Meta-knowledge and Cognition

Avron Barr

Heuristic Programming Project
Stanford University

In AI knowledge representation schemes, structures that describe other structures are said to represent "meta-knowledge." Knowledge about other knowledge can be either about the form of the representation scheme itself (e.g., its syntax) or about the "facts" that are represented (their origin, reliability, importance, etc.). After reviewing the use of explicit meta-knowledge in several systems, some studies of human behavior that indicate people's ability to reason about what they know and about how they reason are described. The concept of meta-level knowledge captures intrinsic, commonplace properties of human cognition that are central to an understanding of memory and intelligence.

The use of meta-knowledge in AI systems like MYCIN, which have reached human-expert-level performance in complex domains, is a key breakthrough in the design of "knowledge-based" intelligent systems. Meta-level knowledge has been used in these systems primarily in the implementation of "introspective" processes: Acquisition of new knowledge and explanation of the system's reasoning to users. The usefulness of meta-level descriptions for these and other functions has prompted proposals for their incorporation in several new general-purpose representation schemes, like KRL, as described in the next section.

But there is more to meta-knowledge than its typical characterization, as knowledge about knowledge, captures. In fact, experiences involving meta-level knowledge and reasoning are an integral part of common, everyday cognitive activity. For example, consider the well-known "tip-of-the-tongue" phenomenon:
You run into someone you have met once before, and you can't remember his name. You remember very well your first meeting at a New Year's Eve party in Oakland, and that he is the brother-in-law of your wife's boss. Then you remember that he has a foreign-sounding name. It rhymes with spaghetti...

You could use all of this knowledge in trying to recollect his name. You would certainly say that you "know his name," even though you can't recall it: You "know that you know it." Conversely, another example of meta-knowledge in everyday experience involves knowing that you don't know something:

If you were asked "Do you know Paul Newman's phone number?", you would know right away that you didn't. You wouldn't check through your mental phone directory or your trivia file on Paul Newman. The reasoning is closer to: "If I knew his phone number, I would know that I knew it. It is notable!"

It is this intuitive knowledge about what we know, and also of how we use what we know, that is the most compelling reason for viewing meta-knowledge as having a central role in human cognition. After examining the use of explicit representations of meta-knowledge in several recent AI systems, we describe some psychological studies which indicate that meta-knowledge may play a much more central role in human memory, reasoning and understanding.
The Representation of Meta-knowledge

The AI systems reviewed here all allow the explicit declaration of meta-knowledge in their representation schemes. In other words, the representation formalisms allow encoding of data-structures that "describe" other data-structures (data or procedures). These meta-level descriptions are of two types:

*Syntactic* meta-knowledge explicates the structure, organization or function of the representation formalism itself. For instance, the *schema-schema* in TEIRESIAS is an explicit representation of the structure of all the schemata in the database.

*Semantic* meta-level structures allow statements about the extent, history (source, age, modifications), relevance, reliability, criteriality, importance, or default values of other statements in the database. For instance, the knowledge that having a trunk is more indicative of elephants than having a tail, might be encoded in a semantic net scheme by labeling the *has-part* link between *elephant* and *trunk* as "criterial."

Of course, these meta-level statements are themselves encoded in syntactic structures, to be interpreted by some process. The kind of interpretive process that is necessary to use the representational structures is, as always, important to keep in mind when the declarative structures are described.

The issues discussed here have come up in many, maybe all, AI systems and are relevant to all representation schemes: predicate calculus, production rules, conceptual dependency nets, semantic nets, procedures, frames, etc. The particular systems described here have attempted to use explicitly represented meta-knowledge.

TEIRESIAS

One area of AI where the use of meta-knowledge evolved naturally is in systems developed in what might be called the "Transfer of Expertise" paradigm. Systems like DENDRAL and MYCIN perform some complex task (chemical structure analysis, medical diagnosis) by using a database acquired from the human experts who are good at the task. The need to give these systems meta-knowledge, knowledge about their structure, developed naturally as part of the effort to endow them with some introspective capabilities, in particular, facilities for automating the acquisition of new knowledge from humans, for doing automatic bookkeeping on the database, and for explaining the system's decisions and reasoning strategies to humans.

