Mss original dated February 28, 1993, revised 8/14/95

Separability Hypothesis[1]

B. Chandrasekaran & Susan G. Josephson[2]
Laboratory for AI Research
The Ohio State University
Columbus, OH 43210

Abstract. In this paper we argue for a Separability Hypothesis as a working hypothesis for AI. There is no a priori equivalence of the terms "mind" and "cognition." Phenomenologically, a mental state has a cognitive content as well as subjectivity and emotions. The Separability Hypothesis asserts that an architecture for intelligence is a homomorphism of the architecture for mind. None of the technical problems that AI is working on today require machines which have anything other than a cognitive content to them. We discuss how this Hypothesis protects AI from a number of its critics, and lets the field get on with its technical concerns.

Is Intelligence Separable from other Mental Phenomena?

In both cognitive science and AI, the terms "mind" and "cognition" (or "intelligence") are generally presumed to refer to one and the same thing. This is not an explicit doctrine of the fields in question, but more of a tacit assumption. Thus, both people within and outside AI take the questions, "Can machines have a mind?" and "Can machines be intelligent?" to be equivalent questions. A point of this paper is to argue that there is no a priori identity of these things. Both methodologically and philosophically it

---

[1] Parts of this paper originally appeared in B. Chandrasekaran and S.G. Josephson, "Architecture of Intelligence: The Problem and Current Approaches to Solutions," *Current Science*, Vol. 64, No. 6, March 1993, pp. 366 - 380, and in *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*, V. Honavar and L. Uhr, editors, Academic Press, pp. 21-50, 1994. The current paper is a much revised version of the earlier discussions of this topic.

[2] Also with Columbus College of Art and Design.

is better to assume some separability between intelligence and the totality of mental phenomena. Further, with this separability as a working hypothesis, technical work in the disciplines of AI and cognitive science can proceed independent of the fate of at least some of the anti-computationalist claims.

Separability Hypothesis

Intelligence and cognition are terms that are associated with the study of thought processes. Basic to this study is the idea of an agent as being in some knowledge state[3], that is, having thoughts, beliefs. Correspondingly, the underlying process of cognition, thinking, is viewed as something that changes these knowledge states.

This sort of knowledge-state talk is meant to be a theory-neutral way to talk about thinking. It is simply a statement about the phenomenology of having thoughts, without committing ourselves to any particular theory as to how those thoughts come about. For example, one sort of theory we can propose about thinking is a computational theory, which posits that the engine of thought is a computational engine of the Turing machine family. However, we could also hypothesize that knowledge states and the changes from one to the next are the result of a purely physical process with no information processing involved. Describing thinking as being in and changing knowledge states does not depend upon there being a representational content literally being processed, but only our being able to consistently *ascribe* representational content to the thoughts that cognitive agents have.

However, besides these knowledge states, mental phenomena, as commonly construed, also include such things as emotional states and subjective consciousness. Under what conditions can these other mental properties also be attributed to the artifacts to which we attribute knowledge states? Is intelligence separable from these other mental phenomena? That is, is it possible that intelligence can be explained or simulated without necessarily also explaining or simulating other aspects of mind? A somewhat formal way of putting this Separability Hypothesis is

---

[3] Actually *belief* states. For our purposes, though, this distinction is unimportant.

that the knowledge state transformation account can be factored off as a homomorphism of the mental process account. That is:

> If the mental process can be seen as a sequence of transformations: $M_1 \rightarrow M_2 \rightarrow ...$, where M_i is the complete mental state, and the transformation function (the function that is responsible for state changes) is F, then a subprocess $K_1 \rightarrow K_2 \rightarrow ..$ can be identified such that each K_i is a knowledge state and a component of the corresponding M_i, the transformation function is f, and f is some kind of homomorphism of F.

A study of intelligence alone can restrict itself to a characterization of K's and f, without producing accounts of M's and F. If cognition is in fact separable in this sense, we can in principle design machines that implement f and whose states are interpretable as K's. We can call such machines cognitive agents, and attribute intelligence to them. However, the states of such machines are not necessarily interpretable as complete M's, and thus they may be denied other attributes of mental states.

