13 The Frame of Reference Problem in the Design of Intelligent Machines

William J. Clancey

Senior Research Scientist, Institute for Research on Learning, Palo Alto, California

CHANGE AND CONFLUENCE IN COGNITIVE SCIENCE

What accounts for the regularities we observe in intelligent behavior? Many cognitive scientists would respond, "Mental structures which are representations, symbols of things in the world." Since at least the mid–70s there has been widespread agreement among cognitive scientists that models of a problem-solving agent should incorporate knowledge about the world and some sort of inference procedure for interpreting this knowledge to construct plans and take actions. Research questions have focused on how knowledge is represented in computer programs and how such cognitive models can be verified in psychological experiments.

But we are now experiencing increasing confusion and misunderstanding as different critiques are leveled against this methodology and new jargon is introduced (e.g., "not rules," "ready-to-hand," "background," "situated," "sub-symbolic"). New robotic research is founded on the idea that knowledge does not consist of objective representations (maps) of the world; conversely, other researchers define rational behavior in terms of an observer's supposedly objective descriptions of a task environment. This diversity of approaches takes us back to fundamental issues about the nature of perception, theories, and system dynamics.

There have been many philosophical arguments posed against cognitive science and AI research over the years; what reason is there to suppose that we are making progress now on these complex issues? Most striking is the convergence of ideas and new approaches over the past five years:

- The long-standing criticism by Dreyfus (1972), for example, has been joined by

In K. vanLehn (Ed.), Architectures for Intelligence: The Twenty-Second Carnegie Mellon Symposium on Cognition (pp. 357–424). Hillsdale, NJ: Lawrence Erlbaum, 1992.

insiders (Clancey, 1987a; Winograd & Flores, 1986), and is reflected in sharply divergent, new approaches by previously staid proponents of AI (Anderson, chap. 1 in this volume; Brooks, chap. 8 in this volume; Cohen, 1989; Rosenschein, 1985);

- Neural net research has reminded us of the extent of the gap between neurobiology and cognitive science models, while new hardware and programming techniques have enabled a resurgence of network modeling (Edelman, 1987; Rumelhart et al., 1986);
- Cognitive science itself has flourished and succeeded in including social scientists within the community, and their methods and analyses often starkly contrast with the AI view of human knowledge and reasoning (Lave, 1988; Suchman, 1987). They place increasing emphasis on representation construction as an activity within perceptual space, organized by social interaction (e.g., Allen, 1988), not something in memory that precedes speaking, drawing, or action in general.

Criticisms of cognitive science and AI may often fail to be effective because they aren't sufficiently grounded in computational modeling terminology and may even appear to be compatible with existing programs. For example, the current buzzword "situated" might just mean "conditional on the input data of particular situations"; hence all programs are situated. Moreover, the discourse of other intellectual traditions may appear incoherent to cognitive scientists; consider for example the claim that "representation must be based on interactive differentiation and implicit definition" (Bickhard & Richie, 1983). Experienced AI researchers believe that an engineering approach is essential for making progress on these issues. Perhaps the most important reason for recent progress and optimism about the future is the construction of alternative cognitive models as computer programs, the field's agreed basis for expressing theories:

- The AI-learning community is focusing on how a given ontology of internal structures—the designer's prior commitment to the objects, events, and processes in the world—enables or limits a given space of behavior (e.g., the knowledge-level analyses of Dietterich, 1986; Alexander et al., 1986);
- New robots ("situated automata") demonstrate that interpreting a map of the world isn't required for complex navigation; instead, maintaining a relation between an agent's internal state and new sensations enables simple mechanisms to bring about what observers would call search, tracking, avoidance, etc. (Agre, 1988; Braitenberg, 1984; Brooks, chap. 8 in this volume; Rosenschein, 1985; Steels, 1989);
- Neural networks, incorporating "hidden layers" and using back-propagation learning, provide a new means of encoding input/output training relationships, and are suggestive (to some researchers at least) of how sensory and motor learning may occur in the brain (Rumelhart, McClelland & the PDP Research Group, 1986).

In essence, this new research lead us to reconsider how the internal states in an agent derive from the *dynamics of a physical situation*, relegating an observer's later descriptions of the patterns in the agent's behavior (what has been called "the agent's knowledge" or a "knowledge-level description") to a different level of analysis. That is to say, this new research suggests that we reclassify many existing cognitive models as being *descriptive and relative to an observer's frame of reference*, and not isomorphic or literally identical to physical mechanisms internal to the agent that cause the observed behavior.

By systematically analyzing these alternative architectures, placing them in ordered relation to each other, we should be able to articulate distinctions that the researchers couldn't accomplish alone. The result will be a better understanding of the diverse approaches to "situated cognition" and "neural networks" research, contrasted against conventional AI architecture research. Thus, understanding a new approach and reconceptualizing the claims behind a traditional approach will arise together.

SCOPE AND OUTLINE OF THIS CHAPTER

When I began gathering notes for this chapter and preparing my symposium presentation a year ago, I had a good notion of how to describe the traditional approach. I believed and still do that knowledge engineering can be fruitfully characterized as a methodology for modeling processes qualitatively, and that this is in fact the main technique and lasting contribution of AI programming, regardless of what we later discover about how the human mind works or how to build intelligent machines. I am particularly concerned that we not lose sight of this modeling methodology as a legitimate, separate discipline, and to this end have recounted the main ideas in a series of papers (Clancey, 1983a, 1985, 1987a, 1989).

It should therefore be clear that the purpose of this chapter is not to undermine or dismiss knowledge engineering, that is, the idea of representing knowledge in computer programs in order to construct useful computer tools. Rather, my analysis aims to improve our understanding of how expert systems relate to human knowledge, so qualitative modeling methods can be more systematically applied, improved, and used appropriately. However, the view that knowledge engineering is fundamentally the development and application of a modeling methodology implies that AI researchers must clearly identify whether their architecture is intended to be a contribution to knowledge engineering, or is to be taken as a general model of an intelligent agent's functional architecture (to use Pylyshyn's terminology [Pylyshyn, 1984]). That is, I draw a distinction between knowledge engineering and the study of intelligence and believe that evaluation of research, and hence progress, hinges on clearly committing to one or the other. It is fine to aim for both, but the respective contributions must be separated out.

One reason for bringing up the knowledge engineering vs. study of intelligence distinction here is to make explicit the knowledge-engineering contributions of GUARDIAN, specifically the ACCORD framework for representing knowledge and control useful for configuration tasks (Hayes-Roth, Hewett, Vaughn, Johnson, & Garvey, 1988), which has no apparent counterpart in PRODIGY, SOAR, or THEO. This curiosity might be stated as the question, What is Hayes-Roth doing? Two other challenges are posed by the papers: To resolve formalization phobia (what is Genesereth doing?) and to integrate the situated automata work with the rest of the field (what is Brooks doing?). Somewhat surprisingly, these are related: We will find that understanding Brooks' insight requires a significant reinterpretation of the observer-relativity of knowledge-level (KL) descriptions, and this in turn places a new primacy on methods of formalization. In brief, we will conclude that KL representations are a theoretician's formalizations and should therefore be stated in a proper mathematical notation; they are not to be automatically identified with physical structures possessed by the agent being studied or designed. Of course, the beauty of our investigative dilemma is that the theoretician is an agent himself, and these representations reflect his or her beliefs. In this recursion we will reconsider the fundamental nature of the understanding process and how it relates to perception, memory, and representation.

This commentary can therefore be viewed from three perspectives:

1. an attempt to integrate the symposium papers;
2. in the style of Pylyshyn, an attempt to lay bare assumptions, take a stand on foundational issues, relate current work to classical philosophical issues, and draw out the implications for the study of intelligence;
3. a relativistic reinterpretation of KL analysis, which synthesizes analyses by Newell, Pylyshyn, and Dennett.

The chapter is organized as follows:

1. I start with a simple overview of superficial distinctions in the papers, emphasizing how the word "architecture" is interpreted: what is viewed as constraining an architecture, and what an architecture is expected to support.
2. I next focus on the work of Brooks and Anderson, which is problematic from the conventional AI paradigm, and present a framework for interpreting their points of view. Related work by Agre, Rosenschein, and Harold Cohen is discussed, focusing on the relation of design specifications to observed, emergent behavior.
3. I then discuss the relation of a KL description to a functional architecture de-

scription: What system is the KL about? What is the relation of the observer-theoretician to this system? How are perception and representation related?

With respect to these questions, I will argue the following about KL descriptions:

- A KL description is about a situated system, not an agent in isolation. That is, the systems level being described is above that of individual agents. Therefore, a knowledge-level description cannot be identified with (isomorphically mapped to) something pre-existing inside an individual head, but rather concerns *patterns that emerge in interactions the agent has in some (social) world*.
- A KL description is always ascribed by some observer, and so is relative to the observer's frame of reference and is inherently subjective. Therefore, a KL description *is only created when an act of perception has occurred* (otherwise there is no observation).
- A KL description must be expressed in some perceivable medium. Knowledge representations are always and only expressed as perceivable statements and drawings, including silent speech and imagined visualizations (mental imagery). Therefore, a theoretician's KL description cannot be identified with (isomorphically mapped to) something pre-existing inside the observer's head, it *physically exists only in the observer's statements, drawings, computer programs, or modeling medium in general*.

Further claims can be made about how KL descriptions affect later behavior. In general, statements about the world constitute an observer-agent's way of adding information about the current situation. This process is *reflective*—objectifying by commenting on ongoing activities—and it is *perceptual* in the way it leads the agent to organize the world in a new way (manifested in a new way of talking about the world). Thus, I follow cognitive science in emphasizing the generation and use of representations, but view them as something that goes on in "perceptual space," not as subconscious manipulations of grammars or other interpreted descriptions, and emphatically not via storage and indexing of preconceptions ("knowledge"). This idea stems from Bartlett's analysis of memory; its detail and nontraditional view takes us too far afield for full treatment in this paper. The idea is briefly introduced and elaborated in a separate monograph.

In summary, a year of considering the symposium papers, supported by extensive reading in fields outside AI, leads me to conclude that it is now possible to integrate views that heretofore were viewed as discrepant or even incoherent. The writers that have influenced me the most (roughly in order of consideration) are Tyler (1978), Newell (1982), Winograd (1986), Bartlett (1977), Bateson (1988), Gregory (1988), Pylyshyn (1984), and Dennett (1988). The fundamental problem I face is that I believe I have a coherent way of reformulating what we are doing, but it is so dramatically different from conventional approaches and so intermingled with many difficult problems, it is impossible to communicate convincingly without stating precisely how it influences the design of computer programs. This chapter is a contribution to the theory that bridges between how my colleagues and I have talked about AI architectures and a future, different way of designing intelligent machines.

The arguments here suggest that we need to dramatically reformulate how we talk about memory, perception, and representation. Current work on situated automata and analyses in other fields suggests that we focus on how simple mechanisms in interaction with a complex environment produce behaviors that are perceived by observers as recurrent patterns. In effect, researchers are rediscovering the heuristic value of Simon's "ant on the beach" metaphor (Simon, 1969), which in its current interpretation suggests the following changes in perspective:

- a better appreciation of the nature and capabilities of AI's qualitative modeling methodology, relative to cybernetics and new statistical approaches,
- viewing computation in terms of self-organizing processes,
- viewing information as perceptual as opposed to objectively defined,
- viewing interaction as dialectic as opposed to linearly causal or objectively characterizable from any one frame of reference,
- viewing science as subjective and socially organized, and hence
- a different view of the observer-theoretician's relation to a machine's design and behavior.

Obviously, this cannot be fully communicated or worked out in one paper. However, as a beginning, I believe that the "design stance" (Dennett, 1988)—specifically, analyzing the design of robots from the frames of reference of designer, machine, environment, and observer—is the approach that will provide the bridge between our current programs and the ideas of the writers I have cited above.

ARCHITECTURAL VIEWS OF COGNITION

> … [We] need to expose what by now have become some stable intuitions which cognitive science researchers share and which to a great extent guide their work, but which have been articulated only in the most fragmentary ways. (Pylyshyn, 1984, p. xix)

It is useful to begin with a simple overview of superficial distinctions in AI architecture research, emphasizing how the word "architecture" is interpreted; what is viewed as constraining an architecture and what it is expected to be able to support. In this way, we can get started on Pylyshyn's central challenge to acknowledge our philosophical commitments. Here I focus on the architectural implications of building knowledge into a program, as emphasized in knowledge engineering, versus the evident emphasis in the symposium papers on generating knowledge (SOAR), expressing it in more primitive forms (THEO), or perhaps avoiding it entirely (Anderson, Brooks). This serves as an introduction to more basic questions about the nature of memory and information.

Architectural Levels

Figure 13.1 organizes the AI architectures described at the symposium, according to the operations that individual researchers assume are "hardwired" or directly supported by the architecture.[1] The nesting of the three levels corresponds to a composition of functions, such that inner functions are invoked to carry out tasks required by the higher levels, GUARDIAN and INSECTS emphasize reasoning functions corresponding to the interaction of the agent with the world. The operations of their architectures are designed to support the program's role in its environment (e.g., monitoring a large amount of changing data or following a wall in a room). The internal operations of the architecture are stated in terms of this interaction between the agent and the world (e.g., the need to monitor a great deal of data simultaneously, the need to maintain a constant reading on a sensor).

More specific capabilities, which we view as lying wholly inside the agent, construct a model of the world in order to take action. It is this level—the process of forming and testing descriptions of processes occurring in the world—that I have sought to make explicit in HERACLES (Clancey, 1987a). These functions correspond to what we commonly call "model building" or "understanding." This basic pattern of inference capabilities is revealed when we abstract domain-specific inference rules and state the control knowledge separately in terms of operators for constructing a situation-specific model of the domain system being reasoned about (Clancey, 1989), such as in HERACLES (diagnostic operators) and ACCORD (configuration operators). Knowledge engineers emphasize this system-modeling level when relating tools, tasks (e.g., diagnosis and configuration), and representations (e.g., classification or causal networks).

Finally, most of the architecture descriptions at this symposium focus on the innermost functional layer, the representation, inference, and control constructs that enable access and search of knowledge representations, including reflection and learning. In many respects, the functional requirements of the outermost layers propagate down to these more basic functions, constituting a machine specification in terms of memory, processing states, transitions between states, scheduling, and interrupt/resume capabilities. Here AI architecture research clearly departs from the concerns of routine knowledge engineering. Researchers are wrestling with foundational issues: What sort of machine could automatically generate the modeling and interactional functions that are constructed by ad hoc means in expert systems? What are the basic memory, reflection, and learning requirements of an architecture that supports intelligent behavior?

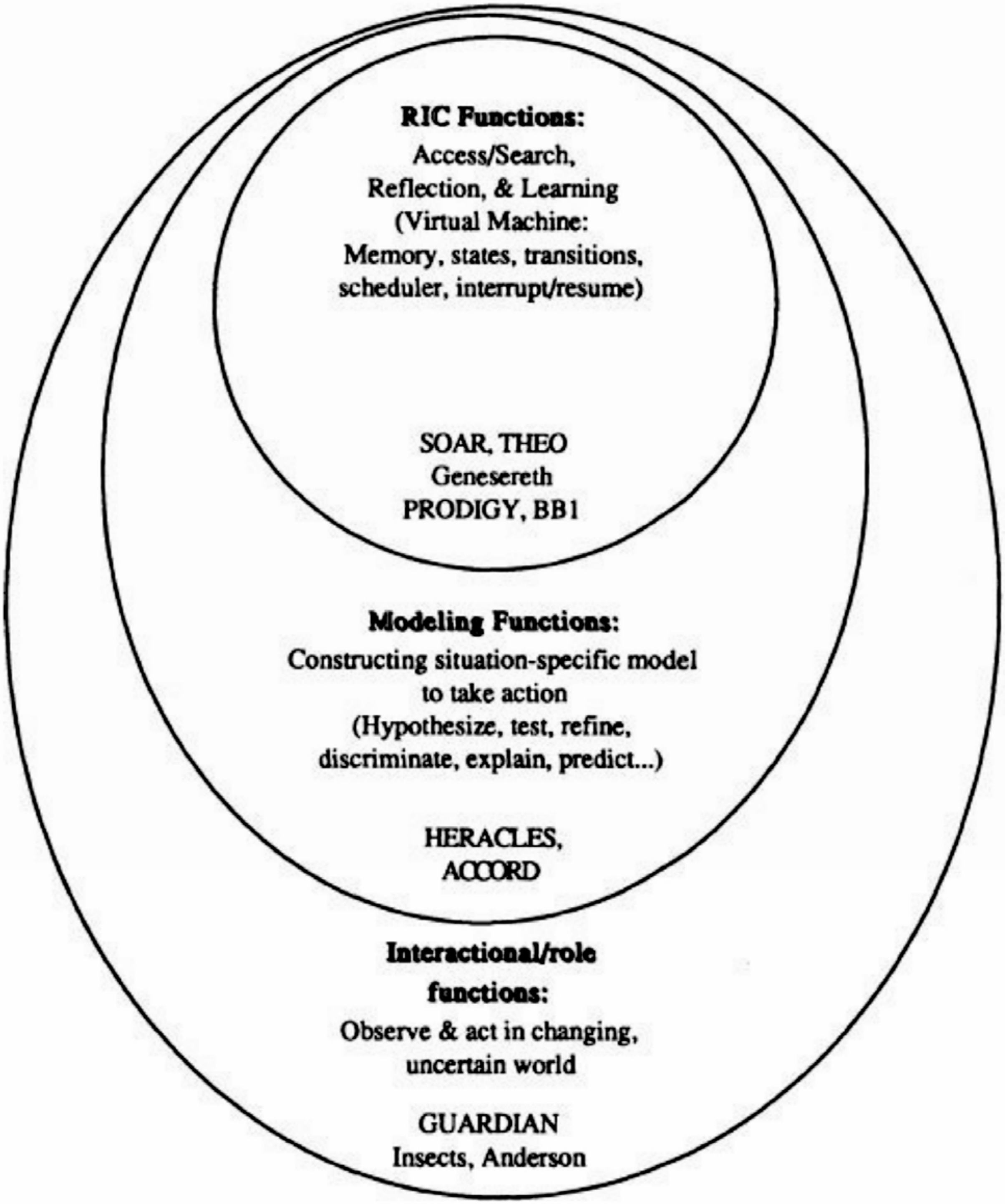


FIG. 13.1. Functional capabilities emphasized by different researchers, shown integrated as layers in a single machine.

The distinction between knowledge engineering and the study of intelligence becomes increasingly important in theories of learning, where we move from *building in* knowledge as structures (the approach of engineering efforts, such as GUARDIAN and HERCULES) to *generating or inferring* knowledge from problem-solving experience (SOAR, THEO, PRODIGY). It is just this tension—wanting to generate rather than build in—that motivates Brooks' design of INSECTS (abandoning the idea of stored knowledge structures entirely) and design trade-offs considered by Anderson and Genesereth (attempting to formally derive how an agent's behavior is constrained by the nature of the environment and task). Similarly, this tension between building in and generating motivates how the search process in SOAR is formalized (indirectly generated by a cycle of operator/operand hypothesizing and testing, rather than a directly-programmed search of static structures in memory). Here also we have competing views of the innermost level of mechanism: In contrast with SOAR'S flat production rule memory, Mitchell hypothesizes that hierarchical structures are directly supported by "schema" primitives for storing knowledge. This distinction between the structure of memory and the representations manipulated during problem solving in fact becomes our central concern when we attempt to relate the knowledge-level descriptions to the functional architecture.

"Generic Systems" Versus Functional Architecture

Pylyshyn provides a simple, useful definition for talking about architectural levels that isn't commonly used by AI architecture designers. By definition, the functional architecture concerns a level of mechanism that is *cognitively impenetrable* (Pylyshyn, 1984). Like the relation between VLSI circuits and computer programs, this level of mechanism is independent of the machine's outward behavior. Programs do not inspect or deliberately change structures at this level, nor is the program's operation dependent on how operations are implemented at this level. To turn it around, an agent's knowledge is assumed to be cognitively penetrable because by definition an agent's beliefs and goals have a direct effect on behavior. The distinction is useful here, because it suggests that operators like "test a hypothesis," which are articulated by an agent—and hence are cognitively penetrable—are not part of the functional architecture, but constitute material ("knowledge") that is represented (learned) and manipulated by the functional architecture. A functional architecture specification must be general, independent of any particular content (Pylyshyn, 1984, p. 36); its purpose is to support a variety of content and behavior, not to implement it directly in hardwired form.

