Metacognition in Computation: A selected research review

Michael T. Cox

BBN Technologies
10 Moulton St.
Cambridge, MA 02138
mcox@bbn.com

Abstract

Various disciplines have examined the many phenomena of metacognition and have produced numerous results, both positive and negative. I discuss some of these aspects of *cognition about cognition* and the results concerning them from the point of view of the psychologist and the computer scientist, and I attempt to place them in the context of computational theories. I examine metacognition with respect to both problem solving (e.g., planning) and to comprehension (e.g., story understanding) processes of cognition.

Introduction

Any intelligent agent with a choice between what to do in the world actually has three very different choices. First it must decide which of several actions to perform is best in its current situation. For example it might decide that spending its lunch money on a new watch offered by a stranger in the parking lot is a good bargain. But secondly it also must decide whether it has thought enough about the decision that the choice is sufficiently informed to warrant commitment to action or whether more thought is in order. That is it must think about its own thinking. Furthermore given that it chooses to save the money for a rainy day and exercise at lunch to drop a few pounds instead, the agent has a third kind of choice when it considers the reasons that led to poor judgement and being mugged by the watch thief. That is, it must decide what went wrong and why in its reasoning process through a bit of introspection and self-criticism. This paper examines the research involved with the latter two types of reasoning. We discuss not only the literature in the computer sciences, but we also review a select portion of the metacognition literature in psychology with the goal of better understanding the computational issues.

In its most basic form, the algorithm-selection problem (Rice, 1976) in computer science represents a classical metacognition task. In such situations the input to a program is a particular problem and a set of algorithms that can compute solutions to that class of problems. The task is to choose the algorithm that will run in the least amount of time. Decisions can be based on not just characteristics of the input problem, but a good choice involves knowledge about algorithm performance. Lagoudakis, Littman, and Parr (2001; Lagoudakis, Parr, and Littman, 2002) illustrate this task with simple sorting problems. Three choices are quick sort, insertion sort, and merge sort. They show that the decision can be formulated as a Markov decision process (MDP) where the state is the size of the input problem, the actions that cause state transitions are the algorithms, and the objective function is estimated by an empirically gathered profile of times it took to perform the sort on past problems. Note that because two of the three algorithms are recursive, the algorithm selection task is repeatedly performed over problems of many sizes during a typical sort. Now using this statistical model of reasoning (where reasoning is sorting in this case), a system can rationally choose the best reasoning process to maximize its utility (here, run-time). The combined solution outperforms any of the individual algorithms.

The distinctions in the metacognition literature are often very subtle, however, and the line between reasoning and metareasoning is sometimes less clear than with sorting. Consider a story understanding task. Figure 1 shows that a story is composed of characters and events that change these characters and the world in which the story takes place. To understand the story requires that some intelligent system reason about the events and states and why characters choose to perform the particular actions the events represent. Such NLP systems will have a representation of the story and will perform certain computations that alter the representations until a satisfactory interpretation of the story is achieved. Like the states and events in the story's domain, this NLP domain (shown in the central part of Figure 1) has mental states (e.g., story interpretations) and mental events (e.g., schema retrieval). Now if these mental states and events are themselves represented in a third domain, they too can be reasoned about as was the story itself. The resulting introspection is therefore reasoning about the NLP reasoning task and hence is a second-order reasoning process called metareasoning or more generally metacognition.

The metacognitive task may be to explain errors in the cognitive task or it may be to select between cognitive "algorithms" to perform the reasoning. In either case, confusion arises when the various levels, processing or representations are conceptually intermixed or left implicit. One of the goals of this article is to examine some of the various research programs related to metacognition in computation and separate these various aspects for the reader.