We have not seen the Separability Hypothesis explicitly recognized or stated as we have done here. The identity of “mind” and “cognition” is an unexamined underlying assumption in and outside the field. Where the issue has been recognized in some form, it has usually been in response to questions such as “Can computers feel pain or anger?” and “Can computers understand?”. People on either side of this question assume that the question is important to settle the fate of the computationalist proposal.

Arguments against Separability Hypothesis

There are two categories of opposition to the Separability Hypothesis. The first category is from the pro-AI camp. The second category is from people who disbelieve in the computationalist approach for mind.

Most people in AI would probably oppose the Separability Hypothesis, because of the widespread belief in the community that computation is sufficient for all aspects of mentality. We have been able to identify three arguments for this position.

The first argument from the pro-AI camp is the claim that emotions and subjectivity are simply “emergent” properties of the appropriate kind of

computer. Intelligence is inseparable from mind because the other aspects of mind are simply emergent properties of certain kinds of complex agents with knowledge states. Once a complex enough agent with knowledge states is built, subjectivity and emotional states will just arise without anything extra being added. If this is the case, the knowledge state account, and with it an account in terms of information processing, will be by itself a sufficient basis for explaining and building minds. Explanation of the phenomena of intelligence and cognition will also turn out to be explanation of the full range of mental phenomena. Artificial agents that can be plausibly interpreted as solving problems, achieving goals, and performing reasoning will also have emotional states and subjective consciousness.

The emergence thesis declares that somehow what appears to be one category, subjectivity of experience, will emerge from another category, computation. Science has of course often shown that phenomena which are in different categories of experience are related: for example, mass and energy, space and time. Nevertheless, without additional arguments or demonstration of the claims of category-jumping emergence, skepticism about the emergence thesis is understandable.

The second argument from the pro-AI camp is the belief that the other aspects of mentality are nothing but information processing, but it acknowledges that we do not at this point know how to design information processors with subjective properties. People who hold this view would agree that their PC's do not have subjective experiences, and even when they print out statements such as, "I am in pain," these people do not really believe that the PC's are in pain. They also do not believe that getting to that point is simply a matter of complexity — they grant that there may be some important additional principle involved that we don't yet know.

We have so far heard the third argument from just one person[4] in the pro-AI camp. Consider the state of being in pain. The argument is that having pain corresponds to the belief state that one is in pain, but it acknowledges that we don't know yet how to organize our information processors to be in such a belief state. But, according to this argument, there is no logical distinction between believing that you are in pain, an

[4] Drew McDermott, personal communication.

information state, and being in pain. Thus, information processing is adequate for explaining other aspects of mentality as well.

In the standard information-processing view, there is usually no distinction made between different types of belief states. A medical diagnosis program which has various elements of knowledge represented in it goes through various hypotheses. As it accepts a hypothesis H as the diagnosis for the case, one simply asserts that the program knows or believes that the patient has the disease H. In this argument, in order for a belief state about emotions to be a real belief state — that is, it actually causes us to believe that we are in pain — something more is being demanded of the representation of belief states. In this instance, what is being demanded, it seems to us, is that the agent have the subjective experience corresponding to being in that belief state. Then it would logically follow that being in the belief state, "I am in pain," entails actually being in pain. If we don't know yet how to make machines be in specific types of belief states, then the problem is not being solved, simply postponed.

One criticism that can be made of this argument is to point out that this argument has the same logical structure as solipsism in philosophy. The reader might remember that solipsism holds that no one exists but ourselves, there is no world external to us, and that all our perceptions about the external world are simply ideas in our mind. While logically irrefutable, solipsism usually loses its attraction to most people after a few weeks.

In any case, a burden of this paper is that AI as a technical discipline does not really need to join this battle. If other components of mental states emerge from information processing, or if we figure out how to make specific types of belief states, so be it, but in the meantime, we have a perfectly well-defined problem of designing a cognitive engine.

