Report 80-01
Stanford -- KSL

The Representation Hypothesis.
Avron Barr,
Jan 1980

card 1 of 1

HPP-80-1 January 1980

The Representation Hypothesis

Avron Barr
Computer Science Department
Stanford University

Talk given at IJCAI-79, Tokyo, in lieu of the paper in the proceedings.

This talk is about "knowledge representation," particularly some of the fundamental assumptions involved in the way we handle knowledge in current AI and cognitive science research. The whole enterprise seems to have abandoned much of what we know about *knowing*, about the way humans know, in favor of a very simplified model of knowledge. Namely, we have agreed to assume that knowledge is something that can be *represented*--that knowing something means having a data structure stored away that stands for what is known. In other words, we treat knowledge as an object, a representable object.

The advantage of this way of looking at things is that there is a very simple relationship between the things we say a person or system knows, and the memories, knowledge, or data structures we say he or it has stored. This *representational correspondence* greatly facilitates our discussions of cognition: when we say the system "knows John's telephone number," we mean that the appropriate bit string is out there on the disk and the program knows how to find it. I think that this object-like conception of knowledge, which I call the *representation hypothesis*, is inadequate for the task of understanding cognition. It is a "simplifying" approximation, in good scientific tradition, but its simplicity is dangerously blinding.

Origins of the Representation Hypothesis

Our current attitude toward knowledge has evolved over the last twenty-five years. Perhaps it all started with early AI programs like the Logic Theorist[10] which manipulated logic formulae that had been transcribed into data structures within the machine. In logic, statements that people make are formalized syntactically so that logicians can examine reasoning in terms of syntactic manipulation of these formulae. But the formulae were not meant to stand for *knowledge*--they were meant to be descriptions of statements. In the earliest days of AI however, there was a change. The formulae became data structures in the computers *memory*, which since von Neuman's earliest work has been viewed as a separate box, distinct from the processor. Furthermore, since the machine's CPU certainly had no knowledge of its own, and since LT and its contemporaries were intelligent programs which evidently *knew* something about the world, the identification of the data structures in memory and *knowledge* was inevitable.

Let us not forget about *the interpreter*. Everybody in AI is trained to pay lip service to the interpreter, insisting that the interpreter must be "kept in mind" when we talk about what a system knows. Which is correct, but an inadequate defense against sloppy thinking. Few of us can resist identifying WHITE (SNOW) or HIT (MARY, JOHN) with the knowledge they "represent." We worry about things like the relative merits of various conflict resolution strategies in production systems, or about how to represent angels standing on the head of a pin in semantic nets. But

representational correspondence, that simplistic equivalence between what is *known* and the data structure that represents it, is always assumed. And in modern theoretical psychology the situation is more serious, because storage and retrieval are traditional ways of thinking about memory (since around 1800). Nowadays psychologists might talk about the transfer of hunks of semantic net from the short-term to the long-term store!

Problems with Representation

During the time when AI was being born, there were some people who talked about the limitations of the representational, knowledge-as-object paradigm. Perhaps the clearest voice was Heinz von Foerster who emphasized that "transfer of information" was not necessarily what was going on biologically--it was at best a convenient fiction.[12] There were other voices. In fact, some important early AI programs were non-representational in the sense that individual data structures did not "stand for" the things the system "knew." (I will talk a bit more about my favorite example, EPAM, later on.) Nonetheless, by the mid-sixties our thoughts about our programs and about cognition were entrenched in representational terms.

Dissatisfaction with the limitations of this view of what it means to know has surfaced on occasion in the AI literature, most notably in what was called the procedural/declarative controversy. These discussions about the relative merits of encoding knowledge implicitly as procedures or explicitly as declarative structures include the most thoughtful work on knowledge representation of the late sixties and early seventies. The entire discussion focuses on what kind of data structures are the best representation medium. The conclusion was that "both are nice," for different kinds of knowledge. This famous non-debate never resolved itself because it presumed the relevant issue--it focussed on how to represent knowledge, not on the more fundamental question of what aspects of knowing are the kinds of thing that can be represented? Throughout the procedural/declarative discussion, in fact through all AI and cognitive psychology research, the fundamental hypothesis of representation has remained deeply seated in the way that we think and talk about cognitive processes: Terms like "retrieval of information," "acquisition of knowledge," and "memory storage" bring to mind pictures of useful cognitive objects being sifted out of the environment, processed, stored, retrieved, and passed on.