A prototype system for incorporating such capabilities into AI programs was designed by Randy Davis for his doctoral dissertation at Stanford, leading directly to the development of techniques for the explicit representation of meta-knowledge (Davis, 1976). The main application of Davis's TEIRESIAS formalism has been to the MYCIN system (Shortliffe, 1976), a production-rule based program that assists in medical diagnosis.
The application of TEIRESIAS to MYCIN centered on increasing the effectiveness of the system's acquisition of knowledge from the experts and of its explanation of how it reached its conclusions to the physicians on the ward who were to use the system in diagnosing cases. These capabilities involved the construction of new data structures in TEIRESIAS that gave it certain kinds of knowledge about the existing MYCIN data structures—i.e., explicit meta-knowledge (Davis & Buchanan, 1977). Here is a summary of the kinds of meta-level structures that were added to manipulate the rules in the database:

**Schemata** are a hierarchically organized, frame-structured representation scheme for the knowledge about the objects that the system knows about, like organisms, lab tests, and medication. Used in acquisition, they maintain bookkeeping information that allows automatic updating of the system's tables when new data objects are added or old ones are modified. The schema-schema is the blank frame for unknown data objects, the progenitive model of all the schemata.

**Function templates** describe the system's predicate functions, including types of arguments, calling sequence, etc. These are (primitive) descriptions of the interpreting process itself, in an explicit, declarative representation structure.

**Rule models** are a classification scheme generated automatically by the system that is used in acquisition as a framework for the description of new rules. The rule models are built up as new rules are acquired by noticing points of similarity among the acquired rules. In particular, by "classifying" the left-hand sides of the rules in the database, the system forms subsets of rules that might be applicable in similar situations. These are used to:

1. Check the completeness and consistency of new rules during knowledge acquisition: "Most rules about bacteremia infection also mention the portal of entry of the organism."

2. Build "expectations" as an aid to top-down processing of natural language input during acquisition.

3. Discuss the system's decision methods: "How do you decide that an infection is pseudomonas?"

4. Pinpoint areas of which the system is ignorant.

**Meta-rules** encode strategies for deciding the "appropriateness" of the many "object-level" rules. By describing the content of the object-level rules (in terms of the rule models), meta-rules can embody heuristics about which lines of inference are most likely to be profitably followed.

TEIRESIAS's various meta-knowledge representation structures are all encoded and used differently within the system, each having its own set of interpreting procedures. The first two listed above encode primarily syntactic knowledge, while the latter two encode semantic knowledge. However, the important point is that all of these structures arose out of the effort to implement some new, introspective abilities in the system.
KRL and KRS

The best known of the new frame-oriented representation languages is KRL, under development at Xerox PARC. In addition to frame-like data structures, the language design stresses ideas like multiple viewpoints of data objects, reasoning by matching, and resource-limited reasoning processes. The explicit representation of meta-knowledge was already an espoused feature of the first implementation effort, KRL-0 (Bobrow & Winograd, 1977). Each slot of a unit, or frame, could be tagged with certain features, selected from a set of predefined meta-level characteristics: DEFAULT, CRITERIAL, and PRIMARY.

DEFAULT descriptors are assumed to be true for descendants unless specified otherwise.

CRITERIAL descriptors are necessary and sufficient in matching one unit to another.

PRIMARY descriptors are the ones from the set of alternative descriptions that should be used for inheritance in generating descendants.

The idea of making the inheritance properties of slot descriptors explicit in the representation, as in KRL's DEFAULT and PRIMARY tags, has appeared in several new frame-based languages, such as FRL (Goldstein & Roberts, 1977) and UNITS (Stefik, 1978). The semantics of inheritance in network-based representation schemes has been a central concern of recent work in representation theory (Woods, 1976; Brachman, 1977; Hendrix, 1976), and making inheritance constraints explicit in the declarative structure appears to be a common solution.

Besides the feature tags, another meta-knowledge construct was proposed in KRL-0, the use of an entire description to describe another description—i.e., as a meta-description. This mechanism was not implemented in KRL-0, or thought out much further, but it is the seed of the layers of KRS.

Recent work on KRL has been heavily influenced by some ideas about the formal semantics of representation languages proposed by Brian Smith at MIT (B. Smith, 1978). Smith's formalism, called KRS, includes meta-level description, called layers, as one of its basic characterizations of a representation. Layer-1 descriptions describe objects outside the process ("in the world"), Layer-2 descriptions describe layer-1 descriptions, and so on.

