The Concept of Intelligence in AI

Ipke Wachsmuth
AG Knowledge-Based Systems (AI)
Faculty of Technology, University of Bielefeld
ipke@techfak.uni-bielefeld.de

Introduction

Prerational intelligence is a new theme tackled by a year-long work of a research group at the Center for Interdisciplinary Research. It assumes that there is something like *rational* intelligence. While examples related to prerational intelligence include most striking yet simple neuronal mechanisms that give rise to astoundingly complex behavior – such as the functioning of the digestive system of a lobster – some behaviors related to human intelligence seem of a distinct quality.

Natural for a human, thinking appears very different from other natural doing like digesting, for example. We cannot be smart or dumb at digesting. Neither do we have a choice of doing it in a conservative or unusual manner, nor is it done deliberately or at will. Many philosophers and psychologists have attempted to find models that view thinking as similar to other inner mechanisms like digesting. But none have arrived at a point satisfactory enough to prevent other researchers from finding their own approaches, including those that contribute to the field of artificial intelligence.

Perception – Reasoning – Action

There have been many ways to define the subject of artificial intelligence. I choose to start with a definition given by Patrick Henry Winston (1992, p. 5):

> "Artificial Intelligence is the study of the computations
> that make it possible to perceive, reason, and act."

Assuming this definition, AI differs from most of psychology because of its greater emphasis on computation, and it differs from most of computer science because of its greater emphasis on perception,

reasoning, and action. One of my points will be that the "reasoning" is particularly essential for higher intelligent functioning. Reasoning involves internal processes that make a subject "think" about what might be the best way of acting before actually moving to act. Cognition about world and *alternate* ways of acting in the world add a quality able to differentiate between a smarter or a dumber individual.

The scientific goal of artificial intelligence is to determine which ideas about representing knowledge, using knowledge, and assembling systems explain various sorts of intelligence (Winston, 1992). Ideas prevalent to the various theoretic approaches to intelligence in AI include that intelligence emerges from the interaction of many simple processes "in concert," and that process models of intelligent behavior can be studied in detail by using the computer.

In this paper, I will try to deal with the question of what concept of intelligence evolves from the work in AI. I will focus on two central paradigms that have found wide recognition, and I will keep short to point out their most significant points. Before doing so, I will give a brief historic account of the origins of the field of artificial intelligence.

A Generative Theory of Intelligence

"Artificial intelligence" is a synthetic term which – due to its suggestive potential – has caused many misunderstandings and false expectations. Its origin can be traced back to the year 1956. This year was important in many respects. For example, the book "Automata Studies" came out, compiling now famous articles in the field of cybernetics (Shannon & McCarthy, 1956). It was also the year when Bardeen, Shockley, and Brattain received the Nobel Prize for the transistor. Noam Chomsky was about to publish his famous paper on syntactic structures, giving way to a theoretic account of natural language (Chomsky, 1957). In 1956, the *Dartmouth Conference* took place in Hanover, New Hampshire. Its promotors were John McCarthy, at that time an assistant professor of mathematics at Dartmouth College, Marvin Minsky, a Harvard junior fellow in mathematics and neurology, Nathaniel Rochester, a manager in information research at IBM Poughkeepsie, NY, and Claude Shannon, who at that time was a mathematician at Bell Telephone Lab. The key sentences in their proposal to the Rockefeller Foundation read as follows (*cf.* McCorduck, 1979, p.93):

> "We propose that a two-month, ten-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

Among the participants of the Dartmouth Conference were Allen Newell of Rand corporation, and Herbert Simon, of Carnegie Tech. Together with John Shaw, they had just completed their "Logic Theorist," a program which could prove mathematical theorems in Whitehead and Russell's Principia Mathematica. This program embodied what one came to call the information-processing approach of modeling. This approach has the underlying position that theories of human voluntary behavior are to be sought in the realm of information processing systems – systems consisting of memories, processors, and control structures, and which work on data structures. The central agreement among researchers starting from this line is that, with respect to intelligent behavior, a human is this kind of a system, being active, autonomous, rule-governed, discrete, with limited structural and resource capabilities.