Metacognition

I would like to describe some commonplace aspects of human cognition that are among those that seem most troublesome when we think of knowledge and memory in terms of storage of information.[1] These abilities, which you will recognize as familiar experiences, are collectively referred to as *metacognition.* I believe that they offer essential insight into the way that humans "know."

Consider, for example, the well-known experience called the tip-of-the-tongue phenomenon: You meet someone at a party; you can't remember his name, but you remember meeting him before; at you sister's house; his name rhymes with spaghetti.... In this situation you would say that you *know* his name, but that you can't remember it--you know that you know it, and you know some things about it. Similarly, you often know that you *don't* know something--like Paul Newman's telephone number. This knowledge about what you know, called metaknowledge, is not easily understood in terms of the representational paradigm. The fact that you know something you can't remember, or that you don't know something you don't even try to remember, are not themselves likely to be "stored in memory". (Nor are they really conclusions based on other stored information. There is experimental evidence that the *knowing not* phenomenon, as it is called, is sometimes more rapid than simple "retrievals.")[8]

Interestingly people's metacognitive abilities develops--for example, children are not as good at knowing and using the feeling that they know something.[7] The scope of metacognitive abilities includes knowing about what you know and don't know, knowing about your cognitive abilities (that you never forget faces, are good with maps, etc.), and knowing about the status of on-going cognitive tasks (that you understand directions that are being given, that you are ready to pass a test, and so on). Far from being anomalies in an otherwise well-understood field of mnemonic behaviors, metacognitive abilities come into play almost all the time that people think. Allan Collins argues that metacognitive abilities involved in reasoning from incomplete knowledge (e.g. knowing that if a certain fact were true you would have heard of it) are central to everyday reasoning.[4]

The Representational Approach to Metacognition

If one thinks of *knowing* in terms of data structures that *represent* pieces of knowledge, then metaknowledge, or knowledge about what one knows, is simply having data structures that represent knowledge about other data structures (or about what they represent). And this is exactly the approach that has been followed in AI research. For example, in order to endow an expert AI system with some metacognitive capabilities, like talking about what they know and explaining their reasoning, Randy Davis was led naturally to the invention of a variety of meta-level data structures in his TEIRESIAS system.[5]

In current knowledge representation languages like KRL, the idea of a meta-level description has been incorporated very tidily into the guts of the representation scheme[3]--allowing meta-level descriptions to be expressed in the same syntax and interpreted by the same interpreter as the other "knowledge." The idea of using meta-level data structures in AI representation languages is relatively new, but there are already several variations on the them (there is not a shortage of new representation languages to try things out in). All, however, are based on using some data structures to *describe* other data structures, in the same sense that the latter describe things in the "real world."

However, as I described earlier, the phenomenology of human meta-cognitive behavior just doesn't look like it stems from meta-level facts that are added to our cerebral database. Meta-knowledge is not just part of what we know, it is part of how we know. Your knowledge that you don't know Paul Newman's telephone number could be an additional fact that you know about your Paul Newman database, or it could be a conclusion after retrieval of an empty slot, but that's not what it *feels* like (and there is empirical evidence based on reaction-time studies that indicates that retrieval is not what's going on here). Rather, you reason that if you knew Paul Newman's telephone number you would *know* that you know it--it is notable.

The problem, as it appeared in the procedural/declarative controversy, is that the representational approach allows too great a dichotomy between "what is known" and the "way that we know." The way that things are learned is an integral part of what we can know. It is not just a matter of storing away information or facts or data structures. If we view knowledge not as a *substance* that we input, store, and eventually output, but rather as a quality that we *ascribe* to systems that behave in certain ways at certain times, then the question of where a certain piece of knowledge resides may not be the right kind of question. Rather we must ask how the system is *structured* to behave in such a clever (human) way. How does the system's organization allow change in the face of a changing environment? How is the *memory* of an appropriate part of the past made efficiently available when it is relevant?

