Architecture of intelligence: The problems and current approaches to solutions

B. Chandrasekaran and Susan G. Josephson
Laboratory for AI Research, The Ohio State University, Columbus, OH 43210, USA

We propose as a working hypothesis a Separability Hypothesis which posits that one can factor off an architecture for cognition from a more general architecture for mind, thus avoiding a number of philosophical objections that have been raised about the 'strong AI' hypothesis. Using a coin-sorting machine as an example, we discuss a range of positions on representations and argue that, for many purposes, the same body of matter can be interpreted as bearing different representational formalisms. We then propose that one way to understand the diversity of architectural theories is to make a distinction between deliberative and subdeliberative architectures. The search for *one* architectural level which will explain all the interesting phenomena of cognition is likely to be futile. There are a number of levels that interact, and this interaction makes explanation in terms of one level quite incomplete.

Dimensions for thinking about thinking

A major problem in the study of intelligence and cognition is the range of—often implicit—assumptions about what phenomena these terms are meant to cover. Are we just talking about cognition as having and using knowledge, or are we also talking about other mental states such as emotions and subjective awareness? Are we talking about intelligence as an abstract set of capacities, or as a set of biological phenomena? These two questions set up two dimensions of discussion about intelligence. After we discuss these dimensions we will discuss information processing, representation, and cognitive architectures.

Dimension 1. Is intelligence separable from other mental phenomena?

When people think of intelligence and cognition, they often think of an agent being in some knowledge state, that is, having thoughts, beliefs. They also think of the underlying process of cognition as something that changes knowledge states. Since knowledge states are particular types of information states the underlying process is thought of as information processing. (We will discuss this in more detail later in the paper.) However, besides these knowledge states, mental phenomena also include such things as emotional states and subjective consciousness. Under what conditions can these other mental properties also be attributed to artifacts to which we attribute knowledge states? Is intelligence separable from these other mental phenomena?

It is possible that intelligence can be explained or simulated without necessarily explaining or simulating other aspects of mind. A somewhat formal way of putting this *Separability Hypothesis* is 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: $M_1 \rightarrow M_2 \rightarrow \ldots$, 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 $K_1 \rightarrow K_2 \rightarrow \cdots$ 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 if they achieve goals. 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.

For example, Searle[1] holds that a computer program that successfully translates from Chinese to English cannot be said to 'understand Chinese', even though it is behaviorally intelligent in this task. In our terminology, we would attribute to the program various appropriate knowledge states. Searle's objection can be formulated as the claim that 'understanding' is a subjective property that goes beyond merely being in the corresponding knowledge state, and thus the

program can be denied that attribute.

However, other researchers claim that intelligence cannot be separated from other mental phenomena. Such a claim is often made from two opposite perspectives. Most people in artificial intelligence (AI) and cognitive science say that intelligence and other aspects of mind are inseparable because the other mental aspects (subjectivity, emotional states, etc.) are simply 'emergent' properties of certain kinds of complex agents with knowledge states. If this is the case, the knowledge state account, and with it an account in terms of information processing, will be a sufficient basis for explaining and building minds. From this perspective, explanation of the phenomena of intelligence and cognition will also turn out to be explanation of the full range of mental phenomena. By the same token, it is assumed that artificial agents that can be plausibly interpreted as solving problems, achieving goals, and performing reasoning will also have emotional states and subjective consciousness attributable to them.

The second perspective from which intelligence is taken to be inseparable from other mental phenomena holds that there is no coherent way to factor off a knowledge state process account from a mental state process account. There is only one mental process. That is, 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 'subprocesses' which only have to do with information processing. In this view, the only way to explain or build an intelligence is to solve the problem of explaining or building a mind. Thus 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.

Edelman[2,3] 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 neural pathways in response to interactions with the world. The processes that underlie this neuronal path selection resemble Darwinian evolutionary processes. Cognitive phenomena, in his view, cannot be separated and understood in information processing terms, since cognitive states are simply aspects of more general brain states, and the basic brain mechanisms are not information processes.

Dimension 2: Functional versus biological

The second dimension in discussions about intelligence involves the extent to which we need to be tied to biology for understanding intelligence. Can intelligence be characterized abstractly as a functional capability which just happens to be realized more or less well by some biological organisms? If it can, then study of biological brains or of human psychology is not logically necessary for a theory of cognition and intelligence, just as enquiries into the relevant capabilities of biological organisms are not needed for the abstract study of logic and arithmetic or for the theory of flight. Of course, we may learn something from biology about how to practically implement intelligent systems, but we may feel quite free to substitute non-biological (both in the sense of architectures which are not brain-like and in the sense of not being constrained by considerations of human psychology) approaches for all or part of our implementation. Whether intelligence can be characterized abstractly as a functional capability surely depends upon what phenomena we want to include in defining the functional capability, as we discussed. We might have different constraints on a definition that needed to include emotion and subjective states from one that only included knowledge states. Clearly, the enterprise of AI deeply depends upon this functional view being true at some level, but whether that level is abstract logical representations as in some branches of AI, Darwinian neural pathway selections as proposed by Edelman, something intermediate, or something physicalist is still an open question.

Newell holds a functional view of intelligence. According to Newell[4], intelligent agents can be abstractly characterized by their goals, their knowledge and the Principle of Rationality. That is, when we attribute intelligence to an agent in some behavior, we are attributing to that agent a goal, a body of knowledge, and a capability, at least in that instance of behavior, of applying knowledge relevant to the goal to decide what to do. It is important to note that all of this is *attribution*. Newell calls a description of an agent in these terms a *Knowledge Level* description. Knowledge Level descriptions view the agent as being in a knowledge state, and the Principle of Rationality as the abstract characterization of how the agent changes knowledge states. (Attributing knowledge and goals to an agent is similar to taking an *intentional stance* towards agents that Dennett[5] has proposed.)

There is no claim that knowledge is internally represented explicitly, and in just the same propositional units as in the attribution, or that explicitly inferential processes are operating. Newell defines the functionality of intelligence as the ability of an agent to realize the knowledge potential inherent in its Knowledge Level description. For Newell the important character of intelligence is the agent's ability to make full use of the knowledge attributed to it, not the amount or the specifics of the agent's knowledge. Even humans are only an approximation to the ideal intelligence so

defined. In this perspective, biological evolution will be seen as operating in the direction of better and better approximation to this sort of intelligence through the evolution of more complex knowledge state representations (of the sort that finds its culmination in human language) which are capable of supporting open-ended deliberation and the application of knowledge to new goals.

So, with Newell, we have a functional characterization of intelligence which is independent of biology. But Newell goes on to propose an architecture which is inspired by one biological instantiation, the human cognitive apparatus. This architecture is similar to the human one in that it has a long-term memory and a deliberative architecture similar to the one that in his view characterizes human cognition. But, because it is a functional architecture, it goes beyond the biological in many ways. For example, the ideal architecture always retrieves the relevant knowledge, unlike the human version which often fails to remember. Further, the functional architecture is based on digital computer-like symbol structures. For Newell, it does not matter if the human brain is literally such a computer. All that matters is that the kind of computer-like symbol structures can support the functionality needed. Further, the architecture that is proposed by Newell as a possible one for AI is just one among many possible realizations of the abstract functional capability specified in his definition of intelligence.