This definition of functional architecture helps explain why the symposium papers don't describe their programs in terms of the task-specific operators that are emphasized in knowledge engineering. Specifically, this would explain why Hayes-Roth didn't mention the ACCORD framework that is built on top of BB1 and included in

GUARDIAN, used to represent the specific knowledge and control structures required by the intensive-care unit problem. Instead, Hayes-Roth tells us about the input-output transducers and reflective monitoring cycle that constitute the most general, knowledge-independent components of her program, that is, the functional architecture.

But to turn this around, we have seen repeatedly in "generic expert system tools" like HERACLES and ACCORD, control structures that construct models of specific physical systems in the world, for diagnosis and for configuration respectively. Viewed a step back, these programs model a general understanding process, strongly guided by beliefs and goals, but described at a domain-general level. For example, the operations of "yoking," "refining," "confirming," suggest that the functional architecture should provide some direct support for achieving coherence, directly driving the problem-solving process. Describing problem solving in terms of impasses and operators may be too low-level. Might the functional architecture directly support what we commonly call "story understanding"? It is noteworthy that the view that representations are not stored but generated (or reinterpreted) freshly for every new problem would put primacy on such a capability for achieving coherence by an automatic process below the knowledge level. The idea that knowledge representations are stored and problem solving involves combining and matching primitive elements may have led researchers to inadvertently minimize this problem.[2]

To summarize, most researchers apparently take the symposium title, "Architectures for Intelligence," to be a charge to describe what Pylyshyn calls the functional architecture. This emphasis on functional architecture follows naturally from the value and priority AI researchers place on structures and processes that are generative. While we all tend to agree that successful problem solving depends on having a lot of knowledge, intelligence per se is to be characterized in terms of the memory, sensory, and learning capabilities that allow this knowledge to be acquired, stored, and accessed effectively. As Pylyshyn puts it, we don't want to find ourselves "mimicking the most frequent behavior rather than inferring the underlying mechanisms" (1984, p. 85). Nevertheless, there is some question about how directly the "explanation" or "modeling" operators in generic systems like HERACLES and ACCORD are to be supported by the functional architecture. Chandrasekaran's group has recently undertaken to redescribe MDX in these terms (Johnson et al., 1989); in general the question is not raised by the symposium papers. As a stable, widely-referenced theoretical level that appears to be between knowledge-level descriptions and theories of memory and learning, this "comprehension process" may be an important clue of how current architectures need to be improved.

Maps and Learning

Given that knowledge is about the world, one way of characterizing the representations used by a program is the extent to which it uses map-like representations of the world, and secondarily whether these are these stored or generated. This is illustrated in a simple way by the following spectrum:

maps—(more learning required) → no maps

"Maps" here refers to the idea of building into the program descriptions of the world. As the program reasons or manipulates objects in the world, it is constantly comparing a situation-specific description to an idealized, map-like description of how the world is supposed to appear or operate. The architectures of both PRODIGY and GUARDIAN assume that a great deal of such knowledge is provided by a person who constructs the program. As stated before, routine knowledge engineering is distinguished from the study of intelligence by not requiring commitments in this respect: Building in maps is just fine, the problem is to do it efficiently.

Aiming for machines with generative intelligence capability and not just built-in knowledge, Brooks (and his colleagues in arms, Rosenschein and Agre) → reject this idea; their models disavow or minimize the agent's use of map-like descriptions of the world. Interestingly, the "no maps" approach has been advocated by people following bottom-up construction strategy like Brooks, as well as by people working top-down from studies of social organizations (Lave, 1988; Suchman, 1987). Thus, in some sense, the emphasis on maps in traditional AI is being squeezed by both the biological and social bands, as Newell labels these perspectives (Newell, 1990).

Most researchers strive for a happy medium, assuming that the problem solver needs maps, but they must be learned. In this respect, we can distinguish between researchers who rely strongly on psychological data (as in SOAR) and those who collect and study knowledge abstractions and later formulate the underlying processes of memory and learning that form such abstractions, exemplified by the generic system approach.

FRAMES OF REFERENCE FOR DESIGNING INTELLIGENT MACHINES

> When we analyze a mechanism, we tend to overestimate its complexity. In this uphill process of analysis, a given degree of complexity offers more resistance to the workings of our mind than it would if we encountered it downhill, in the process of invention.... The patterns of behavior described in vehicles (just illustrated) ... undoubtedly suggest much more complicated machinery than that which was actually used in designing them. (Braitenberg, *Vehicles*, 1984, p. 21)

The work of Brooks and Anderson poses a dilemma for someone trying to understand the current state of AI architecture research. In this section, I provide a framework for interpreting their points of view, but I do this by presenting related work that was not represented at the symposium—research by Rosenschein, Agre, and Cohen. This framework (and the associated Fig. 13.2) will provide the backbone for the discussion in the rest of the paper. The central thesis is that we need a better way of talking about the design of machines. We must make explicit the different roles, point of view, and causal properties of: the designer, the specification of how the machine is to work, processes in the operational environment, and the observer who later describes and theorizes about the machine's behavior. In this respect, it is useful to talk about robots as designed artifacts, not just "intelligent agents," making the frames of reference of the designer, environment, and observer an integral part of our theory. In particular, I believe that this perspective will enable us to restate the rhetoric of "situated cognition" in terms of its implications for AI architectures. We will then be in a position to reconsider how human memory, perception, and learning are different from present-day machines.

My approach here is to characterize the ontological commitments of alternative architectures: What facts about the world are built into each program? Two useful, related questions are: *Who owns the representations* (robot, designer, or observer)? *Where's the knowledge* (in robotic memory, in a designer's specification, or in our statements as observers)? Throughout, I will use the term "robot" to emphasize that we are dealing with designed artifacts intended to be agents in some physical environment. I believe we need to distance ourselves from our programs, so we can better understand our relation to them. Our orientation here is not of philosophical discourse in the abstract, but rather an attempt to find an appropriate language to describe existing robots and the process by which they are designed, so the engineering methods for building them are clear enough to allow us to order, compare, and improve them.

The Problem: The Ontological Commitments of Plans

When we examine the situated automata research of Brooks, Rosenschein, and Agre, we find a striking emphasis on the *nature of planning*, focusing on the precommitments made by the designer of the computer program. These commitments are characterized as *ontological*, that is, they concern the designer's view of the kinds of objects and events and their properties that can occur in the robot's world. The researchers arrive at this focus from distinctly different considerations and objectives. Agre (1988) and Kaelbling (1988) emphasize the resource and information limitations of real-time behavior–deliberation between alternatives must be extremely limited and many details about the world (e.g., will the next closed door I approach open from the left or the right?) can't be anticipated by the designer or by the robot.

Rosenschein (1985) found that formal analyses of knowledge bases are problematic —how can knowledge structures in a computer program be related in a principled way to the world, when their meaning depends on the designer's changing interpretations of what the representations mean? Cohen (1988) was wedged in a designer's conundrum: Since AARON is supposed to be producing new drawings of people standing in a garden, how could he build in a representation of these drawings before they are made? Cohen was face to face with the ultimate ontological limit of traditional cognitive models: Any description of the world that he builds in as a designer will fix the space of AARON'S drawings. How then can a robot be designed so it isn't limited by its designer's preconception of the world? If such limitations are inevitable for designed artifacts, how can the specification process be accomplished in a principled way? Abstracting the work of Rosenschein, Brooks, Agre, and Cohen, here are four perspectives on these questions.

Classical Planning—Knowledge is in the Robot's Memory

In most AI/cognitive science research to date, the descriptions of regularities in the world and regularities in the robot's behavior are called "knowledge" and located in the robot's memory. A robot preferably uses a declarative map of the world, planning constraints, metaplanning strategies, etc. This view is illustrated especially well by natural language programs, which incorporate in memory a model of the domain of discourse, script descriptions of activities, grammars, prose configuration plans, conversational patterns, etc. Aiming to cope with the computational limits of combinatoric and real-time constraints, some researchers are reengineering their programs to use parallel processing, partial compilation, failure and alternative route anticipation, etc. These approaches might incorporate further ontological distinctions (e.g., preconceptions of what can go wrong), but adhere to the classical view of planning.

Knowledge is in the Designer's Specification

Rosenschein (1985) introduces an interesting twist. Besides using efficient engineering (compiling programs into digital circuits), his methodology explicitly views the robot as a designed artifact. He formally specifies robotic behavior in terms of I/O and internal state changes, gaining the advantages of internal consistency and explicitly-articulated task assumptions. The problem of building a robot is viewed as an engineering problem, nicely delineating the designer's relation to the robot and the designed behavior.

Knowledge is not incorporated as data structure encodings; it is replaced by a design description that specifies how the state of the machine and the state of the environment should relate. Thus, knowledge is not something placed in the robot, but a theoretical construct used by the designer for deriving a circuit whose interactive coupling with its environment has certain desirable properties. These "background constraints ... comprise a permanent description of how the automaton is coupled to its environment and are themselves invariant under all state changes" (Rosenschein, 1985, p. 12). Regardless of how program structures are compiled or transformed by learning, the program embodies the designer's ontology. Rosenschein's formal analysis can be contrasted with Brooks' analogous, but ad-hoc constructive approach (functionally–layered, finite–state automata) (Brooks, see chap. 8 in this volume); Brooks assembles circuits without spelling out his ontological commitments to world objects, machine states, and relations among them.

Knowledge is the Capacity to Maintain Dynamic Relationships

Agre (1988) views the ontological descriptions built into his robot as *indexical* and *functional*. That is, descriptions of entities, representations of the world, are inherently a combination of the robot's viewpoint (what it is doing now) and the role of environmental entities in the robot's activity. For example, the term *the–ice–cube–that–the–ice–cube–I–just–kicked–will—collide–with* combines the indexical perspective of the robot's ongoing activity ("the ice cube I just kicked") with a functionally-directed visualization (one role of ice cubes is for destroying bees).

Agre demonstrates that an internal representation of the world needn't be global and objective, in the form of a map, but—for controlling robotic movements at least —can be restricted to ontological primitives that relate the robot's perceptions to its activities. There are two more general claims here: (a) that representations are *inherently* indexical and functional (that is, a rejection of the correspondence theory of truth, which holds that representations are objectively about the world) and (b) that the robot can get by with mostly local information about the activity around it. Agre is showing us a new way of talking about knowledge base representations, and demonstrating that a different perspective, that of "dynamics" as opposed to "objective description," can be used for constructing an ontology. It is arguable that Agre's programs aren't fundamentally different from conventional AI architectures; the use of hyphenation just makes explicit that internal names and variables are always interpreted from the frame of reference of the agent, relative to its activities. The important claim is metatheoretical: All representations are indexical, functional, and consequently subjective.

Knowledge is Attributed by the Observer

Cohen's work nicely articulates the distinction between designer, robot, behavioral dynamics, and observer's perception that Rosenschein, Agre, and Brooks are all wrestling with.

> AARON draws, as the human artist does, in feedback mode. No line is ever fully planned in advanced: it is generated through the process of matching its current state to a desired end state. ... All high-level decisions are made in terms of the state of the drawing, so that the use and availability of space in particular are highly sensitive to the history of the program's decisions. (Cohen, 1988)

Notably, AARON'S internal, general representation of objects is sparse; it doesn't plan the details of its drawings; and it maintains no "mental photograph" of the drawing it is producing. There is no grammar of aesthetics; rather 3-d properties, *as attributed by an observer*, emerge from following simple 2-d constraints like "find enough space." The point is made by Agre, in saying that the purpose of the robot's internal representation is "not to express states of affairs, but to maintain causal relationships to them" (1988, p. 190). The internal representations are not in terms of the "state of affairs" perceived by an observer, but the immediate, "ready-at-hand" dynamics of the drawing process (again, the terms are indexical/functional, e.g., "the stick figure I am placing in the garden now is occluded by the object to its left"). The robot's knowledge is not in terms of an objective description of properties of the resultant drawing, rather the ontology supplied by Cohen characterizes the relation between states of the robot (what it is doing now) and how it perceives the environment (the drawing it is making).

Who Owns the Knowledge?

The above analyses demonstrate the usefulness of viewing intelligent machine construction (and cognitive modeling in general) as a *design problem*. That is to say, we don't simply ask "What knowledge structures should be placed in the head of the robot?" but rather, "What sensory-state coupling is desired and what machine specification brings this about?" Figure 13.2 summarized the elements of this perspective.

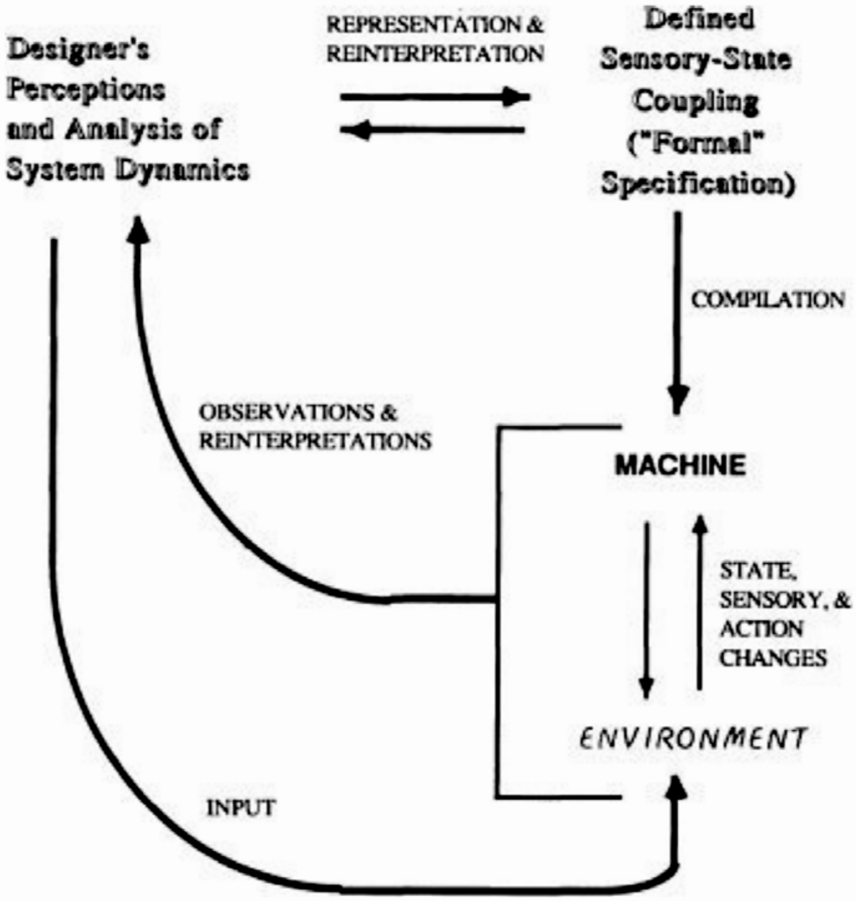

Designer's Perceptions and Analysis of System Dynamics
REPRESENTATION & REINTERPRETATION
Defined Sensory-State Coupling ("Formal" Specification)
COMPILATION
OBSERVATIONS & REINTERPRETATIONS
MACHINE
STATE, SENSORY, & ACTION CHANGES
ENVIRONMENT
INPUT

FIG. 13.2. Relation of designer's theory to machine and coupling.

Briefly, the figure illustrates that a machine specification is a representation that derives from the designer's interpretation of the machine's interaction with its environment. No "objective" descriptions are imputed—how the machine's behavior is described is a matter of selective perception, biased by expectations and purposes. The recurrent behavior attributed to the machine by the observer/designer is a matter of how people talk about and make sense of the world. Furthermore, the specification—usually an external representation in the form of equations and networks—is itself prone to reinterpretation: What the specification means (its semantics) cannot be described once and for all. The validity of the specification lies in the over-all coherence of the designer's goals, the machine's behavior, and what the designer observes.

Cognitive science research has to date not been driven by such metatheoretical analyses. Most researchers have simply assumed that the world can be exhaustively and uniquely described as theories, and that learning itself involves manipulating theories—a correspondence view of reality. But a radically different point of view has played a central role in methodological analyses in fields as diverse as anthropology and physics. For example, one interpretation of Heisenberg's Uncertainty Principle is that theories are true only with respect to a frame of reference. AI and cognitive science research has been based on the contrary point of view that theories (representations and language) correspond to a reality *that can be objectively known* and knowledge consists of theories; consequently, alternative design methodologies have rarely entered the discussion (this is discussed further later on).

To recapitulate the emerging alternative approaches to cognitive modeling: In classical planning, epitomized by present-day expert systems, descriptions of regularities an observer will perceive in the robot's interaction with the world are stored in the robot's memory and interpreted as instructions for directing the robot's behavior. Rosenschein breaks with this idea, instead compiling a state-transition machine from a designer's specification of the desired coupling between machine and environment. Agre's work reminds us that regardless of what compilation process is used, a program still embodies a designer's ontological commitments, and these are fruitfully viewed as indexical and functional with respect to the robot's activity. As an artist, reflecting on the robot's behavior, Cohen reminds us that this indexical, functional theory is to be contrasted with an observer's statements about the robot's behavior. The essential claim is that representations in computer programs are not objective—true because they correspond to the world—but inherently indexical, functional, relationships between the agent and the world that a designer specifies should be maintained. Moving from engineering "knowledge structures" in an agent to designing on the basis of state-sensory coupling constraints is a significant theoretical advance.

However, situated automata research doesn't get to the heart of the matter: Each program still embodies the designer's ontology, which is neither fixed nor objective. Rosenschein, in particular, continues to speak of an objective physical reality, implying that perception is just a matter of processing data on fixed sensors in an axiomatic way (cf. Neisser, 1976). He fails to acknowledge that his coupling specification and background constraints are linguistic entities prone to change under his own interpretation, no less than knowledge structures built into a classical planning system. Formality is not gained by behavioral specification, because these specifications still embody the designer's perceptions of the robot's behavior and theory of the dynamics of the robot's interactions. Compilation into circuits only changes computational efficiency; the resultant physical structures formally correspond to the designer's original formal notations of "world conditions" and "behavioral correlations." And what these notations mean cannot be objectively specified.

Furthermore, while the robot's structural form is fixed after the design process, the coupling can be modified by human intervention. When a person interprets internal structures during the operation of the program (e.g., providing input by responding to the robot's queries), the coupling between robot sensation and action is changed. This interpretation is again an inherently subjective, perceptual process.

Viewing knowledge as relative to an observer/designer's perceptions of dynamic indexical-functional relations between an agent and its environment is indeed a major theoretical reconceptualization of the process of constructing intelligent agents. However, is a more radical stance possible? Further analysis might focus on the nature of the primitive ontology, specifically to restrict it to sensations inherent in the agent's peripheral sensors (if any) or to primitive perceptual structures that arise in the early developmental interactions of the agent and its environment.

From a strict sense, we could claim that the robots described above react to sensors, but never perceive, because they never form new ontologies, new ways of seeing the world. Driving this analysis would be the radical hypothesis that all perceiving is a form of learning and it is dialectically coupled to development of new physical routines. In this respect, it is highly significant that none of the above programs have any learning capability. We must explain how a string like "potentially-attacking-bee" could be created as a new way of seeing the world by the robot itself, rather than being a designed structure that determines its behavior in a fixed, programmatic way. How do we break away from modeling learning by grammatical reshuffling of grammars? In short, situated automata research has laid down the gauntlet: How far can we go in removing the observer–designer's commitments from structures built into the machine?

Implications for the Study of Intelligence

The above discussion is only an introduction to complicated issues that require considerable elaboration. My initial objective is simply to provide a way of organizing this diverse work so we can begin to see a larger picture. In essence we need a much better articulated theoretical framework for talking about computer programs and machine behavior, emphasizing interactional dynamics and the role of human perception and representational acts. Rather than dealing directly with topics like system dynamics and emergence, I will continue my approach of grounding the discussion in the symposium papers and associated computer programs.[3] Three central issues are introduced here, then elaborated in later sections of the paper: the importance of including a formal description of the environment in a KL-theory, the impossibility of exhaustively representing what a symbolic structure means, and the inherent subjectivity of information.