The Non-representational Tradition

Most of you have thought about all of these fuzzy, general issues in the context of our representational paradigm, and have had trouble imagining how to pin down these questions without thinking about specific models like frame matching, hashing through wffs, and so on. But it can be done. As I mentioned earlier, some of the original work in AI was non-representational. I am thinking in particular of the discrimination net à la EPAM, a program that was designed to learn lists of nonsense syllables with the performance characteristics that human learners display.[6] Where is the *knowledge* that FEG follows POB in EPAM's discrimination net? How is this *information* stored and retrieved efficiently? These questions do not make much sense in the context of the discrimination net data structure: At best, you might say that a node in the net *knows* "if the syllable ends in *B* then proceed down the left branch, else right." Clearly we think of EPAM as *knowing* the pair POB-FEG, but the knowledge does not reside in a unique representational data structure within the discrimination net--you can't point to it!

Newell and Simon[11] have defined a paradigm for the study of computation and cognition based on the concept of *physical symbol systems*, which are composed of patterns, called symbols, within some physical medium and which can manipulate those patterns. Both EPAM and representational systems like LT fit within this paradigm. Their differentation into separate sub-paradigms involves the nature of what Newell and Simon called *designation*, the relation between a symbol and that which it symbolizes. In both types of systems, symbols can designate three types of things, other symbolic structures, processes, and *external objects* (in the sense that unique symbolic patterns within the system can be interpreted as external phenomenon). In representational systems, *pieces of knowledge* are included in the class of objects, since unique data structures are taken to *represent* facts or information. EPAM does not support this kind of correspondence between its internal data structures and *information*. (EPAM might be described as a *closed* cognitive system, à la Maturana.[9])

The assumption of representational correspondence makes some things simple. Pieces of knowledge can be pointed out and manipulated within the system in a very intuitive way. But cognition is not a simple problem, and the representation hypothesis precludes some very basic approaches to the structure of cognitive systems. If we are just engineers of intelligent systems, then perhaps we can ignore aspects of knowing that fall outside of the representational paradigm. But most of us are concerned with understanding cognition, either as a subgoal toward intelligent systems design or as an endgoal in itself. We must remember that although we can ascribe knowledge to our systems, we shouldn't necessarily equate it with the data structures in "memory."

References

[1] Barr, A. Meta-knowledge and cognition. Proceedings of IJCAI, 1979, 31-33.

[2] Barr, A., Bennett, J., and Clancey, W. Transfer of expertise--A paradigm for AI research. Working Paper HPP-79-11, Computer Science Dept., Stanford University, March, 1979.

[3] Bobrow, D., and Winograd, T. An overview of KRL, a knowledge representation language. *Cognitive Science*, 1:1 (1977), 1-46.

[4] Collins, A. Fragments of a theroy of human plausible reasoning. Proceedings of TINLAP-2, 1978, 194-201.

[5] Davis, R. Applications of meta-level knowlege to the construction, maintainance, and use of large knowledge bases. Memo AIM-283, Computer Science Dept., Stanford University, 1976.

[6] Feigenbaum, E. A. The simulation of verbal learning behavior. In Feigenbaum and Feldman (Eds.), *Computers and Thought*, New York: McGraw Hill, 1963.

[7] Flavell, J. Metacognition and cognitive monitoring: A new area for cognitive-developmental inquiry. Unpublished manuscript, Psychology Dept., Stanford University, 1979.

[8] Kohlers, P., and Palef, S. Knowing not. *Memory and Cognition*, 4:5 (1976), 553-558.

[9] Maturana, H. Biology of language: The epistemology of reality. In *Psychology and Biology of Language and Thought*, Ithaca: Cornell University, 1976.

[10] Newell, A., and Simon, H. *Human Problem Solving*. Englewood Cliffs: Prentice Hall, 1972.

[11] Newell, A., and Simon, H. Computer science as empirical inquiry: Symbols and Search. ACM Turing Award Lecture. *Communications of the ACM*, 19:3(1976), 113-126.

[12] von Foerster, H. On constructing a reality. In *Cybernitics of Cybernetics*. Biological Computing Laboratory, University of Illinois, 1970.