As a field of academic study, AI reaches to understand intelligence by becoming able to produce effects of intelligence: intelligent behavior. One element in AI's methodology is that progress is sought by building systems that perform: synthesis before analysis. Drastically, it is *not* the aim of AI to build intelligent machines having understood natural intelligence, *but* to understand natural intelligence by building intelligent machines.

Symbolic Representation

It is one special feature of AI that it pays explicit attention to internal symbolic representation and symbol manipulation as a basis of internal processes referred to as "thinking." In this perspective, AI went beyond what was so far the main subject of information processing psychology. While Neisser's (1967) book was seen as a crystallizing point for a new *cognitive* psychology by many researchers, it dealt only with perception and basic processing and ignored higher mental functions such as problem solving, concept formation, or planning. In their work for the

Logic Theorist, Simon and Newell began to perceive what they later called the symbolic-functioning capability of computers. Symbols were conceived as signifying objects with access to meanings – designations, denotations, and descriptions. The symbolic level, as represented in the early work of Newell, Shaw, and Simon (1958) and also Bruner, Goodnow, and Austin (1956), provides notions of plans, programs, procedures, strategies; it also leans on views of rule-governed generative systems (Chomsky, 1957).

Most importantly, the symbolic level was seen to allow studying mind with its functional features and capabilities apart from regarding neural architecture and its processes. As Minsky and Papert (1972) wrote in their AI memo No. 252, "thinking is based on the use of *symbolic descriptions* and description-manipulating processes to represent a variety of kinds of *knowledge* [...] control of the problem-solving process is affected by heuristics that depend on the meaning of events." So for researchers in AI, brain, memory, or recall processes are not the subject of study, but rather the meaning that can be associated with a certain process through symbolic descriptions.

The Knowledge Paradigm

A central and widely recognized paradigm of AI is Newell and Simon's (1972) description of the general intelligent agent. They take an abstract (functional) view of the memory possessions of an individual and its ability to build on it in affecting the world, and they use the term *knowledge* to refer to this functional quality. The agent has sensors to perceive input from a task environment, and actuators to affect the outside world (*cf.* Figure 1). Specific to this view is that "Probehandeln" is possible within the agent: Before acting on the world, an internal representation is manipulated to observe the effect of alternative methods. Methods are selected from an internal method store, and their exploration is guided by general world knowledge also internally inspectable.

The questions leading particularly the work of Newell (1981) in the early 80s were the following:

- What is the nature of knowledge?
- How is it related to representation?
- What is it a system has, when it has knowledge?

The Knowledge Level Hypothesis was put forward in Newell's plenary address at the First National Conference on Artificial Intelligence in Stanford, 1980 (*cf.* Newell, 1981). It assumes a distinct computer systems level, above the program symbol level (and other levels like the register-transfer level, logic circuit level, circuit level, and device level), which is characterized by knowledge as the medium. Representations exist at the symbol level, namely, data structures and processes that realize a body of knowledge at the knowledge level. The connection between knowledge and intelligent behavior is described by the Principle of Rationality which states: *If an agent has knowledge that one of its actions will lead to one of its goals, then the agent will select that action.*