Making the Environment Explicit

First, I claim that Anderson, Rosenschein, Brooks, and Genesereth are converging on "understanding the nature of the problem being solved" by an agent (Anderson, chap. 2 in this volume), "framing" the information processing problem in a way that makes explicit the environment. We can view this work as a reaction against the complexity of AI architectures and an attempt to reground the study of intelligence in the behaviors we are seeking to explain. These researchers are looking more closely and asking, "What is the robot accomplishing from its point of view?" This question serves to refocus the description of behavior on more local, moment-by-moment interactions between the agent and the environment, as opposed to the much more varied and complex patterns of behavior an observer will see over time. The researchers of course have wildly different approaches for developing a new methodology: Anderson throws out any discussion of functional architecture; Brooks throws out discussion of representations(!); Rosenschein goes to the extreme of compiling his specification into an electronic network (as if to disavow any connection to the machine that results from his analysis); and Genesereth just forges ahead with a mathematical analysis of an assumed-objective world.

Symbolic Interpretation as Perceivable Commentary

Second, I will give an example from MYCIN that illustrates the problem Rosenschein is struggling with and serves as an introduction to a more protracted discussion of KL descriptions. In the epistemological study of MYCIN (Clancey, 1983b), I describe a problem called *concept broadening*, in which MYCIN concepts were reinterpreted as new rules were added to the system. Rather than introducing new (intermediate) concepts, knowledge engineers wrote new rules that used existing concepts in a more general way, broadening their meaning. For example, several rules were of the form, "If X then the organism is significant." These rules originally used information about cultures; for example, a positive culture from a normally sterile body site is significant. However, rules added later were based on other patient findings, for example, "If the patient has a high fever, then the organism is significant." After several rules were added, it became clear that there were two categories of evidence that made an organism significant, "evidence for infection" and "evidence of non-contaminated cultures." Thus we reinterpreted what "significant" meant in the *original* rules—what MYCIN knew changed because how we talked about the rules changed. This was Rosenschein's dilemma: How can we say that the program's knowledge causes its behavior to be intelligent when the knowledge changes under our interpretation as designers?

Computer representations are human utterances; they are interpreted with respect to a background, so our interpretation of what they mean is prone to change. In this respect, they are no different than what we say or write anywhere else. From our perspective as designers, representations don't have a fixed, noncontextual, meaning-separated nature. With every use and by every observer they are under interpretation (cf. Agre, 1988). However, the program itself is constrained to manipulating these same expressions formally (literally by their form, syntactically), not their semantic interpretation. So we can say what the term "significant" means, but MYCIN cannot. Even if MYCIN had a well-structured definition of the term, it wouldn't be able to define its primitive elements. And although the designers of MYCIN could define the term, the meaning was never fixed.

This discussion quickly takes us far afield, but I want to introduce the issue early on. The essential claim is that as we speak and interpret representations we are capable of doing something that today's computer programs cannot do. This is because of the nature of human memory, learning, and perception. Put somewhat coarsely, in people these capacities are combined in such a way that, for humans, to speak is to conceive something new each time. Furthermore, the interpretative action is going on in the outward sequence of our behavior: Meaning is never defined or preconceived and then translated into an "output statement"; meaning attribution occurs only in the ongoing commentary of one behavior referring to a earlier one. Indeed, it is our talking about another utterance that makes it a representation, specifically, by providing a context (what we say the representation is about). Crucially, each statement or phrase is generated by direct recombination of processes that generated past behaviors, not from representations that describe or label these processes. Obviously, the ongoing oral and written commentary reorients behavior, but these representations must all be perceived in order to exist and have any effect (silent speech and visual imagination included). I briefly expand on this theory later. My point here is to make clear that I have a definite synthesis I am working towards, for which the present exposition provides one path of support.

I want to underscore that the essential questions facing the study of intelligence today concern the nature of representations. We have not precisely enough described how the representations of designers and observers relate to the representations used by the machine, and indeed today's machines do not create or use representations in the way people do. The key points of my argument are:

- Representations are inherently generated and used in *sequences of behaviors*, as commentary on each other. We may point to a particular drawing for example and say it represents a geometry problem, but the representation is as much in our comment as in the original drawing.

- All semantic interpretation lies outside AI computer programs, just as it lies outside any written text, diagram, or code. We cannot build in a semantic map that definitely relates notations to the world. This is because the world is not an objectively fixed thing and because our experience from which our interpretations are drawn is always changing: There is no final statement, no definitive representation, that could be built into the program and that would say, "This is everything that this program means."[4]

Putting this together, it is apparent that our use of the term *representation* in referring to something in a program has been far too loose. The symbols in programs are representations in the sense that they are statements a person has made about something else seen, heard, etc. But to the program itself, these same statements are nothing more than tokens, forms or marks that are themselves about nothing and only manipulable by their shapes. To the AI computer program, every problem is like assembling a puzzle with the picture-side facing down. In short, my solution to the "semantic interpretation problem" (e.g., see Pylyshyn, 1984, p. 39) or "how can symbols in a computer program refer to the world" is to claim that tokens refer only by virtue of what we say about them. In some crucial sense, symbols don't refer, people do. Semantic interpretation cannot be captured by a map, rather it occurs only in ongoing outward behavior.

It is neither objective nor ever definitive. It is always relative to an observer and the purposes at hand.

Information as Relative

The final issue I want to introduce here concerns the notion of information. The idea of an information-processing analysis, which Anderson wants us to rededicate ourselves to, supposes that information is like a substance that every observer would objectively describe in the same way. As Reeke and Edelman (1988) put it, the AI view is that the "organism is a receiver rather than a creator of criteria leading to information" (p. 153). In describing a functional architecture, Pylyshyn rightly realizes the fundamental problem of separating what is fixed and given by the organism's input "transducers" and what can be attributed to inference. Ultimately, the problem surfaces in explaining the nature and origin of "primitive representations stored in memory" (e.g., Rosenbloom et al., see chap. 4 in this volume): By defining perception as something distinct from cognition and prior to conceptual inference, we have possibly grossly distorted how representations are created. Reeke and Edelman continue, "To place this problem [of finding a representation] in the domain of the designer rather than the designed system is to beg the question and reduce intelligence to symbol manipulation" (p. 147).

Information is relative to a point of view. There is no such thing as "all the information" in a particular situation. Information-processing analyses are strictly observer-relative. This does not detract from their analytic value; we just must be more careful in saying things like "the nature of the problem being solved" that we realize that "the problem" is our description of the situation and the information-processing formalizations are our conceptions as observer–designers, not something that necessarily resides in the head of the subject being studied.

With this preparation, we are now ready for a more detailed reconsideration of the nature of knowledge-level descriptions.

THE RELATIVITY OF KNOWLEDGE-LEVEL DESCRIPTIONS

> If we desire to explain or understand the mental aspect of any biological event, we must take into account the system—that is, the network of closed circuits, within which the biological event is determined. But when we seek to explain the behavior of a man or any other organism, this "system" will usually not have the same limits as the "self" ... mind is immanent in the larger system—man plus environment. (Bateson, *Steps Towards an Ecology of Mind*, 1972, p. 317)

In his response to Dennett's precis of the *Intentional Stance* (Dennett, 1988), Newell asks whether there is "something about the knowledge level that makes it more in the head of the analyst than atomic physics is in the head of the physicist?" (Newell, 1988, p. 521). Why does Dennett refer to the ascription of intentions as a stance, as if it were a maneuver by the observer-theoretician? Aren't intentions an objective property of the subject? Without getting bogged down in the subjective nature of physics (which we will briefly consider later), my objective here is to bring Newell's and Dennett's analysis together, primarily by making a strong interpretation of Newell's claims in his KL paper.

My foil in this discussion is the paper by Rosenbloom, Newell, and Laird presented at the symposium, "Towards the knowledge level in SOAR." Rosenbloom et al. begin by repeating some of the key ideas of Newell's KL paper: that the KL lies above the symbol level, that rationality is defined in terms of using knowledge to accomplish a goal, and that it is the knowledge, not the internal structures, that determine behavior. However, I claim that their ensuing discussion violates Newell's own view that a KL description is an observer's attribution. Rather, the SOAR program itself is referred to as a "knowledge-level system" (as if this where an inherent, objective property of the system in isolation) and the problem reduces to showing how SOAR'S structures and mechanisms provide "direct support" for knowledge. I claim that the essential matter is not "how does the architecture support knowledge," but rather, why would we ascribe knowledge to the behavior produced by such an architecture? How does our frame of reference as observers of the SOAR system, coupled to its environment, lead us to ascribe procedural competence or say that it has episodic knowledge? What are the requirements on our observations, our perceptual processes, and the sequence of behaviors available to be observed that provide the minimum support for such ascriptions?

Thus, "supporting the KL in SOAR" is as much a matter of devising an appropriate functional architecture, as devising an appropriate set of tasks, sequence of observations, and theoretical abstractions that could support our claim that SOAR has knowledge of a certain type. Indeed, we must describe the nature and properties of the perceptual and understanding process by which the theoretician sees patterns, names them, and explains them. Regarding the special kinds of knowledge we ascribe to SOAR, the question is not how to encode or store structures (Rosenbloom et al., chap. 4 this volume) that we would interpret as procedural or episodic. The KL is not a property of the architecture per se. Support for a KL, what makes it possible, comes as much from the environment, including the people who interact with and observe the program, as from the functional architecture. Claiming that knowledge of a certain type is a possible ascription that could be made about a given architecture requires specification of the world, tasks, and observers in which the architecture is embedded.

In what follows, I will characterize knowledge as structures created dynamically in the agent's working memory (and written or aural space in the world), distinguishing this from the KL description that constitutes an observer's claims about these knowledge structures and other aspects of the agent's behavior. The very idea that SOAR has four main kinds of knowledge is a theoretician's claim, which is not to be realized in a purely programmatic way as objective structure storage and retrieval, but as attributions about sequences of agent behavior. As Rosenbloom et al. (chap. 4, this volume) acknowledge at the end of their paper, "The most important missing aspect is the relationship between SOAR'S mechanisms and the principle of rationality." That is to say, the most important missing aspect in their analysis is the relationship between SOAR'S mechanisms, the patterns of behavior an observer will claim SOAR manifests, and how these patterns come to be interpreted as rational by the observer. In short, they have not explained the nature of rationality, because they have not explained the observer's theory-formation process.

Newell's Knowledge Level Reviewed

Newell defines knowledge to be "Whatever can be ascribed to an agent, such that its behavior can be computed according to the principle of rationality." The essential properties of knowledge that I wish to emphasize are as follows:

Knowledge, in the form of an observer's articulated KL description of an agent, is:

- observer-relative, not an objectively defined property or structure;
- external (perceived), not encoded or stored;
- constantly reinterpreted, not fixed or definitive;
- about a social system, not agents in isolation;
- about emergent phenomenon, not linear causal interactions.

An important corollary I will have much to say about later is that to claim another agent has knowledge is to be in a position yourself to be described as having knowledge. That is, the process of making a KL attribution can be studied as an example of how knowledge is created and used. To explain the observer's theory-formation process will be to describe a functional architecture and knowledge creation and use process that human observers and agents share.

Newell's (1982) paper, particularly Section 4.3 on knowledge, is an uncanny performance. As if sleepwalking, Newell negotiates all the difficult turns with ease, yet somehow we cannot believe he would say all of these things were he fully awake, talking directly about SOAR or other AI programs. The paper is a wonderful abstraction, reaching beyond where we are, and always trying to say what we would believe if we had all our wits about us.

Consider for example the following remarks:

- The knowledge level is not realized as a state-like physical structure, "running counter to the common feature at all levels of a passive medium." Knowledge isn't embodied in structures. (p. 105)
- "It seems preferable to avoid calling the body of knowledge a memory." (p. 101) "The total system (i.e., the dyad of the observing and the observed agents) runs without there being any physical structure that is the knowledge." (p. 107)
- "Knowledge of the world cannot be captured in a finite structure." (p. 107) "Knowledge can only be created dynamically in time." (p. 108) "Knowledge is not representable by a structure at the symbol level. It requires both structures and processes." (p. 125)
- Knowledge can only be "imagined as the result of interpretive processes operating on symbolic expressions." (p. 105)
- "Knowledge remains forever abstract and can never actually be *in hand*." (p. 125)
- "One way of viewing the knowledge level is as the attempt to build as good a model of an agent's behavior as possible based on information external to the agent." (p. 109)

Ironically, Newell anticipates in his introduction that our reaction will be, "But that is just the way I have been thinking about knowledge all along" (p. 93). This far into the analysis, steeped in observer-agent relations and non-physical encoding, Newell concludes quite the contrary, "The definition above may seem like a reverse, even perverse way of defining knowledge" (p. 106).

Newell's crucial insight is that a KL description is an abstraction made by an observer. This attribution derives from the observer's projection of himself onto the subject; he adopts the role of the other and considers what he could find out and what he would do (p. 109). Such attribution is a possible and valid prediction because the observer has the same underlying functional architecture. When the observer ascribes correctly, "the agent behaves as if he has knowledge K and goals G" (p. 106).

Dennett uses similar language, referring to the intentional stance as "the strategy of prediction and explanation that attributes belief, desires ... and predicts future behavior from what it would be rational for an agent to do" (Dennett, 1988, p. 495). Newell is disturbed by the suggestion that such a level of description is somehow less real or is more subjective than a description of the symbolic or neurophysiological levels.

The simplest resolution of these points of view is to make clear that a KL description is not about a particular agent in the same way that a description of his circulatory system is about him. Contrary to what Newell says, a KL description is a level higher than the physical body, it is inherently about the individual agent interacting with the world. This is evident in any remark concerning the agent's knowledge: see, believe, know, and hypothesize are always predicates about the world relative to the agent, evident in the most mundane examples, "Pat knows Mike's telephone number" (Newell, 1982, p. 118). They are statements about the situated agent, not a mechanism in isolation.

A KL description for people is a description of a social system, that is, about how people perceive each other and their common activities, which would belie Newell's claim that the KL is analogous to a "register–transfer" description, just a higher level. This would easily explain why KL abstractions have no isomorphic realization as structures in the agent. KL descriptions are about interactions the agent has with its environment. Emergent patterns can result; categories for describing what is happening will lie outside the awareness or control of individual agents. Indeed, this orientation is apparent in Newell's remarks, which incidentally are a perfect summary of Anderson's approach: "Knowledge, in the principle of rationality, is defined entirely in terms of the *environment* of the agent, for it is the environment that is the object of the agent's goals, and whose features therefore bear on the way actions can attain goals" (Newell, 1982, p. 110).

By analogy with the theory of relativity in physics, the laws of rationality will be the same in every frame of reference. That is, observers might attribute different knowledge to the same agent, but coherence relations that hold between beliefs, goals, and actions, and hence the predictive laws, will be the same. Thus, identifying a KL description with an observer-theoretician is not to adopt the folk view that "it's all relative." Instead, as for physics, psychology must drop the idea that observations are objective facts that everyone will agree upon. Rather, observations, perceived patterns, and subsequent knowledge-level descriptions are dependent on the observer's frame of reference: his beliefs, goals, and actions, and more specifically how he interacts with the agent and elicits the behavior he subsequently theorizes about. The laws of rationality are the same because we share the same functional architecture and, through a common theory-formation process, are capable of projecting different perceptions and social norms onto behaviors that would otherwise be discrepant in our culture. This does not mean that our theories will agree, but rather, that we can usually find systematic justifications for behavior that are relative to an agent's presumed beliefs, goals, and knowledge. It is this metatheoretical systematicity, which we term rationality, that is invariant.

KL Representations

The relation between knowledge and representations is no less bewildering than the relation between knowledge and an observer. Newell's key break with the conventional AI view is to say that knowledge representations are the observer's statement of his KL theory. They *represent* the knowledge ascribed to agent, they are not to be identified with the agent's knowledge itself: The map is not the territory. Yet throughout, even if Newell is not ambiguous, his point is far from clear. Readers must pick and choose the statements that speak to their own biases.

The discussions of logic and conceptual dependency provide anchoring points. The simple claim is that "logic is just a representation of knowledge" (Newell, 1982, p. 110). Surely many researchers must be bewildered when Newell then goes on to say, "It is not the knowledge itself, but a structure at the symbol level," for it is AI gospel that knowledge is encoded in the brain as symbolic structures. The way out of this, I believe, is presented in Figure 13.3. A logic statement is an observer's KL description, in the form of symbolically interpretable expressions (e.g., knowledge representations in a computer program, such as predicate calculus statements). Such knowledge representations have the same status as any other representations, such as book chapters—they are written down by an observer and they are interpretable. Crucially, they are in "perceptual space"–they are externalized in a form that can be reflected upon (otherwise they couldn't be perceived and given a subsequent semantic interpretation). (Again, I include silent speech and mental imagery as externalized expressions.) Thus, when Newell says that a logic encoding is a symbolic structure, but not the agent's knowledge, he must mean that what is encoded are the observer's representations, symbols in the observer's perceptual field, which are to be interpreted as being *about* the agent's knowledge. A knowledge representation is not the agent's knowledge, and it's not the observer's knowledge either: It is a representation of the observer's knowledge of the agent's knowledge.

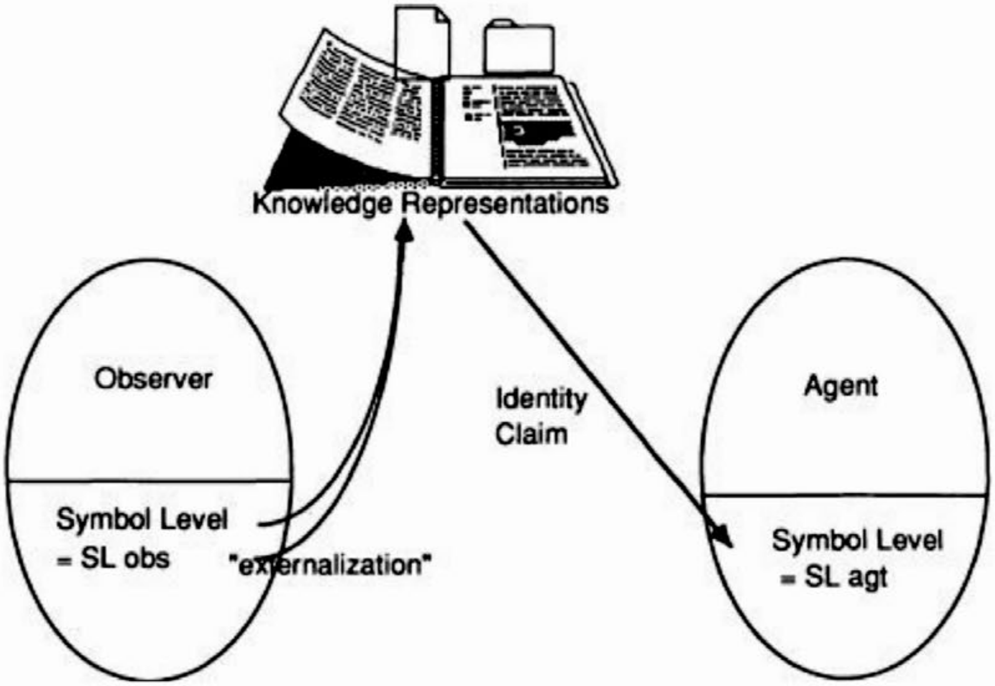


FIG. 13.3. Relation of knowledge representations to observer and agent; adapted from Newell (1982, FIG. 4). This paper criticizes the identity claim.

Newell claims that conceptual dependency is a contribution to the KL because it expands our repertoire as theoreticians, it gives us another formalism for expressing KL theories. Similarly, calculus was a contribution to physics because it provided another means of "encoding knowledge of the world in a representation." Thus, conceptual dependency, like logic, is a theoretician's tool. It is how cognitive scientists (and robot designers) might specify an agent-environmental system in a way that allows predicting the agent's behavior (or specifying what behavior is desired in the designed machine).

Newell's analysis enables him to pinpoint the essential dilemma confusing our enterprise: What is the relation of the observer's knowledge representations to the symbol level of the agent? Some logicists (exemplified by McCarthy and Nilsson) simply assume that "the role of logic … (is) for reasoning by intelligent agents," rather than being "a tool for the analysis of knowledge" (Newell, 1982, p. 118). That is, these researchers identify the theoretician's expressions with physical structures that preexisted in the body of the agent and caused the observed behavior (see Figure 13.3, "Identity claim").