The second category of opposition to the Separability Hypothesis from the anti-computationalist camp is based on the argument that there is no coherent way to factor off a knowledge state process account from a mental state process account. That is, aspects of mind like subjectivity and emotion are not emergent from knowledge states. Further, the knowledge state process cannot be fully explained or simulated without also explaining or simulating the whole mind, including these other aspects as well. From this point of view, the categorical difference between different attributes of

mental states is affirmed, but the Separability Hypothesis is denied. We can talk about knowledge components of mental states, but mental processes have no "sub processes" which only have to do with knowledge state transformations. In this view, the only way to explain or build an intelligence is to solve the problem of explaining or building a complete mind. Only agents which have the totality of what we call "mind" will be able to perform as truly successful problem solvers across the whole range of situations deemed to require intelligence. Subjectivity and so on will not emerge from complex problem solving skills, nor can complex cognitive skills be had independent of subjectivity, consciousness and emotion.

With respect to this second set of arguments from the anti-computationalist camp, we know that it is possible to build machines to perform a number of cognitive tasks. We also know that we have not felt any need to argue that these machines have subjective experiences. AI's critics cannot identify specific cognitive tasks which require machines with subjective experience. In fact, Searle, who is a charter member of the "pandiabolikon"[5] of AI, grants as part of the premise of his Chinese Room argument the existence of what we would call a cognitive machine performing a complex task. Until someone comes up with a clear specification of a cognitive task that *requires* a machine with subjective experience to accomplish it, we can proceed with adopting the Separability Hypothesis as a working guide. We elaborate on this point in the next section.

How the Hypothesis Protects AI from its Critics

Searle (1980) holds that a computer program that successfully answers questions in Chinese cannot be said to "understand Chinese," even though it is behaviorally intelligent in this task. In our framework, we interpret Searle as claiming only that the *subjective* property "being in a state of understanding" is beyond computer programs as computer programs. That is, Searle is denying that a computer program, *qua* computer program, can be a complete mind.

Since almost all of the discussion regarding the Chinese Room has revolved around what it means to "understand," let us consider that term in

---

[5] pan-diabolikos (devil) -on, in analogy with pan-theios (god) -on for pantheon.

some detail. Searle considers a person who does not understand Chinese, and asks us to imagine him following rules from a rule-book (i.e., a program) to respond to queries in Chinese. Suppose the rule book is sufficiently good that his behavior in response to queries is identical to that of a person who understands Chinese. Searle regards it as self-evident that the rule-following person still does not understand Chinese. In fact, if we imagine ourselves to be that person, we would, he says, know that we don't understand Chinese. Thus, Searle concludes, the mind is not equivalent to computer hardware running a program. The most popular objection to this claim of Searle has been that it is not the rule-following person who understands Chinese, but the person plus the rule book. We stop at this point in the argument, since our goal is not to revisit the entire Chinese Room debate, but to focus on one aspect of it, namely, how the debate turns on the term "understand."

It seems to us that Searle uses the term to refer to the subjective feeling that we have when we understand something. To Searle, it seems odd for someone else to say that I really do understand Chinese when it is quite clear to me, I *know,* that I don't understand Chinese. For Searle the subjective sense of the term "understand" seems to be the real meaning of "understand." On the other hand, his adversaries in this argument take the meaning of the term to be behavioral. If some system S behaves as if it understands Chinese, well, then it does. Searle is puzzled that his adversaries cannot see something as obvious as the fact that the rule follower does not understand Chinese, because the rule-follower knows, feels, that he does not understand Chinese. Searle's opponents are puzzled that Searle cannot see something as obvious as the fact that understanding is as understanding does, that we have no more grounds on which to say that a native Chinese speaker understands Chinese than we do about a computer which runs a rule-following program and behaves as if it understands Chinese.

None of the goals of AI *as practiced* depend on the claim that computer programs can have subjective properties. The programs for vision or natural language processing or problem solving do not have as their success criterion that they exhibit subjectivity. Nevertheless, many AI people think Searle's arguments are an attack on AI's technical goals, and so Searle needs to be proven wrong. It seems to us that, even without subjective

properties, it would be a wonderful enough thing to have a machine whose states are consistently interpretable as knowledge states, whose state changes can be modeled computationally, and which can intelligently answer questions in Chinese. In such a case, the sufficiency of computationalism for cognition would have been shown. Such an engine may "know" about emotional states and subjective experiences of other agents so that it may take them into account in its planning. It may even follow information processing strategies corresponding to being in various emotional states, if such strategies confer a net benefit in its goal-achieving behavior. We would, however, not claim that it has mental experiences of subjectivity or emotions.