In general, functional characterizations end up using aspects from very different levels of descriptions of biological mind. For example, the connectionists want to be biological enough to include some of the smooth concept learning done by humans, and an architecture based on some abstract properties of what they take to be the information processing of brains, but their orientation is not so biological as to demand wet neurons and neuronal chemistry. Searle wants to be biological enough to demand the inclusion of the subjective awareness of being in a knowledge state (which is how we interpret his claim that a translator who follows the algorithm does not really 'understand Chinese') that humans have, but he thinks that it is most likely the chemistry of the brain that is responsible for it, and thus a pure information processing account will not succeed. Edelman wants to be biological enough to include the way in which organisms' brains, in his view, do not use pre-made internal labels (which he takes to be the characteristic property of information processing). Since his theory of pathway selection itself is stated as an abstract mechanism, presumably artifacts could be constructed which implement that abstract architecture without any further reference to biology. Connectionists (Rumelhart *et al.*[6]) and Edelman want to be biological enough to understand the common heritage between animals and humans, while traditional AI researchers stop their biological commitment to characterizing intelligence as using knowledge to reason and achieve goals (since they take humans to be doing that). Thus, all such proposals pick out some interesting aspect from biological phenomena. They then proceed to formulate a functional model that includes the selected aspect. After this, real biology is no longer logically necessary. Whether any of these proposals would lead to the production (or explanation) of mentality in total, or almost circularly, produce only those aspects of mentality that are included in the functional definition, is obviously an open question.

Coin-sorters and knowledge states

In this article we will take the Separability Hypothesis as a working hypothesis. At this point, *for all practical purposes AI (and cognitive science) can be considered the study of those regularities of mind that have information-processing explanations.* We will assume that it is a worthwhile enterprise to concentrate on phenomena in which knowledge states of the agent seem to play the central role. Further we will focus on processes that account only for generation and transformation of such knowledge states. Now this might appear to be a commitment to information processing so strong that many interesting theories will be ruled out. However, we will argue that the knowledge state account is very flexible, and can even be applied to situations where there is no explicit information processing in the conventional sense. To illustrate this we will use the example of a coin-sorter for coins of USA.

Analysis of a coin-sorter

Let us suppose that we have a black box coin sorter in front of us, and we want to describe its behavior computationally. All we see is that the coins are put into the top of the coin sorter, and then they come out through one of four slots at the bottom, with all the dimes coming out of the slot designated the dime slot, and the quarters coming out of the slot designated the quarter slot, and so on. Let us assume we have four types of AI theorists: a logician, someone who is committed to algorithms alone as the language in which to formulate AI theories, a connectionist and a physicalist, i.e. one who claims that the appropriate explanation of the coin-sorter should be in terms of its physics, not representations.

Logic system coin sorter. The logician proposes that the machine's behavior can be understood in terms of

four logical axioms, one for each coin. A set of measurements is made on each of the coins. Perhaps diameter, weight and thickness are the coin's important features for this purpose. Each coin type is characterized by a logical formula of predicates involving the measurements. For example, the axioms for each of the four types of coins will indicate what combination of weight, thickness and diameter characterize that coin type. The logician claims that the behavior of the machine can then be characterized by a theorem-proving decision procedure that attempts to prove each of the theorems for each coin that is inserted, followed by a mechanism that places the coin into that slot corresponding to the theorem that was proved.

Note that this language enables us to argue about different theories about what is being measured by the sorter. Someone could watch the behavior of the coin sorter and assert that the machine is not using information about the weight and diameter of the coins, but rather about, say, its color and metallic content. They could propose an alternative axiom system in terms of color and metallic content. Each such axiom system is a different content theory expressed in the logic formalism.

Further, the formalism can be used to evaluate these alternate theories and test them experimentally. We can use logical inference to draw out the consequences of each proposal. One hypothesized content theory might predict that a given foreign coin, say an Indian rupee, will come out of the quarter slot, while another might predict that the rupee will come out of the penny slot. We can then test to see which hypothesized content theory most accurately describes the decision-making process within the black box by putting the rupee in and seeing which slot it is placed at.

Notice that the usefulness of the logic formalism has two levels. On one level, we can use the formalism to describe different content theories, e.g. the theory that the coins are being sorted by color versus one that they are being sorted by weight. We can use the inference machinery that comes with logic to derive consequences of different axioms and test one theory of representational content against another. For this purpose, there is no need to commit oneself to how the insides of the sorter work in any detail, except that information of certain types is being used to make decisions of certain types. We are simply using logic to reason about the agent, much as it is used in computer science to reason about the correctness of a computer program written in some other language than logic. We are using logic to give a *Knowledge Level* description of the system.

The second use of logic may be to model, or carry out, internal processing. For example, the coin-sorter might actually have dedicated Prolog chips inside actually implementing the theorem provers. The coin-sorter might literally work by actuating an arm that places the coins in the slots as soon as the results of the theorem provers are available.

Decision tree coin-sorter. The second theorist observes the coin-sorter and announces that its behavior can be described by a decision tree. In a decision tree machine, there is an initial decision made between two groups, e.g. between the group consisting of the nickel and the quarter, and the group consisting of the dime and the penny. For each of the groups, at the next point in the tree, an additional decision is made to make a choice among subgroups, and this is repeated until each leaf node corresponds to one of the elements of the original group. We now have a decision tree. In the coin-sorter example, we would only need two levels in the tree. The criteria for the decisions at each node are given in the form of rules involving values of measurements made on the coin.

Again, we can use the formalism as a descriptive device, or as a commitment to a certain internal processing. For example, as a descriptive device, the decision tree still enables us to propose different content theories, not only about what aspects of the coins are measured as in the logic case, but also about what sets of decisions are made before what decisions. In this sense, what was left as a feature of internal processing in the use of logic for external description, namely some aspect of control strategy, is actually now made part of the external description of the device. The axiom system made no commitment to control. This expresses the difference between a Knowledge Level account and a program level account.

On the other hand, similar to the logic case, one can imagine microprocessors actually implementing the decision tree algorithm, using the measurements to make the choices in the tree, and activating the coin-placing mechanism appropriately when a leaf node is reached.

Connectionist network coin-sorter. The connectionist claims that what is really going on in the coin sorter involves the same features, diameter, color, or whatever, as the other theories assumed, but these evidences are 'weighted' and combined as in a connectionist network. Different theories of representational content could still be represented by identifying the nodes with different types of measurement. How the information is used can be described by means of different weights and thresholds in the network. Intermediate abstractions may be captured by hidden units. The intermediate abstractions are combined with other intermediate abstractions and again weighted and higher level decision units are constructed. A specific output node is

identified for each coin. The 'energy' at the output nodes will be a function of how much evidence is coming through for the coin for which it stands. The output node corresponding to the largest activation will be chosen as the decision node.

Pretty much all the points we made about logic and decision trees can be repeated for this account as well. The connectionist framework can be used to describe content theories about what information is used, and to give an account of what evidence is combined in what proportion with what other evidence. Inferences about different content theories can be made and tested. At this level, no commitment needs to be made that the inside of the sorter is literally a connectionist machine. On the other hand, the connectionist network can be used as the internal information processor as well.