The identity claim encapsulates the idea of *strong equivalence* (described by Pylyshyn, chap. 7 in this volume). It claims that our models of agents are isomorphic to structures in the agent. But the frame of reference perspective suggests that this couldn't possibly be the case. A KL description isn't even about an individual, isolated agent, let alone physical processes that are unobservable.

The argument might then take another form. Perhaps these knowledge representations are to be identified with symbolic structures lying inside the *observer?* This would be a curious move to make; it would equate "pictures, physical views, remembered scenes, linguistic texts, utterances" (Newell, 1982, p. 112) with physical structures in the head of the observer. We might be confused about many things here, but we surely know better than to say a book made out of paper is *identical* to something inside the skull. Instead, we are interested in the processes by which an observer perceives these representations and "extracts knowledge from them" (Newell, 1982, p. 113). A theory that relates an observer's KL description to the observer's symbol level (and hence his behavior) must account for this extraction or interpretation process.

KL as Reinterpretation

The issue of semantic interpretation is probably the riskiest quagmire on the terrain of the philosophy of AI. However, it is on these grounds that the whole "symbol processing" view rests, and it is here that we must follow Pylyshyn in taking a definite stand. Given what I have just said, it is worth restating my position: Semantic interpretation goes on in our "outward" behavior, in our ongoing sequence-cycle of perceiving a representation and commenting on it through another utterance, gesture, drawing, image, or notation. Thus, symbols are interpreted semantically, but not in a hidden way, not subconsciously (dreaming aside), but always in the space of our perceptual field.

In short, it is perfectly fine to describe the mind as operating upon symbols that are interpreted semantically. However, these symbols must be perceived in order to be interpreted. Or to put it a better way: It is the perceptual process and our subsequent activity that gives a token symbolic status. Since this is something the person is very well aware of, we prefer to say that "*the person* has symbols, rules, and representations," just like we would say "the person has a book," and would feel very strange to say "the mind has a book." Representations are not stored in memory, rather they are constantly created and interpreted by the person. It is the process of perception and conceptualization that we must explain.

Pylyshyn realizes very well the dilemma of semantic interpretation. The internal symbolic codes[5] "must carry all the relevant aspects of the interpretation as part of their intrinsic and functional form" (p. 66). The symbol system does not have access to the interpretations; "only the theory provides that." We must therefore distinguish between two types of interpretation:

1. A program syntactically interprets a representation by relating its formal properties to rewrite rules for creating/modifying structures.

2. A human observer semantically interprets a perceived representation by commenting on it.

The key distinction is that it is the human's comment that gives the structure representational status. Similarly, we say that the computer program "has" representations, but this is only because they are meaningful to us. Strictly speaking, MYCIN interprets our representation of "a significant organism" by relating the formal properties of the token SIGNIFICANT to rewrite rules for creating and modifying other token structures. A human semantically interprets this same representation ("(IF (SAME CNTXT SIGNIFICANT) (CONCLUDE CONTXT INFECTION TALLY 300))") by supplying a context, relating it to other concepts in which it can be viewed as an instance, a cause, etc. For example, I might say, "This is one of a dozen rules that interpret culture results as possible evidence of disease."

A key claim is that this commentary process cannot be reduced to application of syntactic interpretation rules; rather, it is inherently a conceptualization process. Each time I talk about the SIGNIFICANT rule, I am prone to discover something new about it. My representations aren't stored in memory; they are in my speaking and what I have written down. The content of my representations is in commentary about them. Representations don't have an inherent content: they are prone to new interpretations by every observer on every next occasion (cf. Agre). Our essential problem is to explain (a) how this conceptualization process works (in functional architecture terms); that is, what is the process by which I create representations out where I can perceive them? and (b) how does commenting on a representation organize my ongoing sequence of behaviors? Like Rosenschein, we must reject the idea that formal structures could be formally linked to their interpretations: There is no such thing as "the content" or "the meaning" of a representation.

At the very least, viewing conceptualization as a process of storing and manipulating labels on internal structures is misguided because it puts perception into the peripherals, viewing it merely as input to a conceptual processor. To understand how a different mechanism is possible, we must show how perception/commentary organizes processes directly. Our behavioral processes aren't internally tagged and matched against rules that describe how they are to be assembled. Again and again, we find that it is impossible to take a new stand on these issues without overturning the entire cart: memory, perception, meaning. Everything must go at once. Before tackling this further, I want to say more about the KL and how it is and isn't related to the functional architecture.

KL as Descriptive Theory

It is helpful to consider again the nature of the statements we make in our computer programs. Knowledge representations are statements about regularities. They are descriptions of and relations between *patterns of behavior*. They are not necessarily descriptions *of mechanisms* that create such behaviors in people. Indeed, the possibility and usefulness of such explanatory, but non-mechanistic descriptions is one reason why we distinguish between a KL description (an observer's theory of a social, interactive system, in terms of individual beliefs, goals, and activities) and the functional architecture (a theory of physical mechanisms implemented in a neurobiological system).

A simple example is the metarule in NEOMYCIN (paraphrased), "Generalize an inquiry, rather than request the specific finding of interest." This rule is surely part of the mechanism of the program. However, its relation to people is different. It is a description of a regularity in observable human behavior. Such rules should be viewed as *grammatical* characterizations. They are perfectly fine ways of summarizing and abstracting observations made over a variety of problems, agents, and domains. Knowledge-level attributions are therefore similar to natural language grammars. They are a theoretician's way of stating regularities; they are descriptive; they are generative; they have predictive power. However, as explanations, they aren't to be taken literally as structures and processes that are encoded in the heads of the subjects we are studying. Their explanatory power isn't (necessarily) mechanistic. As Dennett says, "We should not jump to the conclusion that the internal machinery of an intentional system and the strategy that predicts its behavior *coincide* ..." (p. 497).

Once again, we find ourselves on familiar, controversial terrain in the philosophy of AI. The surprise is that so many people have identified every explanation with mechanism, ignoring the nature of grammatical, descriptive theories that state regularities, abstracting behavior. Such descriptions can surely be related to lower level processes of memory, learning, and perception, but they cannot be reduced or replaced by better mechanistic models. Again, to quote Dennett (1988), "There are patterns of 'behavior' ... that are describable only from the intentional stance ... there are no 'deeper facts' to resolve the outstanding questions of belief attribution" (p. 497).

A KL description is surely a model, but it is not a model of the physical mechanism. While we surely want to know what the functional architecture is, this KL description is not to be viewed as inferior. It is a legitimate level of explanation which has no isomorphic embodiment in physical mechanisms: It summarizes patterns and states principles that arise in the agent's interaction with the world, the theoretician's interaction with the agent, and the theoretician's perceptual process, goals, and beliefs.

Thus, we have restated the situation in Figure 13.3: A KL description is like a grammar. Just as for natural language, we must decide whether these grammars are literally encoded in the head of the agent and thus were the structures that caused the agent's behavior. The identity claim is tempting because (a) such structures really are part of the causal mechanism of computer programs, and (b) we know that articulating such representations has an effect on the agent's behavior (i.e., telling me a new grammar rule can alter how I speak). I have argued that KL descriptions should not be identified with causal mechanisms in the agent because:

- they are *attributions made by an observer*, involving his own selective interactions with the agent, his own perceptions, and his point of view;
- they *abstract a sequence of behaviors*, not single, moment-by-moment responses;
- they *characterize a social system*, not processes within an individual agent;
- the interpretation of such representations, which itself is claimed to cause behavior, is constantly changing, dependent on the observer of the representation, and in any case is always made in perceptual space;
- such interpretations patently occur only through outward behavior, and there is no evidence that the agent, despite being a theoretician of his own behavior, has any such notations (e.g., see (Stucky, 1987) for a related analysis from a linguist).

Indeed, it now seems perverse to think that a theoretician's KL descriptions (what I say about an agent after watching over time) could have caused the agent's behavior. To say that an agent follows a pattern is not to say that the pattern is necessarily a thing inside the agent.[6]

From the perspective of explanatory theory, the KL can be viewed as necessary because we need to express generalities that cannot be reduced to mechanisms at a lower level (Pylyshyn, 1984, p. 35). Pylyshyn relates this approach to the idea that there can be constraints on behavior that are above the level of "actual performance." This is consistent with my claim that the KL describes the interaction of individuals within the social environment. It follows also from the introduction of an observer's point of view and the desire to describe behavior temporally, in terms of sequences of behavior. Thus, environment, observation, and time supply the context for KL descriptions. The direct ramification for the design of intelligent machines is that the KL level is for specification of a design; it will use representations that don't causally enter into the machine's behavior, rather *they will describe what the behavior will look like* (e.g., "rational," like someone imagining a 3-D world, like someone avoiding obstacles, like someone trying to be efficient).

Furthermore, just as the existence of a KL is brought into being by our theoretician-designer's perspective, so is the very idea of regularities in the agent's behavior. When we talk about "regularities that need explaining" (Stefik, 1989, p. 242), we must keep in mind that regularities aren't substances and patently aren't objective. They are an observer's statements with respect to some behaviors perceived in some frame of reference. A regularity is not a property of an agent so much as a perception that arises in the interaction of the observer, agent, and environment.

Understanding the nature of KL descriptions therefore requires considering the nature of perception.

KL as Perceptual, Emergent, Interactive

Imagine placing a cafeteria chair near Brooks' robot—what will happen? Suppose it jams under a rung. We must ask Brooks whether this is what he intended. Suppose the robot starts wheeling around the fat leg of a chair; Brooks says, "Ha, it thinks the leg is a wall!" There is no end to the games we can play with the robot, introducing new elements to its environment and watching what happens. In the end, how we characterize the robot's beliefs, goals, and knowledge will depend on what obstacles are placed in its path and even whether we were watching when something occurs. After awhile, we might feel confident that we fully understand the robot's capacities and foibles. But this is just part of the stable order of our own purposes and social organizations. Someone might coat the floor with jello next week to see what happens.

A number of related issues surface here, some of which we have considered previously:

- There is no such think as "all the data" or "all the information" in a situation—this is an observer-theoretician's analysis, dependent on the measuring devices (consider for example how viewing a video in fast-forward can reveal new patterns[7]);
- Perceptual interactions among people are dialectic—such that my interpretation of your response to what I did biases my interpretation of my original intentions; thus, we define the present by interactively constructing the past (indeed, articulating intentions after the fact places the agent in the role of KL-theoretician);
- Teams of people are not merely "cooperating agents," not merely "distributed"—activities are not transmitted or conceived or planned completely (i.e., fully anticipated in all its particulars) by any individual, but arise through mutually constrained perception.

These are the kind of claims you will find in the work of Agre, Lave, and Suchman.

The bottom line (again) is that the functional architecture mechanisms supporting what we call human memory, learning, and perception are different in kind from our KL theories.[8] Furthermore, a better model of the functional architecture (or simply not equating a KL description with a FA) will explain the plasticity of human behavior. By not building in grammatical descriptions of how people can behave, we discover a mechanism (remember AARON and PENGI) that has the potential to generate a wider range of behaviors and is more robust (capable of responding without having predefined situation types) than any KL description.

Redefining the idea of "information" will turn out to be pivotal, because it so strongly affects our ideas of perception, memory, and learning in turn. The strong claim is that information is dialectically defined by the interaction of processes in the agent and processes going on in the environment. "Dialectic" here is to be contrasted with the idea of "conditioned" or "dependent on the data," in which "data" is something objective and given (like tokens put in the slot of the machine) and conditionality is reflected by conversion of the data to internal codes that serve as labels or tags, which are stored in memory and referenced by rules. Thus, identifying information with data is the same as claiming that "situations" are enumerable (in terms of constellations of input, precisely the formal analysis Genesereth strives for).

My claim is that we will be able to relate the analyses of Agre, Lave, Suchman, et al., to SOAR by translating their discussion of "social situation" and "dialectics" into a different view of information, and hence perception. Furthermore, we should look for examples in which the social band is not a phenomenon manifested only in changes of behavior over days, as characterized by Newell (1990, p. 338), but is manifested, for example, in conversational interactions, mutually constrained over seconds (a good example to start might be patient–psychiatrist dialogues involving the transference effect). Again, this social orientation is not just a claim that our goals, beliefs, and desires are pervaded by the social organization (and hence KL descriptions are *about* a social system). Rather, the more important implication for us will be the changes it requires in how we view the functional architecture, by realizing that the KL is a description of interactive, dialectic phenomena—a description of *the result* of processes interacting within and outside individual agents.

Finally, we should relate this orientation to my earlier emphasis that representations are in the perceived environment. The cycle by which information is perceived involves a sequence of creation and interpretation of representations that are in the environment (plus imagination). Therefore symbol manipulation[9] is inherently a coupled phenomenon of the agent to its environment. Symbolic reasoning is not merely conditional on the current situation (i.e., influenced by supplied "data"), and it is not an invisible, cognitively impenetrable process. Rather, symbol manipulation —a KL characterization—is a characterization of the result of how the agent interacts with its environment. To say it more directly: Symbol manipulation is what agents do in their outward activities, not something happening to physical structures inside the mind (silent speech and visualization aside). People manipulate symbols. This is why social theorists place so much emphasis on the actual materials surrounding agents and how they are moved around, pointed to, and modified (e.g., see Allen, 1988; Suchman, 1987).

KL as Sense Making

One of the most valuable twists in our analysis is to view the observer-theoretician as an agent and characterize his behavior as indicative of what agents do: Agents make observations, they articulate theories (representations), they generally attempt to understand the world in which they live. To properly characterize and improve KL descriptions, we need, as Dennett puts it, to explain the power of folk psychology.

The fact that we ascribe knowledge to agents, as well as the kinds of theories that are satisfying to us, reveals as much about the nature of our understanding processes as about the agent or system being studied. In fact the entire cognitive science enterprise reflects what agents do. We are agents after all.

Hence our use, as theoreticians, of the KL reflects back the very property we wish to study-the nature and origin of beliefs and their influence on behavior. Examining our own activity, we discover that:

- Beliefs are expressed as representations in perceivable, shared media (most notably speech and writing). Representation construction involves changes to the environment, which is then interpreted and modified by other agents.
- Theoreticians, as humans, need to view their beliefs and goals as coherent. Hence, KL descriptions reflect humans in their ordinary process of making sense: the world is described in terms of law-like statements. The complexity of many specific observations is abstracted, categorized, and parsed by grammar-like theories.

In short, KL descriptions exist because we, as agents, need to construct a coherent story about why agents interact the way they do. Put another way, a semantic level exists in our theories of intelligence because we need to explain why this system of interacting agents is meaningful. The concepts of a KL description are our categories. They need be articulated only in our theory, they exist apart from the neurobiological processes going on internally in individual agents. Of course, individual human theorizing about social behavior has profound effect on how members of the community act (i.e., intention and social theory is cognitively penetrable). This itself makes the point: The intentions ascribed are relative to each agent's point of view and there are multiple explanations, depending on where you stand, what you saw, and what you are trying to do.

As an aside, we now see how ironic it is to say that logic "is therefore a candidate ... for the representation to be used by an intelligent agent" (Newell, 1982, p. 121). The fact is that McCarthy, et al. are intelligent agents and they indeed use logic as a representation! As I have said, their symbol manipulation is going on out where we can see it; there is no doubt that they use logic representations. This itself tells us very little about their functional architecture, except maybe we will want to account for this emphasis on tidiness and elegance in terms of how the processes inside are organized.

Similarly, it is ironic to say that conceptual dependency "made relatively little contribution to the symbol level" (Newell, 1982, p. 120), given that it provided a convenient symbolic language for expressing theories. Here again, the "symbol level" as we know it is what's going on in our perceptual space. As agents, this is how we state what we believe; conceptual dependency, like the predicate calculus, provides a more disciplined means of articulating and sharing theories.

The primary difference between a KL description and any other theoretical statement is that it is an explanation of agent behavior vs. inanimate phenomenon such as meteorology. As an aside, it is interesting to consider the attributes that human behavior and the weather share:

- Both meteorological and social systems can be described in terms of loosely coupled, interacting processes;
- Local interactions over short periods of time can be described and predicted fairly well, but emergent macro effects, while statistically correlated, are difficult to predict;
- The laws and principles describing behavior (e.g., rationality in people) describe the results of dynamic interactions, not mechanisms that are internal to the components. (For some reason, this is more obvious in physics, where we wouldn't say Newton's laws are "known" by the planets, who then changed their minds when Einstein came around.)

In general, a KL description, just like any other theoretical statement,

- arises in making sense (thus, it exemplifies the centrality of human storytelling and comprehension);
- is based on observations from a frame of reference (i.e., it is highly dependent on the measuring devices and sampling period);
- is attributed to the system being observed (i.e., it is viewed as a property that belongs to the system, not the observer);
- is realized in some perceivable form (a representation).

Putting this together, stating a KL description is creating information. A KL description, like any representation, is not translated from a prior internal representation. Thus, when we speak we are not translating from a description of what we are going to say. By this view, speaking itself is a conceptual process that is only realized in the actual physical activity of uttering a phrase. The oddity of such a strong claim is that it makes perception and activity one process. What is perceived–conceived–stated can then be perceived and commented on.

It is in this respect that semantic interpretation is going on in our outward behavior; the semantic relation is embodied in the reference one external representation makes to another. That is, semantic relations are realized in our ongoing behavior, not encoded or predescribed in some internal form. This is also why semantic relations cannot be reduced to laws or representations: Semantic relations are inherently processes realized in our interaction with the environment. As such, an observer can comment on them, but they cannot be reduced to single point-of-view, objective descriptions.

Of considerable interest is how such commentary orients and controls subsequent behavior, so the whole ongoing sequence of sense making is providing an accumulating orientation, which is composing new sequences of behavior. These ideas are again too non–traditional to be fully elaborated and digested here; I later attempt a simple presentation; a more complete story is told in Clancey (in preparation).

To summarize, KL descriptions are crucial for the sense making of any agent and appropriate for designers of intelligent machines. But they are inherently subjective and realized only in an observer's representations. We have been very confused about this because as AI researchers we have adopted mentalist arguments, specifically that representations of the world and descriptions of our own behaviors are stored in memory, indexed, and manipulated as data structures. Any description is relative to a frame of reference, so what is inside cannot be the descriptions made by the theoretician. Any descriptions (knowledge, representations, theories) are articulated with respect to and about an interactional, observational space; they presuppose a frame of reference, are perceptual in character, and inherently subjective. KL descriptions are about agents, but they belong to the theoretician. They constitute his knowledge, his beliefs.

Adding to our analytic difficulty is our observation that agents being studied do make statements about what they know, and such statements have an effect on subsequent behavior. This has led theoreticians to assume that their own attributions of beliefs and desires, which explain behavior, must be actual statements encoded in the agent, which are causing the behavior. However, beliefs and desires are expressions that only an observer can make. A theoretician–observer could attribute beliefs and desires to us, but they aren't in this sense our beliefs or desires.

Following Bartlett, expressions of beliefs and desires are reflective constructions that for the agent serve to resolve an impasse in behavior by providing a new orientation (Bartlett, 1977). Thus beliefs and desires must be stated to have effect; their causality is towards the future, not as a mechanistic account of what has already occurred. Thus, we must distinguish between:

1. an agent's expression of his own beliefs and desires (where these come from);
2. an observer's grammatical attribution of beliefs/desires to others;
3. the effect reflecting on #1 has on the agent's subsequent behavior.

Items 1 and 2 arise by the same "understanding process," the production of a KL description. Hence, for the agent as observer, we have the interesting process by which a KL account of what happened in the past has implications at the level of the functional architecture in changing future behavior. As in the conventional AI picture, a representation is causing behavior. The interesting twist is that the representation is constructed in order to create new information ("a way of turning round on its own 'schemata'" (Bartlett, 1977, p. 202)), as a reflection on past behavior, serving as an orientation (way of perceiving) that allows a present impasse to be resolved (and behavior to proceed again automatically, without reflective construction of representations).

