our mind in resonance with it, and thus it does not teach us anything about the mind of artificial system, failing humans as well as machines. 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, biological person? 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’, ‘run the instruction’ in response to questions.

Although the Chinese Room argument is fatally flawed it may be twisted a bit to show that articon really understands. Learning in natural environment each articon system is individual, unique, has exponentially large number of possible internal states. Therefore you cannot be in charge of the flow of articon dynamical states; there is no set of rules that can reproduce the dynamics of such a system. Suppose that articon can answer questions in Chinese and that you can observe its working memory, receiving all information that appears there in form of iconic images (perhaps less abstract than Chinese characters). Your instruction book contains explanations of all incoming patterns, referring it back to the observations 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 our brains may start to ‘resonate’ with the flow of some of the observed patterns and develop the feeling that we understand what the conversation is about. We may have a glimpse of the first-person view of the articon’s internal world. It is impossible to enter fully someone else internal world, that is 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, memory resonances and associations that the incoming information elicits. Although the difference between real brains and articon systems may be large, some understanding should be possible. If the articon will

pass the Turing test, and if it will also pass 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.

So, **what is it like to be someone else?** Questions about subjectivity are frequently asked to show that it is impossible to understand consciousness. 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 experts 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.

Experiential understanding between two minds requires certain ‘resonance’ between brain states responsible for the contents of these minds. 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 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} and A_{2i} and transitions between states within each network should be roughly similar. If both brains are in dynamical states $A_1 \sim A_2$ that in the network of their relational states correspond to each other a resonance may occur, and the experiential understanding established. We are able to share such mind states to a high degree with our family members, with other members of the same culture, to a somehow lesser degree with members of different cultures and to even lesser degree with animals, since not only their minds are formed by very different environment but 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.

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, human brain is not sufficient. Nevertheless a fairly detailed description of bat’s internal states may be formed, and some intellectual understanding achieved through modeling of 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 the behavior of bats. There is no deep mystery in the celebrated question of Thomas Nagel (1974) “how is it like to be a bat”. Nagel himself admits that perhaps all robots complex enough to behave like a person, would have experiences. His main objection is not to the physicalism itself, but rather to the lack of “the beginnings of a conception of how it might be true”. This is precisely what I have tried to show here, although it is not just the complexity, but specific organization of the robot’s brain that is important.

Another famous thought experiment concerns **Mary, the colorblind neuroscientist**, who gains color vision and learns about red color (Jackson, 1982). There are inner facts that are over and above 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 solution (Dennett, 1996) 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 bright blue banana and immediately recognizes the 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, stimu-

lated for the first time by color light, assumed new dynamical states, so she must have felt it as a new 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 were born blind and gain their sight after they are grown-up 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. Articons support similar flow of inner states as real brains and seem to be immune to philosophical critique that applies to computers.

CONCLUSIONS.

Instead of trying to define consciousness I have tried to show that systems based on brain-like computing principles will not only have inner life, but will also claim to have qualia and be conscious of them. I have proposed minimal architecture of a system called articon that should 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 fundamental way from states of a Turing machine. They include peripheral 'dressing' components that lead to subtle variation of the interpretation of the meaning of the state. The articon will recognized 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) inner experiences will 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 et al., 1996).

Taylor (1998) has described in some details possible neural underpinning of phenomenal experience characterized in terms of transparency, presence, unity, intentionality and perspective. There is no reason why articon qualia could not have the same properties as human qualia, providing that organizational principles of information processing in the artificial system are sufficiently similar to that of real brains. Claims of qualia are necessary consequence of brain-like organization of computations, in particular the ability to comment upon physical states of the architecture carrying out these computations. They may have a wide range of structural properties, depending on the complexity of the artificial system, its sensors, modalities, grounding 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, providing initially only a rough functional approximation to the 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 for different species, and it will obviously be rather different for artificial system capable of sustaining an internal state with pain-like characteristics, resulting for example from temperature sensor overheating. Burned sensor may send signals disturbing normal flow of inner states and thus requiring attention. 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 the words is in sensorimotor actions (Harnad, 2003). Perhaps similar consensus will slowly grow for the idea that qualia are physical continuous states of the brain. Weiskrantz (1997) analyzing blindsight and amnesia patients came to the conclusion that the ability to render a parallel acknowledged

commentary is indispensable for consciousness. On the engineering front Heikonen (2003) looks for consciousness in winner-takes-all associative memory circuits, while Holland and Goodman (2003) concentrate on robots with internal models.

