

DRAFT CHAPTER

ARTIFICIAL MINDS AND CONSCIOUS MACHINES

Pentti O A Haikonen
Nokia Research Center

Introduction

Machine cognition; artificial abilities like the understanding of scenery, speech and text as well as imagination, reasoning, planning and learning will have an important role in future information technology and robotic applications. However, practical machine cognition has proven to be notoriously difficult by traditional computational means.

Common sense says that we humans can understand and the machine cannot because we can think and the computer cannot, we are conscious and the computer is not. We are capable to mindful actions and the computer is not because we have a mind and the computer doesn't have. Therefore, to create useful machine cognition we should create true thinking and conscious machines, ones with artificial minds. These machine minds should imitate the human mind in all its expressions and ways of operation.

The human mind is characterized by the flow of inner imagery, inner speech, sensations, emotional moods and the awareness of these. The human mind is imaginative, creative and intelligent. The human mind possesses intentionality; it operates with meanings and significance, it understands what it is doing. The human mind seems to unify effortlessly past experience, present multisensory information, the expected and desired future, the needs, drives and goals as well as any triggered emotional states. Our minds have contents that make us what we are; our personal history, fortunes and misfortunes, needs, values, secret desires, yearnings and hopes, our way of perceiving and acting upon the world. Mind and consciousness go together; consciousness provides us the instantaneous focus of awareness and the ability to report our mental content to ourselves and others.

Are artificial minds possible? The philosophy of mind has provided us with loose descriptions like the one above. Philosophy has also pinpointed some relevant questions like those related to the mind-body problem. However, an engineer wishing to create an artificial mind needs a more practical approach than most of those so far provided by philosophy. This kind of approach has been proposed recently by the author.

The Problems to Be Solved

The fundamental issues of artificial minds and conscious machines are, as the author sees it:

Visions of Mind

1. The mind-body problem; the apparent immaterial nature of the human mind and the implications of this to the possibilities to construct artificial minds.
2. Symbolic processing with meaning and significance in the human sense. This is against the tradition of rule-based symbolic processing in computers and also the conventional artificial neural networks.
3. Perception process.
4. A cognitive architecture that facilitates the unification of sensory information from multiple senses, experience, the present status, needs and goals of the machine and supports inner imagery, inner speech, system reactions and emotions.
5. The consideration and creation of the mind-machine as a system with system reactions and emotions.
6. Consciousness; supervisor, soul or only a subjective inner appearance? Testing machine consciousness; how can we know if a machine is conscious?
7. Learning and training. A mind, human or artificial is not a neural network or a collection of cognitive functions, instead it is the content. The hardware must be designed to support the mind, but the actual mind will only arise when the system learns about itself and its environment.

In the following these issues are discussed and some solutions are proposed.

The Immaterial Mind–Body Problem

Our mind seems to be immaterial. This is our everyday experience and it cannot be overlooked. Our mind seems to be about external and internal, abstract and concrete entities effortlessly and without any perceived material machinery. In fact, we perceive the external world to be out there; tables, chairs, books, whatever, directly without the awareness of any neural activity, intermediate symbols or representations.

Any serious attempt towards the design of artificial minds must first address this apparent immateriality of mind. If mind was indeed immaterial then obviously any material construction effort would fail and our work would be futile. Therefore we have to show that mind is indeed based on carrying material symbols and representations, but how could this be so given our everyday perception of the contrary?

What proof do we have about the assumed immateriality of the mind? We have only the perceived immaterial appearance, the subjective perception of the non-existence of any material basis of the mind, but does this constitute solid proof? The point is: For the observer self the appearance of perception "without the awareness of any carrying material symbols or processes" and "without any carrying material symbols or processes" can be the same. Therefore, the missing awareness of material carrying symbols or processes does not prove that these would not exist. Yet, for centuries common people and philosophers alike have made this logically unsound conclusion, which has led to the idea of an immaterial mind. This in turn leads to the mind-body problem; how an immaterial mind can interact with material body. By the

Visions of Mind

very definition of immaterial and material substances this is a problem without solution, yet philosophers have tried and tried to find one. However, it can be seen that the mind-body problem evaporates as soon as its unsound foundation is realized. Instead, the question remains: Why can't we perceive the material machinery behind our mind? Even this question is distorted and we should rather ask: What would it take for us to be able to perceive the machinery, the neural firings, etc.? The answer is: Whatever it takes we do not have and the material machinery remains beyond our introspective powers.