To bring a few points together in a different way, consider that I might attribute beliefs and desires to a cat chasing a bird: "The cat believes the bird can't hear her." This takes us up short because we have never heard a cat state any beliefs and desires, yet we recognize it as a valid KL explanation of what the cat is doing. It reminds us that such attributions are the observer's statements and representations and not of the agent being studied. This case is clearer because of point #3 above, namely a cat's behavior, as far as we know, is not affected by KL descriptions.

Hence, we're studying at least two phenomenon:

1. The nature of the KL as a social system-level description, necessary to predict and explain interactional behavior. (Or, in its degenerate form, the case of a robot in isolation, this reduces to just a physical system constituting the robot and its physical environment.)

2. The effects such articulations by agents themselves have on their future behavior, what forms such explanations take, and why they seek such understanding (specifically, its origins in both the interactional environment and in the neural processes constituting the functional architecture that maintain formal relational consistency in perceptions and behaviors).

KL as Physical Theory

To carry the analysis of the relativity of the KL a step further, it is worth making explicit connections to the philosophy of physics. The interpretation of the Uncertainty Principle has caused much debate, not surprisingly since it involves many of the same issues and complications AI is struggling with. Maybe most surprising is that Bohr and Heisenberg themselves wrote about the relativity of knowledge and the origin of theories. For example, Heisenberg said, "What we observe is not nature itself, but nature exposed to our method of questioning." Bohr wrote, "It is wrong to think that the task of physics is to find out how nature is. Physics concerns only what we can *say* about nature" (both quoted in (Gregory, 1988)). Heisenberg and Bohr saw that the metaphysical implications of quantum theory involved viewing theories as relative to a frame of reference, whose appropriateness depended on the measuring devices and purposes of the experimenter. Gregory summarizes this well: "We interact with the world and create interpretations of what this interaction means" (Gregory, 1988).

Figure 13.4 provides a summary of the interactional and perceptual aspects of theory formation, characterizing the process by which representations develop. The world ("reality") is viewed here as an undifferentiated, continuous field. Organized subsystems interact and interfere with one another. The dynamics of interactions result in locally–stable configurations, due partly to exchange of energy (equilibrium) and partly to the conserving effects of memory (e.g., in genes and neurons). The measurement devices we use (including our visual system) constitute another situated subsystem, which both enable observations by their interactions with the containing environment and distort these observations through their own character. Thus, human perception is biased by both the external interactive process (e.g. our use of a camera with certain recording properties) and our orientation as we interpret and make sense of stimuli.

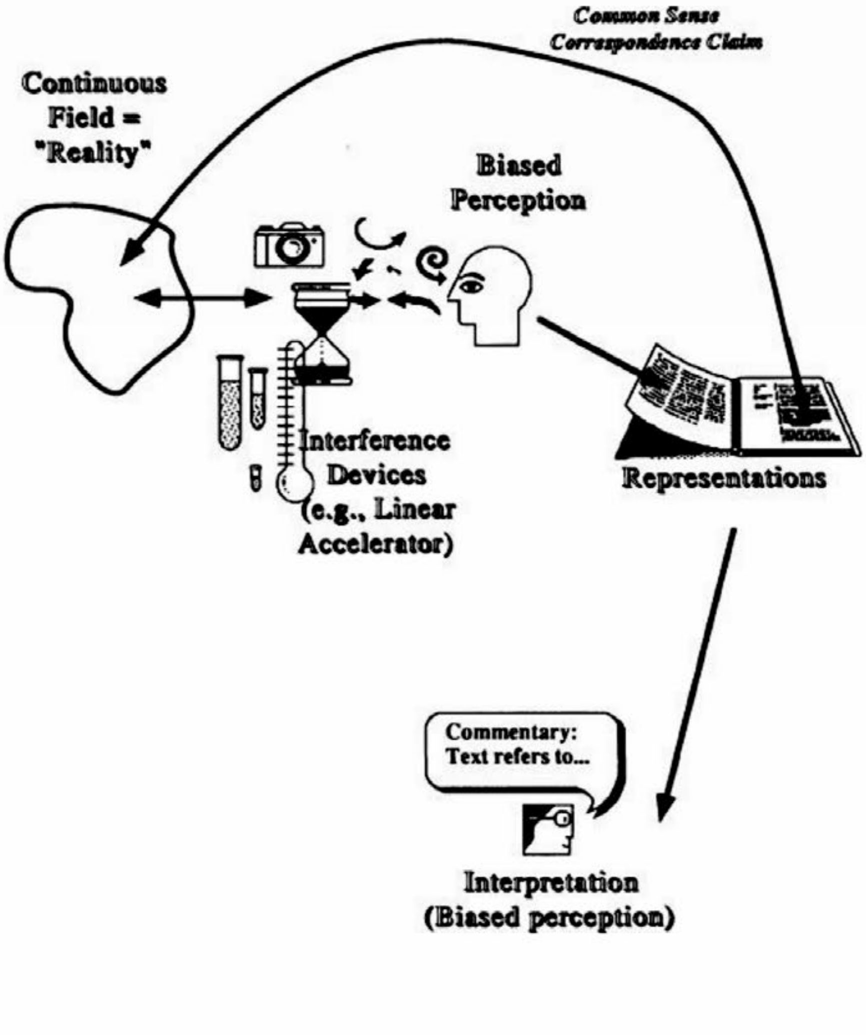

Common Sense
Correspondence Claim
Continuous Field = "Reality"
Biased Perception
Interference Devices (e.g., Linear Accelerator)
Representations
Commentary: Text refers to...
Interpretation (Biased perception)

FIG. 13.4. How representations correspond to reality, illustrating first order relativity of representation construction and second order relativity of representation interpretation.

Our representations (e.g., our written theoretical statements in some notation) are part of the dynamics of our own interaction with the environment we are studying. One way of saying this is that representations emerge as perceptions, whose construction accomplishes and maintains coherence (in the mapping between perception and action) against the dynamic interaction with the environment. Futhermore, as I have emphasized throughout this paper, these representations are subsequently interpreted by another perceptual process (e.g., performed by the reader of the scientific text), in a process of commentary which creates a secondary conceptualization.

The upper arrow in the figure illustrates the common sense view that a scientist's statements are objective descriptions of reality. The *correspondence theory of reality holds* that what we know and say is about the world (reality) and is only more or less accurate. This is the commonsense view that scientific models are only approximations of reality. This is also the commonsense view that words refer to things and events "out there." The correspondence theory underlies most theories of truth and has been the subject of volumes of philosophical debate (most notably in AI, the work of Brian Smith (1987)). In contrast, the strong claim (supported by Bohr and Heisenberg) is that there are no objects and events to be described independent of the interference effect of making an observation (which is why whether the electron is a wave or a particle depends on the experimental process that interferes with it). Indeed, the very notion of objects, categories, and time arises in our perception and language (Gregory, 1988; Tyler, 1978). The most immediate implication for AI research is that we must abandon the search for some kind of "formal semantics" that could be used to formalize the correspondence between programs (*qua* representations) and "the world." There is no independently knowable, objective world that our representations can be mapped to. This is another way of saying that the meaning of a representation depends on the frame of reference of the observer.

In short, understanding the nature of KL theories is tantamount to understanding how any theory corresponds to the phenomenon it is about. There are two forms of relativity: First-order perceptual relativity (a KL theory is no more objective than a physics theory), and second-order representational relativity (KL descriptions are always open to interpretation). To relate this to AI architectures, a knowledge base represents theories about the world. Every knowledge base contains models of one or more systems in the world (e.g., a model of an electronic circuit, a model of the physiological processes in the human body). This model constitutes the beliefs of an agent about the system in the world. As described above, these beliefs are inherently a product of how the agent has interacted with the world and the purposes at hand (how the model will be used). We say that this model constitutes the agent's knowledge.[10]

Figure 13.5 summarizes how a KL description of agent behavior can be viewed as a causal theory.

The left side of Figure 13.5 shows how I have characterized KL theories. A KL theory is something that an observer knows about an agent, who in turn has knowledge of some system in the world. The system in the world that the KL theory is about is the system containing the agent, with which it interacts, not the agent itself, in isolation. Thus, the properties of subjectivity and relativity that hold for physics (left inner box, "agent knows domain causal theory") hold as well for the observer's KL description of the agent (right box, "observer knows KL-causal theory"). This is not surprising or especially profound, but it does point out that to take a stand on the nature of the KL is to take a stand on the issues raised by Bohr and Heisenberg.[11]

It is not surprising that the same confusion about system levels has cropped up in expert systems research. Many people believe that all qualitative models can be reduced to function-structure blueprints. According to this view, the only reason physicians deal with disease taxonomies is because they don't understand how diseases are caused. After medical science improves, we would be able to describe every disease exclusively in terms of abnormal states and processes within the body. Classification models are inherently inferior, these researchers suppose; real scientists work with hardware diagrams.

However, this interpretation is false, for the same reason that the KL cannot be identified with processes within an individual person. In fact, diseases are descriptions of the result of a pattern of interaction between an individual person and his environment. Consider for example tennis elbow. This syndrome cannot be causally explained in terms of processes lying exclusively within the person or within the environment. Rather it is a result of a pattern of interaction over time. As for any emergent effect, it can't be predicted, explained, or controlled by treating the person in isolation, or even by studying the person-environment system over short periods. It is a developmental effect, an adaptation in the person that reflects the history of his or her behavior. The same claim can be made about the entire taxonomy of medical diseases—trauma, toxicity, infection, neoplasms, and congenital disorders—they are all descriptions of the agent after a history of recurrent interactions. Similar examples can be drawn from computer system failures; faults cannot reduced to changes in a blueprint, but in fact the space of possible etiologies is constantly changing with the dynamics of interaction with an open environment. For example, a favorite story at SUMEXAIM at Stanford is how system crashes were caused every fall when the first October rains wet the phone lines going to Santa Cruz, swamping the computer with spurious control–C input attempting to get its attention.[12] Such problems aren't fixed by swapping boards.

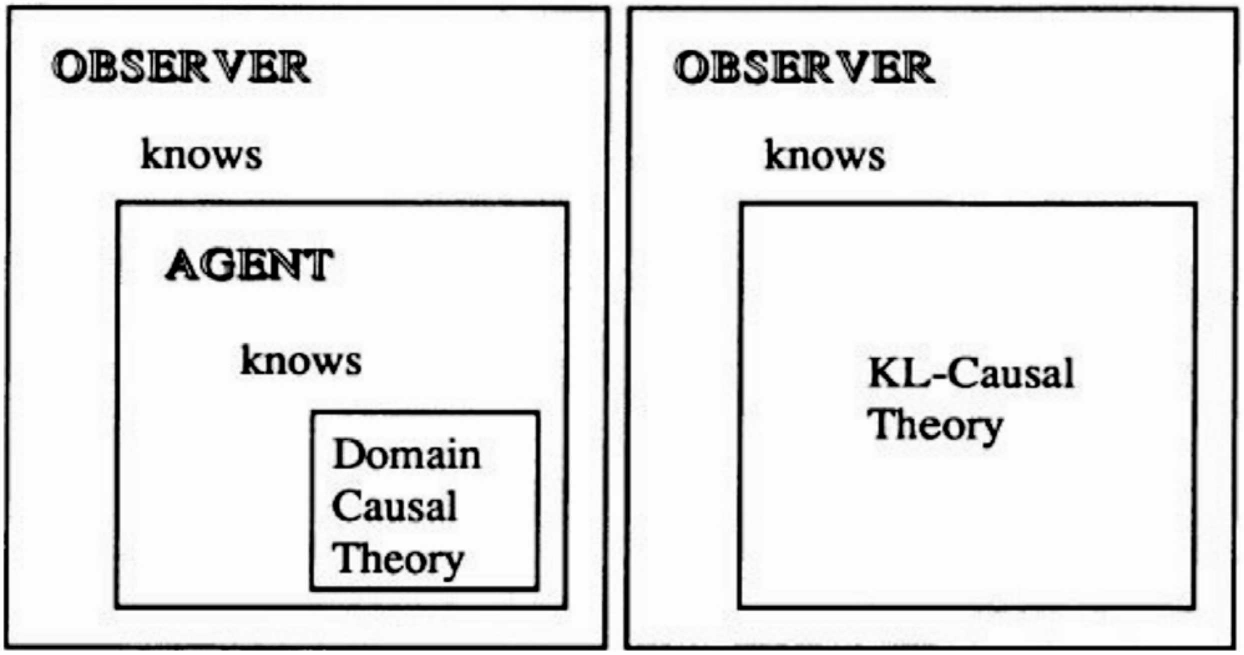


FIG. 13.5. KL descriptions viewed as causal theories.

It is worth restating how the KL is like any other causal theory. First, it is relative to an observer's frame of reference; it has space-time characteristics. As previously stated observers might attribute different knowledge to the same agent, but from their point of view the coherence relations (why agent behaviors appear meaningful) will be based on the same principles (concerning how the agent's goals and beliefs account for his behavior). The law of rationality holds because human observers more or less share a functional architecture that produces a sense-making process with similar characteristics:

- representation use and generation, by commentary about perceived structures;
- a process of combining past behaviors in way that maintains and achieves coherence in every action;
- plasticity, adaptability, learning rates, perceptual capability that leads to a more or less common capacity to generate new information and develop a shared domain of discourse in collaborative work.

Finally, it is intriguing to reach for an analogy with the General Theory of Relativity. In particular, an observer can't tell whether an agent's behavior is generated by a perception (influence) we cannot see (and might be just an imagined state of affairs) or an ongoing change in the environment (i.e., to maintain equilibrium, the agent's behavior is evolving with the changing dynamics of the interaction with the environment). As for inertial frames of reference, the organism responds to perceived *changes* in the environment (i.e., acceleration). Constant movement in the environment is adapted to by constant behaviors.[13] Attempts to apply dynamic systems theory to the study of intelligence must be framed in terms of development with respect to the history of environmental interactions, as opposed to local responses. Put another way, memory crucially changes the study of dynamic systems, just as onto- and philogenetics separates biology from physics.

McCarthy has said that AI research needs a few Einsteins. Applying Einstein's (and Bohr-Heisenberg's) ideas about relativity and the nature of scientific theories to the study of cognition might be a good start.

Summary: Relation of KL-description to Functional Architecture

I have argued that a KL description is necessary and useful, but it is not to be identified with the physical processes that cause behavior. Knowledge is an observer's characterization, not something that the agent owns. In cognitive science and AI, we have heretofore taken our *selective* (based on limited interaction with the system being studied/modeled) *perceptions* (necessarily abstractions, generalizations) and placed these *grammar-like descriptions* (formal, expressed as rewrite rules) in the heads of our subjects, claiming not only that they aren't ours, but as representations they existed before we created them, and they even existed as descriptions of what the agent was going to do, inside working memory, before he or she behaved. The strong claim is that representations do indeed play a crucial role in human behavior, but they are created fresh, out where they can be perceived; they are not manipulated, indexed, and stored by hidden, inaccessible processes.

Table 13.1 contrasts the mentalist, cognitive science or AI architecture claim against the contrary point of view that I have synthesized from a variety of fields. A KL description is about a situated, social system, the result of interacting internal and external processes, and is an interpretation by an observer; it is neither objective nor a property of individual agents.

As Newell says, the KL is a real level of description, as much as any description is real. It is a system-level description, but attributable to social systems, not individual agents. We need such a level because the behaviors we observe are emergent in interactions, and as such they could not be preconceived or predescribed in the individual agents. Of course, agents themselves can predict what will happen and this can enter into their deliberate planning about what to do. The fact that agents have their own KL descriptions of themselves and this does affect their behavior greatly complicates our analysis. We can't tease this apart without resolving longstanding issues in psychiatry, and we should recognize that's the domain we're dealing with.

TABLE 13.1
Opposing View of the Nature of the Knowledge Level Descriptions.

Cognitive Science	*Anthropology, Philosophy, Linguistics, Sociology, Physics*
subject's knowledge	observer's theory
stored in memory	expressed, stated, written on paper
pre-existing plan, determining behavior	product of selective interaction and perception
objective, corresponding to reality	subjective, relative to a frame of reference
fixed, causally determining behavior	continuously interpretable (a representation, not the mechanism itself)
reflection = examining internal data structures	reflection = objectifying own activity, perceiving and commenting about a sequence of behavior

The most critical distinction between my analysis and Newell–Pylyshyn's is that I claim the three levels are not views of the same system, the individual agent, "bearing an implementation relationship" (Newell, 1982). The KL can't be reduced to (implemented as) structures in an individual. Furthermore, I claim that this is the essential insight that distinguishes traditional AI from the evolving view of situated cognition research. Figure 13.6 illustrates this difference.

This idealized diagram shows the theoretician and agent occupying one environment (or social system). For robotic design, it may be practical to view the agent as being in an idealized, closed world, and hence, not the environment of the theoretician. They share an environment at least in the sense that the theoretician has some way of observing the agent's behavior. KL descriptions are shown as being part of the environment, in a space that other agents can access. This incorrectly leaves out silent speech and mental imagery. Strictly speaking there is a private perceptual space for each agent and a shared space of sensations. As I have stressed, the whole analysis is recursive—as for any agent behavior, KL descriptions themselves are emergent, arising through the interaction of the observer–agent with his environment.

Prior to Newell's KL paper, we might have shown the mind with everything "below the line" (encoded in a program or human memory):

```
            ------------------------------------------------
Knowledge  =  symbolic  structures  in  the  program
```

Newell separated the symbol level into two parts (Newell, 1982, p. 99):

```
Knowledge   =   observer's   description
------------------------------------------------
  symbols   (internal   to   the   agent)
```

The effect is to redefine knowledge as an ascription made by an observer and to begin to view representations in a new way: as perceivable forms that are used by an observer to articulate his or her beliefs and theories. My move has been to claim that this is how it is for the agent as well—all symbol manipulation is going on above the line, in the agent's behavior. Thus, I move the symbol level fully to the top, leaving only "syntactic" relations between neural processes inside. I claim that a KL description is just an instance of the general sense-making process by which agents produce and manipulate representations of their world and themselves:

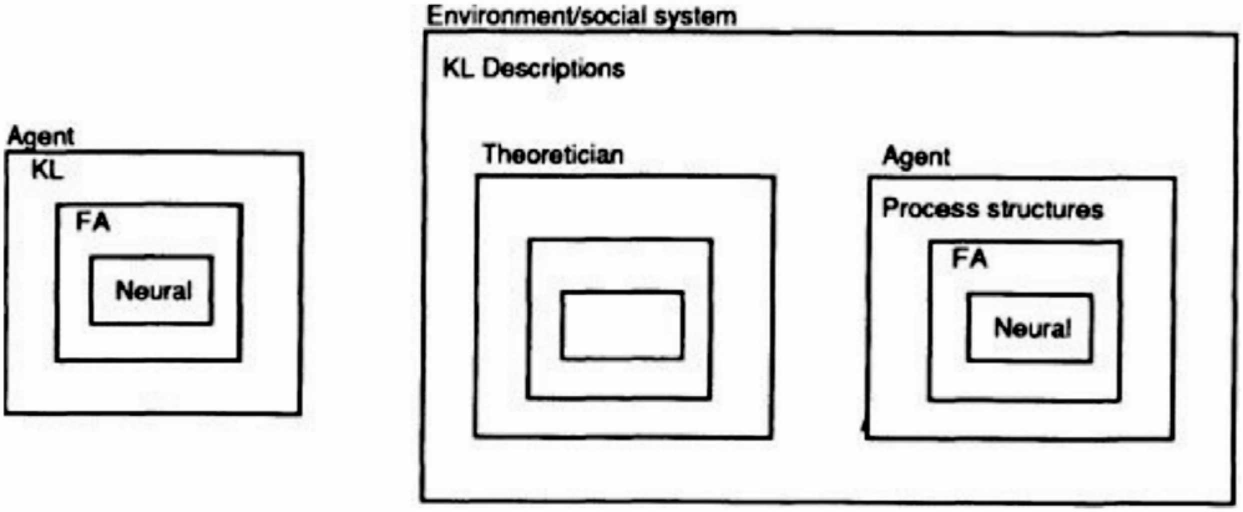
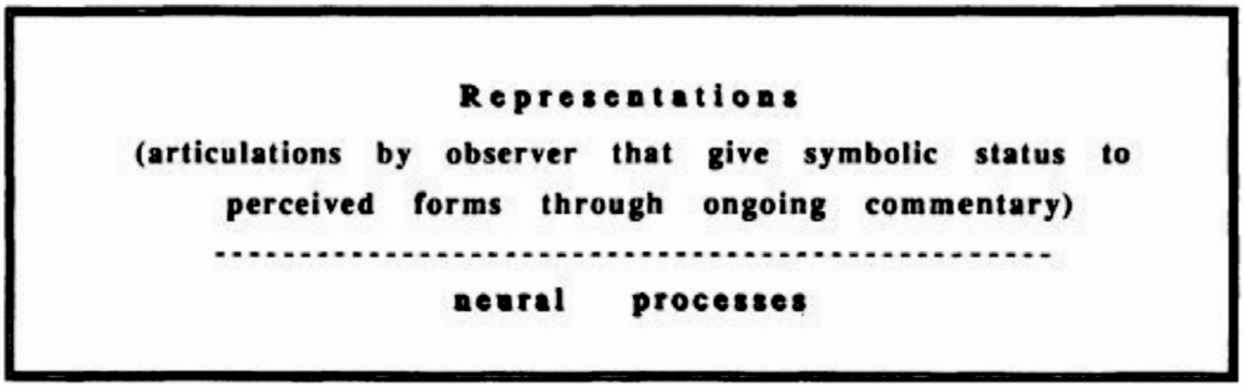


FIG. 13.6. Newell and Pylyshyn's Knowledge-Level (left) opposed to situated, interactive, relativistic view (FA = Functional Architecture).