Voila: Levers and holes! Let us now open the coin-sorter and look at its inside. We see that as you put a coin in, it passes through levers and holes, all cleverly arranged such that the coin makes its way to the right slots. Clearly, the different weights and the sizes of the coin have different effects on the levers and the holes. There are no prologue chips or microprocessors or connectionist networks inside the black box, just mechanical parts. The physicalist, the one who does not believe in representations, smiles.

Does the sorter have a knowledge state interpretation?

In response to the question, 'How did the quarter end up in the slot named "quarter"?', two kinds of answers, both correct, can be given. In one, the answer would be physicalist: an account of the coin's movement through the inside of the sorter following the physical laws. In the other, the answer would be in terms of how the levers and holes 'use' information about the diameter and the weight and how the sorter 'decides' about the coin's direction of movement. Clearly, whoever designed it designed the sorter in such a way that there is a close mapping between the information story and the physical story. Because of this mapping, one can talk about the sorter being in various knowledge states. Of course, if the sorter that works by levers and holes has a consistent interpretation in terms of knowledge states, then certainly any sorter that actually had a chip proving theorems or implementing the decision tree algorithm or the connectionist network will also have a similar interpretation. That is, the knowledge state and information processing talk is applicable to all devices whose behavior has a decision-making interpretation, irrespective of how they actually work.

We can see that the logic account, the decision tree algorithm account and the connectionist account are all alternative languages in which to couch the information processing account. While all three frameworks can be used to describe information representation and processing, they are not all equivalent. Connectionism enables one to talk about 'softer' combination of information using real numbers, while logic enables us to talk about variables and quantification, and the language of algorithms enables us to talk about control strategies. However, our main point here is that they can all be used as frameworks for describing information representation and processing, and also for implementing information processing. In Newell's language, they can be used both as languages for the Knowledge Level and for the Symbol Level.

The coin-sorter is a simple device, but it illustrates the issues with respect to understanding biological brains. People take a whole range of stances on whether the brain is actually doing information processing on representations. Strong materialists argue that representationalist accounts of such systems are wrong, and the only scientifically acceptable causal story is at the level of the matter that composes the brain. Edelman is also against the information processing account, but his counter-proposal is in terms of an abstract pathway selection account, which is still an abstract functional architecture (i.e. no appeal to physical laws is made), though not an information processing one. Among those who agree that there is a causal story to be told at the level of representations, there are many divisions, but broadly, we can distinguish between connectionist style representations and discrete symbol structure representations. The moral of our analysis of the coin-sorter is that for explaining behavior which itself is couched in informational terms, the information processing account is useful as a stance to describe the biological brain.

Much of the argument in the field is a result of a confusion between two senses of being an information processor using representations. In one sense, when we ask whether the brain processes information we are really asking whether it is appropriate to ascribe informational activity to the brain and in the other sense we are literally describing what the brain or device actually does. Ascribing information processing is to take an information processing stance. For example we might ascribe information processing activity to the visual system on the grounds that it produces information about the world. This is the sense of information-processing we are using when we stand outside the brain and look at behavior and ascribe an information-processing structure to the behavior that we see. When we look at a black box coin sorter as a decision maker and work out a model of its input/output behavior, we are ascribing information processing to it.

However, taking an informational stance whereby we

ascribe information processing to a device (or brain) does not commit us to that device literally processing information, or using representations, in the specific medium in which the description is made. There *is* a fact of the matter about whether the information processing is being done in one medium or another. At some point the behavior of the sorter which employs a Prolog theorem prover will be different from that based on levers and holes. When the latter sorter malfunctions, the explanation may be given in terms of physical properties, such as a lever being jammed, while in the case of the former type of sorter, the explanation might be in terms of an error in the program in the chip or some hardware failure in the chip. (And in the case of the brain, in addition to the problem of failure modes, there are other issues where the medium becomes relevant: properties related to learning, are one example.) But for most purposes where people think that the issue is the medium of representation, the issue often turns out to be one that can be formulated at the Knowledge Level.

We can certainly ask similar questions about the brain. It is a matter of fact whether the brain is an information processor of the 'physicalist' type, one of the connectionist variety, or one that has Turing machine-like symbols. (Putnam[7] has argued that even whether a piece of matter is a Turing machine is just a stance, but we think that the consensus is that Putnam's argument does not really work, and that not all pieces of matter can be interpreted as a given Turing machine.) But as long as we are interested in aspects of the organism's behavior that have an informational flavor (such as decision-making), talk of information and its use is necessary in the analysis, just as it was in the case of the coin-sorter. Much of the criticism of the information processing view (from Edelman, e.g.) of information processing is based on a narrow view of what the information-processing talk commits one to. Conversely, many proponents of information processing explanations are also committed to such a narrow view, making far more commitments about internal processes than necessary.

In the rest of the article, we will adopt this broad sense of information processing or knowledge state account as a stance that is useful in describing agents to which we ascribe cognitive capacities.

Architectures for intelligence

We now move to a discussion of architectural proposals within the information processing perspective. Our goal is to try to place the multiplicity of proposals into perspective. As we review various proposals, we will present some judgements of our own about relevant issues. But first, we need to review the notion of an architecture and make some additional distinctions.

Form and content issues in architectures

In computer science, a programming language corresponds to a virtual architecture. A specific program in that language describes a particular (virtual) machine, which then responds to various inputs in ways defined by the program. The architecture is thus what Newell calls the fixed structure of the information processor that is being analysed, and the program specifies a variable structure within this architecture. We can regard the architecture as the *form* and the program as the *content*, which together fully instantiate a particular information-processing machine. We can extend these intuitions to types of machines which are different from computers. For example, the connectionist architecture can be abstractly specified as the set $\{\{N\}, \{n_1\}, \{n_0\}, \{\xi_i\}, \{w_{ij}\}\}$, where $\{N\}$ is a set of nodes, $\{n_1\}$ and $\{n_0\}$ are subsets of $\{N\}$ called input and output nodes respectively, $\{\zeta_i\}$ are the functions computed by the nodes, and $\{w_{ij}\}$ is the set of weights between nodes. A particular connectionist machine is then instantiated by the 'program' that specifies values for all these variables.

We have made a distinction between an architecture, the form in which the architecture will accept content (the programming language form) and the content of the representation itself. When we explain specific types of cognitive phenomena, we will end by coming up with a complex budget of credit allocation: some aspects will be explained by the properties of the architecture (perhaps some timing phenomena, and also some aspects of learning), some will be explained by the sort of information that is involved in the content. Credit allocation in this manner is a tricky analytic task.

We also need to make an additional distinction between micro- and macro-architectures, a distinction that is especially useful for cognition. A bank of information processors of identical type connected in some way has a macro-architectural description in terms of the modules and their connections, while the entire system has a uniform micro-architectural description.

Many AI and cognitive science theories are really theories about the content of knowledge, or types of knowledge, needed for some task of interest, with minimal commitment to the architecture. Many debates in the field, which are ostensibly about the architecture, turn out to be about the types of knowledge. For example, Dreyfus[8] talks about 'What computers cannot do'. It turns out that he is opposed to the idea that intelligence can come out of a system that has a knowledge base which explicitly and exhaustively

represents world facts and relationships in some logical form. However, there are people within computational AI who have been making this point as well. For example, Schank[9] has argued that our knowledge is not in the above form of abstract facts at all, but rather in the form of experiences indexed and abstracted in various ways. Thus the issue, at least based on Dreyfus' arguments, is not what computers cannot do, but what certain kinds of knowledge representations cannot do. It may turn out that the kind of information that Dreyfus sees as necessary cannot be represented in computers either, but he does not make the arguments for this position.