The conclusion: There is no real proof about the immateriality of mind, instead it is more likely that mind is grounded to material carrier processes that take place in the brain. Accordingly it can be assumed that equivalent carrier processes may be realized in artificial systems so that artificial minds can arise. However, it would be fair to demand that any artificial cognitive system should perceive its mind as immaterial just like we humans do; this is an important design requirement and a test for the success of the design.

Representation, Meaning and Signal Arrays

It is known that in biological systems information is carried by neural signals. The initial meaning of a sensory signal is derived from the corresponding sensor; each of these signals represents a small fraction or feature of the total information available from that sensor. This leads to the idea of distributed representations; sensed entities may be represented by large signal arrays where the meaning of each individual signal is grounded to a particular property of the sensed entity. Likewise we can assume that all motor acts could consist of orchestrated elementary movements, motor primitives, each controlled by a single signal. Thus each composite motor act would be governed by a controlling signal array, a distributed representation that switches motor primitives on and off as needed. In this way distributed representations could cover both sensory percepts and motor responses. Thus, when we sense visually an object we get a certain array of signals that correspond to the perceived pattern; when we hear a sound we get a certain sequence of signal arrays that corresponds to the auditory pattern; when we think of moving our hand we initiate a certain sequence of signal arrays that correspond to the sequenced motor activity. This kind of distributed representation by large arrays of signals can be easily implemented by electronic means.

However, while all this is rather trivial and sufficient for simple stimulus-response systems, it will not explain or allow higher cognition and thinking. A heard word would be useless if it could be taken as a sound pattern only; a seen letter would likewise be useless if it could be taken as a visual pattern only. Words and letters must depict something beyond their sensed appearance; higher cognition and thinking can only arise if entities can stand for something that they are not. Thus we must be able to associate secondary meanings with the signal arrays that are initially generated by sensory processes as a response to sensed entities. Only via this association these signal arrays will become useful tokens by which complicated matters can be described and imagined; this is a necessary prerequisite for the eventual generation of inner speech and inner imagery and the emergence of mind.

Thus association between representations, i.e. signal arrays and sequences of signal arrays would seem to be a necessary mode of processing. This association

Visions of Mind

would allow the evocation of a given representation by the associated representation. Especially the following associative linking capacity would be needed:

signal array – signal array
sequence of signal arrays – signal array
signal array – sequence of signal arrays
sequence of signal arrays – sequence of signal arrays

Distributed representations, arrays of signals, are not monolith symbols. Instead they have an inherent fine structure and this allows also partial representations to evoke a response. This leads to automatic classification; representations that are similar enough will evoke the same response representation. Here no extensive learning is needed, only one example with associated response will already do and all further representations that are close enough to the original one will evoke the same response.

What kind of hardware could support these basic associative processes with distributed representations? Several possibilities may exist. The author has experimented with artificial associative neurons and neuron groups that are able to learn the required associations via modified Hebbian process. Here the associative connection between two signals is established if these signals appear together couple of times. Time-averaging is also included; if the signals fail to coincide, the eventual associative connection is impeded. Thus random coincidences will not accumulate and become learned associations. This mechanism allows also e.g. the labeling of a property that does not appear separately. A large number of these neurons can be used to learn associations between signal arrays. Additional short-term memory elements are needed for the manipulation of sequences.

The actual realization of the circuitry is of no consequence here. It suffices that we define an artificial neuron group that is able to learn and produce the above-mentioned associations between signal arrays and sequences of signal arrays.

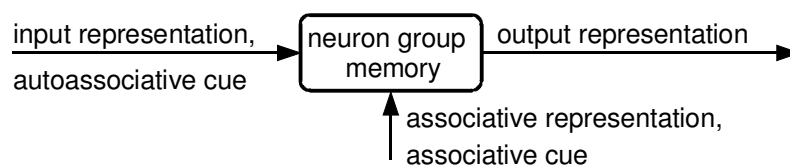


Fig. 1. The defined neuron group

The neuron group of fig. 1. associates the associative representation with the input representation and the input representation with itself. Thus a part of an input representation may evoke the full representation as the output and likewise, a part of an associative representation may evoke the associated input representation or sequence as the output. In this way the generic neuron group works also as an associative memory that responds also to partial evoking representations. Winner-Takes-All principle can be applied; if competing evoking representations occur, the strongest evoking representation will win and its response will emerge as the output.