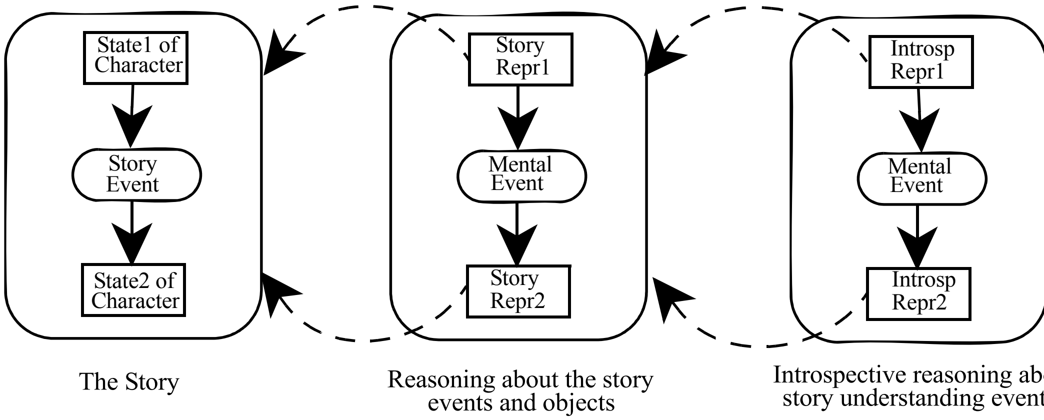


Figure 1. Metacognition entails two levels of reasoning and representation but three sets of states and events (dashed arrows indicate what is being represented).

The 21st century is experiencing an interest in computational models of higher order reasoning analogous to the kinds of metacognitive activity exhibited by humans. In addition to the recent 2005 AAAI Spring Symposium on Metacognition in Computation (Anderson & Oates, 2005), the AI community has conducted several similar workshops including the AISB 2000 symposium on How to Design a Functioning Mind, April, 2000 (Davis in press); the St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence, April, 2002 (Minsky, Singh, and Sloman 2004); the DARPA Workshop on Self-Aware Computer Systems, April, 2004 (McCarthy and Chaudri 2004); the NDIST Workshop on Self-Reconfiguring Software Systems, December, 2004; and the LEMORE05 Workshop: Learner Modelling for Reflection to Support Learner Control, Metacognition and Improved Communication between Teachers and Learners held at the 12th International Conference on Artificial Intelligence in Education in Amsterdam. The excitement associated with these developments can especially be seen in Brachman (2002). However, many of the foundations for this work were formulated at the beginning of artificial intelligence and in some cases earlier.

Metacognition research encompasses studies regarding reasoning about one's own thinking, memory and the executive processes that presumably control strategy selection and processing allocation. Metacognition differs from standard cognition in that the self is the referent of the processing or the knowledge (Wellman, 1983). In most interpretations (e.g., Hayes-Roth, Waterman, and Lenat 1983; Kuokka 1990), meta-X can be translated to "X about X:" Thus metaknowledge is knowledge about knowledge, and metacognition is cognition about cognition. But often metaknowledge and metamemory (memory about one's own memory) are included in the study of metacognition, because they are important in self-monitoring and other metacognitive processes. Thus in much of the literature, the term metacognition is broadly construed to apply to all self-reflective facets of cognition.

Artificial intelligence certainly does not have a monopoly of interest concerning metacognition. Philosophers and observers of the human condition have been fascinated by the subject for a very long time. Around the turn of the 16th century in *De Trinitate*, Augustine (1600/1955) asks "What then can be the purport of the injunction, know thyself? I suppose it is that the mind should reflect upon itself."[1] Mathematicians and philosophers have realized since at least the time of Socrates the problems associated with self-referential sentences such as the liar's paradox represented by the statement "This sentence is false." (Epstein and Carnielli 1989; see Perlis (in press) for a treatment of some of these metalanguage problems).

More recently, Hofstadter (1979/1989) convincingly argues that the concept of reflection, or an object turning in upon itself (i.e., his concept of "Strange Loops"), is a common and powerful theme, in and outside of science. Strange Loops can be found in mathematics with the proofs of Gödel, in art with the painting of Escher, and in music with the compositions of Bach. But with few exceptions (e.g., Lyons 1986, Pollock 1989a), AI and cognitive psychology present the most thorough mechanistic explanations for such phenomena. Many of the roots of metacognition in computation are influenced by the large body of work in cognitive, developmental, and social psychology, cognitive aging research, and the educational and learning sciences. This paper examines a selection of these research areas as well as those in computer science.[2] Initially I limit this history to the 20th century, starting first with the formative metacognition research in the human psychology literature and then with related research in computer science. Research in the 21st century is summarized toward the end of this paper.

1.Cited in Lyons (1986, p. 1).
2.I deliberately exclude cognitive neuroscience research from this review. I also do not address the considerable body of research on consciousness. But see the selected bibliography on consciousness in philosophy, cognitive science and neuroscience (Metzinger and Chalmers 1995) and also Chalmers' online bibliography at `consc.net/biblio.html`

Psychology, Metacognition, and Human Behavior

The psychological literature on metacognition and metamemory[3] provides a wide array of influences that bear on metacognition in computation. Here I examine specific studies that emphasize cognitive self-monitoring, the importance of explicit representation, higher-order problem-solving, the function of understanding one's own memory system, and data demonstrating a person's ability to assess (or not) the veracity of their own responses and learning. I end this section on a note of caution with some caveats.