We are now ready to give an overview of the issues in cognitive architectures. We will assume that the reader is already familiar in some general way with the proposals that we are discussing. Our goal is to place these ideas in perspective.

Intelligence as just computation

Until recently the dominant paradigm for thinking about information processing has been the Turing computer framework, or what has been called the discrete symbol system approach. Information processing theories are formulated as algorithms operating on data structures. In fact AI was launched as a field when Turing proposed in a famous paper that thinking was computation (the term 'artificial intelligence' itself was coined later). A natural question in this framework would be whether the set of computations that underlie thinking is a subset of Turing-computable functions, and if so, how the properties of this subset should be characterized.

Because of the technological nature of much of AI, only a small number of researchers have been concerned with intelligence in general. Most of the work consists of algorithms for specific problems that seem to require intelligence and that are practically important. Algorithms for diagnosis, design, planning, etc. are proposed, because these tasks are seen as important for an intelligent agent. But as a rule no effort is made to relate the algorithm for the specific task to a general architecture for intelligence. While such algorithms are useful as technologies and to make the point that several tasks that appear to require intelligence can be done by certain classes of machines, they do not give much insight into intelligence in general.

Architectures for deliberation

Historically most of the intuitions in AI about intelligence have come from introspections about human consciousness, specifically about what people perceived to be the relationships among conscious thoughts. We are aware of having thoughts which often follow one after another. These thoughts are mostly couched in the medium of natural language, but sometimes thoughts include mental images as well. When people are thinking for a purpose, say for problem solving, there is a sense of directing thoughts, choosing some, rejecting others, and focusing them towards the goal. Activity of this type has been called 'deliberation'. Deliberation, for humans, is a coherent goal-directed activity, lasting over several seconds or longer. For many people thinking is the act of deliberating in this sense. Activities in this time span should be contrasted with other cognitive phenomena, which, in humans, take under a few hundred milliseconds: real-time natural language understanding and generation, visual perception, being reminded of things, and so on.

Different kinds of theories about the architecture of the cognitive machine have been proposed depending upon what kinds of patterns among these thoughts the researchers have been struck by. Two groups of proposals about such patterns have been influential in AI theory-making: the *reasoning* view and the *goal–subgoal* view.

Deliberation as reasoning. People have for a long time been struck by logical relations between thoughts and have made the distinction between rational and irrational thoughts. Remember that Boole's book on logic was titled 'Laws of Thought'. Thoughts often have a logical relation between them: we think thoughts A and B, then thought C, where C follows from A and B. In AI, this view has given rise to an idealization of intelligence as rational thought, and consequently to the view that the appropriate architecture is one whose behavior is governed by rules of logic. In AI, McCarthy is most closely identified with the logic approach to AI, and ref. 10 is considered a clear early statement of some of the issues in the use of logic for building an intelligent machine. Researchers in AI disagree about how to make machines which display this kind of rationality. One group proposes that the ideal thought machine is a logic machine, one whose architecture has logical rules of inference as its primitive operators. These operators work on a storehouse of knowledge represented in a logical formalism and generate additional thoughts. For example, the Japanese Fifth generation project came up with computer architectures whose performance was measured in (millions of) *inferences* per second. The other group believes that the architecture itself (i.e. the mechanism that generates thoughts) is not a logic machine, but one which generates plausible, but not necessarily correct, thoughts,

and then knowledge of correct logical patterns is used to make sure that the conclusion is appropriate.

Historically rationality was characterized by the rules of deduction, but in AI, the notion is being broadened to include a host of non-deductive rules under the broad umbrella of 'non-monotonic logic'[11] or 'default reasoning', to capture various plausible reasoning rules. There is considerable difference of opinion about whether such rules exist in a domain-independent way as in the case of deduction, and how large a set of rules would be required to capture all plausible reasoning behaviors. If the number of rules is very large, or if they are context-dependent in complicated ways, then logic architectures would become less practical.

At any point in the operation of the architecture, many inference rules might be applied to a situation and many inferences drawn. This brings up the control issue in logic architectures, i.e. decision about which inference rule should be applied when. Logic itself provides no theory of control. The application of the rule is guaranteed, in the logic framework, to produce a correct thought, but whether it is relevant to the goal is decided by considerations external to logic. Control tends to be task-specific, i.e. different types of tasks call for different strategies. These strategies have to be explicitly programmed in the logic framework as additional knowledge.

Deliberation as goal–subgoaling. An alternate view of deliberation is inspired by another perceived relation between thoughts and provides a basic mechanism for control as part of the architecture. Thoughts are often linked by means of a *goal–subgoal* relation. For example, you may have a thought about wanting to go to New Delhi, then you find yourself having thoughts about taking trains and airplanes, and about which is better, then you might think of making reservations and so on. Newell and Simon[12] have argued that this relation between thoughts, the fact that goal thoughts spawn subgoal thoughts recursively until the subgoals are solved and eventually the goals are solved, is the essence of intelligence as a mechanism. More than one subgoal may be spawned, and so backtracking from subgoals that did not work out is generally necessary. Deliberation thus looks like search in a problem space. Setting up the alternatives and exploring them is made possible by the knowledge that the agent has. In the travel example above, the agent had to have knowledge about different possible ways to get to New Delhi, and knowledge about how to make a choice between alternatives. A long term memory is generally proposed which holds the knowledge and from which knowledge relevant to a goal is brought to play during deliberation. This analysis suggests an architecture for deliberation which retrieves relevant knowledge, sets up a set of alternatives to explore (the problem space), explores it, sets up subgoals, etc.

The most recent version of an architecture for deliberation in the goal–subgoal framework is Soar[4]. Soar has two important attributes. The first is that any difficulty it has in solving any subgoal simply results in the setting up of another subgoal, and knowledge from long term memory is brought to bear in its solution. It might be remembered that Newell's definition of intelligence is the ability to realize the knowledge level potential of an agent. Deliberation and goal subgoaling are intended to capture that capability: any piece of knowledge in long term memory is available, if it is relevant, for any goal. Repeated subgoaling will bring that knowledge to deliberation. The second attribute of Soar is that it 'caches' its successes in problem solving in its long term memory. The next time there is a similar goal, that cached knowledge can be directly used, instead of searching again in the corresponding problem space.

This kind of deliberative architecture confers on the agent the potential for rationality in two ways. With the right kind of knowledge, each goal results in plausible and relevant subgoals being setup. Second, 'logical rules' can be used to verify that the proposed solution to subgoals is indeed correct. But such rules of logic are used as pieces of knowledge rather than as operators of the architecture itself. Because of this, the verification rules can be context- and domain-dependent.