Perception; Acquisition of Information and the Tokens of Cognition

A cognitive system uses perception processes to access information about its environment and its own physical states via sensors. The interpretation of the sensory information to represent one object and not another, perhaps equally or even more probable from the sensory point of view, depends on the experience and contextual state of the cognitive system. This requirement calls for feedback from the system's inner processes. An outline of a perception system with feedback, the sensory perception loop, is depicted in fig. 2.

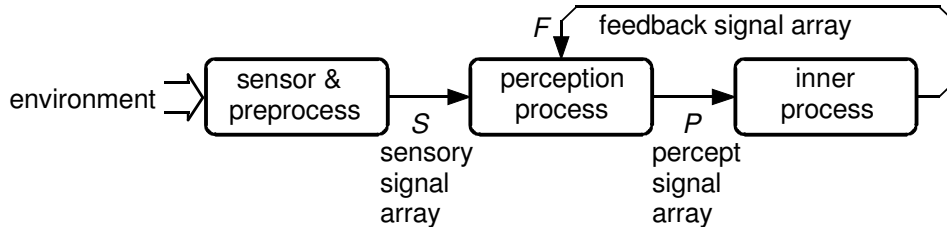


Fig. 2. Sensory perception loop

In fig. 2. the perception process starts with a sensor. Initial preprocessing of sensory information may be necessary so that a suitable distributed representation sensory signal array can be had. Perception process combines the effect of sensory signal array S and internally generated feedback signal array F . The resulting signal array P is called percept signal array as it will now be the "official" output from the perception process and as such will be forwarded to the system's inner processes.

Initially when the system is in unlearned state, there is no feedback and the sensory signal array passes through the perception process as such and will become the percept for the system. However, in learned system feedback signals may be generated. The percept will now be a function of the sensory signal array and the feedback signal array. The feedback may now amplify matching sensory signals and if thresholds are applied at various points within the system then only the thus amplified part of the sensed signal array will pass and have effect.

Sensory match, mismatch and novelty conditions can be defined here as follows:

- $S \approx F$ \Rightarrow match condition
- $S \neq F$ \Rightarrow mismatch condition
- $S \rightarrow \text{no } F$ \Rightarrow novelty condition

The evoked feedback signal array may constitute a prediction for the sensory signal array, in that case the match/mismatch/novelty condition would indicate the successfulness of the prediction. Sometimes the evoked feedback signal array may depict an entity to be sought. In that case match/mismatch would indicate the success status of the search. It is obvious that the system should strive towards match-condition. This it can do by attention control. The desired effect of match/mismatch/novelty condition on attention can be summarized as follows:

- Match condition \Rightarrow sustain attention
- Mismatch condition \Rightarrow refocus inner attention
- Novelty condition \Rightarrow focus attention

The feedback mechanism facilitates also introspection. Introspective perception of inner imagery or other inner representations takes place when the percept is due to the feedback signals only, when there is no sensory input or the input is subdued.

Cognitive Architectures; Platforms for Artificial Minds

A prerequisite for an artificial mind is the unification of sensory information from multiple sensory sources and machine's own knowledge as well as the system's present status, needs and goals. This is a system level requirement and can be solved by a suitable system architecture. This cognitive architecture must integrate the following functions into one system; multisensory perception, attention, short-term and long-term memories, learning, the flow of inner speech and inner imagery, judgement, deduction, reasoning, planning, motivation, response generation, system reactions and emotions. The system architecture must also be able to combine visual, auditory, touch and body position information so that a consistent view of the environment can be achieved, one that allows e.g. the fluent navigation around and reaching of objects.

Obviously in a system with multiple sensory modalities, like visual, auditory, tactile etc. each modality would have its own perception loop. Also the system would include a motor response module. It is also obvious that each sensory modality would have to have access to the motor response module; motor responses could be needed for visual, auditory, tactile etc. stimuli independent of each other. Associative connections between the perception loops would be needed, too. A simplified outline of this kind of a system architecture with two sensory modalities is depicted in fig. 3.