Cognition and Metacognition

Since Flavell's (1971) coining of the term metamemory, and especially since the seminal metacognition research of Flavell and Wellman (1977), many have investigated the phenomenon surrounding cognition about cognition.[4] Of all research on the modern-day concept of metacognition, the child development literature (i.e., how cognitive function develops during childhood) has perhaps the longest history (see, for example, Yussen 1985). Moreover, developmental psychology has reported the most positive evidence for the importance of metacognitive strategies and monitoring (see Schneider 1985; Wellman 1983). Researchers interested in learning disabilities have studied the metacognitive components of such pathologies. For example, *Part II: Macrolevel Cognitive Aspects of Learning Disabilities* (Ceci 1987) contains a number of papers relevant to this class of investigations. Research examining the relationship between metacognitive skills and educational instruction have made significant progress. For example, Forrest-Pressley, MacKinnon, and Waller (1985) and Garner (1987) report successful instruction procedures related to both problem solving and reading comprehension (see also Ram and Leake 1995, for a related discussion from computer/cognitive science). Most of these works concentrate on applications relevant to teaching in general school environments, although some address specific instruction of the learning disabled. Finally, the social psychology and philosophical communities have all taken considerable interest in individuals' beliefs about their own beliefs and beliefs about others' beliefs (e.g., Antaki and Lewis 1986; Metcalfe 1998b; Pollock 1989a, 1989b).[5]

Wellman (1983; 1985; 1992) views human metacognition, not as a unitary phenomenon, but rather as a multifaceted theory of mind. Metacognition involves several separate but related cognitive processes and knowledge structures that share as a common theme the self as referent. Such a theory of mind emerges during childhood from of an awareness of the differences between internal and external worlds, that is, from the perception that there exist both mental states and events that are quite discriminable from external states and events. This theory encompasses a number of knowledge classes considered by Wellman to be psychological variables: *person variables* that deal with the individual and others (for example, cognitive psychologists can recall many facts about cognition, whereas most people cannot), *task variables*, which concern the type of mental activity (for example, it is more difficult to remember nonsense words than familiar words), and *strategy variables* that relate to alternative approaches to a mental task (e.g., to remember a list it helps to rehearse). Finally, Wellman's theory includes a self-monitoring component, whereby people evaluate their levels of comprehension and mental performance with respect to the theory and the norms the theory predicts.

Nelson and Narens (1990/1992) present a general information-processing framework for integrating and better understanding metacognition and metamemory. This framework is illustrated in Figure 2. Behind it lie three basic principles: 1. Cognitive processes are split into an object-level (cognition) and a meta-level (metacognition); 2. The meta-level contains a dynamic model of the object-level; and 3. A flow of information from the object-level to the meta-level is considered monitoring, whereas information flowing from the meta-level to the object-level is considered control. Monitoring informs the meta-level about the state of the object-level and thus allows the meta-level's model of the object level to be updated. Then depending upon the state of this model, control can initiate, maintain, or terminate object-level behavior. Object-level behavior consists of cognitive activities such as problem solving or memory retrieval.

Nelson and Narens address knowledge acquisition (encoding), retention, and retrieval in both monitoring and control directions of information flow during memory tasks. Monitoring processes include ease-of-learning judgements, judgements of learning (JOLs), feelings of knowing (FOKs) and confidence in retrieved answers. Control processes include selection of the kind of processes, allocation of study time, termination of study, selection of memory search strategy, and termination of search. Both acquisition and retrieval of memory items have computationally explicit decompositions in their paper. Although the framework is directed at memory related performance rather than inference-based problem-solving, the distinctions between monitoring and control and the information processing perspective is highly compatible with the views presented in the computational sciences. Their framework has been widely used in psychology to integrate disparate research and we will summarize some of that here. We will also use it to frame some of the research topics in computer science and AI.

3.I also will not discuss the extensive literature on metamemory here. For a general review see Dunlosky (2004) or Metcalfe (2000).
4.Brown (1987) notes that the relationship between text comprehension and metacognitive activities has been studied since the turn of the century, but under the guise of other technical terms.
5.Pollock in particular (1989b) distinguishes between knowledge about the facts that one knows and knowledge about one's motivations, beliefs and processes.