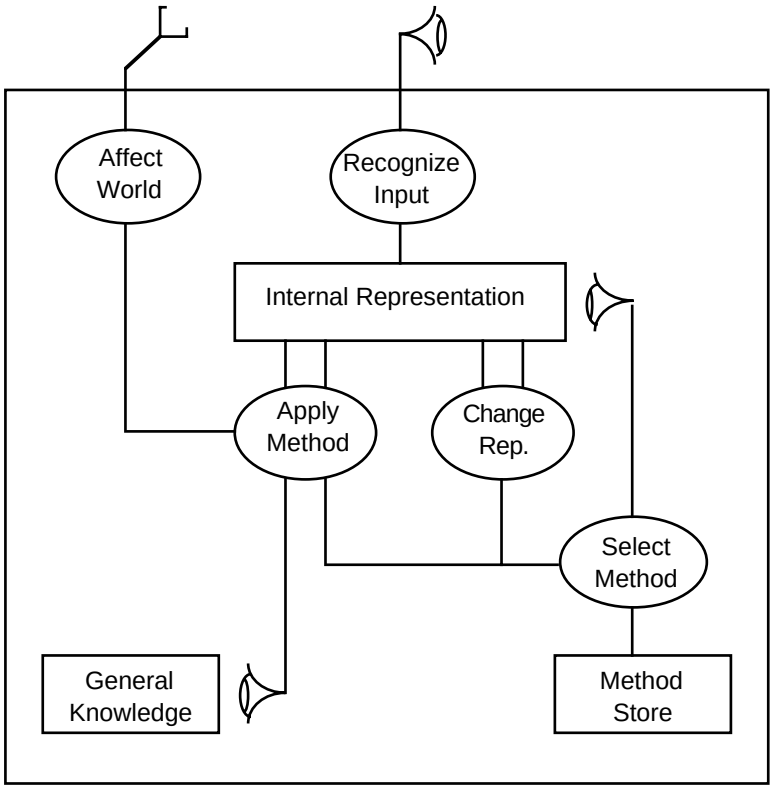

Fig. 1 Functional diagram of general intelligent agent (after Newell & Simon 1972, p.89; Newell 1981, p.2)

Newell views as *knowledge* whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality. From this perspective, knowledge serves as the specification of what a symbol structure should be able to do. More importantly, this conception views knowledge as a competence – a potential for generating action – and as an abstract quality that can never be actually *in hand.* According to Newell, it must be coupled with a symbol-level representation and symbol-manipulating processes to become workable. Newell and Simon (1972) assume a physical symbol system to be present in any intelligent agent.

One of the main points in Newell's approach is that *logic* is a fundamental tool for analysis at the knowledge level, and that logic formalisms with theorem-proving procedures can be used as a representation in an intelligent agent. The knowledge-level view in AI is, so to say, an attempt to mathematize certain aspects of intelligence (apart from considerations of its symbol-level realization), particularly those that deal with rational behavior and logical reasoning in problem solving. Logic-based formalisms have been used for expressing an explicit set of beliefs for a rational agent. Such a set of beliefs – expressed in some representation language – is what is typically meant by the term *knowledge base*.

The logic-oriented (knowledge level) view has helped to clarify and settle many debates that were going on up to the late seventies (Brachman, 1979). For example, different ways of describing knowledge such as semantic networks, frames structures, slot-assertion notation, and slot-and-filler notation were found to be notational variants with respect to what can be expressed and what can be inferred (Charniak & McDermott, 1985). The prominence these alternative notations still have in many application fields derives from their "object-centeredness" which gives the convenience of indexing knowledge by the entities that the knowledge is about, and the whole area of object-oriented programming grew up with it.

Deficiencies of a Purely Logical View

The attempt to reconstruct knowledge of the world in a set of logical formulas (first-order predicate logic, in general) faces some crucial problems that have long been known; for instance, when deciding on

the immediate effects of an event, given an open-ended list of ways it might be modified by context (the *Qualification Problem*; *cf.* McCarthy & Hayes, 1969). This fact is illustrated in the following example (Brewka, 1993):

> Turning the key results in starting the car except when
> • the battery is dead
> • the wires are loose
> • somebody has stolen the engine
> • the key breaks off
> • the tank is empty...

Even worse is the situation when specifying what does *not* change when an event occurs (the *Frame Problem*; McCarthy & Hayes, 1969). For example (*cf.* Brewka, 1993), we may want to describe that it is true that Fred is in the kitchen in a present situation "105" and that the color of the kitchen is red. If we consider a new situation resulting from Fred going to the bathroom, we might describe this in the following set of formulas:

```
Holds(in(Fred,kitchen),situation105)
Holds(color(kitchen,red),situation105)
situation106 =
  result(go(Fred,bathroom),situation105)
```

We might have a general formula accounting for the fact that if someone `x` has gone to a place `y` in a situation `s` that it is now true that `x` is in `y`:

```
forall x,y,s.Holds(in(x,y),result(go(x,y),s))
```

Now it is possible to derive that Fred is in the bathroom in situation 106. But what is now the color of the kitchen? There is no way to derive that the color of the kitchen is still red in situation 106. If we want to be able to do so, we need to write a "frame axiom" that assures so:

```
forall x,y,v,w,s.Holds(color(x,y),s)  →
Holds(color(x,y),result(go(v,w),s))
```

Given any more complex world, the work of writing frame axioms will hardly ever be finished.

