1
Introduction

1.1 GENERAL INTELLIGENCE AND CONSCIOUS MACHINES

Suppose that you were going to see a conscious machine, perhaps a robot. What would you expect to see? Recent advances in robotics have produced animal and humanoid robots that are able to move in very naturalistic ways. Would you find these robots conscious? I do not think so. As impressive as the antics of these artefacts are, their shortcoming is easy to see; the lights may be on, but there is 'nobody' at home. The program-controlled microprocessors of these robots do not have the faintest trace of consciousness and the robots themselves do not know what they are doing. These robots are no more aware of their own existence than a cuckoo clock on a good day.

Artificial intelligence (AI) has brought chess-playing computer programs that can beat grand masters and other 'intelligent' programs that can execute given specific tasks. However, the intelligence of these programs is not that of the machine; instead it is the intelligence of the programmer who has laid down the rules for the execution of the task. At best and with some goodwill these cases of artificial intelligence might be called 'specific intelligence' as they work only for their specific and limited application. In contrast, 'general intelligence' would be flexible and applicable over a large number of different problems. Unfortunately, artificial general intelligence has been elusive. Machines do not really understand anything, as they do not utilize meanings.

Robots do not fare well in everyday tasks that humans find easy. Artificial intelligence has not been able to create general intelligence and common sense. Present-day robots are definitely not sentient entities. Many researchers have recognized this and see this as a shortcoming that must be remedied if robots are ever to be as versatile as humans. Robots should be conscious.

Exactly which features and abilities would distinguish a conscious robot from its nonconscious counterpart? It may well be that no matter how 'conscious' a robot might be, some philosophers would still pretend to find one or another successful argument against its consciousness.

Robot Brains: Circuits and Systems for Conscious Machines Pentti O. Haikonen

An engineering approach may bypass the philosophical mess by defining what the concept of 'machine consciousness' would involve. However, care should be taken here to consider first what natural consciousness is like so that the cautionary example of artificial intelligence is not repeated; AI is definitely artificial but has somehow managed to exclude intelligence.

Folk psychology describes human consciousness as 'the immaterial feeling of being here'. This is accompanied by the awareness of self, surroundings, personal past, present and expected future, awareness of pain and pleasure, awareness of one's thoughts and mental content. Consciousness is also linked to thinking and imagination, which themselves are often equated to the flow of inner speech and inner imagery. Consciousness is related to self, mind and free will. Consciousness is also seen to allow one to act and execute motions fluently, without any apparent calculations. A seen object can be readily grasped and manipulated. The environment is seen as possibilities for action.

Folk psychology is not science. Thus it is not able to determine whether the above phenomena were caused by consciousness or whether consciousness is the collection of these phenomena or whether these phenomena were even real or having anything to do with consciousness at all. Unfortunately philosophy, while having done much more, has not done much better.

A robot that could understand the world and perceive it as possibilities for action would be a great improvement over existing robots. Likewise, a robot that would communicate with natural language and would even have the flow of natural language inner speech could easily cooperate with humans; it would be like one of us. It can be seen that the machine implementation of the folk psychology hallmarks of consciousness would be beneficial regardless of their actual relationship to 'true' consciousness, whatever that might eventually turn out to be.

This can be the starting point for the engineering approach. The folk psychology hallmarks of consciousness should be investigated in terms of cognitive sciences and defined in engineering terms for machine implementation. This would lead to an implementable specification for a 'conscious machine'.

This 'conscious machine' would be equipped with sensors and perception processes as well as some means for a physical response. It would perceive the world in a direct way, as objects and properties out there, just as humans do, and perceive itself as being in the centre of the rest of the world. It would be able to process information, to think, in ways seemingly similar to human thinking, and it would have the flow of inner speech and imagery. It would also be able to observe its own thoughts and introspect its mental content. It would perceive this mental content as immaterial. It would motivate itself and have a will of its own. It would judge the world and its own actions by emotional good–bad criteria. It would be aware of its own existence. It would be able to move and execute actions as freely and readily as humans do. It would emulate the processes of the human brain and cognition. It would appear to be conscious.

This leads to two questions: first, how do the human brain and mind actually work and, second, how could the workings of the brain be best emulated in an artificial

system? Unfortunately the exact answer is not yet known to the first question and it may well be that a definite answer can be found only after we have learned to create successful artificial minds. Thus, models of machine cognition may also help to model human cognition.