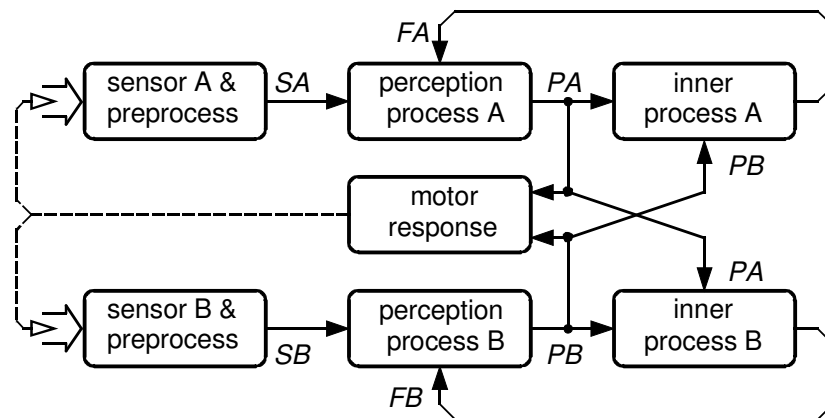


Fig. 3. A cognitive architecture based on cross-coupled perception loops

The associative cross-coupling between sensory modalities will allow the association of percepts from different modalities with each other so that later on one can be evoked by the other. In biological systems this kind of cross-modality association is advantageous already in primitive cases. It is useful to associate the taste of edibles with the visual percepts of the same so that later on visual search for these edibles could be possible. It is also useful to associate the sound of a predator with its visual counterpart and perhaps with the percept of pain so that the predator could be recognized and evaded also in darkness. Along these lines this associative mechanism will allow the generation of consistent auditory, visual and motor "landscapes".

Thus various percepts may be associated with and evoked by others. The evoking percept does not have to be generated by sensory stimuli, instead it may consist of feedback from the inner processes. Thus sequences of "imagined" percepts may be generated; possible motor responses for these may or may not be acted out. If the motor response is able to modify the external world that generates the input stimuli, another closed feedback loop is formed. The resulting action is not necessarily trivial and complex stimulus-response sequences may arise.

Visions of Mind

The cross-association process allows the association of any entities that are in temporal conjunction. Thus percepts that do not have any natural connection may nevertheless be associated with each other. This apparent flaw has unexpected consequences; certain originally unrelated percepts may become to represent and symbolize other entities. This is a basic prerequisite for language; arbitrary sound patterns become words with meanings.

System Reactions and Emotional Significance

A true cognitive system must be able to evaluate the significance of an event like "this is good", "this is bad", "this is dangerous" etc. and base its further operation on the results of these evaluations. This evaluated significance should guide attention, learning and memorization. Biological cognition bases its significance evaluation on elementary sensations like taste, smell, pain and pleasure. These elementary sensations are also often generalized to apply more abstract matters. Elementary sensations may also evoke basic system reactions as immediate short-cut responses to the situation.

I propose here that machine evaluation of significance should be based on elementary sensory information originating from suitable sensors. These sensors include smell and taste as well as pain and pleasure. Even though a robot may not need to accept or reject things by their smell and taste, an artificial sensor for good and bad value and system reaction initiation could still be provided. Likewise pain and pleasure system reaction initiation may be based on artificial inputs. In robotic applications sensors for physical damage should be used as pain sensors. These inputs could then also be used to punish and reward the system.

These sensors are called here elementary sensors as their outputs need very little in the way of cognitive interpretation. Instead, in biological systems sensations from these sensors have direct and automatic system reactions. Some of these are listed here:

Good taste, smell	⇒ accept, approach
Bad taste, smell	⇒ reject, withdraw
Pain in general	⇒ demand attention
Pain; self-inflicted	⇒ withdraw, discontinue on-going action
Pain, caused by others	⇒ aggression, retaliation
Pain, overpowering	⇒ submission
Pleasure	⇒ sustain on-going action, approach

It can be seen that these sensations have built-in judgement criteria; they dictate whether something is good or bad, whether some activity should be continued or not. They also offer short-cut reactions as immediate responses. In human cognition these criteria and reactions are also generalized to apply to other sensory percepts and situations. Thus things and situations in general may acquire emotional significance; things and situations will be pleasant and desirable or bad and to be avoided, will evoke anger, etc.

In the context of associative signal array processing emotional significance should be implemented in a way that allows the focussing of attention on the emotionally significant signal arrays and also the evocation of system reactions that correspond to the specific emotional value. These functions can be added to the basic system model in the way depicted in fig. 5.