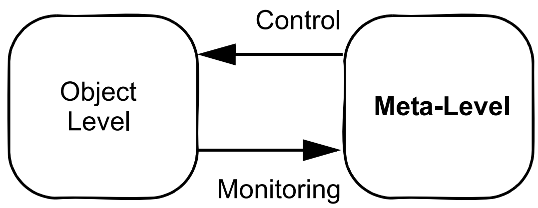


Figure 2. Metacognitive monitoring and control of cognition

Problem Solving and Metacognition

Problem solving is one area where a natural fit exists to computational theories in AI. Concepts such as executive control and monitoring are important to problem solving in order to manage problem complexity and to evaluate progress towards goals. Here much leverage for metacognitive knowledge could be gained by humans. But although Flavell (1976) represents the first reference with metacognition and problem solving in the title, relatively few psychological studies have examined this phenomena explicitly since then. Some are described here.

Dörner (1979) reports the earliest experiment on the effects of cognitive monitoring on human problem solving. The experimental design categorizes subjects into one of two conditions according to how they perform protocols after problem solving. In the introspective condition, subjects reflect out loud about their own reasoning during problem solving (at the meta-level), whereas subjects in the statistical-control group discuss their solution to the problem in terms of the hypotheses they developed (at the object level). The experiment itself involves a complicated machine with three lights. Each light can be turned on in four different colors. There are eight push-buttons on the machine with which subjects control the lights and their colorations. The subjects solve ten problems during the experimental trials. Problems consist of an initial state in which the lights of the machine begin operation and a goal state consisting of a different light configuration. Dörner reports that the experimental group performs significantly better than the control group after the third trial. Moreover, Dörner claims that introspective subjects exhibited improved performance during transfer tasks of subsequent experiments, although the details of many of the experiments are lacking and no replication of these results appear in the literature.

Derry (1989) offers a comprehensive model of reflective problem solving for mathematical word problems inspired by John Anderson's ACT* (Anderson 1983) and PUPS (Anderson and Thompson 1989) theories of general cognition. Based on such a theory, Derry and her colleagues developed a computer-based instructional system to teach word problems to military servicemen. Prior to the development of this application, Derry performed the following experiment on groups of college students and military personnel. Given an assumption that general problem solving behaviors, such as reasoning from the goal backwards to the solution and means ends analysis, form the bases for human problem solving, the experimenter gathered subject protocols during solution of mathematical word problems. The protocols were classified into 27 categories falling into four basic phases of problem solving: clarifying a problem, developing a strategy, executing a strategy, and monitoring/checking performance. The surprising result was that neither group performed problem solving in a linear fashion, and that most protocols were classified into clarifying and execution phases. The strategy-development and monitoring/checking phases lacked significant protocols.

Delclos and Harrington (1991) report that both subject conditions with general problem-solving skill training and those with problem-solving coupled with metacognitive skill training demonstrate equal performance on a problem solving task. With greater task complexity, though, subjects with the problem-solving/metacognitive training perform better than either a control group or the problem solving training alone group. Also, Swanson (1990) claims to have established the independence of general problem aptitude from metacognitive ability. Subjects with relatively low aptitude, but high metacognitive ability, often use metacognitive skills to compensate for low ability so that their performance is equivalent to high aptitude subjects.

Finally, Davidson, Deuser, and Sternberg (1994) present results from a series of studies that show the use of metacognitive abilities correlate with standard measures of intelligence. In their experiments on insight problem-solving they report that, although higher IQ subjects are slower rather than faster on analyzing the problems and applying their insights (not surprising if more processing is being performed), their performance is higher. They argue that the difference in performance is due to effective use of metacognitive processes of problem identification, representation, planning how to proceed, and solution evaluation, rather than problem solving abilities *per se*.

Dominowski (1998) reviews many such studies (particularly those that require talking aloud protocols) and concludes that although some conflicting evidence exists, subjects in metacognitive conditions generally do better on problem-solving tasks. The reason for the difference is not just that subjects are verbalizing their thoughts. Silent thinking and simple thinking out loud perform equally well. The difference is that problem-focussed attention of subjects improve local problem-solving behavior, whereas metacognitive attention allow subjects to be flexible globally and thus have a greater chance of finding a more complex and effective problem-solving strategy.