There are more problems. In predicate logic, we assume certain rules of inference together with a set of axioms that constitute what is assumed to be true in a domain of discourse. And we assume that a reasoning process (theorem prover) has global access to all that. If an agent knows a set of logical expressions $\{L_i\}$, then this is by far no means equivalent to what can – in principle – be inferred from $\{L_i\}$. What a particular person (an intelligent agent) knows, is somehow stored in memory and has to be retrieved within an enormous body of knowledge before it can be reasoned with. Aspects of computational efficiency and of resource-limited processing come in when an action has to be reached under restricted time (Konolige, 1983).

Even when the logical intellect could, in principle, be compared with logical reasoning, further essential differences exist. The "axioms" of an individual are not acquired in one leap but distributed over a long time. Many knowledge items are specific to particular domains of discourse (domain-specific). They are commonly tied to context, and there is hardly a global view on what one knows. The cultural way of passing knowledge on from teacher to student by way of instruction involves that one may take new facts for granted without checking their consistency with the previously known, i.e., there may be inconsistencies. However, consistency of logical expressions is needed for a logical reasoning process to be sound. Apparently, people are able to carry out logical reasoning soundly within a given context.

Partitioned Knowledge

My research in the 80s has focussed particularly on domain specificity, accessibility, and consistency of knowledge. Based on findings from an empirical study with mathematics students lasting more than one year, I came to postulate grouping effects in large knowledge bodies, and principles in how knowledge is accessed. I found strong indication that performance differences between subjects depended not only on the soundness of the rules ("axioms") they were able to apply but also on the way their rules appeared to be accessible in context. Further findings showed that the presence of certain concept words in a situation description may trigger access to specific domain knowledge and further vocabulary associated with this knowledge. Details of the study and its findings are described in (Wachsmuth, 1989).

The findings were elaborated to describe principles of how knowledge elements are *organized* and *accessed* within a large body of knowledge (Wachsmuth, 1989; see also Wachsmuth & Gängler, 1991). Knowledge organization refers to how a collection of knowledge elements can be structured (partially ordered) by way of set containment. This pertains to different degrees of specificity as well as to different aspects in which a specific body of knowledge may branch to extend in competitive sub-bodies. Access conditions describe how knowledge is made available to the processing system and how access is changed in the course of a task situation. By way of partitioned knowledge bases, the approach gives a logical account of competitive and locally consistent knowledge.

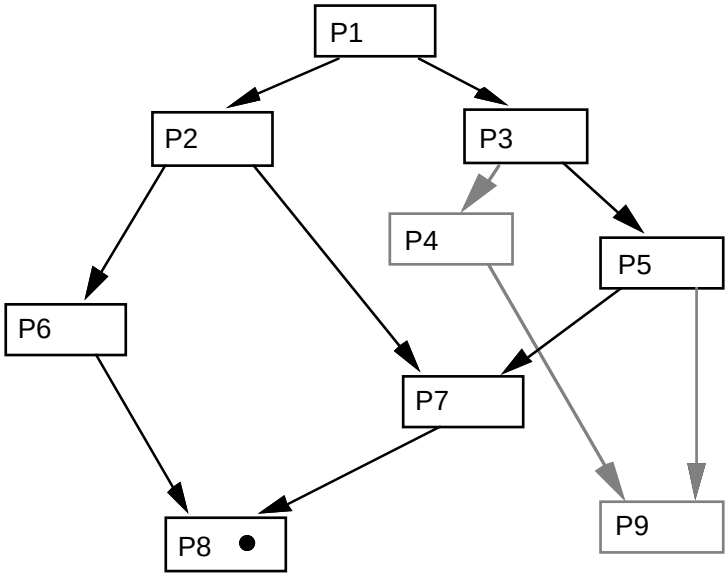

Fig. 2 Access graph of a partitioned knowledge base with competitive axioms. In the access condition shown, only such axioms associated with focussed node P8 and nodes above P8 are eligible for the processing system (from Wachsmuth, 1989, p.157).

In Figure 2, the access graph of a partitioned knowledge base is shown. A node as P1 contains a subset of all expressions in a knowledge base. Lower nodes contain more specific knowledge than the higher ones. Knowledge associated with nodes shown bold is eligible for reasoning.

The collection of knowledge elements eligible in each reachable access condition must be *locally* consistent while the total knowledge base may be inconsistent. The way a body of knowledge is structured embodies a kind of local control in that it restricts the succession in which access to knowledge is attempted in the fulfillment of particular goals.