Representations
(articulations by observer that give symbolic status to perceived forms through ongoing commentary)
neural processes

I call this the *externalization move*. In the most general sense, for example in the study of humans, the KL can't be reduced to (or restated exclusively in terms of) internal neural processes, because it's about the interactions that occur between an environment and neural processes. These interactions are manifested by agents who create and use representations. Because these representations are only interpreted within ongoing sequences of behavior and themselves emerge in an interaction between internal processes and the world, we cannot say that agent behavior is ever strictly caused by representations; rather, the sequence of behavior and changing representations are arising together, dialectically. Agent behavior is not conditional on objective characterizations of a situation, in the sense that representation creation and use could be reduced to (implemented as) rulelike mechanisms that match objectively-defined inputs.

Following Newell's lead, we have explored the hypothesis that knowledge is not stored as structures. I have brought forward arguments raised outside mainstream AI research by people in anthropology, philosophy, linguistics, and physics, specifically trying to account for the "situated" perspective. As such, this is an attempt to bridge the gap between different points of view, while remaining grounded in AI terminology as much as possible (cf. Newell, 1988, p. 521). My approach has been to retell the situated story from the perspective of designing a robot that interacts with its environment, in which the design and the observer's attributions are found to be on different levels: one a physical mechanism (the machine's design in terms of physically interacting parts), the other a description that names and accounts for patterns in how the robot and environment interact (the KL description).

The next step is to propose and implement a different kind of functional architecture, or better, show how SOAR should be modified. The trick is that we must change our understanding of three pivotal concepts: the nature of representations (what gives something symbolic status), memory (what is retained from previous activity), and perception (how input from fixed transducers is organized to constitute a new conception). The key move we must accomplish is to integrate perception into the reflective, conceptual process, rather than making it a peripheral process. I have argued that this requires a subjective notion of information, in fact, viewing information as the product created by the agent, externalized as representations, as opposed to something objective and supplied.

One approach is to now investigate the relations in a sequence of changes to representations, for example, the study of situated design (Allen, 1988). Another approach is to return to the functional architecture and ask what neural processes would support the production of (what appear to the observer as) recurrent, coherent phrases of behavior. Could we account for the practice effect and sense-making in a way that is consistent with the strong claims that place representations only in perceptual space? This is the research program I claim we are now faced with. I sketch out one possible approach in the following section.

A FUNCTIONAL ARCHITECTURE THAT MANIPULATES PROCESSES

> All the psychological schools and trends overlook the cardinal point that every thought is a generalization; they all study word and meaning without any reference to development. (Vygotsky, *Thought and Language*, 1934)

As Pylyshyn points out, heretofore there has been "only one nonquestion-begging answer" to the dilemma of how the semantics of representations could cause behavior, given that "only the material form of the representation could be causally efficacious," namely that "the brain is doing exactly what computers do when they compute numeric functions ..." (Pylyshyn, 1984, p. 39). I have argued that semantic attributions occur only in an ongoing sequence of behavior, in the form of commentary one representation (articulation) makes on another. That is, semantic attribution is an observer's statement about how one behavior relates to another. My objective in this section is to point the way to a new conception of the functional architecture that could in particular account for *how such representations in a sequence (the evolving commentary) are related to one another*. Specifically, I want to describe a functional architecture that could account for how phrases of behaviors are constructed, retained, and recombined (remaining fully aware that what constitutes a phrase is relative to an observer's frame of reference).

Having ruled out a functional architecture that manipulates descriptions of processes, we are left with the possibility of a mechanism that manipulates processes directly. That is to say, memory is the capacity to reenact a phrase of behavior and perception/learning, the process by which phrases are composed in an apparently hierarchical manner. Obviously, you won't find a program listing at the end of this paper that does this. Our field is at the intermediate stage of theorizing about what could be possible, using KL descriptions of the behavior the functional architecture must support. For this new beginning, special emphasis should be placed on highly-interactive behaviors like jazz improvisation, drawing, speaking, and ensemble performances of all kinds. These all place a premium on a cycle of movement, perception/reflection, and incremental modification that comments on what has come before, composing a coherent new form. Following Agre and Chapman's analysis, research should shift from geometry and algebra problem solving to examples of developing and never fully-definable routines, dialectically coupled to the agent's changing perceptions of its own interactions with the environment (recall AARON). By this, cognitive science would move from building in ontologies (however hyphenated or compiled) to finding ways that a process-oriented memory would embody and create (rather than describe) recurrent interactions the agent has with its world.

What follows is obviously speculative, but it makes the point that there is another way of talking about symbols, memory, and conceptualization, that gives perception a central place and avoids the semantic-attribution problem (how internal, unperceived representations could relate to what they are about). I start by discussing a simple example, elaborate upon Vygotsky's idea that speaking is conceptualizing, and finally attempt to pin down how phrases of behavior could be related to the neural processes that create them.

Tokens, Symbols, and Reference

Perlis states an essential question for AI research: "If a system employs symbols, in what sense are they symbolic, of what are they symbolic, and in what sense is it the system that makes them symbolic?" (Perlis, 1987). Many discussions of symbols, meaning, and reference in computer programs are based on a misconception about the nature of symbolizing and representing in human reasoning. A token (mark, sound, or anything perceived) becomes a symbol by virtue of comments people make about it, rather than being a property inherent in the identified thing itself. Today's computer programs combine tokens according to an observer's grammar-like descriptions of how the resultant behavior will appear from outside, while human behavior proceeds at some level directly from the remembered history of the processes that directly generate physical movements. The key changes in perspective called for here are:

- Human memory is not a store for things, but rather the functional capacity for creating and recombining phrases of activity.
- Representing, such as speaking, is a mode of perception, of claiming a new distinction, adding information that achieves coherence in the memory of processes.

My running example is how people talk about the Sydney Opera House:

> The New South Whale, they called it, the Operasaurus, a pack of French nuns playing football, an opera house with eight sheets to the wind. Then it was finished. The London *Times* said it was "the building of the century," and the Aussies shut up, looked again, and saw a pearl–pale sculpture glowing suspiciously like a national symbol on their waterfront. (Godwin, 1988, p. 75).

Consider the Sydney Opera House as an example of a token (albeit larger than most). The commentary about it as it was built is precisely how any observed object or event becomes a representation. Each comment (e.g., "a pack of French nuns playing football") provides a context for interpreting the structure, for viewing it in a new way. Most strikingly, notice what happened after the Opera House was finished —it glowed "suspiciously like national symbol." That something is declared a symbol *after its creation as a thing* contradicts the typical stance of AI research. In fact, this is how it always is. Something becomes symbolic by virtue of what people say about it, not for something inherent in the thing itself. This supplied context, coming after the occurrence and observation of the token itself, is what gives it meaning (Langer, 1958). I call this the *commentary model of cognition*.

The common sense point of view is that language refers to reality (Berger & Luckmann, 1967). For example, we say that a drawing or picture refers to something other than itself and this is what makes it a representation. However, this is backwards. Meaning or reference is not in the token or in my head prior to my speaking. In speaking, I am perceiving the token in a new way, providing a context with respect to which it can be interpreted as being meaningful. Viewing the Opera House as a ship, I might say it resembles eight sails. But that's just one of many interpretations. It isn't inherently a representation of any one thing, just what someone says. Certainly, we care a great deal about the designer's interpretation, but nevertheless, even the designer's statements are apart from the structure itself and prone to change.

Two aspects of reference need to be distinguished here. Suppose I observe the Opera House (the token, OH) then make a statement *about it*, calling it the New South Whale (the context, NSW, a pun about Sydney's location in New South Wales). The *formal* aspect of reference is the pact of mentioning or pointing to OH as a thing-in-itself when saying NSW. The *symbolic* aspect of reference is the act of saying what OH is about, what it refers to, by my comment NSW. Thus, we have a reciprocal action (Figure 13.7).

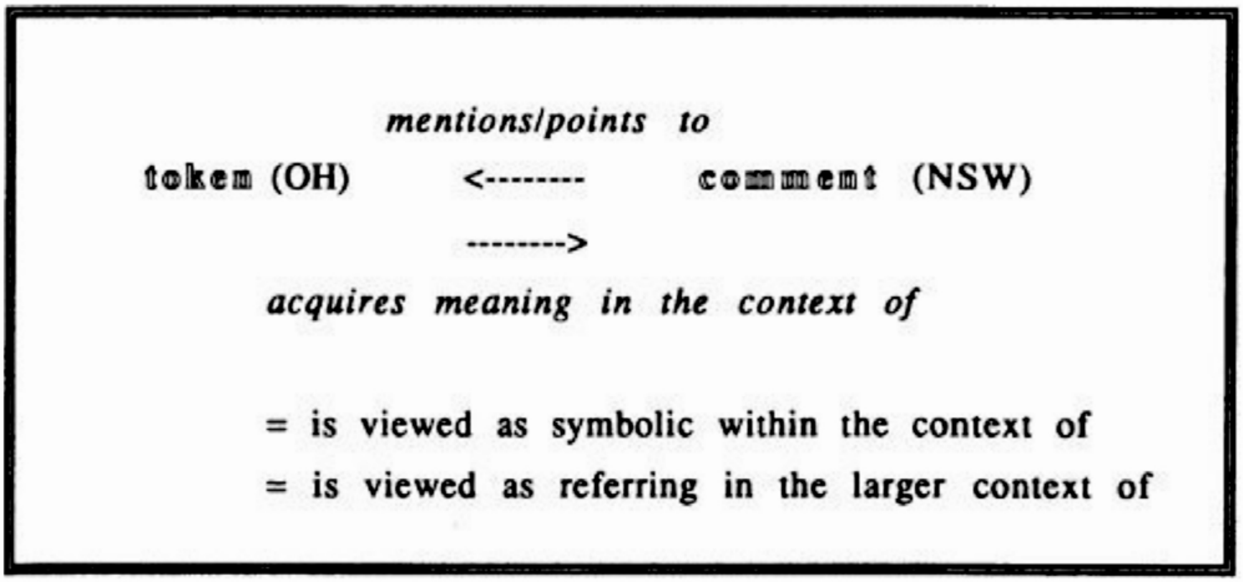

mentions/points to
token (OH) <-------- comment (NSW)
-------->
acquires meaning in the context of
= is viewed as symbolic within the context of
= is viewed as referring in the larger context of

FIG. 13.7. Relation of token to a comment that gives it symbolic status.

Representing (claiming that something symbolizes something for me) is what I do when I comment on (symbolic aspect) what has gone before (formal aspect). Our computer programs are of course full of symbols, because and only because of how we talk about them.

The phrase "attaching meaning to symbols" (Sloman, 1985) is incoherent—something becomes a symbol by virtue of our "attaching a meaning" to it. There isn't first a symbol, then the meaning. There is first something I refer to formally, as a token, in a remark, which supplies a context indicating what it is about, thus endowing it with symbolic status. We routinely say things like "the sentence implies a great deal," but such a way of talking disguises where meaning comes from, suggesting that it resides in the sentence itself.

Perhaps the most difficult "hidden matter" (Bateson, 1988) confusing our analysis is the idea that words refer to the world, that there is such a thing as objective knowledge (Tyler, 1978). This is an integral component of everyday understanding: There is something external to speakers that their meaning is about; language is not tautological; people are always trying to work through and define the contexts that shape them and that they must work within. But this orientation of common sense and science alike is not a good description of the activity of speaking itself and what people are doing when they use words. In a related analysis Perlis says, "There is a *presumption of an external object of thought*, something we take as real. Expressions or other internal forms (even images) do all the work, but at least one is momentarily taken as the thing-in-itself. We have no other way to refer, no casting our mind forward to external things" (Perlis, 1987).

By this definition, today's computer programs do not engage in symbolic reasoning or use or create representations—for any internal "interpretation" of their tokens is always grammatical and thus bounded by the axioms of combination. People speak and draw from the unformalized processes of their memory. Thus, to answer Sloman, who has struggled over these same issues, there is "a real distinction between understanding and mere manipulation." A quotation system, according to Perlis, allows a system to use "symbols" to refer: "the system itself has both symbol and symboled at hand." However, this is purely formal, grammatically defined reference. The program has no way to jump out (cf. Winograd & Flores, 1986). Human reference doesn't proceed from axioms of what kind of references are possible, that is formally, from a preclassification of behavior, but by directly recombining (the processes that generated) past sequences of behavior.[14]

The fact that we can relate to one of today's program as if it understands demonstrates our capacity to ascribe meaning, not something inherent in the program itself. Token-producing acts by a machine, just like your speaking to me or drawing a picture, or writing a note, are open to interpretation, a matter of what an observer says about them (Agre, 1988). A robot appears to obey commands, answer questions, and teach (cf. Sloman, 1985) because the observer says so. In this respect, a human user's responses to computer inquiries, for example during a consultation, creates *a combined system that is doing symbolic reasoning*, with the computer program playing a role no different from a numeric calculator as a manipulator of notations.

How then do representing actions follow from our experience, what do they do for us, how are they organized by the representations themselves?

Speaking as Conceptualizing/Perceiving

The quote by Vygotsky which opens this section, that all speaking is generalizing, contains a crucial insight. Speaking is not an act of translating a concept but of creating one. Speaking is an act of grasping, encompassing, taking in, contextualizing. It does so by pointing or mentioning. To mention is to include, to create a composed form. This is representing: creating a new order, perceiving a higher level of organization (Bateson, 1988).

Through the act of commentary, a token is seen in a new way. It acquires a larger meaning, which is to say that it is seen no longer as just a thing-in-itself (the form of the Opera House) but as part of a larger context (notion, idea, concept). This is what it is for something to have *meaning* or to be a *symbol*.

Of course, it's not "having a meaning" in the sense of a static property of a thing, but a matter of how it is perceived, by virtue of the context supplied, the comment. This context can't be defined, or rather isn't supplied as a definition. It is known tacitly, and is changed as much by the act of pointing at the token as the token is seen in a new light by the context (the idea of Australia is changed by including the Opera House). Thus, the token under interpretation becomes an "instance" of the context, a manifestation of it, tacitly changing the meaning of the context itself. This view is critically at odds with the AI view of representation and reasoning, in which the Opera House would be subsumed under an existing category as an example or instance. The commentary model holds that the act of subsumption is not a matter of matching a subsuming category, but of changing the category so that it includes the example, while changing the example so that it is included by the category.

Asking, "What is the meaning of the comment? To what does it refer?" is incoherent unless you didn't understand the comment itself. For example, it is only meaningful to ask, "What is the meaning of the statement that the Opera House is the New South Whale?" when you don't understand the connection. For you, I didn't succeed at supplying a context that gives the token meaning. Too often we assume in constructing formal knowledge representations that every question about meaning is meaningful, just because it is suggested by a formal calculus. Philosophers and linguists have argued that many questions about reference and meaning have been derived from formal analyses such as diagrams with links between words, spawning fruitless, impossible searches for a semantic calculus (rules that could generate the space of meaningful statements [Tyler, 1978; Wittgenstein, 1958]). But, to tie this to the larger themes of this paper, reference and meaning are incoherent without a notion of time and a selective, calibrating agent who perceives by naming. Time and agent constitute a frame of reference for the laws of rationality. Thus, the study of semantics is the study of the KL, of sense-making, of the conditions for generating and using representations.

If necessary, a context-supplying statement can be pinned down further by another comment, until finally the listeners (including the speaker, who is also hearing these ideas for the first time) are in a state of feeling that nothing further needs to be said, or more specifically, nothing more remains to be *done*. According to Bateson, this state is tauto-ecological—an interaction of relational consistency to what has been experienced before (the tautological processes of the brain) and the demands of the ongoing activity in which the comments are made, tacitly supplying meaning to them (the ecological processes of interaction with an environment).

The logic of an utterance is relative to how the words have been used before, the phrases they have been part of, the relations they have borne to other phrases. These are past processes of mentioning and referring, and having been generated by the combination of them, the current utterance bears an analogical relation to them. Schemas (recurrent phrases of behavior) emerge by the coherent recombination of past speaking processes, themselves composed of other processes. In this respect, every act of speaking is an act of conceptualization, of stepping towards an understanding, of composing a story, of being coherent.

Each act of speaking is a complete act of perceiving in itself. No further act of representation is needed. However, because each act of perceiving adds new information, something more might need to be said to re-establish the tautology of our understanding, to complete the story.[15] The distinction between recognition and generation is that one perceives meaning in a given thing and the other creates the thing (such as a program, drawing, or paragraph) iteratively, reflecting and commenting on each statement and the evolving whole.

In commenting on a sequence of behavior, we unify it, making it an item. We view it as a whole, perceiving one form. *The essence of representation is converting process* —both our memory of past activities and the ongoing activity we are engaged in— *into pattern* (Bateson, 1988). Speaking is a mode of perceiving (drawing distinctions, seeing forms). Converting process into pattern involves sampling, counting, defining bounds, claiming discontinuities in an inherently continuous world.

Bateson draws on cybernetics and genetics to help us understand what happens when a digitally randomized stochastic process (e.g., a genetic process) develops by interacting with a continuously randomized stochastic process (the environment). Regularities can be perceived in the structures that result, which biologists call the phenotype of the organism. Similarly, if we take the neural processes of memory as a conserving mechanism, similar to the effect of the genes, we can understand the regularities psychologists perceive in behavior as the product of development resulting from the interaction of two stochastic processes. We call these regularities *homologies* (Bateson, 1988); they constitute our law-like statements of how people behave, our KL descriptions.

From the perspective of the agent, the tautological recombination of past processes of memory, in a developmental process of interacting with an environment, is manifested in the grammatical appearance of everything people do, in the conceptual forms of speaking as well as the routines of skilled behavior. Every statement and action is a claim that the world is regular, a new generalization, and hence should be viewed as an ongoing attempt to reduce reflectively-constructed behaviors to routines. The space of resultant behaviors can be characterized in terms of analogies, of schemas, which are neither discrete nor continuous, neither fully coherent and definable nor arbitrary, but *constantly adapted* to the history of what we have done before and the ongoing demands of our interactions with the world around us. Bateson has characterized this as our satisficing nature. In a perverse, ever-changing world, in which no routine will work, it is experienced as the double bind of schizophrenia.

The Neural Processes of the Functional Architecture

Here I sketch some specifications deriving from the above discussion and contrast these with Pylyshyn's and Dennett's analyses.

Pylyshyn says that to implement the cognitive science approach, we will need "a system of transformations that preserve the semantic interpretations of the codes" (Pylyshyn, 1989). This is true. However, it is not required by the commentary model because interpretation occurs only by a sequence of perception and expressed commentary. The semantic relation is always an observer's post-hoc comment about such a sequence, and there are no internal codes that *describe* either the relation or process by which this meaning is created.

The question remains, what are the internal transformations that accomplish this creation of new conceptions and account for our reaching a state of satisfaction that we understand something? More mechanistically speaking, how is a phrase of behavior constructed from previous phrases? How is a new sequence "chunked" into a remembered phrase? What accounts for the compositional character by which impasses are resolved by perceiving higher orders of organization (Bartlett, 1977; Bateson, 1972)?