Berardi-Coletta, Buyer, Dominowski, & Rellinger (1995) illustrate this difference in a task where subjects must deal out a deck of cards such that alternating cards are placed either on a table face up or on the bottom of the deck. Thus to deal out the cards 1, 2, 3, and 4, the deck must be arranged as 1, 3, 2, and 4. Berardi-Coletta et al. identified five possible subject strategies in this task that range from simple guessing or swapping incorrectly dealt cards, to more complex approaches such as differentially representing the difference between "up" and "bottom" cards. Subjects in the metacognitive verbalization condition answer out loud questions such as “How are you deciding what went wrong?” and “How are you deciding on a way to work out the order for the cards?” Subjects in a problem-focussed group answer question such as “What is the goal of the problem?” and “What cards do you have in order so far?” They discovered that subjects in the metacognitive group never guess, and, although some may use swapping at first, they abandon it to pursue the more complex reasoning approaches.

This section has illustrated some of the findings that describe how humans introspect about their cognitive performance (processes) when solving problems and how this ability can lead to improved performance. Although the findings are mixed, and no researcher claims that humans are inwardly omniscient, the results support the relevance of metacognitive theories for modeling intelligence and high-level reasoning. The careful monitoring of cognitive activities allows humans to control not only search for a problem solution but search for an effective problem-solving strategy.

Computational Models

Finally a number of psychologists have also built computational models that represent various aspects of human performance related to metacognition. Lynn Reder and her colleagues have an interesting model of metacognitive awareness of one’s own knowledge implemented in a computational model called SAC (Sources of Activation Confusion) (Reder and Schunn 1996). As a spreading activation model of declarative memory, it accounts for fast FOK judgements by activation of a problem node at the intersection of two or more semantic nodes triggered by terms in a given question. It successfully predicts whether or not subjects will use a memory retrieval or compute from scratch strategy to answer the question based on such judgements. Although a highly contentious proposition in the cognitive psychology community, the model also supports the notion that much of metacognition is an *implicit* process not subject to verbal reports.

Chi (1995; Chi, Bassok, Lewis, Reimann, and Glasser 1989) reports that improved learning is correlated with human subjects who generate their own questions during reasoning and *explicitly* explain the answers themselves (see also Pressley and Forrest-Pressley 1985). This is the so called *self-explanation effect*. This strong and positive effect has been modeled computationally by VanLehn and colleagues (VanLehn, Jones and Chi 1992; VanLehn, Ball and Kowalski, 1990). Note that this effect refers to explanations of self-generated questions about problems and not necessarily explanations about the self.

In relation to Chi and VanLehn’s research, Recker and Pirolli (1995) have shown that a Soar-based model of learning called SURF can explain individual differences exhibited by human subjects while learning to program in LISP using instructional text. The difference that accounted for much of the variability was self-explanation strategies. Those students who explained problems to themselves during comprehension of the instructions performed well on a subsequent performance task consisting of LISP programming exercises. The students who did not exhibit this behavior were not as likely to excel in the LISP task. The SURF model predicted such differences. The model took into account only domain-related elaborations; however, subjects exhibited other self-explanations that the model did not cover. In particular, some subjects seemed to exploit metacognitive feedback, like comprehension monitoring, in order to judge when to learn (Pirolli and Recker 1994). If self-reflection on the states of a subject’s comprehension of the instruction indicated an understanding failure, then this was sometimes used as a basis to form a goal to learn.

Caveats and the Relation of Psychological Research to Computational Research

Research concerning introspection has long been controversial (e.g., see Boring 1953; Nisbett and Wilson 1977 for objections to such research). Around the turn of the 19th century, trained introspection was assumed to be the proprietary scientific tool of the psychologist when “objectively” studying the mind.[6] The behaviorists tried to erase all scientific association with introspection by claiming not only that learning should be examined without the use of such introspective methods (e.g., Watson, 1919), but moreover that learning should be explained without reference to any intervening mental variables whatsoever (e.g., Skinner, 1950, 1956). Under the banner of metacognition research, however, interest returned to the study of introspection, second-order knowledge, and their roles in cognitive activities.

Yet, to believe that metacognition is a kind of psychological or computational panacea is a deceptive assumption. Wilson and Schooler (1991) have empirically shown that conditions exist under which introspection actually degrades specific performance (e.g., preference judgements). In the context of story understanding, Glenberg, Wilkinson, and Epstein (1982/1992) reported that human self-monitoring of text comprehension is often illusory and overestimated, especially under the conditions of long expository text. In general, people are overly-confident in cognitive tasks such as question answering (Fischhoff, Slovic, and Lichtenstein 1977). Furthermore recent studies specifically about metacognition have emphasized the fragility of people’s knowledge concerning themselves and their own reasoning processes.