Another point to note is that one of the results of this form of deliberation is the construction of special purpose algorithms or methods for specific problems. These algorithms can be placed in an external computational medium and as soon as a subgoal arises that such a method or algorithm can solve, the external medium can solve it and return the results. For example, during design an engineer might set up the subgoal of computing the maximum stress in a truss, and invoke a finite element method running on a computer. The deliberative engine can thus create and invoke computational algorithms. The goal-subgoaling architecture provides a natural way to integrate external algorithms.

In the Soar view, long term memory is just an associative memory. It has the capability to 'recognize' a situation and retrieve the relevant pieces of knowledge. Because of the learning capability of the architecture, each episode of problem solving gives rise to continuous improvement. As a problem comes along, some subtasks are solved by external computational architectures which implement special purpose algorithms, while others are directly solved by compiled knowledge in memory, while yet others are solved by additional deliberation. This cycle makes the overall system increasingly more powerful. Eventually, most

routine problems, including real-time understanding and generation of natural language, are solved by recognition. (Recent work by Patten *et al.*[13] on the use of compiled knowledge in natural language understanding is compatible with this view.)

Deliberation seems to be a source of great power in humans. Why is not recognition enough? As Newell points out, the particular advantage of deliberation is distal access to and combination of knowledge at runtime in a goal-specific way. In the deliberative machine, temporary connections are created between pieces of knowledge that are not hard-coded, and that gives it the ability to realize the knowledge level potential more. A recognition architecture uses knowledge less effectively: if the connections are not there as part of the memory element that controls recognition, a piece of knowledge, though potentially relevant, will not be utilized in the satisfaction of a goal.

As an architecture for deliberation, the goal–subgoal view seems to us closer to the mark than the reasoning view. As we have argued elsewhere[14], logic seems more appropriate for justification of conclusions and as the framework for the semantics of representations than for the generative architecture.

AI theories of deliberation give central importance to human-level problem solving and reasoning. Any continuity with higher animal cognition or brain structure is at the level of the recognition architecture of memory, about which this view says little other than that it is a recognition memory. For supporting deliberation at the human level, long term memory should be capable of storing and generating knowledge with the full range of ontological distinctions that human language has.

Is the search view of deliberation too narrow?. A criticism of this picture of deliberation as a search architecture is that it is based on a somewhat narrow view of the function of cognition. It is worth reviewing this argument briefly.

Suppose a Martian watches a human in the act of multiplying numbers. The human, during this task, is emulating some multiplication algorithm, i.e. appears to be a multiplication machine. The Martian might well return to his superiors and report that the human cognitive architecture is a multiplication machine, but we know that the multiplication architecture is a fleeting, evanescent virtual architecture that emerged as an interaction between the goal (multiplication) and the procedural knowledge of the human. With a different goal, the human might behave like a different machine. It would be awkward to imagine cognition to be a collection of different architectures for each such task; in fact, cognition is very plastic and is able to simulate various virtual machines as needed.

Is the problem space search engine that has been proposed for the deliberative architecture such an evanescent machine? One argument against it is that it is not intended for a narrow goal like multiplication, but for all kinds of goals. Thus it is not fleeting, but always operational.

Or is it? If the sole purpose of the cognitive architecture is goal achievement (or 'problem solving'), then it is reasonable to assume that the architecture would be hard-wired for this purpose. What, however, if goal achievement is only one of the functions of the cognitive architecture, common though it might be? At least in humans, the same architecture is used to daydream, just take in the external world and enjoy it, and so on. The search behavior that we need for problem solving can come about simply by virtue of the knowledge that is made available to the agent's deliberation from long term memory. This knowledge is either a solution to the problem, or a set of alternatives to consider. The agent, faced with the goal and a set of alternatives, simply considers the alternatives in turn, and when additional subgoals are set, repeats the process of seeking more knowledge. In fact, this kind of search behavior happens not only with individuals, but with organizations. They explore alternatives, but we do not see a need for a fixed search engine for explaining organizational behavior. Deliberation of course has to have the right sort of properties to be able to support search. Certainly adequate working memory needs to be there, and probably there are other constraints on deliberation, but it does not have to be exclusively a search architecture. Just like the multiplication machine was an emergent architecture when the agent was faced with that task, the search engine is the corresponding emergent architecture for the agent faced with a goal and equipped with knowledge about what alternatives to consider. In fact, a number of other such emergent architectures built on top of the deliberative architecture have been studied earlier in our work on Generic Task architectures[15]. These architectures were intended to capture the needs for specific classes of goals (such as classification).

The above argument is not to deemphasize the importance of problem space search for goal achievement, but to resist the identification of the architecture of the conscious processor with one exclusively intended for search. The problem space architecture is still important as the virtual architecture for goal-achieving, *since it is a common, though not the only,* function of cognition.

Of course, that cognition goes beyond just goal achievement is a statement about human cognition. If we take a design perspective and seek to specify an architecture for a function called intelligence which itself is defined in terms of goal achievement, then

clearly we are free to design an architecture best suited for that purpose. A deliberative search architecture working with a long term memory of knowledge certainly has many attractive properties for this purpose as we have discussed in this section.

Architectures below deliberation

We made a distinction between cognitive phenomena that occur in under a few hundred milliseconds and those that evolve over longer time spans, and covered the latter under deliberation. We will call the architecture that handles the former phenomena *subdeliberative*. In deliberation, we have access to a number of intermediate states in problem solving. After you finished planning the New Delhi trip, I can ask you what alternatives you considered, why you rejected taking the train, and so on, and your answers to them will generally be reliable. You were probably aware of rejecting the train option because you calculated that it would take too long. On the other hand, we have *generally no clue about how the subdeliberative* architecture came to any conclusion. If you recognize someone after not having seen him for twenty years, and that person expresses surprise by asking, 'I have changed a lot in twenty years. How did you recognize me?', you may come up with something like, 'I bet it is your nose!', but you cannot be sure. You have no access to how your perception system actually recognized that person. Similarly, you may have your own theory of why you were reminded of something, but you have no special access to what went on during that reminding. Freud's model of mind had complicated unconscious processes working, and in fact, in this view, consciousness was often misled about the real content of these unconscious processes.

Many people in AI and cognitive science feel that the emphasis on complex problem solving as the door to understanding intelligence is misplaced, and that theories that emphasize rational problem solving only account for very special cases and do not account for the general cognitive skills that are present in ordinary people. This group of researchers focus almost completely on the nature of the subdeliberative architecture. There is also a belief that the subdeliberative architecture is directly reflected in the structure of the neural machinery in the brain. Thus, some of the proposals for the subdeliberative architecture claim to be inspired by the structure of the brain and claim a biological basis in that sense.

Alternative proposals. The various proposals differ along a number of dimensions: what kinds of tasks the architecture performs, degree of parallelism, whether it is an information processing architecture at all, and when it is taken to be an information processing architecture, whether it is a symbolic one or some other type.

With respect to the kind of tasks the architecture performs, we already mentioned Newell's view that it is just a recognition architecture. Any smartness it possesses is a result of good abstractions and good indexing, but architecturally, there is nothing particularly complicated. In fact, the good abstractions and indexing themselves were the result of the discoveries of deliberation during problem state search. Being smarter, from the Newell perspective, is done by converting more and more deliberative problems into stored recognition patterns through chunking. The real solution to the problem of memory, for Newell, is to get chunking done right: the proper level of abstraction, labeling and indexing is all done at the time of chunking. Theories of memory representation (such as Schank's) are in this sense content theories of indices and labels, not architectural theories. Such content theories of memory are not really in conflict with the Newell theory of deliberative architecture, since the latter merely gives a way for the content to come to be the way it is.