1.2 HOW TO MODEL COGNITION?

Presently there are five main approaches to the modelling of cognition that could be used for the development of cognitive machines: the computational approach (artificial intelligence, AI), the artificial neural networks approach, the dynamical systems approach, the quantum approach and the cognitive approach. Neurobiological approaches exist, but these may be better suited for the eventual explanation of the workings of the biological brain.

The computational approach (also known as artificial intelligence, AI) towards thinking machines was initially worded by Turing (1950). A machine would be thinking if the results of the computation were indistinguishable from the results of human thinking. Later on Newell and Simon (1976) presented their Physical Symbol System Hypothesis, which maintained that general intelligent action can be achieved by a physical symbol system and that this system has all the necessary and sufficient means for this purpose. A physical symbol system was here the computer that operates with symbols (binary words) and attached rules that stipulate which symbols are to follow others. Newell and Simon believed that the computer would be able to reproduce human-like general intelligence, a feat that still remains to be seen. However, they realized that this hypothesis was only an empirical generalization and not a theorem that could be formally proven. Very little in the way of empirical proof for this hypothesis exists even today and in the 1970s the situation was not better. Therefore Newell and Simon pretended to see other kinds of proof that were in those days readily available. They proposed that the principal body of evidence for the symbol system hypothesis was negative evidence, namely the absence of specific competing hypotheses; how else could intelligent activity be accomplished by man or machine? However, the absence of evidence is by no means any evidence of absence. This kind of 'proof by ignorance' is too often available in large quantities, yet it is not a logically valid argument. Nevertheless, this issue has not yet been formally settled in one way or another. Today's positive evidence is that it is possible to create world-class chess-playing programs and these can be called 'artificial intelligence'. The negative evidence is that it appears to be next to impossible to create real general intelligence via preprogrammed commands and computations.

The original computational approach can be criticized for the lack of a cognitive foundation. Some recent approaches have tried to remedy this and consider systems that integrate the processes of perception, reaction, deliberation and reasoning (Franklin, 1995, 2003; Sloman, 2000).

There is another argument against the computational view of the brain. It is known that the human brain is slow, yet it is possible to learn to play tennis and other

activities that require instant responses. Computations take time. Tennis playing and the like would call for the fastest computers in existence. How could the slow brain manage this if it were to execute computations?

The artificial neural networks approach, also known as connectionism, had its beginnings in the early 1940s when McCulloch and Pitts (1943) proposed that the brain cells, neurons, could be modelled by a simple electronic circuit. This circuit would receive a number of signals, multiply their intensities by the so-called synaptic weight values and sum these modified values together. The circuit would give an output signal if the sum value exceeded a given threshold. It was realized that these artificial neurons could learn and execute basic logic operations if their synaptic weight values were adjusted properly. If these artificial neurons were realized as hardware circuits then no programs would be necessary and biologically plausible artificial replicas of the brain might be possible. Also, neural networks operate in parallel, doing many things simultaneously. Thus the overall operational speed could be fast even if the individual neurons were slow.

However, problems with artificial neural learning led to complicated statistical learning algorithms, ones that could best be implemented as computer programs. Many of today's artificial neural networks are statistical pattern recognition and classification circuits. Therefore they are rather removed from their original biologically inspired idea. Cognition is not mere classification and the human brain is hardly a computer that executes complicated synaptic weight-adjusting algorithms.

The human brain has some 10^{11} neurons and each neuron may have tens of thousands of synaptic inputs and input weights. Many artificial neural networks learn by tweaking the synaptic weight values against each other when thousands of training examples are presented. Where in the brain would reside the computing process that would execute synaptic weight adjusting algorithms? Where would these algorithms have come from? The evolutionary feasibility of these kinds of algorithms can be seriously doubted. Complicated algorithms do not evolve via trial and error either. Moreover, humans are able to learn with a few examples only, instead of having training sessions with thousands or hundreds of thousands of examples. It is obvious that the mainstream neural networks approach is not a very plausible candidate for machine cognition although the human brain is a neural network.