Note that the Separability Hypothesis is not a restatement of the Strong Vs Weak AI distinction made by Searle. The hypothesis proposes that the computational cognitive engine is not just a simulation of the real thing — it is itself an instance of the real thing. However, the cognitive engine is not necessarily the complete mind engine.

Newell and Simon (1976) only talk about intelligence, not mind, when they propose the Physical Symbol Hypothesis (PSSH). Specifically, their notion of intelligence is the intuitive notion we have of it as mental activity oriented towards goal achievement and problem solving. There is no claim that PSSH is going to be sufficient for all aspects of mentality, even though there is no explicit denial of that claim either. Whatever the fate of Searle's arguments about whether computationalism is sufficient for mind as a whole, PSSH is protected from Searle-like claims. In fact, it is worth noting that, in a interview that Newell gave to one of us (Newell, 1993) a few months before he died, he agreed that it was hard to see the primitives for emotion in information processing. Whatever the fate of Searle's arguments about whether computationalism is sufficient for mind as a whole, PSSH is protected from Searle-like claims.

Now Searle and his supporters are likely to raise another objection in response to talk of cognitive engines being in "knowledge" or "information" states. They might say that terms like "knowledge" and "information" are really only applicable when applied to complete minds, and that talk of machines or other such entities being in "knowledge states" or "information states" is to take what started out as a convenient short hand for derived intentionality into a philosophical absurdity. But this is an operationally

inconsequential point. Suppose we grant them their point. Nothing changes about our ability to build such machines. The Separability Hypothesis does not require that the states of the machine be knowledge states *for* the machine. All that we need is that there is an isomorphic mapping between the knowledge states of minds and their state transitions, and the states of our machine and its state transitions. If the intentionality of these machines is declared to be a derived one, fine. If such machines prove mathematical theorems, solve problems, control complex systems, make and execute plans, and, yes, respond to Chinese queries, AI and cognitive science can be very proud sciences nevertheless, since they would have captured an extremely important and impressive set of regularities.

Let us call a task a cognitive task if it can be formulated as one of transforming an information-bearing representation (the initial state) to a desired information-bearing representation (the goal state). Not all cognitive tasks are accomplishable (there may simply be no transformation of any kind from the initial to final representation) or Turing-computable (the transformation may not be computable). However, let us restrict ourselves to the cognitive tasks that are within human capabilities, i.e., that are accomplishable. If the Separability Hypothesis is true and computationalism is correct for the knowledge state machine, then all cognitive tasks which are within human capabilities are also within the capacity of some knowledge state machine, and further the knowledge state machine is a Turing Machine.

If one can demonstrate that there are purely cognitive tasks that are accomplishable by humans, but are beyond the ability of computers, that would be another way of undermining the Separability Hypothesis and of course computationalism. Penrose (1989), another critic of AI, claims to be able to show just this. He argues that human minds are not computer programs. If they were, they would be formal systems which would be limited by Godel's Theorem in their ability to prove certain propositions about themselves. He thinks that human mathematicians, however, would be able to "see" the truth of these propositions. Note that the task of proving propositions is a purely cognitive task in our sense of the term.

What could this argument mean in the framework of our discussion? One possible implication is that the Separability Hypothesis might still be true, but that the knowledge state transformation machine is not a Turing

Machine. The functions that need to be computed by this machine are not Turing-computable. If this implication is correct, the AI project — interpreted not as making a complete mind, but making a Turing Machine-based cognitive agent — will find severe limits to what it can accomplish. We may have to look for other kinds of computation for making cognitive agents, intelligent machines with human capabilities. In fact, Penrose seems to propose that new families of computers that take advantage of quantum-mechanical phenomena may be able to compute functions which are not Turing-computable and could form the basis for implementing cognitive agents.