In contrast to the recognition view are proposals that see relatively complex problem solving activities going on in subdeliberative cognition. Minsky[16] originally proposed a specific architecture for memory based on frames, which are organized as a network of concepts, each of which contained prototypical information about the concept. Relatively complex procedures were embedded in these concepts. More recently, he has outlined a Society of Mind[17] architecture for cognition. Cognition in this picture is a communicating collection of modular agents, each of whom is simple, but capable of some degree of problem solving. For example, they can use the means-ends heuristic (the goal-subgoaling) feature of deliberation in the Soar architecture).

Deliberation has a serial character to it. Almost all proposals for the subdeliberative architecture, however, use parallelism in one way or another. Parallelism can bring a number of advantages. For problems involving similar kinds of information processing over somewhat distributed data (like perception), parallelism can speed up processing. Some problems that require explicit search if done serially can be done without search in a parallel architecture. For example, perception problems often involve evaluating a number of alternative interpretations and choosing the best. These alternatives can be simultaneously assessed in parallel and the best picked. Ultimately, however, additional problem solving in deliberation may be required for some tasks.

Within the school that views the subdeliberative architecture as representation-processing, there has been a debate about the medium in which information is represented. Turing computational architectures have

been the representational frameworks of choice for modeling deliberation. For subdeliberation, the same framework was used until connectionism came along. Connectionism replaced the explicit processing of symbolic tokens with a specific type of analog computation. The original connectionist proposal of the PDP type[6] were in some ways less powerful than Turing machines. For example, it had to face the criticism that that kind of computation cannot account for the systematicity and generativity of natural language which requires variable binding and symbols of some type[18], requirements which the Turing-computational framework can handle well. A number of ways of enlarging the connectionist frameworks to give them these capabilities have been proposed. Some involve using explicit symbols in connectionist representations (see for example, ref. 19), while others involve representations that have some of the properties of symbols without being symbols in the Turing-computational sense (see for example, ref. 20). In any case, most of these connectionist proposals are actually implemented and simulated in digital computers, and none of the functions that they compute are outside the Turing framework. The problem does not really seem to be with Turing computation *per se*, but rather the way in which Turing computation has been used in AI and cognitive science, namely as applications of inference on axiomatically represented world knowledge.

Connectionism has been evolving in a number of directions. A proposal that has been gaining currency is that the information processing of the brain is a dynamical system[21] defined by complex nonlinear differential equations. It has been claimed, for example, that chaos may be useful as a creative device for new states in a search[22], and that dynamic systems at criticality have the unbounded dependencies characteristic of context-sensitive grammars[23].

Edelman argues strongly against information processing theories of cognition on the ground that they require a prelabeled world of objects and relations, whereas biological organisms in his view discover patterns as regularities in their interactions with the world rather than start with prelabeled information. He also argues against connectionism since he thinks they require some form of prelabeled information as well. His architectural proposal is not couched as computation on representations, but as one in which successful neural pathways are selected in a process similar to Darwinian evolution. The selection is done in response to the physical interaction of the organism with the external world. This process results in neural structures which categorize the organism's interaction with the world, but these are not fixed logical categories, but flexible, constantly changing ones, to reflect the organism's continuing interaction. Edelman has proposed additional mechanisms by which these structures develop higher and higher order categorizations and coordinations.

The motivation behind connectionism and its offshoots is generally couched as opposition to symbolic computation, and Edelman argues against information processing, but, as we have argued earlier, the real opposition seems to be to the idea of a representational repertoire that corresponds to the theories of the external world of objects and relations that we conceptualize in our conscious models of the world. There is a widespread suspicion that AI and cognitive science have confused the externally visible constructions of mind (explicit knowledge of the world, grammars, etc.) as the raw material of mind. In this view, just because we seem to be using pieces of knowledge in our deliberation does not mean that this knowledge was represented in that form in memory. The phrase 'information processing' has been too closely associated with the view that what is inside the mind is much like what we seem to have in our consciousness. The opposing view is that whatever is inside us is not in the form of abstract statements of facts about the world, but rather is concretely tied to our interaction with the physical world, flexible, open-ended, and constantly changing with each interaction.

With this proviso accepted, we can take a representational stance towards connectionist networks as well as Edelman's selection machine. In that sense of attributed information or knowledge that we argued for in our discussion of the coin-sorter, Edelman's organism has knowledge and information. We can, from outside, watch an Edelmanian brain at some point in its evolution, and say things like, 'This organism knows about x, but not about y.' In the broad sense of information processing that we have been advocating, Edelman's organism is an information processing agent and its neural pathways represent knowledge. If knowledge of the world can be in the form of on-going abstractions of experience, which at the Knowledge Level, can be interpreted as partial, but increasingly more veridical, knowledge of the world, then all these approaches qualify as information processing theories.

Is there a 'right' architectural theory of subdeliberation? Later in the article we discuss how to place the various alternative proposals in useful relations to each other.

So far we have talked about the micro-architecture of the subdeliberative system. A few brief comments on macro-architecture are relevant. Fodor[24] has proposed the *Modularity Hypothesis* which asserts that there are separate modules for each of the perceptual modalities, the language modality and central cognition. That is, there is relatively little interaction between them until the perceptual and language modules have completed

their interpretation tasks. These interpretations are available in the working memory of deliberation. There is some debate about how much information flow is there from one modality to another during recognition, but there is general consensus that the degree of intermodality information flow is small in comparison with the information processing within each module.

Situated cognition. Real cognitive agents are in contact with the surrounding world containing physical objects and other agents. A new school has emerged calling itself the *situated cognition* movement which argues that traditional AI and cognitive science abstract the cognitive agent too much away from the environment, and place excessive emphasis on internal representations. The traditional internal representation view leads, according to the situated cognition perspective, to excessive amounts of internal representation and complex reasoning using these representations. Real agents simply use their sensory and motor systems to explore the world and pick out the information needed, and get by with much smaller amounts of internal representation processing. At the minimum, situated cognition is a proposal against excessive 'intellection'. In this sense, we can simply view this movement as making different proposals about what and how much needs to be represented internally. However, there are more radical versions of the movement in which any internal representation is denied. Specifically, the movement rejects the idea that knowledge is represented in the brain and retrieved as needed, but instead holds that knowledge is constructed by the agent in a complex interaction between neural processes and the external situation. '[Representations] are the *product* of interactions, not a fixed substrate from which behavior is generated'[25]. The reader will recognize that this view is close to that of Edelman. This constructivist view of knowledge is a major dividing line between traditional 'knowledge representation' view in AI and the situated cognition view. To take an example, schema theories in psychology and frame theories in AI have held that memory is organized in terms of schemas, stereotyped concepts or events. The newer view would hold that such schemas are actually constructed in response to the situation, not units of memory representation and organization[26].