Are articons kinds of computers? Yes, in the same sense as physical processes are doing computations, for example gravitational forces in the Solar system solve the N-body problem. Rules and computations are not good replacement 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 rules, therefore any approximation to experiential understanding based on experts systems shuffling symbols is not likely to converge to similar behavior and to reach a high level of competence. Articons definitely cannot be implemented using the von Neumann architecture of ordinary computers. 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 rule-based expert system. The data flow in the SOAR architecture (Newell, 1981) is a bit similar to the data flow in articon, and the development of SOAR in form of a rule-based expert system shows what can be achieved in artificial intelligence at the symbolic level. Processing of expert systems rules does not lead to states which could be characterized by qualia. Silicon models of analog neurons already exist, capable of sustaining dynamical states. An open question is to what extent digital technology can imitate such processes.

It remains to be seen whether there is something more about the experience to explain. The illusion that someone inside us is the real subject of experience, that homunculus exists, is very strong, but it is possible to go beyond it. 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 everybody makes casual observations on his/hers state of mind – a few recent exceptions include the neurophenomenology of Varela (1996; see also

Shanon, 1998; Shear, 1996). Ancient Indian philosophy, especially Buddhist philosophy, was based on introspection and critical reflection (Novak, 1996). When the mind learns how to focus attention it sees that “all skandhas are empty”, as one may read in the *Heart Sutra* (Conze, 1978) written more than 16 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. All these are called “empty” because they do not have permanent, independent existence, everything arises as activations of brain modules. If we look deeply enough 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.

In Theravada Buddhist philosophy mind and body are on equal footing. Brain provides the substrate, and through excitations of its modules enables the inner world. Associative memory and various neural structures shape the possible states of this inner world, forming minds. Complexity of the brain, time it takes to grow in natural environment, makes it the most valuable object in the Universe. Mind contents and mind events are based on neurodynamics (Duch, 1997), but brain processes are only the substrate for the inner world. 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 the memory. Psychological processes admit more fruitful analysis if minds are considered on their own footing.

REFERENCES

Atienza, M., & 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 (1-2),* 151-160.

Baars, B. J. (1988). *A Cognitive Theory of Consciousness.* Cambridge, MA: 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 Biol. 6, 1087-1088.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *J. 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. *J. of Consciousness Studies, 4*, 3-46.

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

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

Dennet, D. C. (1996). Facing Backwards on the Problem of Consciousness. *J. of Consciousness Studies 3,* 4-6

De Quincey, C. (2002). *Radical Nature: Rediscovering the Soul of Matter*. Invisible Cities Press.

Duch, W. (1997). Platonic model of mind as an approximation to neurodynamics. In: *Brain-like computing and intelligent information systems*, ed. S-i. Amari, N. Kasabov (Springer, Singapore 1997), chap. 20, pp. 491-512.

Frank, T. D., Daffertshofer, A., Peper, C. E., Beek, P. J., & 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., 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(1-2),* 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.

Harnad, S. (2003). The Symbol Grounding Problem. *Encyclopedia of Cognitive Science*. Nature Publishing Group/Macmillan.

Haikonen, P. (2003). *The Cognitive Approach to Conscious Machines*. Exeter, UK: Imprint Academic.

Holland, O., & Goodman, R. (2003). Robots with Internal Models: A Route to Machine Consciousness? In: *Machine Consciousness*. Holland, O (ed.), Exeter, UK: Imprint Academic.

Humpden-Turner, C. (1981). *Maps of the mind*. MacMillian.

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.

Messiah, A. (1976). *Quantum mechanics*. Amsterdam: North Holland.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 4,* 435-50.

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

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

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

Ramachandran, V.S. (1999). Consciousness and body image: lessons from phantom limbs, Capgras syndrome and pain asymbolia. *Phil. Trans. R. Soc. Lond. B 353,* 1851-1859.

Reggia, J. A., Ruppin, E., Berndt, R. S., eds. (1996). *Neural Modeling of Brain and Cognitive Disorders*. Singapore: World Scientific.

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

Searle, J. R. (1984). *Minds, Brains, and Science*. Cambridge: Harvard University Press.

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

Shear, J. ed. (1997). *Explaining Consciousness: The Hard Problem*. Cambridge: MIT Press.

Silvers, R. B., Eds. (1995). *Hidden Histories of Science*. New York: New York Review.

Taylor, J. G. (1998). Cortical Activity and the Explanatory Gap. *Consciousness and Cognition, 7,* 109-148

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

Weiskrantz, L. (1997). *Consciousness Lost and Found: A Neurophysiological Exploration*. Oxford: Oxford University Press.

Wilson, D. A., & Stevenson, R. J. (2003). Olfactory perceptual learning: the critical role of memory in odor discrimination. *Neuroscience & Biobehavioral Reviews, 27(4),* 307-328.

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