Metcalfe (1998a) surveys a variety of cognitive tasks in which humans over-estimate their actual performance and exhibit a wide range of false expectations. For example they will think that they can solve particular problems when they cannot; they become very confident that they are about to generate a correct answer when they are actually on the

6.Titchener and others took great pains to develop a rigorous method of introspection and attempted to equate it with objective inspection (observation) as practiced in physics. For example, Titchener (1912) claims that “Experimental introspection, we have said, is a procedure that can be formulated; the introspecting psychologist can tell what he does and how he does it.” (p. 500). This remarkable statement is at the same time naïve and arrogant, given the hindsight of history.

verge of failing; they think they have answers on the tip of their tongue when an answer actually does not exist; and most amazingly they insist that they did give correct answers when provided evidence to the contrary. Such data make suspect earlier more simple interpretations of metacognition such as Dörner's.

Likewise, computational introspection is not effective under many circumstances given the overhead associated with it, and, given the demonstrated limitations of human introspection, computational theories should try not to overstate its scope. One must be cautious, however, when dismissing metacognition simply because of computational overhead costs. Doyle (1980, p. 30) warns that to disregard the introspective component and self-knowledge in order to save the computational overhead in space, time, and notation is discarding the very information necessary to avoid combinatorial explosions in search.

Research regarding metacognition processes in humans is relevant to metacognition in computation in at least two ways. First, and foremost, is the emphasis on cognitive self-monitoring for control. This behavior is the (limited) human ability to read one's own mental states during cognitive processing and use the information to influence further cognition. Thus, there exists some insight into the content of one's mind resulting in an internal feedback for the cognition being performed and a judgement of progress (or lack thereof). Garner (1987) has argued that metacognition and comprehension monitoring are important factors in the understanding of written text. Reading comprehension is therefore considered to be chiefly an interaction between a reader's expectations and the textual information.[7] Psychological studies have also confirmed a positive correlation between metamemory and memory performance in cognitive monitoring situations (Schneider 1985; Wellman 1983). This evidence, along with results from the studies above linking problem-solving performance with metacognitive abilities, directly supports the conviction that there must be a second-order introspective process that reflects to some degree on the performance element in an intelligent system, especially a learning system involved in understanding tasks such as story understanding.

Second, much of AI theory (especially GOFAI, or "good old fashioned AI," a term coined by Haugeland, 1985) places a heavy emphasis on explicit representation. Trains of thought, as well as the products of thought, are represented as metaknowledge structures, and computation is not simply the calculated results from implicit side-effects of processing. This emphasis is echoed in Chi's (1987) argument, that to understand knowledge organization and to examine research issues there must be some representational framework. Although diverging from the framework suggested by Chi, the following section describes specific research in the computer sciences that represent knowledge about knowledge and knowledge about process. It also surveys many other important theories and implementations that bear on the phenomena discussed in the current section.

Artificial Intelligence, Metareasoning, and Introspection

The AI community has long considered the possibility of providing machines with metacognitive faculties. In the 1980s and 1990s, researchers organized a number of conferences and symposia to explore some of the issues that relate to this concern: the Workshop on Meta-level Architectures and Reflection held in Alghero, Italy, during October, 1986 (Maes and Nardi, 1988); the International Workshop on Machine Learning, Meta-Reasoning and Logics held in Sesimbra, Portugal during February, 1988 (Brazdil and Konolige 1990); the IMSA-92 Workshop on Reflection and Metalevel Architectures held in Tokyo, Japan, during November, 1992; the AAAI Spring Symposium on Representing Mental States held at Stanford University during March, 1993 (Horty and Shoham 1993); the AAAI Spring Symposium on Representing Mental States and Mechanisms held at Stanford during March, 1995 (Cox and Freed 1995); and the Second International Conference on Meta-level Architectures and Reflection held in Saint-Malo, France during July, 1999 (Cointe 1999). In general, the loci of related research efforts has tended to focus the logic community on belief representation and introspective reasoning about such beliefs; the expert system community on metaknowledge and the control of rules; the decision-making and planning communities on search control and the choice of reasoning actions; and the model-based and case-based reasoning community on reasoning about reasoning failure, representations of process, and learning. This section presents a brief sketch of these trends.