The functional architecture appears to have the following mechanisms for manipulating neural processes:

- direct "playing" of a previously enacted sequence,
- subsumption of phrases in hierarchies of substitutable actions,
- continuous substitution and recombination by integrating substitutions from multiple perspectives (corresponding to multiple parents in orthogonal hierarchies),
- interruption when recombination cannot proceed automatically,
- a reflective process by which a detail (usually an image) becomes the starting point for a new perception that constitutes a new, compositional organization (Bartlett, 1977),
- a memory for sequences created in this way (by interruption and attentive reflection) so a new phrase is created from the constructed sequence.

The key idea of impasse and rationalization based on a perceived detail comes from Bartlett. He also believes that emotional experiences during an impasse, or more generally an individual's "attitude" towards his current state, is a manifestation of the capability to come to terms with himself in a global way, an aspect of how the functional architecture can get things moving again by creating a new organization of processes. Another important idea is "feed forward" (emphasized by Pribram (1971) and Minsky (1985)), by which a process once begun is enhanced by propagation above to subsuming processes, which then enhance the activation of the included process. Combining these ideas, it is significant that an emotional attitude moves forward in memory of sequences (that is, it is remembered earlier when the sequence recurs), capturing the way in which an attitude serves to orient behavior. Further theorization might focus on the discrete nature of these processes, particularly how processes are manipulated at the level corresponding to phrases of behavior (where the temporal extent of a phrase is bounded practically only by the demands of the changing environment—a symphony is a phrase to the conductor, another phrase may be how you live a typical day).

The relations among processes reflect semantically acceptable sequences of commentary (from some observer's perspective). Figure 13.8 shows a sequence of actions (e.g., statements) A and A′, generated by physical processes P and P′.

The functional architecture must account for the production of process P after P′. The commentary model suggests that the primary relation is that of composition, so P′ subsumes P when A′ is a comment on A. The functional architecture is the underlying process that maintains coherence in the organization of neural processes (what Bateson called the "tautological relations" (Bateson, 1988)). Note again that process P doesn't represent action A—it generates it. Nor does process P manipulate any codes that describe A, just as you won't find a gene that describes a part of the body or a physiological process. The semantic relation an observer claims holds between A and A′ is not pre-encoded, but rather reflects the logic of the construction of process P′ from P, achieved by the functional architecture's maintaining coherence with respect to previously constructed processes. The functional architecture's transformations do not so much "preserve semantic coherence" (Pylyshyn, 1984, p. 249), in the sense of adhering to rulelike descriptions of what is possible, as create/achieve/accomplish coherence in ongoing construction of processes that maintain a subsumption relationship through (or during) their production of sequential actions.

Perception plays a key role here. The commentary model suggests that moving from A to A′ involves constructing a process such that the perception that leads to statement A′ subsumes the perception that led to statement A. For example, when someone first constructed the pun that the OH was NSW (A), a subsequent explanation that the OH was in New South Wales (A′) provides a way of seeing the first action, so that A′ comments on A, giving it meaning. Notice that these "semantic relations" are clearly not static relations among these words, but rather are more properly characterized as an agent's sequence of perceptions. While it might appear that the explanation A′ is what generated the pun A, it is just an observer's restatement of the speaker's perception when saying A.

$$\begin{tikzcd} A \arrow[r] & A' \\ P \arrow[u] \arrow[r] & P' \arrow[u] \end{tikzcd}$$

FIG. 13.8. Two physical actions (A and A′) generated by neural processes.

Another example may be helpful. Consider what happened when a colleague came up to me after a talk and said that I should reverse the order in which I presented my slides. This forced an impasse; now I couldn't see my talk as being coherent. What could be wrong? Proceeding from the idea of ordering, I viewed what as I had done and posed it as the question, "Why did I do X (show theory) before Y (examples)?" I then formulated a rule: "I usually present material in the order I conceive it." From this, I reabstracted X and Y so a new generalization would subsume what I had done in the past that worked: "Show concrete before theory." I then realized that usually I develop a theoretical description after writing programs; this was a different experience, the theory came first. That is, I went back to explain why the new rule would work in the present case—what's odd about this case?

Notice how this is related to explanation-based learning (EBL), but distinctly different. Like EBL, the orientation is entirely procedural, tracking back through a particular example to explain why a different behavior wasn't produced. However, the rule, "present material in the order conceived" was not required to generate past behaviors, it was generated after the fact, as a rationalization that describes what I do. Broadly speaking, it was a reflection over a range of individual behaviors, abstracting them in order to answer the question, "What do I generally do?" Applied to the current context, the rule fails to produce the desired behavior, so the entire sequence must be reabstracted, in a way that would produce a different behavior in the current case.

Thus, the abstraction, "Show concrete before theory" is a new way of talking about the past, subsuming those activities *as if it generated them*. It provides a useful new way of seeing the overall activity of talk preparation; that is, it provides a way of organizing the activity. In this respect, the new perception "Show concrete before theory" is about the past activities; it sees them in a new way, commenting on them, composing/grouping them in a new way, representing them. In this respect, a way of seeing or talking about behavior creates a new order, subsuming old behaviors and providing a means for organizing subsequent ones. We say, "Perception subsumes action." Ways of seeing/talking orient the particulars of what we do by directly enabling them as processes.

In the future, this new rule is likely to be remembered when making slides because a new routine has not developed yet and the process of making slides is now subsumed by this commentary. The new perception will "move forward" (it originally came after the process of making slides and giving talks) because it is about the process of making slides. That is, it will be activated by upwards propagation when the lower-level activity is engaged in (by subsumption of processes).[16] Summarizing the relation to EBL, we find that the ideas of impasse, reminding, and reasoning about cases are central, but there are no internal representations of behaviors, just the processes themselves, which are activated and recreated, and their actual or imagined results commented upon. This comment is not saved as a rule that generates behavior directly. Rather, the comment has the immediate effect of reorganizing processes it is about and its articulation in the future will provide a representation for more deliberately stepping through these processes.

A few related observations: When we as observers say that A′ bears an analogic relationship to A, we must recreate for ourselves the perceptions and hence the underlying organization of processes such that P′ subsumes P for us. This provocatively suggests that the functional architecture provides us with the capability to start up multiple processes and hold them active as we attempt a new organization that could subsume them. This is in essence the capability we require to deal with impasses, and as Bartlett suggests, is why consciousness is useful. Furthermore, the creation of underlying coherence involves a cycle of perception (reflection on what has been said or done before) and a new physical behavior. The only way of moving forward is incrementally, by doing something, and commenting on that. Reflection is inherently a process of actual behaving, not hidden cogitating (though you needn't talk out loud). There appears to be a connection between deliberate reflection (e.g., creating the sequence of A followed by A′) and the creation of A-A′ as a new phrase of behavior. We should remember that such learned "chunks" are not substances, but processes, implemented as strengthened connections between the neurons (Edelman, 1987).

All of this maintains the traditional view that symbols are semantically interpreted structures. However, by the externalization move—moving semantic relations out into the perceptual space of an observer—we can talk more coherently about relation of neural processes that produce behaviors and the relation of perception to repetition and creation of new processes. That is, there is reason to believe that cognitive science and AI research has identified enough KL phenomena so we can reasonably look to the neural process level for functional architecture mechanisms that could support the conceptualization, learning, reminding we have described at the KL.

Table 13.2 summarizes the tri-level view according to Pylyshyn-Newell, Dennett, and the descriptions given here. To bring together a few key ideas:

- The relation of the levels is not "implementational"; however, in our explanatory account it is generative; it enables predictions to be made.
- The claim that there is a semantic relation between A and A′ is a KL description and is always made by an observer, relative to his or her frame of reference. Such statements (for example, that "A → A′ is a rule") are representations made by an observer and are not structures that physically cause the agent's behavior (rather, that is what happens when moving from P to P′).
- The semantic level concerns the result of interactions that occur as the functional architecture maintains internal coherence relative to its activities in the world; that is, the agent's resultant behavior is situated.
- Semantic or KL descriptions are expressed as categories and laws of behavior, thus they express the interaction between beliefs, goals, and activities as abstractions and rewrite rules. That is, a KL theory is analogous to a natural language grammar; more specifically, a natural language grammar is one aspect of a complete KL theory.
- We need the semantic level in order to give principled explanations of why, of all possibilities that the functional architecture allows, certain behaviors are favored. These constraints, which lie outside the machine's functional architecture, are the result of emergent effects from its developmental interaction with the environment. Most notably, psychiatry often requires consideration of the agent's social organization in order to explain behavioral impasses. Thus, the principles that "prevent semantically deviant states from occurring" (Pylyshyn, 1984, p. 37) lie in the combined system, not within the individual agent.

TABLE 13.2
The Tri-Level Architectue (Knowledge, Symbolic, and Physical Levels) from Different Perspectives.

Newell-Pylyshyn	*Dennett*	*Clancey*
representational "semantic" (Knowledge Level)	Intentional	observer's description and interpretation of perceived regularities in behavior of an agent-world system
symbols & symbol-manipulation rules	Subpersonal	self-organizing neural processes (subsumption & sequence relations

(Symbol Level)	between past processes)
physical, neural transducer processes (Physical Level)	processes that create macro neural processes (feed forward & upward activation; composition & sequence creation)

- A KL description is an example of the perception of information, of conception in general. "Information content" is thus relative to an observer. Specifically, "input information" is not objective, but relative to the agent's perception. Without a perceptual process to create information, there is only data processing. In conclusion, placing perception prior to manipulation of representations has the process inside-out: perception creates representations.

The mistake of cognitive science has been to place observer-relative and environment-relative regularities in the machine, as pre-existing descriptive structures. I have described here how a different view of representation, memory, perception, and even language itself suggests a simpler functional architecture, with far more flexible capabilities for symbolic reasoning.

SYMPOSIUM PAPERS RECONSIDERED

The symposium papers make a clear attempt to establish a foundation from which intelligent behavior can be characterized and generated. The authors suggest that we should:

1. Situate the program in the world; view interactive and real-time processing as the primary constraint (e.g., GUARDIAN, INSECTS).
2. Develop a formal framework of representational primitives for grounding learning or rationality; attempt to relate uniformity, expressiveness, reflection, flexibility, learning, efficiency, responsiveness, etc. (e.g., PRODIGY, THEO, SOAR, Genesereth).

My critique of this work is that it fails to address directly the frame of reference issues in the modeling of dynamic systems. Following from my proposal that a new kind of mechanism should be sought, the question naturally arises, how could we tell that an agent has a functional architecture that directly manipulates processes, versus one of today's AI programs, in which processes are labeled, stored, and grammatically recombined? This is a tricky question, for it presupposes that we could use grammatical methods to construct a program that would be so good as to resemble human capability and fool us. Of course, nothing today comes close. A better form of the original question is to ask, what precisely are the capabilities that today's programs lack?

In what follows, I characterize the contributions of the various projects, focusing on how the theoretical framework could be improved and thereby improve the capabilities of the programs.

Mixed Architectures

As claimed in the introduction, this research falls into two categories, knowledge engineering and the study of intelligence. The knowledge engineering contributions, exemplified by PRODIGY and GUARDIAN illustrate how a KL formalization of reasoning and learning can be used to produce a useful program. Both systems make a contribution to KL theories, but they exemplify especially well how these theories can be integrated into a complex system that can control complex mechanisms in real-time (e.g., the satisficing cycle of GUARDIAN), as well as assist theorists in improving the KL descriptions (e.g., the EBL process in PRODIGY). Both systems are *mixed architectures* (Newell, 1982); the researchers make little distinction between the functional architecture and the knowledge level. For example, Hayes-Roth (this volume, chap. 11) describes "backlog monitors" and "new-focus monitors," without making clear whether these are KL descriptions or to be viewed as distinct physical mechanisms. Similarly, in PRODIGY there are both search control rules and domain-schema rules; it is not clear how these KL descriptions map onto mechanisms in the functional architecture (e.g., are there two separate memories?).

Formal "Objective" Analysis

In the formal frameworks presented by Anderson and Genesereth, we find no mention of the observer-relativity of KL descriptions. Genesereth says, "There is a symbol for every state of the agent's environment, every percept, and every action," suggesting that an environment can be described objectively or that the agent's perceptions can be exhaustively predefined in terms of primitive symbols. (Of course, these are tokens, not symbolically interpreted representations.) In essence, Genesereth starts with the idea of a machine as a calculator operating on non-numeric tokens (a grammar calculator) and provides a formal analysis of tradeoffs in compilation (which gives speed) versus runtime processing (which gives flexibility). This is a contribution to computer science and could justify design decisions in an architecture like GUARDIAN'S. The analytic techniques being developed here might later prove useful for describing a mechanism with self-organizing processes.

Anderson's paper can be viewed as a reaction against the complexity of AI architectures. He attempts to reground the study of intelligence in a study of the "information processing requirements" posed by the task and environment. The idea of incorporating a formal description of the environment is completely consistent with the view of the KL I have provided. However, Anderson is wrong to suggest that this formalization is objective, that the world somehow is given to us in predefined categories that we only have to discover and name. Information is not objectively-supplied data; Anderson is confusing the theoretician's observations with the subject's constructive acts of perception and representation (as Simon says, "It is the organism that constructs a problem space and strategy to deal with the task environment" (Simon, 1988)[17]).

However, I believe that Anderson has several valid points that Simon skips over. First, when Anderson contrasts the description of a behavioral function to mechanism, he means physical mechanism, the functional architecture. Thus his paper can be viewed as calling us to separate our theorist's perspective ("focusing on the information processing problem") from what is going on in the agent ("the information processing mechanism").

Second, Anderson strives for a more general theory, above the level of specific KL attributions, to characterize task demands in a way that could *frame* the information processing problem. However, following my analysis, we would want to focus not on an objectively-defined environment, but on *interactional* aspects of behavior. That is, following Agre, we would describe task demands in terms of how dynamic aspects of the environment constrain the use of representations and provide opportunities and resources for, or work against, the evolution of routines.

Putting this together, to frame the information-processing problem we need to consider the interaction of the observer-theorist, the functional architecture, and the environment.

Memory

Memory is a clearly a central issue in AI architecture research. Three of the papers in particular can be viewed as attempts to take a strong stand on what memory is. Brooks rejects the idea of maps of the world, that is, static data structures that describe things in the world and are apart from the processing mechanism.[18] Rosenbloom et al. have steadily moved towards the idea that representations are generated in a perceived space ("working memory"), but they retain the idea that memory consists of retrievable descriptive structures and that "knowledge is stored." Mitchell, apparently in response to a perceived weakness in SOAR, provides direct support for hierarchical organization of concepts; thus, THEO'S memory is a representation of a classification of concepts. In short, Brooks sweeps a theorist's KL descriptions out the door, placing them outside the functional architecture, while Rosenbloom et al. and Mitchell still try to find clever ways of encoding an observer's descriptions inside the machine. Brooks attempts to build a robot, while the others continue to tell us how such a machine might appear.

To combine these ideas, we might follow Brooks by doing away with the idea of a separate memory store. To account for conceptualization and learning, we could find some way to dynamically reconfigure a subsumption architecture, such that prior configurations are marked in some way and more prone to reconstruction. Furthermore, higher processes would not only control how the lower processes occur, they would control the network configuration process itself. Thus subsumption of processes would support hierarchical conceptualization, memory, and learning directly.

The one weakness that is most glaring in these programs is that they never conceive of anything. The world is precarved by the designer and these elements are grammatically recombined (recall the metaphor of the inverted picture puzzle). Chunking apparently models an important aspect of how new processes are created and reenacted in human memory, however it doesn't account for the compositional aspect of process creation and control (which THEO models in KL terms). This compositional process I have claimed is at the heart of symbolic interpretation, of making sense, of conceptualization. In essence, we must return to basic issues in natural language comprehension. Recharacterizing "reminding" in terms of Bartlett's impasse-rationalization model would be a good start.[19]

CONCLUSIONS: A SCIENCE IN TRANSITION

In this chapter I have sharply called into question our analytic techniques for specifying architectural requirements for the design of an intelligent machine. We have ignored emergent interactional effects and the observer status of our theories. The knowledge-level patterns and processes we describe are partly an artifact of our own sense-making (any theory must state regularities; it's a property of language) and partly a result of routines that have evolved in the agent's interactions with its environment. We have ignored the dynamic and selective aspects of perception. My claim is that the foundation of AI research is faulty. Our ideas of memory, perception, and learning have been distorted because we have viewed knowledge as objective substance, as structures that can somehow statically capture meaning and store it. In contrast, I have argued that semantic interpretation exists only as ongoing commentary, through a process of creating representations in our speaking, gestures, and notations.

The arguments given here strongly build on and emphasize the idea of intelligent behavior as symbol manipulation, however these symbols are moved outside to where they can be perceived, in what I call the externalization move. Memory is not a storage for symbols, or any kind of *place* at all, rather it is a capacity to recreate and recombine processes that have previously related perceptions to actions. By the composition of these processes, perceptions organize behaviors, and hence ways of speaking (concepts) can be *about* what we do. Through primitive capabilities to compose hierarchically and sequentially, the functional architecture creates new routines so that behavior can proceed automatically, without conscious reflection and conceptualization that must occur when impasses arise.

The arguments given here retain the materialist view that intelligent machines can be built. However, my strong claim is that we have an inadequate understanding of the phenomenon to be replicated and (very likely) an inadequate theoretical understanding of the mechanisms that would provide engineering tools for building such machines. We should take a lesson from lasers, holography, VLSI, molecular genetics, etc. that striking advances in the design of machines are built upon fundamental discoveries about microlevel processes; some crucial properties of neural-level processes may remain to be found. The entire notion of computation must be broadened beyond the idea of a stored program operating on data structures.

The guts of our robots are too rigid because we have supposed that the mechanism must operate on descriptions of how the behavior should appear, rather than focusing on simpler mechanisms that would directly respond to and organize stimuli in an immediate way, without intervening descriptions of what is about to occur.

The main argument of this chapter is a rejection of the idea that the functional architecture should "directly support knowledge," which the paper by Rosenbloom, et al. focuses on. Rather, building from Newell's KL paper, I have shown that knowledge is not physically realized (stored) in the structures of the machine; it is "never in hand." Knowledge, in the form of representations about something, only exists in interpretive comments, in ongoing claims about the nature of the world, which themselves are only classified and interpreted as having semantic import by an observer. In this respect, gestures exemplify the nature of representational acts. Gestures are semantically interpretable, but generally exist (are produced) without being perceived this way, at least with the same level of attention and commentary we give to what a person is actually saying.[20] Everything we call a representation (a spoken phrase, a written word, a drawing, an equation, etc.) is generated with the same immediacy as a gesture (not translated from an internal description of it, but created for the first time in the movement itself). The difference is that we generally pay attention to the ongoing sequence of representations, trying to interpret, and immediately respond with another comment. (In this respect, gestures are produced like dreams– coherent, interpretable, but not observed by agents and not commented upon.)

All knowledge–level descriptions are relative to an observer's frame of reference. "Relative" means not just that "different agents know different things" or even "different agents disagree about the world." But rather KL descriptions are:

- *Relative to an observer's view* (a perceived pattern, the result of processes interacting over time) versus an individual participant's view of moment-by-moment interactions.

- *Emergent from the dynamics of interactions*, not ascribable to an individual's action or planning. It is not just that the task environment determines behavior; rather, our law-like models describe the historical, developmental product of the interaction, not mechanisms in the agent that generated his or her moment-by-moment responses. Hence, Brooks, Cohen, Agre, and Rosenschein et al. characterize the functional architecture in terms of the dynamics of movement and internal state, characterizing perception as part of ongoing activity in some changing, interactive environment. (Indeed, Edelman and others claim that without movement or change relative to a point of view there is no perception.)
- *Interpreted in "every next use"* (Agre, 1988). The meaning of a representation can't be characterized by a static structure, rather it is recomposed, reconceived, and reperceived with every new expression.

The essence of this analysis that we should view Figure 13.2 as the framework for the study of intelligence. We should continue to develop our KL theories; for example, the work in explanation-based learning should continue to provide a useful competence model that can focus the design of a functional architecture. However, more work like Brooks' INSECTS is needed, in which the agent is a robot and sensation/movements are produced without building in maps of the world or how the robot's behavior will appear. That is, more researchers should come forward with strong claims to the effect, "My machine does not work by interpreting a KL description of its behavior." In this respect, there appears to be an opportunity to combine SOAR and the INSECTS work, throwing out the idea of a production rule memory.