In our discussions so far, we have presented two different views on internal representations. On the one hand, we have representations in the traditional AI sense of explicit encoding of facts and so on, and on the other hand, we also said that one can often take an external Knowledge Level stance towards the content of knowledge that is implied by an agent's behavior. The situated cognition perspective clearly rejects the former view with respect to internal (sub-deliberative) processes, but accepts the fact deliberation does contain and use knowledge. Thus the Knowledge Level description could be useful to describe the content of agent's deliberation. But the perspective emphasizes the issues relevant to the nature of the neural level descriptions and the processes which work with the external situation to construct the representations in deliberation.

The movement raises many important issues, but the solution to the problem of what sort of neural processes exist and how the interactive process constructs representation is still in the future.

Integrating the perspectives

An integrated view of problem solving. We briefly outline how the major components of the cognitive architecture work together in the solution of complex problems. The agent is embedded in the physical world, receives sensory information, and acts on the world. Deliberation is the central co-ordinating architecture, and its working memory can contain both symbolic and imagistic data, constructed out of long term representations in response to the goal at hand, as the situated cognition movement proposes. Memory can be viewed at the Knowledge Level as containing this information, but this talk should not mislead one into thinking that the information that is in working memory was in that form in long term memory (see our discussion on situated cognition). The agent also has action repertoires which can be thought of as a form of memory, but information representational talk is much less appropriate for describing them.

The degree of abstract problem solving required depends on the kind of goal. Many goals can be simply solved by means of one or more of the action repertoires, with little mediation from anything that one might call problem solving in the sense of manipulation of representations of choices in a search space. The goal-action-sensory system triple is highly evolved and integrated to carry out, in a goal-driven way, such action sequences.

When such action sequences are not immediately available for the goal, there are a number of options. Working memory may contain abstract representations of problem space alternatives. The problem space and the operators available may have not only abstract symbolic components, but imagistic components as well. Working memory may also contain previously developed sequences of solutions or pointers to external methods, algorithms, or models. Some of the subgoals are best accomplished by action sequences, some by operators that are specific to the image modality (e.g. reasoning with mental images), some by application of abstract knowledge operators, and some by invoking

external agents and models. Many of the subgoals can be accomplished just by interacting with the world or sensing the world rather than by reasoning on complex representations. A common way of avoiding complex reasoning is to leave representational markers in the physical world, and use action and sensory operators to 'read off' the information.

The above description emphasizes how much of real problem solving is dominated by the fact that the agent is situated in the world, and how artificial a pure symbolic representation manipulation view can be for many problems. At the same time, the above picture is admittedly schematic. A number of important issues remain unsolved. We already referred to the problem of the mechanisms by which knowledge in working memory is constructed in response to goals. How the sensor-action system is integrated with deliberation in an abstract sense requires many details to be worked out, but it sets a research agenda that is different from that of traditional AI.

Content-driven AI and microstructural accounts are both needed. In a strange way, the perspective we just outlined validates both traditional AI and the new emphasis on microstructure. Traditional AI, with its emphasis on knowledge and the distinctions needed to express it, has tried to wrestle content down. It has been able to do this pretty well up to a point, but because it is not embedded in a theory with appropriate microstructure and environmental interaction, ends up *over-idealizing* content and missing the form in which knowledge really emerges. The microstructural accounts have potential to explain the genesis and evolution of knowledge, and, to the extent that they base themselves on some aspects of biological neural systems, can explain aspects of continuity in cognition between higher animals and humans. It is also often hoped that the content problem in AI can be solved by AI systems that learn from scratch or with little initial knowledge. That is, the hope is that learning will obviate the need to develop knowledge level distinctions. That seems highly unlikely for reasons of complexity, both in time and in the environmental specification, but also due to the need for specifying appropriate initial states. It is more likely that the learning theories will give broad insights about content that might place useful constraints on knowledge level theories. Thus the content-driven AI picture and the microstructure-driven new architectural views need to work side by side for quite a while, hoping to meet in various ways and places for mutual benefit.

Hierarchy of leaky architectures and cognitive explanations. We have mentioned connectionism, dynamical systems, and Edelman's selection machine as three contending proposals for the subdeliberative architecture, and no doubt there will be many others over time. But to look for a 'correct' answer to the cognitive architecture may be to commit an error in reification, in believing that there exists one architecture that can be factored off the physical brain in such a way that the architecture corresponds to and only to cognition (or more generally mentality). In the introductory section on dimensions for thinking about thinking, we discussed the problems associated with factoring off a cognitive architecture from a mental architecture. A similar issue arises in the belief that a mental architecture can be factored off the physical brain or the body, and that a clearly defined set of functionalities can be identified to define mind. What we have in the brain is a biologically evolved complex piece of matter working at many levels, informational, chemical and electrical. Certainly different stances can be taken towards it for different analytical purposes, but believing that there exists a separable architecture called the mental, especially one that has a description at one level, may be Platonism run amok.

If this view is right, then we can see the contending proposals for the subdeliberative architecture as approximate descriptions, at somewhat different levels, of a physical reality called brain, which in turn is the basis for a host of behaviors that have a mentalistic description.

Consider the mathematical description of an economy in a human society. It would be strange to regard the economy as the reality which just happens to be implemented on humans. Description of an economic model is an approximate description of certain types of activities in human society. This is the analogy that we would like the reader to keep in mind as we describe our view of hierarchy of cognitive architecture descriptions.

In this view, the Edelman selection machine is a convenient and approximate description of a machine which is really a complex chemical machine. At a higher level, dynamical systems provide another approximate description, with connectionist descriptions providing yet another level of description. We think that when the selection machine organizes itself to perform some task, say perception, it should be possible to see in it a description of evidences being combined, the language in which connectionism works. At the top level we have the knowledge level description of the agent in terms of knowledge and goals. Each of these descriptions captures some aspects and functionalities, but misses others.

However, this picture of virtual machines all lined up vertically, the deliberative architecture on top of the recognition architecture on top of, say, a dynamical systems architecture which in turn is on top of

something else and so on all the way down to chemistry and physics, might give a false picture of perfect implementations of a higher level by a lower level. Biological brains do not really have cleanly lined up architectures in the way that computers do. In artifacts like computers, we as designers have conceptualized a pure information processing machine and have created a complete one-to-one correspondence between the elements of that and the elements of a physical machine. Except when the machine malfunctions we never have to worry about the lower level machine. In computer software design each level of architecture, each virtual machine, sits cleanly upon the one beneath it without the one beneath it showing through at all. Each level is smooth and closed and separate with respect to other levels of the architecture. This sort of architectural arrangement has guided much of our thinking about human cognitive architecture.

However, in a biologically evolved object like the human brain such a clean separation between levels of architecture and between software and hardware is impossible. This is because, first of all, these architectures we have been describing are all 'leaky' virtual machines. That is, when the surface structures are stressed, or under certain situations, the underlying machine shows through. There are layers of representational structures and representations from other layers peak through at any given layer. Like in the case of vision, where in certain optical illusions the physical structure of rods and cones shows through the interpretive architecture, the architecture of the underlying machines literally shows through in certain circumstances. The cognitive phenomena are thus not all going on at one level of architecture. Secondly, these layers of architectures are not complete, i.e. each level of description does not fully account for all the phenomena of interest. Given some complex mental activity, explanation of some aspects can be given by the Knowledge Level, for some we will need to appeal to the properties of the connectionist architecture, for some to the properties of the selection machine, and for others we may simply need to appeal to chemistry and other physical properties.