In general, the conception of partitioned knowledge is meant to explain how one is able to act on the basis of an enormous repertory of knowledge elements without the confusion of facing all of them most of the time. The empirical findings suggest that a critical feature of human intelligence lies in a dynamic partitioning of the total knowledge in *visible* and *invisible* parts such that the visible part is normally small enough to be tractable. By way of extrapolation, this observation can contribute to understanding the general problem-solving ability of human beings, namely, by their ability to access appropriate subbodies of knowledge based on clues from a task situation.

Modularity

In contrast to Newell's principle of rationality which assumes that an agent can take advantage of any and all information at its disposal, Fodor (1983) has called attention to the fact that the mind could have modular subsystems. At least, this seems to be true for perception and motoric action (*cf.* Garfield, 1987). According to Fodor, the mind comprises a number of modular systems dedicated to sensory and linguistic input analysis and to linguistic and motor output. The operation of these systems – for which he assumes specific, dedicated neural architectures – is mandatory, i.e., they perform automatic functions when given triggering stimuli. They are domain-specific in the sense that, for example, dedicated modules operate only upon acoustic signals taken to be utterances and which are different from those which effect the perceptual analysis of auditory nonspeech.

Modular processes are informationally encapsulated in the sense that they have access only to the information represented within the local structures that subserve them. While domain specificity has to do with the circumstances in which a module comes into use, encapsulation has to do with the information that can be mobilized in the course of that use. The speed observed with such processing is, in Fodor's view,

accounted for by the mandatoriness, the domain specificity, and the encapsulation of modules.

Debates in the line of Fodor's (1983) proposal have questioned whether modularity extends to central processes which appear to be optional and deliberate (*cf.* Garfield, 1987). Yet the above discussion of partitioned knowledge raises the question whether intelligent behavior crucially depends on a global view of everything one knows, or whether there could be a kind of functional modularity in reasoning that occurs on the basis of restricted access to knowledge.

The Society of Mind Paradigm

A paradigm deviating from the notion of general intelligent agent has been taken in Minsky's (1986) perspective of "society of mind." It is an attempt to explain intelligence as a combination of many simple processes which he refers to as agents. Agents that work together can perform a task – as an "agency" – without each agent knowing anything about the task; in total, intelligent behavior derives (*cf.* Fig. 3). Special to Minsky's reach for a theory of intelligence is that it tries to span all the way from seemingly senseless simple mechanisms up to higher intelligent functioning. Some agents do not do much more than turn other agents on and off, and resolve conflicts by simply switching among alternatives. Minsky also tries to point out more complex ways for agencies to interact, by describing how agents could use cooperation and compromise.

This paradigm assumes that intelligence is distributed among many interacting smaller systems, thereby adressing the question how intelligence could emerge from nonintelligence. Minsky's themes include very basic questions about agents like function, embodiment, interaction, and competence (How do agents work? What are they made of? How do they communicate? How can groups of agents do what separate agents cannot do?), and they reach up to high-level questions like selfness, meaning, sensibility, and awareness (What gives agents unity or personality? How could they understand anything? How could they have feelings and emotions? How could they be conscious or self-aware?)

Minsky's ideas have influence on the work of a growing number of researchers in AI. Clearly, it leads to a quite different concept of intelligence. Regarding technology, recent attempts to develop larger and more complex knowledge-based systems have revealed shortcomings of centralized, single-agent architectures and have acted as a springboard for research in Distributed AI (DAI; *cf.* Adler et al., 1992; Müller, 1993). Multi-agent systems emphasize the aspect of task-related cooperation of independent (autonomous) agents. An open spectrum of agent types has been considered (Müller & Siekmann, 1991), reaching from primitive (sensor-driven, reactive) agents through to "social agents" with a "conscious" ability of interaction. Higher agents can have knowledge of other agents and their skills. No agent has a global view of the total problem to be solved, that is, there is no central control.