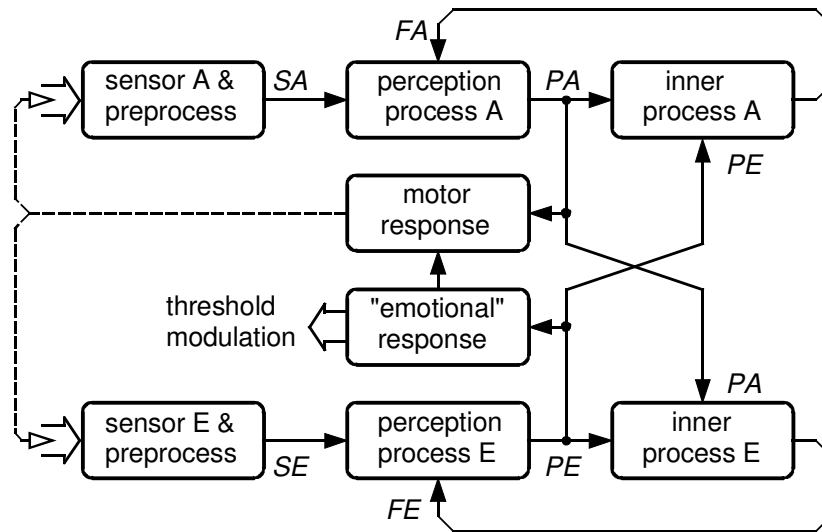


Fig. 5. "Emotional" control of a cognitive system

Fig. 5. depicts the outline of the principle of emotional control in artificial cognitive system. Here *SE* designates a sensory signal array from an elementary sensor such as pain, pleasure, good or bad. The corresponding percept *PE* may initiate a respective "emotional" response that will control attention by modulating signal intensities and threshold levels for inner processes. The "emotional" response affects also the execution of motor responses; the direction, force and speed.

According to this model also percepts from other sensory modalities may become associated with "emotional" significance if these percepts appear together with percepts of pain, pleasure, goodness or badness (*PA* at inner process E). Thereafter this percept, sensed or imagined, may trigger the associated "emotional" significance and the related response. This process may even be abstract as the emotional percepts (*PE*) may be evoked by inner causes only or even by verbal description. Thus, for instance, the system may be verbally taught that "this is bad" etc.

How about machine emotions? Typical human emotions like fear, anger, hate, envy, desire etc. suggest each certain ways of responses; should something be approached or avoided, should aggression or submission be used, etc. In this way emotions function as kinds of model templates for action. This template may not be unequivocal, though, as emotions often involve conflict and confusion. Desire may be sustained because the object of the desire cannot be approached, astonishment involves indecision between approach and withdrawal, etc.

Emotions function also as important motivational factors. We do things out of curiosity, fear, anger, envy, jealousy, guilt, expectation of pleasure or punishment, etc.

Human emotions have also a subjective aspect; they feel like something and have physiological effects.

It is proposed here that the functional aspects of emotions could be modeled as combinations of system reactions caused by actual and generalized elementary sensations. In this way we could for instance get:

- curiosity = novelty + approach
- astonishment = mismatch + approach + withdraw
- caution = novelty + withdraw
- desire = pleasure + good + approach
- anger = bad + aggression

Visions of Mind

sadness = bad + submission
etc.

Emotional significance can be used as a motivational factor in cognitive machines, too. For instance, the emotional significance of pleasure may be associated with the execution of a given task. Thereafter the system would focus its attention to the execution of that task at the slightest cues, by the mechanism explained earlier, whenever the environment would allow it. Likewise the system could be made to avoid actions that have the emotional significance of displeasure or pain.

Machine emotions would involve instant judgement by emotional value, system reactions and direction of action as well as motivational effects. In principle these would be useful functions but sometimes might counteract more appropriate rational responses. Quick emotional short-cut reactions in a dangerous situation may save the day for the robot, but a robot in emotional rage would be no good for most purposes. It is up to the designer to find a proper balance here.

Consciousness in the Machine

What is consciousness? Is it an immaterial soul, not so immaterial system level supervisor or only a subjective inner appearance? I think we can reject the immaterial soul here and discuss consciousness in more practical terms.

What is the difference between being conscious and being unconscious? The snappy answer is: We are conscious when we know that we are conscious. More seriously: Consciousness involves the ability to report oneself about one's percepts of environment, own physical status and one's own mental content. Do we need memory function for this? Maybe not, but if no memories are made then we will not be able to report afterwards about what we experienced a moment ago and thus we will not be able to determine whether we have been conscious or not. Thus, for continuous conscious experience episodic memory is needed. This kind of short-term episodic memory is also necessary for practical reasons because we have to know what we are doing, what we have already done and how they relate to each other and to actions to be done later on.