What description we use to account for the phenomena depends upon our goals. The cognitive phenomena we are looking at are not going to admit of any single level of explanation. They are very multi-dimensional, and for some purposes we can account for the behavior by referring to the deliberative machine, but for other purposes that will not do, and we will have to account for the behavior by reference to a lower level of the architecture. This means that the information processing architectures that we see underlying human cognitive behavior are architectures that we have abstracted for certain classes of purposes.

This is not to espouse a form of relativism, however. Not everything counts. There are lots of machines that could not be brought up as virtual machines by the brain. Interestingly, all the virtual machines that we considered, from Soar to connectionist systems to Edelman's path selection machines, have a special feature: they all are oriented towards adaptation and learning. Thus, there is a relationship between learnability and being capable of being a virtual machine of interest. There are facts of the matter to be investigated and discovered. We can ask of a proposed virtual machine, what work does it do? How is it useful as level of explanation? We can also ask of a particular task how is it being done? What sort of architecture is being used to accomplish it? Although we can potentially model each individual function of cognition, there may be no abstract platonic engine which accounts for all and only cognitive, or all and only mental, behavior. There may well be just various cognitive functions and various machines that can be used to explain those functions.

Concluding remarks

We started by asking how far intelligence or cognition can be separated from mental phenomena in general. We also suggested that the problem of an architecture for cognition is not really well-posed, since, depending upon what aspects of the behavior of biological agents are included in the functional specification, there can be different constraints on the architecture. That is, it is not clear that, from an architectural perspective, the idea of a cognitive architecture is a natural kind. Nevertheless, we said, we can talk about cognition as a coherent phenomenon of interest if we think of it as that behavior in which we ascribe knowledge states to the agent. Newell's Knowledge Level view of an agent is based on a similar point of view about a cognitive agent.

We reviewed a number of issues and proposals relevant to cognitive architectures. The computer metaphor has had its day, but, we argued, the information processing language has significant explanatory powers left. We ended with the position that the search for an architectural level that will explain all the interesting phenomena of cognition was likely to be futile. Not only are there many levels each explaining some aspect of cognition and mentality, but the levels interact even in relatively simple cognitive phenomena. Ultimately even physics will account for some mental phenomena.

By treating mentality, not to speak of its cognitive component, as ultimately not fully separable from the physical substrate, we are not being pessimistic about

the prospects for cognitive science and AI, just being careful about what one might expect. In one sense, this view reinforces the arguments for the need for grounding[27], and being and growing as real humans, as the ultimate requirement for achieving the kind of mentality that we have. On the other hand, explanations of all sorts of mental phenomena can come at various levels. We can build problem solvers, perceivers, cognizers and so on, and depending upon their physics they may have their own version of mentality. There is no need for AI or cognitive science to insist on the various Separability Hypotheses being true in all details for getting nearer and nearer to the goals of explanation and simulation of mind.

1. Searle, J. R., Minds, brains and programs, *Behav. Brain Sci.*, 1980, **3**, 417–424.
2. Edelman, G. M., *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books, New York, 1987.
3. Edelman, G. M., *The Remembered Present: A Biological Theory of Consciousness*, Basic Books, New York, 1989.
4. Newell, A., *Unified Theories of Cognition*, Harvard University Press, Cambridge, MA, 1990.
5. Dennett, D., *The Intentional Stance*, MIT Press/Bradford Books, Cambridge, MA, 1987.
6. Rumelhart, D. E., McClelland, J. L. and the PDP research group (eds.), *Parallel Distributed Processing: Essays in the Microstructure of Cognition, Foundations*, MIT Press/Bradford Books, Cambridge, MA, 1986.
7. Putnam, H., *Representation and Reality*, MIT Press/Bradford Books, Cambridge, MA, 1988.
8. Dreyfus, H., *What Computers Cannot Do: The Limits of Artificial Intelligence*, Harper and Row, New York, 1972.
9. Schank, R. C., *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*, Cambridge University Press, New York, 1982.
10. McCarthy, J. and Hayes, P. J., Some philosophical problems from the standpoint of artificial intelligence, *Machine Intell.*, 1969, **6**, 133–153.
11. McCarthy, J., Circumscription: A form of non-monotonic reasoning, *Artif. Intell.*, 1980, **13**, 1–2, 27–41.
12. Newell, A. and Simon, H., *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ, 1972.
13. Patten, T., Geis, M. and Becker, B., Toward a theory of compilation for natural language generation, *Comput. Intell.*, 1992, **8**(1), 77–110.
14. Chandrasekaran, B., Roles of logic in Artificial Intelligence, *Vivek: A Quarterly in Artificial Intelligence*, National Centre for Software Technology, Bombay, 1991, **4**(2), 13–15.
15. Chandrasekaran, B., Generic tasks in knowledge-based reasoning: high-level building blocks for expert system design, *IEEE Expert*, 1986, **1**(3), Fall, pp. 23–30.
16. Minsky, M., A framework for representing knowledge, *The Psychology of Computer Vision* (ed. Winston, P. H.), McGraw Hill, New York, 1975, pp. 211–280.
17. Minsky, M., *The Society of Mind*, Simon and Schuster, New York, 1986.
18. Fodor, J. A. and Pylyshyn, Z. W., Connectionism and cognitive architecture: A critical analysis, *Cognition*, 1988, **28**, 3–71.
19. Shastri, L., Connectionism and the computational effectiveness of reasoning, *Theor. Ling.*, 1990, **16**(1), 65–87.
20. Pollack, J. B., Recursive distributed representations, *Artif. Intell.*, 1990, **46**(1), 77–105.
21. Pollack, J. B., Review of Unified Theories of Cognition, in *Artif. Intell.*, 1993, in press.
22. Skarda, C. A. and Freeman, W. J., How brains make chaos in order to make sense of the world, *Behav. Brain Sci.*, 1987, **10**, 161–195.
23. Crutchfield, J. P. and Young, K., Computation at the onset of chaos, in *Computation, Entropy and the Physics of Information* (ed. Zurek, W.), Addison-Wesley, Reading, MA, 1989.
24. Fodor, J. A., *The Modularity of Mind: An Essay on Faculty Psychology*, MIT Press/Bradford Books, Cambridge, MA, 1983.
25. Clancey, W. J. and Roschelle, J., Situated cognition: How representations are created and given meaning, Technical report, Institute for Research on Learning, Palo Alto, CA 94304, USA, 1991.
26. Iran-Nejad, A., The schema: A long-term memory structure or transient functional pattern, in *Understanding Readers' Understanding: Theory and Practice* (eds. Tierney, J. *et al.*), 1980, Lawrence Erlbaum, Hillsdale, 1987.
27. Harnad, S., The symbol grounding problem, *Physica*, 1990, **D42**, 335–3466.

ACKNOWLEDGEMENTS. B. Chandrasekaran's work in the preparation of this paper was supported by US Defense Advance Research Projects Agency via contract F-49620-89-C-0110, monitored by Air Force Office of Scientific Research. We thank Tom Bylander, John Josephson and Jordan Pollack for their comments on a draft of this paper, and Prof. Narasimhan for the invitation to write this article.