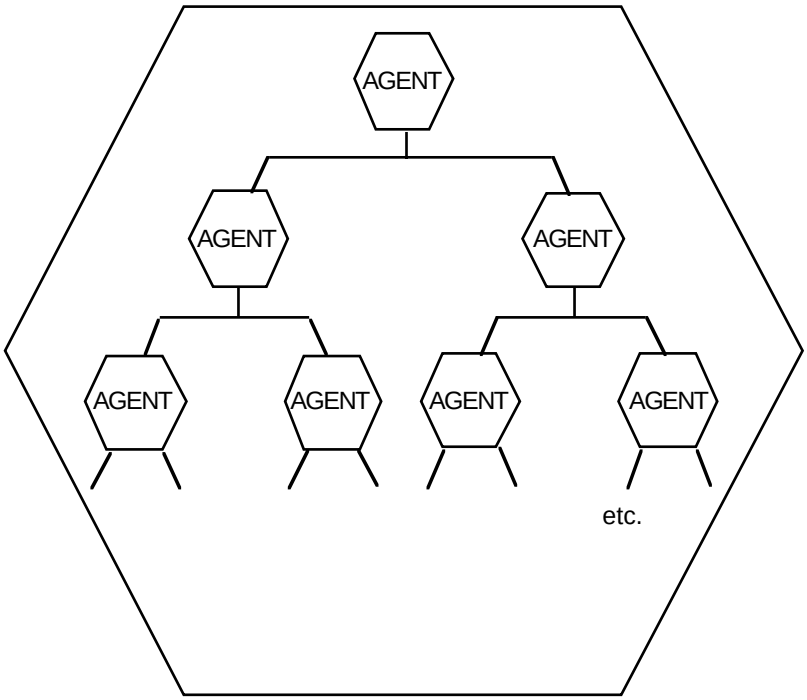

Fig. 3 Seen by itself, an agent is just a simple process that turns other agents on and off. Seen from outside, as an agency, it does whatever all its sub-agents accomplish, using one another's help (after Minsky, 1986).

Situatedness

One of the most crucial problems in AI's previous conceptions of intelligence, as well as in many technical applications, is the fact that world knowledge at one's disposal can never be complete. This is due to contextual variation and the multitude of situations to which an intelligent agent could be exposed. The real problem for an agent is to be able to act and survive in a changing world by way of mapping external input to internal schemata, adapting schemata, and acting without having a full internal description of the outside world.

New lines of AI research have given notice to the fact that the actions of an intelligent agent may decisively depend on its involvement in an actual situation. A situated agent integrates aspects of perception, action, and communication in one system in order to succeed in a situation without having a complete model of it (*cf.* Brooks, 1991). The term "situatedness" refers to the ability of an intelligent system to exploit the actual situation, to the extent possible, as a source of information in perceiving and manipulating its environment and communicating with cooperating partners. And it is crucial for Situated AI to deal with embodied systems that are able to modify their internal processing while they are coupled to their environment by way of sensors and actuators.

Some Concluding Remarks

In this paper, I tried to focus on some of the main ideas that the field of Artificial Intelligence has to offer about the concept of intelligence. Special to AI is that it views intelligence as computational, as based on the processing of information by means of symbols. Computational models of intelligent behavior allow the study of the implications of theoretic assumptions through experimentation. That is, AI theories are generative in the sense that they seek synthesis before analysis. Due to the early subjects in AI which placed a heavy focus on problem-solving and reasoning, *rational* intelligence was mainly the subject of study. Accordingly, an extremely rational perspective was taken by Newell and others building on the knowledge level hypothesis.

It is also clear that the concept of intelligence in AI is grasped in different ways by different authors, and that the field is now in a process

of changing paradigms. There was a recent move to contrast a "global" account of intelligent behavior with more simple interacting systems that may have different internal representations (*or no representation at all*) in the field of Multi-Agent Systems and Distributed AI. This work has extended to the study of embodied systems which are coupled in their environment by way of sensors and actuators in Situated AI.

So far, no single approach has offered a perspective for reproducing or explaining all features of intelligence as it was set for a program of research at the Dartmouth conference (*cf.* McCorduck, 1979). The 1993 november issue of Scientific American quotes Minsky stating that "the mind is a tractor-trailor, rolling on many wheels, but AI workers keep designing unicycles." While there is evidence that more and different "wheels" are presently under consideration, many questions remain to be solved.