There are two important aspects of the representation of meta-knowledge in KRS and KRL-1 that should be pointed out. First, all descriptions at the various layers are encoded in one uniform syntax and are interpreted by the same process, unlike TEIRESIAS'S four different kinds of meta-level structures. Thus the KRS formalism allows the same expressive power for meta-knowledge as for all other knowledge. Second, Smith insists on including the interpreting process in the semantics of the representation:

When we build a representation of the world, we must also build a model of how we understand the world. The representation of knowing.
FOL

Its explicit representation of its own reasoning mechanisms makes the FOL proof-checking system (Weyhrauch, 1979) of interest in this discussion. FOL makes direct use of meta-knowledge in a first-order logic representation scheme. The first key idea in FOL is the concept of *simulation structure*, a "computable semantic model" that is "attached" to the language in which the system reasons (first-order logic): Constants in the language are attached to objects (data structures) in the simulation structure; function symbols (like "+ ") are attached to procedures which execute them (like LISP's PLUS); and predicate symbols are attached to procedures that determine their truth value (like LESS_THAN X Y). The simulation structure can then be used to "evaluate" ground terms in expressions that are being manipulated. FOL's evaluator interleaves *simplification by evaluation* (in the simulation structure) and *syntactic simplification* resulting in a very powerful and general first-order logic expression evaluator.

FOL is a conversational system that talks about "language/simulation-structure" pairs. From postulated statements in the language, L, FOL derives, or checks, new statements, F, using "rules of logic" (syntactic manipulation of expressions in the language) and semantic attachment (evaluation in the simulation structure, SS):

![Diagram](https://example.com/fol_diagram.png)

Language/Simulation-structure pairs in FOL

The second key idea is that, since FOL is a program composed itself of data structures, it is the natural simulation structure to "attach" to a theory of language/simulation-structure pairs, a theory of reasoning. In such a theory one could reason about (prove theorems about) any particular language/simulation-structure pair by using general theorems about them, meta-theorems.

FOL's theory of language/simulation-structure pairs is called META. Its language part is built up of constants for things like WFFs, TERMS, derivations, simulation structures, theorems, etc.; functions that act on these structures (i.e., substitute one term for another in a WFF); and predicates that decide questions like "Is this expression a WFF?" or "Is this a proof?" The semantic attachments of the functions and predicates in this axiomatization are the procedures in the FOL proof checker that do those things.

![Diagram](https://example.com/meta_diagram.png)

The Theory META
Although FOL is a very powerful proof-constructing/checking system in its own right, it is uniquely of interest in this discussion because of the neat way that meta-level reasoning fits into the formalism. Weyhrauch (1979) stresses his belief that much of human reasoning is carried on at the "meta-level," and he is interested in exploring this aspect of human reasoning in his system. One particularly engrossing idea is that, since META is a theory of language-simulation-structure pairs and also is one, it is a complete theory of itself--it can "self-reflect."

The State of the Art

Explicit declarations of the form of the system's representation schemes were necessary in TEIRESIAS to implement "introspective" capabilities like explanation. The form of the meta-level representations in TEIRESIAS was ad hoc, but their use was clear. On the other hand, the representation of meta-knowledge as feature tags in KRL-0 was more uniform, but the ideas about how to use this knowledge were incomplete, involving rather vague ideas of inheritance and matching. Both KRS and FOL have elegant ideas about how to represent meta-level knowledge in their representation schemes, but have not actually specified yet how this kind of knowledge is to be used.
Meta-knowledge in Human Cognition

The psychological studies reviewed here deal with three rather diverse areas of cognition: memory, plausible reasoning, and cognitive development. However, the concept of meta-knowledge is essential to understanding these results. It will be argued that much remembering and reasoning is best described as meta-level activity, that at the core of these mental processes people use knowledge about their own cognitive ability, style and experience, and about the extent, origin, and certainty of their knowledge.

The cognitive behaviors described here are not the results of trick questions or contrived experimental situations. The phenomena that are described are an intrinsic part of human cognition, from remembering to everyday inference making. Their interpretation will require an understanding of the role of computation in cognition in which meta-level structures play a central part.