Penrose's discussion brings together two elements that we treat as distinct. There is the issue of humans being able to compute specific functions that are non-Turing-computable. Then there is the issue of subjective experiences. One way of interpreting Penrose's claim about the human mathematician being able to "see" the truth of a mathematical proposition is that he or she is in a subjective state of understanding, and further that being in such a subjective state is causally necessary in reaching new and relevant knowledge states. Penrose's argument seems to imply that these two issues are related. In his view, human thinking has this subjective character, which further somehow plays a role in the human mathematician being able to compute the non-Turing-computable function. Thus an alternative implication is that, in addition to computationalism being false for the knowledge state machine, the Separability Hypothesis is false as well. In our terms, Penrose would be claiming that we cannot factor out component K, knowledge state, from M, complete mental state. Thinking is only possible in the context of complete mental state, M. Subjective understanding is necessary even for knowledge state transformations of certain kinds. The Separability Hypothesis would then be false.

This is not the place for a detailed consideration of the Penrose's argument. Nevertheless, it is fair to say that his argument that Godel's Theorem implies that humans are capable of arriving at certain mathematical truths which forever remain outside the capabilities of computing machines has not found wide acceptance within the mathematical logic community. Even people who reject the computationalist project, e.g., Putnam, find Penrose's argument less than

compelling. Their main objection is that Penrose has not shown that human mathematicians are not also limited in essentially similar ways as machines, and they reject the claim that humans have any access to truth status of mathematical propositions other than actually generating a proof. In any case, other than deciding to work on quantum computers, an AI researcher cannot draw operationally useful consequences from Penrose's claims. Suppose I am working on a very hard problem, say, face recognition. How am I to decide whether the essential difficulty of the problem arises from the Godelian limitations, or from not having enough good ideas within the traditional AI framework?

Edelman (1987) has argued that information processing is not the appropriate way to talk about cognition. Instead he proposes that the basic mechanisms of the brain are the selection of successful neuronal groups in response to interactions with the world. The processes that underlie this neuronal group selection resemble Darwinian evolutionary processes. He contrasts this with what he takes to be the information-processing view in both cognitive science and AI. The basic mechanisms are group selection mechanisms rather than information-processing mechanisms.

Relating Edelman to the Separability Hypothesis framework is not clear-cut. Edelman's position, on the face of it, appears to constitute a rejection of the Separability Hypothesis since he would resist our talk of knowledge states transformation machine as separable from the mind machine. However, one can argue, as we do in Chandrasekaran and Josephson (1993) that the neuronal circuits that result from the selection processes can be interpreted as representation processing systems and that a knowledge state account can still be given. It is just that his mechanisms are not the algorithms familiar in AI. There is no logical contradiction between a machine being a product of evolutionary selection processes operating over connections between implementation units and the evolved machine being an information processing machine.

In his writings Edelman makes much of his theory's potential to explain consciousness, but the technical work of his group considers problems familiar to AI research: for example, perception and robotics. As we have mentioned several times before, there is no evidence that solutions to these problems *require* other components of mental states such as subjective experience. In this sense, Edelman can be viewed as proposing

mechanisms for the cognitive engine that are different from the ones that AI and cognitive science have traditionally looked at. Perhaps Edelman's research program can benefit from subscribing to the Separability hypothesis as well.

The Separability Hypothesis is a working hypothesis for building intelligent machines. As a practical guide to doing research in AI and cognitive science, the goal of formulating the Separability Hypothesis is to save the energies of the researchers from fighting battles that are irrelevant to the technical problems that we face every day in our research, under the erroneous assumption that the attacks on AI, from Penrose, Searle or Edelman, need to be refuted for the real research goals of AI to be valid. None of the problems that we currently work on in AI depend upon our machines having subjective experiences or pain or pleasure. Building robots that have autonomy in the physical world, computer programs that prove theorems in some area of mathematics, respond appropriately to natural language, to images, to sounds — none of these problems have been *shown* to require the capacity for subjective experiences or having emotional states. If in fact it turns out empirically that some apparently cognitive tasks require the machine to have consciousness, subjectivity, or emotion, then that would be fine, and we can reject the Separability Hypothesis at that point. But as working hypothesis, it would have done its job well. In the meantime, it would have protected AI researchers from battles they need not join. (AI philosophers may, however, feel free to enter the fray, but without implying that the research program depends on slaying the dragons.)