This chapter has also briefly introduced the *commentary model of cognition*, which has the following implications for the design of a functional architecture:

- Reformulate the nature of reflection. The construction and use of representations is inherently a process of commentary and revision in the agent's behavior; it is not "inspecting" or "reading unperceivable structures."
- Reformulate chunking as a means of re-enacting any activity, clarifying how memory is nonrepresentational and all behavior, including reflection and commentary itself, becomes regularized.
- Reformulate conceptualization, exemplified by rationalization at an impasse, as a recomposition of ways of perceiving and behaving, in the form of incremental commentary, by which sequences of behavior are viewed as a unit and hence formed into a new chunk; conceptualization is not reading out, translating, or recombining preconceived descriptions in memory.
- Adopt a more comprehensive view of understanding (making sense) as a primary

high-level function that is directly supported by the hardware (relates to diagnosis, dreaming, remembering, etc. as story understanding).

A great deal of reading and synthesis underlies the arguments in this paper. Making progress requires borrowing ideas from many different fields and selectively reinterpreting what people have said. The most important works, which I strongly recommend to anyone working in this area, are those by Bartlett (the nature of conceptualization and comprehension), Tyler (relation between language, thought, and formal theory), Gregory (the subjective nature of physics), Braitenberg (ways of composing simple mechanisms to get complicated, dynamic behaviors), Agre (the open nature of representations), Newell and Pylyshyn (how to talk precisely about mechanisms and architecture), Winograd and Flores (the concepts of social commitment, breakdown, and unarticulated background), Bateson (development of interacting systems), and Wittgenstein and Ryle (the original, commonsense analysis that inspired Tyler, Gregory, and Bateson).

Much of this work has been ignored (Bartlett's being the most glaring example) because it is at odds with the cognitive science view of mind. If nothing else, I hope that my discussion here has convinced the reader that there are other perspectives —other frames of reference—that can prove useful for constructing intelligent machines. Strikingly, much of this work directly relates to our interests, yet Braitenberg isn't cited by Brooks, and Bartlett's contrary results are simply ignored by everyone. Indeed, just to realize how AI is historically related to non–linear programming, cybernetics, and general systems theory would seem to be the most basic requirement for any beginning student. If my analysis is right, the study of dynamic systems pioneered by others will soon become of central concern to AI, and most of this earlier work and its current developments (e.g., Prigogine & Stengers, 1984) will be reintegrated into the field.

ACKNOWLEDGMENTS

I am grateful for the support and encouragement of my colleagues at the Institute for Research on Learning, without which I might not have attempted to step outside prevailing views about cognition. John McDermott and Kurt VanLehn struggled through the first working draft and helped me produce something understandable to our AI colleagues. In particular, John suggested that I look more carefully at Newell's analyses. I would especially like to thank Allen and Kurt for inviting me to be a discussant at the symposium in May 1988, and hence giving me an excuse to spend a year reconsidering the foundations of AI. Many of the ideas were also developed for a keynote address presented at AI-88, Adelaide, Australia, in November, 1988. The section entitled "Frames of reference for designing intelligent machines" originally appeared as “The frame of reference problem in cognitive modeling,” presented at the Cognitive Science Annual Meeting, August, 1989. The sections entitled “Tokens, symbols, and reference” and “Speaking as conceptualizing/perceiving” have been circulated as “Symbols and Computer Programs.” Paul Duguid provided useful comments on this version. This research has been supported in part by ONR Contract N00014–85K-0305 and gifts from the Digital Equipment and Xerox Corporations.

REFERENCES

Agre, P. E. (1988). *The dynamic structure of everyday life*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Alexander, J. H., Freiling, M. J., Shulman, S. J., Staley, J. L., Rehfuss, S., & Messick, M. (1986). Knowledge level engineering: Ontological analysis. *Proceedings of the National Conference on Artificial Intelligence*, pp. 963–968.

Allen, C. (1988). *Situated design*. Unpublished dissertation for Master of Science in Design Studies. Carnegie Mellon University, Department of Computer Science.

Bartlett, F. C. (1977). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press. (Reprint of original 1932 edition.)

Bateson, G. (1972). *Steps to an ecology of mind*. New York: Ballentine Books.

Bateson, G. (1988). *Mind and nature: A necessary unity*. New York: Bantam.

Berger, P. L., & Luckmann, T. (1967). *The social construction of reality: A treatise in the sociology of knowledge*. Garden City: Anchor Books.

Braitenberg, V. (1984). *Vehicles: Experiments in synthetic psychology*. Cambridge, MA: MIT Press.

Bickhard, M. H., & Richie, D. M. (1983). *On the nature of representation: A case study of James Gibson’s theory of perception*. New York: Praeger.

Chandrasekaran, B. (1986). Generic tasks in knowledge-based reasoning: High-level building blocks for expert system design. *IEEE Expert, 1*, 23–29.

Clancy, T. (1986). *Red storm rising*. New York: Berkley Books.

Clancey, W. J. (1983a.) The advantages of abstract control knowledge in expert system design. *Proceedings of the National Conference on Artificial Intelligence*, pp. 74–78.

Clancey, W. J. (1983b). The epistemology of a rule-based expert system. *Artificial Intelligence, 20*(3), 215–252.

Clancey, W. J. (1985). Heuristic classification. *Artificial Intelligence, 27*, 289–350.

Clancey, W. J. (1987a). From Guidon to Neomycin and Heracles in twenty short lessons: ONR final report, 1979–1985. In A. vanLamsweerde (Ed.), *Current issues in expert systems* (pp. 79–123). London: Academic Press. Also *The AI Magazine, 7*(3), 40–60, Conference, 1986.

Clancey, W. J. (1987b). Review of Winograd and Flores’s “Understanding Computers and Cognition.” *Artificial Intelligence, 31*, 232–250.

Clancey, W. J. (1988). Acquiring, representing, and evaluating a competence model of diagnosis. In M. T. H. Chi, R. Glaser, & M. J. Farr (Eds.), *The nature of expertise* (pp. 343–418). Hillsdale, NJ: Lawrence Erlbaum Associates.

Clancey, W. J. (1989). Viewing knowledge bases as qualitative models. *IEEE Expert 4*, 9–23, Summer.

Clancey, W. J. (in preparation). Interactive control structures: Evidence for a compositional neural architecture.

Cohen, H. (1988). How to draw three people in a botanical garden. *Proceedings of the Seventh National Conference on Artificial Intelligence*. Minneapolis-St. Paul, pp. 846–855.

Cohen, P. (1989). *Why knowledge systems research is in trouble, and what we can do about it.* COINS Technical Report 89–81. University of Massachusetts.

Dennett, D. C. (1988). Precis of “The Intentional Stance.” *Behavioral and Brain Sciences 11*, 495–546.

Dietterich, T. G. (1986). Learning at the knowledge level. *Machine Learning, 1*, 287–316.

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. New York: Harper & Row.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection*. New York: Basic Books.

Godwin, J. (1988). *Frommer's Australia on $30 a Day*. New York: Simon & Schuster.

Gregory, B. (1988). *Inventing reality: Physics as language*. New York: Wiley.

Hayes-Roth, B. (1985). A blackboard architecture for control. *Artificial Intelligence, 26*, 251–321.

Hayes-Roth, B., Hewitt, M., Vaughn M., Johnson, T. R. & Garvey, A. (1988). *ACCORD: A framework for a class of design tasks*. KSL Technical Report 88–19, Computer Science Department, Stanford University.

Heisenberg, W. (1962). *Physics and philosophy: The revolution in modern science*. New York: Harper & Row.

Johnson, T. R., Smith, J. W., & Chandrasekaran, B. (1989). Generic tasks and SOAR. LAIR Tech. Rep. Ohio State University.

Kaelbling, L. P. (1988). Goals as parallel program specifications. *Proceedings of the Seventh National Conference on Artificial Intelligence*. Minneapolis-St. Paul pp. 60–65.

Langer, S. (1958). *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. New York: Mentor.

Lave, J. (1988). *Cognition in practice*. Cambridge: Cambridge University Press.

Minsky, M. (1985). *The society of mind*. New York: Simon and Schuster.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York: W. H. Freeman.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87–127.

Newell, A. (1988). The intentional stance and the knowledge level. *Behavioral and Brain Sciences, 11*, 520–522.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard.

Perlis, D. (1987). How can a program mean? In J. McDermott (Ed.), *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan (pp. 163–166). San Mateo, CA: Morgan-Kaufmann.

Prigogine, I., & Stengers, I. (1984). *Order out of chaos*. New York: Bantam Books.

Pribram, K. H. (1971). *Languages of the brain: Experimental paradoxes and principles of neuropsychology*. Monterey: Brooks/Cole.

Pylyshyn, Z. W. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge, MA: The MIT Press.

Pylyshyn, Z. W. (1989). On "Computation and Cognition: Toward a Foundation for Cognitive Science," a response to the reviews by A. K. Mackworth and M. J. Stefik. *Artificial Intelligence, 38*, 248–251.

Reeke, G. N., & Edelman, G. M. (1988). Real brains and artificial intelligence. *Daedalus, 117*, (1) Winter, "Artificial Intelligence" issue.

Rommetveit, R. (1987). Meaning, context, an control: Convergent trends and controversial issues in current social-scientific research on human cognition and communication. *Inquiry, 30*, 77–79.

Rosenfield, I. (1988). *The invention of memory: A new view of the brain*. New York: Basic Books.

Rosenschein, S. J. (1985). *Formal theories of knowledge in AI and robotics*. SRI Technical Note 362.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing*. Cambridge, MA: MIT Press.

Ryle, G. (1949). *The concept of mind*. New York: Barnes & Noble.

Searle, J. R. (1984). *Minds, brains, and science*. Cambridge, MA: Harvard University Press.

Simon, H. A. (1969). *The sciences of the artificial*. Cambridge, MA: MIT Press.

Sloman, A. (1985). What enables a machine to understand? In A. Toshi (Ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, (pp. 995–1001). San Mateo, CA: Morgan-Kaufmann.

Smith, B. (1987). Two lessons of logic. *Computational Intelligence, 3*, 218.

Steels, L. (1989). Cooperation through self-organisation. In Y. DeMazeau and J. P. Muller (Eds.), *Multi-agent systems*. Amsterdam: North-Holland Publishers.

Stefik, M. J. (1989). Review of "Computation and Cognition: Toward a Foundation for Cognitive Science, by Z. W. Pylyshyn." *Artificial Intelligence, 38*, 241–247.

Stucky, S. (1987). *The situated processing of situated language* (Tech. Rep. No. CSLI-87–80). Stanford University, Stanford, CA.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge: Cambridge Press.

Tyler, S. (1978). *The said and the unsaid: Mind, meaning, and culture*. New York: Academic Press.

VanLehn, K. (1988). Student modeling. In M. Polson & J. Richardson (Eds.), *Foundations of intelligent tutoring systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Vygotsky, L. S. (1966). *Thought and language*. (E. Hanfmann & G. Vakar, Trans.). Cambridge, MA: MIT Press. (Original work published 1934)

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex.

Wittgenstein, L. (1958). *Philosophical investigations*. New York: Macmillan.

[1] interpretation of abbreviations: SOAR (Rosenbloom et al., chap. 4 in this volume), THEO (Mitchell et al., chap. 12 in this volume), PRODIGY (Minton et al., chap. 9 in this volume), BB1 (Hayes-Roth et al., 1985), HERACLES (Clancey, 1987a), ACCORD (Hayes-Roth et al., 1988), GUARDIAN (Hayes-Roth, chap. 11 in this volume), INSECTS (Brooks, chap. 8 in this volume). Author names refer to papers from this symposium: (Genesereth, chap. 10 in this volume) and (Anderson, chap. 1 in this volume). RIC is an abbreviation for representation, inference, and control.

[2] Newell discusses comprehension operators in SOAR/UTC (Newell, 1990). My interest here is not to argue the details of the case, but just to raise the issue and how it relates to foundational concerns about memory.

[3] For an elegant reformulation and extension of Brooks' work, which models cooperating agents in terms of self-organization and dissipative structures using quantitative optimality criteria, see Steels (1989).

[4] Here I am to clarify Winograd and Flores's analysis. Representations (e.g., a set of terms and relations) are fixed as visual and syntactic entities within a program, but they are not fixed in their meaning to us. Thus, a fixed nature is not inherent in representations, just the notations themselves and what a given program can make of them. After reading Winograd and Flores, I wrongly thought that a representation (e.g., something I say) is a fixed, unchanging statement about the world. It has a momentary property of fixing attention within a sequence of utterances, but otherwise there is no fixed meaning attached or attachable to any human statement. The meaning is always what I say it is to me (or what you say it is) in every next interpretation. Thus, people do not have to cope with a "fixed, meaning-separated" aspect of representations; that's the problem that computer programs have. Nevertheless, notations are fixed forms in themselves, and this unchanging nature may invite rote interpretations. Tyler reminds us how representations separate us from experience: "Speaking is the alienation of thought from action, writing is the alienation of language from speech, and linguistics is the alienation of language from self" (Tyler, 1978, p. 17).

[5] I strongly resist calling something internal that cannot be perceived a "symbol," for these postulated tokens/forms can never be commented on and thus can never be given symbolic status by the person. Since they aren't themselves semantically interpreted, we can't call them "codes" either. "Symbolic code" is Pylyshyn's terminology.

[6] This became clear during the development of NEOMYCIN. We abstracted metarules such as "when testing a hypothesis, first seek evidence of enabling conditions" from a physician's initial questions when confirming disease hypotheses. For viral meningitis, he asked if a flu was going around; for fungal meningitis, he asked where the patient had traveled recently; for neisseriameningitis, he asked if the patient had been living in a crowded environment (Clancey, 1988). If probed, a physician might formulate the rule that one should first seek evidence of exposure to infectious diseases, but the abstraction to causal enabling conditions is a knowledge-engineer's theoretical statement. It is intriguing to consider how such routines might evolve by mimicry of other physicians and attempts to be more efficient when interviewing a patient, without conscious formulation of a generalized rule.

[7] "Not many people appreciated the importance of the videocassette recorder .... by using fast forward and fast reverse, the radar data could be used to show not only things were going, but also where they had come from. Computer support made the task easier by eliminating items that moved no more than once every two

hours—thus erasing the Russian radar lures—and there it was, a brand-new intelligence tool" (Clancy, Red Storm Rising, 1986, p. 392).

[8] Indeed, it was on a related point that Ryle introduced the idea of a category mistake: The total interactive system (in Ryle's example, the mind as university) is not to be found on the lower level of agent behavior and internal mechanisms: "But where is the University? I have seen where the members of colleges live, where the Registrar works, where the scientists experiment and the rest.'... The University is just the way in which all that he has already seen is organized" (Ryle, The Concept of Mind, 1949, p. 16). It is no coincidence that Ryle's examples (a University, division, team-spirit) are all descriptions of emergent social phenomena, a level above individual agents, as perceived by an observer. It is not just a matter of using words incorrectly, as Ryle was want to emphasize, but of not understanding the nature of situated systems, emergent effects, and frames of reference.

[9] Here a symbol is something that is semantically interpreted not just a tokin. I make this distinction here because I want to preserve the cognitive science insight that reasoning can be characterized in terms of symbol manipulation, while arguing that it happens in a cyclical process of perception and re-representation commentary.

[10] Here it should be apparent why I have emphasized that we need to view AI programming as a modeling methodology. We need to realize that every knowledge base contains models, specifically that classifications are models. Thus, the term "qualitative reasoning" covers what all AI programs do; for all contain qualitative models (primarily non-numeric relational networks representing causal, temporal, and spatial characteristics of processes in the world). The term "qualitative simulation" is to be contrasted with the kind of classification model in MYCIN. The most prevalent form of qualitative simulation is a behavioral description of internal states and processes, commonly called a "causal-associational network." See Clancey (1989) for elaboration.

[11] It is tempting to get caught in a conundrum: If "everything is relative" then this description itself can't be an objective, absolute description of the nature of reality. If the theory is right, then it must be wrong. Because it is relative to our purposes, not everyone will agree. But then, conversely, the theory must be right, for our explanation of its failure restates the theory itself (that what you perceive depends on your point of view). Trying to work your way out of this is tantamount to wanting a representation that has a defined meaning, that wears its interpretation on its sleeve. The theory says that it is open to interpretation; if that sounds like a fixed, objective statement, it is just because of the illusion that for the moment you think you know what "open" and "interpretation" mean.

[12] The same analysis applies to the controversy in student modeling research between misconception models or "bug libraries" and the assumed superior form of simulation or generative error models (by which errors are produced from a grammar during the course of problem solving, rather than being pre-enumerated) (VanLehn, 1988). Although a generative model is advantageous because it is more general and supplies an explanation of the cause of misconceptions, in fact bug libraries, like disease hierarchies, cannot be avoided. Misconceptions are KL descriptions that reflect developmental interactions in the student's and observer's experience; they cannot be replaced by more objective descriptions of the student or the environment viewed in isolation to each other. However, viewed as a mechanism in the brain, generative models point in the right direction because they don't treat knowledge as something that is stored, but as inherently manifested in problem-solving interaction. As always, treating knowledge, misconception or not, as a *thing* is where the problem begins.

[13] This is analogous to the discovery that data storage requirements are substantially reduced for video images if one stores changes between sequential images, rather than full pictures. By analogy, for repeating routines the brain may only need to have the capability to recognize and generate changes, not descriptions of moment-by-moment appearances. Such embedded (relative to the particulars of the current context, but not declaratively describing them), process-oriented computation could obviate the need for maps and plans, as Brooks' INSECTS suggest.

[14] An obvious connection can be made to Searle's Chinese Room (Searle, 1984). The entire question about whether rule-like manipulation of symbols inside the room constitutes intelligent behavior or not is misguided. There are no symbols being manipulated by hidden processes inside the brain, rule-like or not.

[15] Regarding the tautology-preserving mechanism inherent in realizing that something needs to be said, see Winograd and Flores on "breakdown" and Bartlett on impasses/reflection.

[16] An interesting question is whether the perception should be viewed as a "node"—a separate process that subsumes actions—or whether perception is the process by which an organization of actions is recreated. Under this latter interpretation, the perception is itself the organization of the neural processes. Perhaps experiencing an organizing perception, for example, articulating a rule of behavior, is what enables the organization. Or perhaps an articulation process (saying this rule) subsumes the new organization (way of seeing talks) and is simply activated (at a later time, as a "reminding") by upwards propagation prior to the application of the subsumed actions it is about.

[17] Reeke and Edelman put this well, "[T]he start of the chain of deductions … which for AI justify the notion of the brain as a computer, is the assumption that information exists in the world—that is just there to be manipulated. There is also the idea that the organism is the receiver rather than a creator of criteria leading to information. Once the prior existence of such external information is conceded, it is entirely natural to proceed without further ado to the business of programming the rules to deal with it" (Reeke & Edelman, 1988, p. 153).

[18] This provides an intriguing resolution of the "frame problem," the problem of how changes in the environment are to be noticed and stored without a time-consuming and useless combinatorial explosion of inferences. The frame problem is an artifact of viewing perception as input to cognition, suggesting that input is predigested and exists apart from the process of behaving, and that memory is a special storage for descriptions of the world which are matched against rules that describe behaviors. The frame problem is an artifact of the idea that there must be internal, unexpressed representations (maps) of the world that the organism must keep up-to-date. Indeed, the frame problem is one reason for arguing that the representational view of memory is hopelessly wrong.

[19] Contra Schank, Bartlett argued that a "reminding" occurs when a failure-impasse occurs, in the form of a conceptualization of the past, not a literal retrieval of what happened; the later memory of this failure is secondary. A failure needn't be an emotionally dramatic quandry, but perhaps just a momentary pause in the otherwise automatic flow of activity.

[20] To see this, watch someone's gestures and relate them to what the person is saying. Notice how often they precede the person's words. Notice how your description in terms of a visual language is radically different from the usual way in which you pay attention to gestures. Can you categorize the gesture-concepts in a given person's repertoire? Could he or she formulate these categories without looking in a mirror?