From the very early days of AI, researchers have been concerned with the issues of machine self-knowledge and introspective capabilities. Two pioneering researchers, Marvin Minsky and John McCarthy, considered these issues and put them to paper in the mid-to-late 1950's. Although first exchanged among colleagues, and then printed at conferences at the turn of the decade in preliminary form,[8] reprints of these papers were refined and gathered together in the seminal collection of early AI articles entitled *Semantic Information Processing* (Minsky 1968b). Minsky's (1968a) contention was that for a machine to adequately answer questions about the world, including questions about itself in the world, it would have to have a executable model of itself. McCarthy (1968) asserted that for a machine to adequately behave intelligently it must

7.A special relation exists between metacognition, question asking and text understanding (see Gavelek and Raphael, 1985; Pressley and Forrest-Pressley, 1985). In effect, human learners use question-asking and question-answering strategies to provide an index into their feeling of comprehension of a given piece of text. This metacognitive feedback helps readers find areas where their understanding of the story is deficient, and thus where greater processing is necessary. As a final tangent, not only is metacognition important in language understanding, it is also important in language generation (i.e., in metalinguistic development; see Gombert 1992).

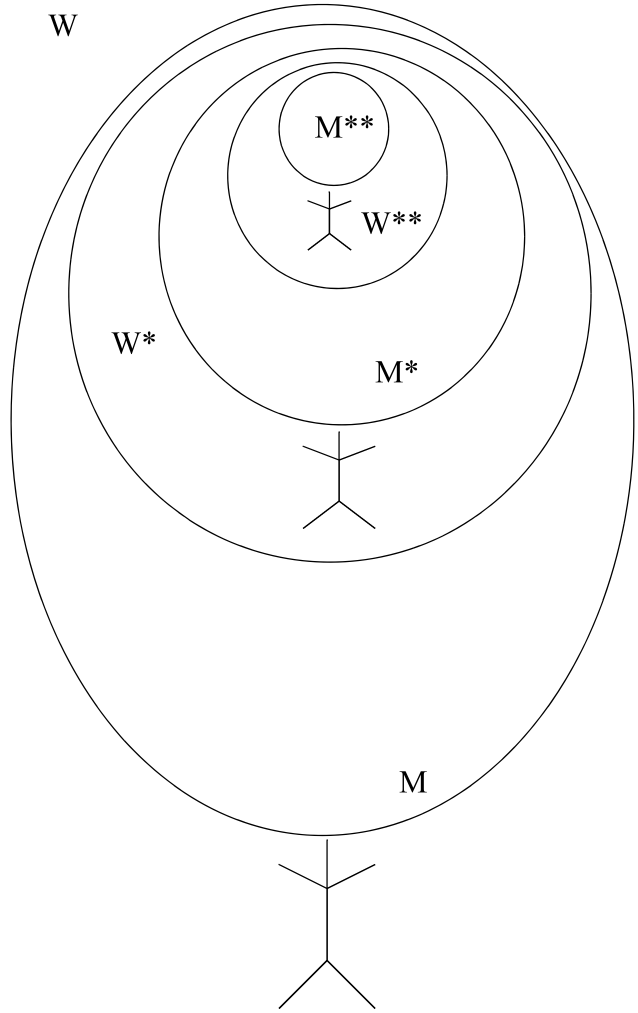
8.Minsky notes that he had been considering the ideas in this paper since 1954. It first appeared as Minsky (1965), although the concluding two pages of Minsky (1961/1963) address exactly the same issue. A significant portion of McCarthy's ideas was first published as McCarthy (1959).

declaratively represent its knowledge. These two positions have had far-reaching impact.

Roughly Minsky's proposal was procedural in nature while McCarthy's was declarative. Minsky believed that an intelligent machine must have a computational model of the outside world from which a simulated execution could answer questions about actions in the world without actually performing any action. He argued that if a machine uses models to answer questions about events in the world and the machine itself is in the world, then it must also use a recursive self-model or simulation to answer questions about itself, its own dispositions, and its own behavior in the world. This was a very early prototype of a mental model that became a precursor to similar research in both problem solving and understanding (e.g., Bhatta 1995; Bhatta and Goel 1992; Johnson-Laird 1983;[9] de Kleer and Brown 1983/1988; McNamara, Miller and Bransford 1991). In the spirit of Minsky's original theme, some very novel work has also been performed to enable a machine to procedurally simulate itself (e.g., Stein and Barnden 1995).

As a four and one half page discussion of the mind-body problem and the idea that human understanding is essentially the process of executing some model of the world, Minsky's paper is most interesting because it includes the modeling of not only the world, but the self (the modeler) as well (see Figure 3). Thus, there is W, the world, and M, the modeler who exists in the world. The model of the world is referred to as W*. W* is used to understand and answer questions about the world. So to answer questions about oneself in the world, it must also be the case that there exists within the model of the world, W*, a model of the modeler, termed M*. One should conceive of W* simply as the agent's knowledge of the world, and likewise, M* as the agent's reflective knowledge of itself in the world.