Notwithstanding its possible use as a working hypothesis in the absence of proof to the contrary, why might one believe or disbelieve in Separability? One reason to believe in it is that no convincing argument has been given for why one would need intermediate states that are more than information states for any task that can be defined completely in terms of information representations. This is not to deny that AI has many problems which remain hard after years of work. After years of work, face recognition and multi-speaker continuous speech recognition, for example, remain beyond the grasp of current AI technology. It is certainly plausible that our information-processing frameworks are insufficient. But no reason has been

given to believe that the difficulty comes about as a result of the inability of our machines to have subjective experiences or emotional states.

A final point before we leave this section: Subscribing to the Separability Hypothesis does not mean subscribing to the adequacy of Turing computationalism for the cognitive engine. Any framework that is compatible with representation and its transformation — from logic to connectionism to dynamical systems to frameworks currently unthought of — is a candidate.

Separability of Mind from the Physical Substrate

The separability issue does not end with the relationship between intelligence and mind. We can formulate a similar Separability regarding the relationship between Mind and the physical substrate of the biological brain:

> If the physical brain process can be seen as a sequence of transformations: $P_1 \rightarrow P_2 \rightarrow ...$, where P_i is the complete physical state, and the transformation function (the function that is responsible for state changes) is F_p, then a subprocess $M_1 \rightarrow M_2 \rightarrow ...$ can be identified such that each M_i is a mental state and an abstraction of a part of the corresponding P_i, the transformation function is f_M, and f_M is some homomorphism of F_p.

Of course this separability question is one version of the traditional mind-body problem. AI and Cognitive Science largely take this Separability as a working hypothesis as well, with some well-known exceptions. For example, the work of Brooks (1991) in robotics and Beer (1990) in modeling insect behavior can be interpreted as a rejection of this version of the Separability Hypothesis. In the theory of mind, eliminative materialism would be a member of this camp. The latest arrival in this camp is the claim for dynamical systems (Port and van Gelder,1995) as directly modeling physical states of brains. Whether these alternative proposals are as representation-free as their proposers claim can be debated (Chandrasekaran, et al, 1988; see also Vera and Simon, 1993, and discussion of that paper in the same issue of *Cognitive Science*). Work that purports to show that physical talk is all that is needed and we can eschew all mental talk typically focuses on phenomena closer to motor behavior and simpler problems in perception. These proposals still have

a way to go before generating complex cognitive behavior without any form of representation. Until that is accomplished, both forms of the Separability Hypothesis are practical necessities for AI researchers.

Concluding Remarks

The seed for our formulation of the Separability Hypothesis was planted as a result of the personal response of the first author to the outrage among AI researchers to the criticisms of Searle and Penrose. He was and is an active AI researcher working on the problems of causal understanding and diagrammatic reasoning. For one, he noted that the technical problems that he faced had no connection to the criticisms, i.e., whether or not Searle and Penrose were right or wrong didn't really matter to progress on the specific issues. It was not as if a different technical approach was being proposed to the class of problems he was working on. This perception enabled both of us to be less emotional in response to Searle's arguments in particular, in comparison with most other AI researchers who reacted as if their research programs would have to be consigned to the dustbin of history if Searle were right.

Until we reach technical problems that we can plausibly see as requiring a solution to the problem of how to give our machines subjective experiences, as a technical discipline, we in AI don't have too much to worry about from Searle's critique. It seems to us that by insisting that the current ideas will necessarily lead to machines with subjective experience or that emotions will simply emerge from information processing, AI researchers are making claims that are largely unnecessary for us as a technical field. The Separability Hypothesis is a way, we think, to protect AI from its own ideologues.

The reader will note that we are not giving strong arguments to believe in the truth of the Separability Hypotheses. Doing so would take us back to the kind of debates that we are urging AI researchers to avoid. In fact, the Hypotheses may well be false, but the only way to find out is by accepting them as a working hypotheses and doing the empirical work.