A core question asks where prerational, adaptive intelligence leads into rational, reflective intelligence. Human information processing can be inflexible and automatic as well as flexible and controlled. Presumably, it is largely by the use of symbols that we achieve voluntary control over our thoughts. Thus it seems sensible to keep with the insights found with rational intelligence and knowledge level abstraction, and observe progress in the understanding of prerational intelligence. It is my impression that models of multiple agents can help to find insights at the borderline.

References

Adler, M., Durfee, E., Huhns, M., Punch, W. & Simoudis, E. (1992). AAAI Workshop on Cooperation Among Heterogeneous Intelligent Agents. *AI Magazine*, Vol 13 (2), 39-42.

Brachman, R.J. (1979). On the epistemological status of semantic networks. In N.V. Findler (ed.): *Associative Networks: Representation and Use of Knowledge by Computers* (pp.3-50). New York: Academic Press.

Brewka, G. (1993). Nichtmonotones Schließen. In G. Görz (ed.): *Einführung in die künstliche Intelligenz* (pp. 55-85). Bonn: Addison-Wesley.

Brooks, R.A. (1991). Intelligence without reason. *Proceedings IJCAI-91*, 569-595.

Bruner, J.S., Goodnow, J.J. & Austin, G.A. (1956). *A study of thinking.* New York: Wiley.

Charniak, E. & McDermott, D. (1985). *Introduction to artificial intelligence*. Reading, MA: Addison-Wesley.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Feigenbaum, E.A. & Feldmann, J. (1963). *Computers and Thought*. New York: McGraw-Hill.

Fodor, J. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Garfield, J.L. (ed.) (1987). *Modularity in knowledge representation and natural language understanding.* Cambridge, MA: MIT Press.

Konolige, K. (1983). A deductive model of belief. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (pp. 377-381). Karlsruhe.

McCarthy, J. & Hayes. P.J. (1969). Some philosophical problems from the standpoint of artificial intelligence. In D. Michie & B. Meltzer (eds.): *Machine intelligence 4*. Edinburgh: Edinburgh University Press, 463-502.

McCorduck, P. (1979). *Machines Who Think.* San Francisco: Freeman.

Minsky, M. & Papert, S. (1972). MIT AI memo 252.

Minsky, M.L. (1986). *The Society of Mind*. New York: Simon & Schuster, Inc.

Müller, J. (1993) (ed.). *Verteilte Künstliche Intelligenz – Methoden und Anwendungen*. Mannheim: BI Wissenschaftsverlag.

Müller, J. and Siekmann, J. (1991). Structured social agents. In W. Brauer and D. Hernández (eds.): *Verteilte Künstliche Intelligenz und kooperatives Arbeiten.* 4. Internationaler GI-Kongreß Wissensbasierte Systeme. Heidelberg: Springer.

Neisser, U. (1967). Cognitive Psychology. New York: Appleton – Century – Crofts.

Newell, A. (1981). The Knowledge Level. *AI Magazine 2*(2), 1-20. Republished 1982 in *Artificial Intelligence 18*(1), 1-20.

Newell, A. & Simon, H.A. (1956). The Logic Theory Machine. *IRE Transactions on Information Theory*, September. Reprinted in (Feigenbaum & Feldmann, 1963).

Newell, A., Shaw, J.C. & Simon, H.A. (1958). Chess playing programs and the problem of complexity. *IBM Journal of Research and Development 2*(4). Reprinted in (Feigenbaum & Feldmann, 1963).

Newell, A. & Simon, H.A. (1972). *Human Problem Solving*.

Shannon, C.E. & McCarthy, J. (1956). *Automata Studies*. Princeton, NJ: Princeton University Press (Annals of Mathematics Studies No. 34).

Winston, P.H. (1992). *Artificial Intelligence* (3rd edition). Reading, MA: Addison-Wesley.

Wachsmuth, I. (1989). *Zur intelligenten Organisation von Wissensbeständen in künstlichen Systemen*. Stuttgart/Heidelberg: IBM Deutschland Wissenschaftliches Zentrum (IWBS Report 91).

Wachsmuth, I. & Gängler, B. (1991). Knowledge Packets and Knowledge Packet Structures. In O. Herzog & C.-R.Rollinger (eds.): *Text Understanding in LILOG: Integrating Computational Linguistics and Artificial Intelligence* (pp. 380-393). Berlin Heidelberg: Springer (LNAI 546), 1991.