Dynamical systems were proposed as a model for cognition by Ashby (1952) already in the 1950s and have been developed further by contemporary researchers (for example Thelen and Smith, 1994; Gelder, 1998, 1999; Port, 2000; Wallace, 2005). According to this approach the brain is considered as a complex system with dynamical interactions with its environment. Gelder and Port (1995) define a dynamical system as a set of quantitative variables, which change simultaneously and interdependently over quantitative time in accordance with some set of equations. Obviously the brain is indeed a large system of neuron activity variables that change over time. Accordingly the brain can be modelled as a dynamical system if the neuron activity can be quantified and if a suitable set of, say, differential equations can be formulated. The dynamical hypothesis sees the brain as comparable to analog

feedback control systems with continuous parameter values. No inner representations are assumed or even accepted. However, the dynamical systems approach seems to have problems in explaining phenomena like 'inner speech'. A would-be designer of an artificial brain would find it difficult to see what kind of system dynamics would be necessary for a specific linguistically expressed thought. The dynamical systems approach has been criticized, for instance by Eliasmith (1996, 1997), who argues that the low dimensional systems of differential equations, which must rely on collective parameters, do not model cognition easily and the dynamicists have a difficult time keeping arbitrariness from permeating their models. Eliasmith laments that there seems to be no clear ways of justifying parameter settings, choosing equations, interpreting data or creating system boundaries. Furthermore, the collective parameter models make the interpretation of the dynamic system's behaviour difficult, as it is not easy to see or determine the meaning of any particular parameter in the model. Obviously these issues would translate into engineering problems for a designer of dynamical systems.

The quantum approach maintains that the brain is ultimately governed by quantum processes, which execute nonalgorithmic computations or act as a mediator between the brain and an assumed more-or-less immaterial 'self' or even 'conscious energy field' (for example Herbert, 1993; Hameroff, 1994; Penrose, 1989; Eccles, 1994). The quantum approach is supposed to solve problems like the apparently nonalgorithmic nature of thought, free will, the coherence of conscious experience, telepathy, telekinesis, the immortality of the soul and others. From an engineering point of view even the most practical propositions of the quantum approach are presently highly impractical in terms of actual implementation. Then there are some proposals that are hardly distinguishable from wishful fabrications of fairy tales. Here the quantum approach is not pursued.

The cognitive approach maintains that conscious machines can be built because one example already exists, namely the human brain. Therefore a cognitive machine should emulate the cognitive processes of the brain and mind, instead of merely trying to reproduce the results of the thinking processes. Accordingly the results of neurosciences and cognitive psychology should be evaluated and implemented in the design if deemed essential. However, this approach does not necessarily involve the simulation or emulation of the biological neuron as such, instead, what is to be produced is the abstracted information processing function of the neuron.

A cognitive machine would be an embodied physical entity that would interact with the environment. Cognitive robots would be obvious applications of machine cognition and there have been some early attempts towards that direction. Holland seeks to provide robots with some kind of consciousness via internal models (Holland and Goodman, 2003; Holland, 2004). Kawamura has been developing a cognitive robot with a sense of self (Kawamura, 2005; Kawamura *et al.*, 2005). There are also others. Grand presents an experimentalist's approach towards cognitive robots in his book (Grand, 2003).

A cognitive machine would be a complete system with processes like perception, attention, inner speech, imagination, emotions as well as pain and pleasure. Various

technical approaches can be envisioned, namely indirect ones with programs, hybrid systems that combine programs and neural networks, and direct ones that are based on dedicated neural cognitive architectures. The operation of these dedicated neural cognitive architectures would combine neural, symbolic and dynamic elements. However, the neural elements here would not be those of the traditional neural networks; no statistical learning with thousands of examples would be implied, no backpropagation or other weight-adjusting algorithms are used. Instead the networks would be associative in a way that allows the symbolic use of the neural signal arrays (vectors). The 'symbolic' here does not refer to the meaning-free symbol manipulation system of AI; instead it refers to the human way of using symbols with meanings. It is assumed that these cognitive machines would eventually be conscious, or at least they would reproduce most of the folk psychology hallmarks of consciousness (Haikonen, 2003a, 2005a). The engineering aspects of the direct cognitive approach are pursued in this book.

1.3 THE APPROACH OF THIS BOOK

This book outlines an engineering approach towards cognitive and conscious machines. These machines would perceive the world directly, as objects and properties out there, and integrate these percepts seamlessly into their cognitive processes. They would have the flow of inner speech and imagery. They would observe and introspect their mental content and perceive this mental content as immaterial. They would judge the world and their own actions by emotional good–bad criteria. They would be self-motivated agents with a will of their own. They would be able to move and execute actions as freely and readily as humans do. They would produce the hallmarks of consciousness.

The requirements for the direct perception of the world and the seamless integration would seem to call for a special way of perception and information representation. Here the traditional way of information representation of the artificial intelligence, the representation of information as statements in a formal language, is rejected. Instead, associative nonnumeric neural networks, distributed representations and cognitive architectures are investigated as the solution.