Knowing Not

The experience of meta-knowledge by humans is addressed directly in a paper by Kolers and Palef (1976). They point out that the "knowing not" phenomenon is a very common characteristic of human cognition; this is simply where, as in the example about Paul Newman's telephone number, people often know rapidly and reliably that they do not know something. Furthermore, these researchers point out that this trait is not easily captured by current "searching" models of memory.

For instance, college students were asked whether they knew certain words well enough to be able to use them in a substantive sentence, and their "yes" or "no" response was timed. The results showed that items that were unfamiliar were identified faster than the more-or-less familiar items (although the common English words were recognized fastest of all). In these reaction time studies, the mean negative response was faster and more accurate than the shortest positive response, for some types of questions. Kolers and Palef (1976) conclude from these results that "we have evidence that some negative knowledge can be accessed more rapidly than some positive knowledge.... People can report that they do not know with greater speed and accuracy than are allowed for by scan and match models of memory."

The experimental demands of these studies involved an inference on the part of the subject: "Can you use this word in a substantive sentence?" Therefore the results must be viewed as not only involving memory, but also inference and reasoning. The studies by Allan Collins and his colleagues directly address the question of the role of meta-knowledge in reasoning.

Meta-knowledge and Inference

Collins, Warnock, Aiello and Miller (1975) discuss "reasoning from incomplete knowledge," that is, what one can conclude from the fact that one doesn't know something. Since the prerequisite to this kind of reasoning is awareness of not knowing some fact, these inferences relate directly to the "knowing not" studies, and to meta-knowledge. The most thoroughly discussed strategy for making inferences based on incomplete knowledge is the "lack-of-knowledge" inference, first discussed by Carbonell & Collins (1973).
The lack-of-knowledge inference is illustrated by asking the question "Is Moldavia one of the United States?" A person may not know what Moldavia is, and may not be able to list the 50 states off the top of his head, but he still might know that if Moldavia were in the United States, he would have heard about it! The lack-of-knowledge inference involves both the "experience of not remembering" the required fact and the inference that if it were true, he would remember it.

Gentner and Collins (1978) suggest two factors used in making the lack-of-knowledge inference. First, one estimates the (relative) "importance" of the fact; the more important it is, the more the fact that you don't remember it implies that it "ain't so." Second, one's own expertise in the topic area is estimated--the more one knows about the area, the more likely not remembering something implies it isn't true. In making a judgment about the truth of some statement, the lack-of-knowledge inference is combined with other inferences, such as induction about the a priori likelihood of the occurrence.

Gentner and Collins ran an experiment testing the influence of these factors on the lack-of-knowledge inference, comparing four conditions where the expertise of the knower and the importance of the knowledge were varied. Their results confirmed the hypothesis that "importance" and "expertise" judgments influence, and have the predicted effect on, lack-of-knowledge inferences. Collins (1978) stresses that much of human plausible reasoning is based on meta-level reasoning about what one knows, and what one would know if some fact were true. These judgments are quite subtle, and our ability to make them is, of course, learned.

The Development of Meta-cognition

John Flavell, a developmental psychologist at Stanford, has been exploring for some time the idea that children develop increasingly accurate "feelings" about their performance on cognitive tasks involving learning, remembering, and understanding (Flavell, 1977). For instance, in a study by Flavell, Friedichs, and Hoyt (1970), two groups of children, preschool and elementary-school age, were told to study a list of words until they could remember every word on the list. The elementary-school age children usually had perfect recall once they said they were ready, but the younger children said they were ready and usually were not. Flavell and Wellman (1977) call this ability "meta-memory," one's knowledge or "feeling" about mnemonic behavior.

In another study, Markman (1977) examines children's feelings about their performance in a cognitive task involving "understanding." Children in the first through third grades were presented with simple instructions for a game or magic trick, but some part of the information required was omitted from the instructions--they were ambiguous and incomplete. The third graders indicated, by asking appropriate questions right away, that they knew that they didn't understand the instructions; whereas the younger children gave very little indication of being aware that they didn't understand, even though the instructions were blatantly incomplete.