Does consciousness have a function and is it related to cognition and the control of actions? Indeed, ideas towards this effect have been proposed by e.g. Baars. According to Baars consciousness is needed to make associations, detect priorities, create access to unconscious resources, facilitate decision making and execute control, detect errors and edit plans, create access for the self (Baars 1997, pp. 158 – 164). Is consciousness really an agent responsible for these tasks? Or would it be more plausible to assume that these actions can be explained by the cognitive functions of perception, learning, attention, deduction, introspection etc., without any reference to consciousness? This, I think, is the case and consciousness remains only as the subjective appearance of the cognitive information processing style.

How can we know if a machine is conscious? Consciousness is a subjective property so initially it would seem that only indirect tests could be applied. The Turing test is one of these and is supposed to indicate whether a machine is able to think and even be conscious; if the machine is able to fool us into believing that it has these properties, then it does have them. It should be obvious that this test is logically unsound and should be rejected.

Rigorous tests for machine thinking and consciousness can be devised. Human thinking is characterized by the inner speech and inner imagery; a thinking machine

Visions of Mind

should have these, too. As the designers of the machinery we will know whether the machine has these or not, we will also be able to monitor these in the actual running system. Thus it would be very easy to see and test whether the machine thinks in a human way, has the flow of inner speech and imagery and correctly grounded symbol utilization.

Consciousness involves one's awareness of one's own mental content; therefore if the machine is able to report its own inner speech and imagery, then obviously one important indicator of consciousness is there. If, on the other hand, the machine were not able to report these, not even in principle, then the machine would not really be conscious.

It may be argued that true consciousness involves a kind of immaterial spookiness that cannot possibly be captured in a machine. This argument reflects the misunderstood mind-body problem and is not valid. A properly designed cognitive machine will not perceive the information carrying signals or machinery as such and these remain transparent to it. Therefore the actual "machine mind" will seem to be about entities directly and will appear as immaterial to the machine itself.

Artificial Minds

Suppose that we have built a thinking machine along the ideas that I have presented here and are ready to switch it on. Does it have a mind now? I don't think so. At the moment when it is switched on for the first time it will possess the faculty of perception, a collection of cognitive functions and the "conscious style" of operation, but, unless vast a priori information is provided, it will not have much of a mind. As I see it, a mind, human or artificial, is not the hardware, not a neural network nor a collection of cognitive functions, instead it is a content-level phenomenon. A mind will only arise when the system learns about itself and its environment, learns to seek to satisfy its needs and drives, acquires its beliefs and values and adjusts its behavior accordingly. True, there can be no mind without a supporting hardware, but the hardware and the cognitive functions provided by it will only be half of the story. Problems with the hardware may cause problems with the mind, but mind-related problems might arise also in good hardware.

Thus, the creation of an artificial mind calls for the creation of a supporting machinery and also the training of the system. Some parts of this training may be provided by self-learning, some parts may require a teacher. Good/bad values will be necessary, reward and punishment may have to be utilized. How exactly will this be done and how much effort will be needed? Should we leave this to engineers or should we hire an elementary school teacher? This remains to be seen, but fortunately, once one system has been successfully trained subsequent copies may be reproduced by conventional electronics manufacturing processes.

References

Aleksander Igor, *How to Build a Mind*, Weidenfeld & Nicolson, Great Britain, 2000

Baars Bernard J., *In the Theater of Consciousness*, Oxford University Press Oxford New York, 1997

Crick Francis, *The Astonishing Hypothesis*, A Touchstone Book, New York, 1994

Visions of Mind

Edelman Gerald M., Tononi Giulio, *Consciousness*, Penguin Books Ltd, London, 2000

Haikonen Pentti O: *The Cognitive Approach to Conscious Machines*, Imprint Academic, UK, 2003

Rantala Arto, Haikonen Pentti: "An associative neuron group microchip", in *Proceedings of the 20th Norchip Conference*, Copenhagen, DK, 11-12, Nov. 2002, Technoconsult. Copenhagen (2002), pp. 335 - 340

Searle John R., *Minds, Brains & Science*, Penguin Books Ltd., London England, 1984

Sloman Aaron, "Introduction: Models of Models of Mind" in *Proceedings of the AISB'00 Symposium on How to design a functioning mind*, 17th - 20th April, 2000, pp. 1 – 9

Taylor John G., *The Race for Consciousness*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, 1999

Turing Alan M., "Computing Machinery and Intelligence" in *Mind* LIX, no 2236 Oct. 1950, pp. 433 – 460