The book chapters proceed in logical order and build on previously presented matter; thus they should be studied sequentially. Some matters may appear trivial to advanced readers. However, in that case the reader should pay attention as apparently familiar matters may also have novel twists and meanings in this context. The last chapter, 'Machine Consciousness', should also be accessible to enlightened nonengineer readers who are interested in machine consciousness research. The contents of the book chapters are summarized here.

Chapter 1, 'Introduction', is the introduction in the present chapter.

Chapter 2, 'Information, Meaning and Representation', notes that humans operate with meanings while computers normally do not. It is argued that the brain is not a digital computer and therefore a cognitive robot should not be one either. Instead of digital representation of information a nonnumeric information representation by

large signal arrays, vectors, is proposed. This method is used in the approach of this book.

Chapter 3, 'Associative Neural Networks', presents neural networks that can be used for the associative processing of signal vectors. With these networks vectors can be associated with each other and can be evoked by each other. The neural network that executes this operation is called an associator. The limitations of the traditional neural associator are pointed out and enhanced associators that remedy these limitations are introduced. These associators are based on artificial associative neurons and here one specific execution of these, the Haikonen associative neuron, is presented. The proposed associators are not statistical neural networks and do not utilize learning algorithms such as 'backpropagation'.

Chapter 4, 'Circuit Assemblies', presents some example realizations and definitions for the basic circuit assemblies that are used as building blocks in the systems that are introduced later on in this book. These circuits include the basic associative neuron group, a 'Winner-Takes-All' circuit, an 'Accept-and-Hold' circuit, associative predictors and sequencers, timed sequence circuits and others.

Chapter 5, 'Machine Perception', describes the principles and circuits that would allow direct perception of the world, apparently as such in the machine. The associative neuron-based perception/response feedback loop is introduced as the fundamental building block that is applicable to every sensory modality, such as kinesthetic, haptic, visual and auditory perception. The purpose of this chapter is not to give an exhaustive treatment on the issues of perception, which would be many; instead the purpose is to present the basic ideas that are necessary for the understanding of the cognitive system that is outlined in the following chapters.

Chapter 6, 'Motor Actions for Robots', describes the integration of motor systems into the associative system. According to this approach a robot will be able to execute motor actions in the directly perceived environment without any numeric computations. Hierarchical control loops allow the planning and evocation of action in associative terms and the control of the actual execution in motor control terms. Simple control examples are described.

Chapter 7, 'Machine Cognition', argues that reactive processes are not sufficient for an autonomous robot and deliberative processes are needed. These processes call for higher-level cognition and the use of mental models. The issues of understanding, memories, perception of time, imagination and reasoning in the cognitive machine are considered.

Chapter 8, 'Machine Emotions', describes how the functional equivalents of pain, pleasure, good and bad can be evolved into a system of values, motivations, attention and learning control, and emotional decision making. The Haikonen System Reactions Theory of Emotions is applied to the cognitive machine.

Chapter 9, 'Natural Language in Robot Brains', outlines how natural language can be used and understood by the machine using the associative neural networks. The author's 'multimodal model of language' that seamlessly integrates language, multisensory perception and motor actions is utilized here. The grounding of word meaning is explained. The robot utilizes natural language also as 'inner

speech', which is a running self-commentary of the moment-to-moment situation of the robot.

Chapter 10, 'A Cognitive Architecture for Robot Brains', summarizes the presented principles in an outline for the system architecture of the complete cognitive machine, the 'robot brain'. This machine is not a digital computer; instead it is a system based on distributed signal representations that are processed by associative neuron groups and additional circuits. This system is embodied, it has sensors and motor systems and it utilizes meanings that are grounded to the percepts of environment and the system itself, both the body and the 'mental content'. This system is also self-motivated with values.

Chapter 11, 'Machine Consciousness', summarizes the essential issues of machine consciousness and explains how the approach of this book may produce machines that can be called conscious. The immateriality of mind, the reportability aspect of consciousness, the internal interaction aspect of consciousness, qualia, self-consciousness, free will, testing of consciousness and other issues are considered. It is argued that the presented approach can produce machines that exhibit most, if not all, of the folk psychology hallmarks of consciousness. The chapter concludes with some legal and moral issues that may arise when cognitive robots are introduced.

Please note that the circuit diagrams and other construction examples in this book are presented only for illustrative purposes; they are not intended to be taken as actual product construction guidance. Anyone trying to replicate these circuits and systems should consider the safety and other requirements of the specific application.