In a recent paper, Flavell (1979) discusses these results in terms of metacognitive knowledge and metacognitive experience. Metacognitive knowledge is knowledge about people as cognitive systems, about the cognitive tasks they face, and about the strategies they employ to accomplish them. Metacognitive experiences are the conscious realizations about
some aspect of a cognitive enterprise, particularly how well it's going. For example, a person may feel:

That he "doesn't understand" what someone has said, or that he understands someone's directions perfectly;

That he has "mastered" the material for an exam, or that he is going to flunk; or

That he has "a long way to go" on a problem that he is working on, or that he has just been given an "easy" problem.

The "feelings" involved in the meta-cognitive experience can be quite subtle but remarkably informative. Consider the following experiment by Wellman (1977). Children (grades K, 1, and 3) were individually presented with pictures of a variety of objects and asked to name them. For any object a child could not name, he was asked if he felt he knew its name anyway and would be able to recognize it from a set of alternative names. Wellman found that these "feeling-of-knowing judgments" were much more accurate in the older children: Children develop increasing ability to predict later recognition of the names of objects that they cannot recall.

Clearly, the kind of knowledge these children had about their memory of the items whose names they couldn't recall was not as simple as just knowing that they knew or didn't know them: They developed an ability to predict accurately their successful recognition in another memory task. The point of these studies for the current discussion is that they indicate that the knowledge we have about our knowledge and memory is not simply that we "know that we know X" or that we "believe that adults know what 2+2 is." The meta-level knowledge that appears to be useful to cognitive processes like learning, remembering and understanding, or what Flavell (1979) calls generally metacognition, covers the full range of descriptions of other knowledge.

The Phenomenology of Meta-knowledge

The psychological phenomena reviewed here illustrate the pervasive role of meta-level activity in human cognition. The studies by Kolers and Palef deal with some fundamental properties of human remembering, namely, that one often seems to "know" rapidly and reliably that one doesn't know something, without going through any sort of "memory-scanning" process. In other words, people seem to have "at their fingertips" an idea of the extent of their knowledge in some kinds of tasks. Collins's studies indicate that this kind of meta-knowledge may be an integral part of much of people's everyday reasoning. Flavell's work shows that this kind of meta-knowledge develops gradually in children and that certain meta-cognitive tasks, like estimating how hard a problem will be or knowing when one has understood directions, are performed surprisingly poorly by young children. Once again, what is developing here is not the child's knowledge of the world, but his knowledge about his own cognitive performance as well as about what he knows.
Meta-knowledge and Computation

The first observation that must be made is that there is a big difference between human capabilities, like "knowing not," "meta-memory," and the "lack-of-knowledge inference" that have been described here in terms of meta-knowledge, and the current uses of meta-level representational structures knowledge in AI systems. In particular, viewing meta-knowledge as additional "facts" or "rules" describing the object-level knowledge does not completely capture its essential characteristics. It must be viewed also in terms of the "process" of knowing.

Typical (pre-meta-knowledge) AI programs can achieve expert performance in their domain and yet be unable to answer questions like "Why did you do this?" or "How do you know that?", which a human expert would naturally be able to answer. It was an attempt to implement these very abilities in TEIRESIAS that led Randy Davis directly to the use of meta-knowledge. Humans acquire, as a natural, integral part of their development and training, knowledge about their own reasoning processes as well as knowledge about what they know. The psychological studies of meta-cognitive behavior are so important because they deal with commonplace human cognitive abilities, like "knowing not" and "meta-memory," that are very difficult to understand in terms of "storage and retrieval" models of memory. The capability of this kind of "knowledge" fundamentally changes the nature of the cognitive systems involved. One is led to agree with Restle (1974) that "...for the concept of 'memory' to be made the basis of a theory of cognition, it must be stretched entirely beyond its natural compass."

When the data structures in an AI program encode knowledge about themselves that is in turn is used in the system's reasoning, the program takes on new properties. In particular, when the meta-level structures are interpreted by the same process as other knowledge (called meta-embedding in Barr, 1977), the "knowledge" that the system "has" and its "reasoning processes" become indistinguishable, not only in form, but also in function. Everything that the system knows and learns is part of its reasoning machinery, and vice versa. The models of computation that are currently used by cognitive scientists, i.e., "information-processing models" based on our "von Neumann" machines with CPUs and "addressable memory stores," do not adequately capture such "self-reflective" processes.
References