The last issue that we would like to discuss is about how biological we need to be in our approach to the problem of designing intelligent machines. Underlying the Separability Hypothesis is the assumption that there is a coherent phenomenon called cognition or intelligence that we are trying to

capture. But characterizing this phenomenon functionally in such a way that everyone's intuition is satisfied is hard to do. Within AI and cognitive science, different proposals are motivated by tacit but definite differences about how faithful to the biological phenomena we want to be. Doyle (1991) defines the task of an abstract intelligence as ideal rationality. The connection of this view to biology is rather minimal — it could be seen as an idealization of some aspects of human cognitive behavior. Newell (1991) defines the task as one of being able to achieve goals by acquiring and applying knowledge. This is an abstract formulation, but his proposed solution, Soar, is based on an abstraction from human deliberation, a step closer to biology, but he is willing to idealize memory as an abstract perfect storage. Schank (1982) characterizes intelligence essentially as memory organization, with the defaults and other computational tricks associated with memory taking center stage. This view is even closer to the biological instantiation of intelligence. Connectionists wish to include phenomena such as smooth concept learning and graded responses, and also things that are common to both higher animals and humans, getting closer to the biological version. Closer still is Beer (1990) who wants to start with modeling how insects move around. We have come quite a distance from Doyle's ideal rationality to Beer's insect motion, all with the ostensible goal of building an intelligence!

This variety in functional characterization is due to different ways of carving out the phenomena — placing this in the box for idealized intelligence and placing that in the box for incidental features of biological implementation. Depending upon how this carving is done, rather different machines might be built. Instead of separating one cognitive machine, we can separate many from the mind, or even more generally from the mind's physical substrate. It is unlikely that there is one architectural level that will explain only and all of cognition. We may have to settle for different machines corresponding to different idealizations. But that should not be a problem. Each would be a legitimate cognitive machine.

Acknowledgments

We thank John Josephson for discussions on the topic and comments on the drafts.

References

Beer, R.D. (1990). *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology*, Academic Press.

Brooks, R. A. (1991). “Intelligence without representation,” *Artificial Intelligence*, **47**(1-3), 139-159.

Chandrasekaran, B., Goel, A., and Allemang, D. (1989). "Connectionism and Information Processing Abstractions: the Message Still Counts More Than The Medium," *AI Magazine*, **9**:4, pp. 24-34, Winter 1989. An abbreviated version with the same title appears in *Behavioral and Brain Sciences*, **11**:1, 26-27, 1988.

Chandrasekaran, B. and Josephson, S.G. (1993) "Architecture of Intelligence: The Problem and Current Approaches to Solutions," *Current Science*, Vol. 64, No. 6, March 1993, pp. 366 - 380. A revised version appears in appears also in *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*, V. Honavar and L. Uhr, editors, Academic Press, pp. 21-50, 1994.

Doyle, J. (1991). “ The Foundations of Psychology: A Logico-Computational Inquiry into the Concept of Mind” *Philosophy and AI, Essays At the Interface*, edited by Robert Cummins and John Pollock, MIT Press, Cambridge, Mass, 1991, pg. 39-78.

Edelman, G. M. (1987). *Neural Darwinism: The Theory of Neuronal Group Selection*. New York: Basic Books.

Edelman, G. M. (1989). *The Remembered Present: A Biological Theory of Consciousness*. New York: Basic Books.

McCarthy, J. and Hayes, P. J. (1969). “Some philosophical problems from the standpoint of artificial intelligence”. *Machine Intelligence*, **6**, 133-153.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*. Oxford University Press.

Port, R. and van Gelder, T. (1995), editors, *Mind as Motion: Explorations in the Dynamics of Cognition*, MIT Press.

Rumelhart, D. E., J. L. McClelland and the PDP research group, eds (1986). *Parallel Distributed Processing: Essays in the Microstructure of Cognition, Vol. I, Foundations*. Cambridge, MA: MIT Press/Bradford Books.

Schank, R. C. (1982). *Dynamic memory: A Theory of Reminding and Learning in Computers and People*, New York: Cambridge University Press.

Searle, J. R. (1980). "Minds, brains, and programs", *Behavioral and Brain Sciences*, **3**, 417-424.

Vera, A. H., and Simon, H. A. (1993). "Situated action: A symbolic interpretation," *Cognitive Science*, **17**, 1, 7-48. Discussion of this paper and the authors' reply appear in the same issue.