Furthermore, as Minsky notes, one must have a model of one's model of the world, or W**, in order to reason about and answer questions concerning its own world knowledge. Although Minsky does not label it as such, the kind of knowledge embodied in this model is typically referred to as metaknowledge. Finally, M** represents the agent's knowledge of its self-knowledge and its own behavior, including its own thinking. Within M** one might include most metacognitive knowledge of person variables (at least concerning the self). It would have a semantic component like "I am good at general memory tasks," as well as episodic components such as knowledge gained through monitoring (e.g, "I just solved a problem by remembering a similar past solution."). Again, although Minsky does not refer to it as such, M** represents introspective knowledge.



W= World
W*= World knowledge
W**= Meta-Knowledge
M= Modeler
M*= Self (Reflective) Knowledge
M**= Introspective Knowledge

Figure 3. A taxonomy of knowledge

Minsky elaborates on his ideas at the end of his book *Society of Mind* (Minsky 1986).

In the following subsection, I explore McCarthy's proposals and their local impact on the logic community and their more global effect on the tone of research into a computational explanation of metacognition. The second subsection then looks at additional varieties of research in the expert-system and decision-making communities. Finally, the last subsection relates some of the relevant research from the case-based reasoning and model-based reasoning communities to the research presented here.

Logic and Belief Introspection

A logical belief system can answer queries about the world given axiomatic facts (a knowledge base) and a logical inference mechanism. Furthermore a logical agent can

9.Johnson-Laird (1988, p. 361) explicitly takes issue with the suggestion that Minsky's concept of a self-model was in such a form that it could correspond to a human's capacity for self-reflection. He claims that Minsky's formulation is equivalent to a Turing machine with an interpreter that consults a complete description of itself (presumably without being able to understand itself), whereas humans consult an imperfect and incomplete mental model that is somehow qualitatively different. However, this argument appears to be extremely weak because the two positions are so similar and closely related.

determine what action to take in a given situation by *proving* that the action achieves some goal; that is the action necessarily follows from what it knows. Model-theoretic reasoning maintains the set of possible worlds consistent with the knowledge base. Logical resolution makes this kind of reasoning practical (e.g., using PROLOG).

As mentioned above, McCarthy (1968) not only established a manifesto for AI (i.e., knowledge representation is foundational, especially in declarative axiomatic form), but suggests that machines can examine their own beliefs when such beliefs are explicitly represented.[10] This suggestion is developed in McCarthy and Hayes (1969) and made explicit in both Hayes (1979/1981) and McCarthy (1979). A system requires such a metacognitive capability if it is to reason fully about the correctness of its knowledge. This is especially useful because beliefs are subject to retraction in the face of new information (i.e., knowledge is nonmonotonic). But beyond any technical details, McCarthy also wonders what it means for a machine to have a mental life. McCarthy (1979) enumerates six reasons why attributing mental qualities to programs and machines is a useful exercise. Among them, he claims (as does Dennett's 1978 essay on the *intentional stance)* that humans can more quickly and more easily understand a program, its behavior, and its intended function by ascribing beliefs and goals to the machine than by analyzing and explaining it in the language of program code and computer states. But most interestingly, McCarthy takes the business of understanding and simulating a machine's mental life beyond a mere practical metaphor. He questions what it means for a machine to have consciousness and to introspect about its mental world. Furthermore, he realizes that "introspection is essential for human level intelligence and not a mere epiphenomenon." (McCarthy 1995, p. 89) Thus, he is keenly interested in the relation between machine and human metacognition.

McCarthy (1979) defines introspection as a machine having a belief about its own mental states rather than about propositions concerning the world. This position has focussed much of the logic community, especially researchers such as Konolige (1985; 1988) and Moore (1985; 1995), on reasoning about knowledge, belief, and internal states, rather than reasoning about process and computation (but exceptions exist such as Genesereth's (1983) MRS system that reasons about the correctness of logical proofs).

Konolige (1986) represents a belief system with a deductive model rather than a possible worlds model. A *deduction structure* is a mathematical abstraction of many types of belief systems, especially expert systems (see the next section). The structure contains a knowledge base of facts and a finite set of inference rules. Although the model assumes that all possible deductions are made by a belief system, it does not assume that all possible logical consequences of the particular facts will be made, because the inference rules the system actually has may be incomplete due to the domain abstraction chosen by the designer. Regardless if a bounded belief system or machine, M, uses an introspective machine, IM, to answer queries concerning itself, the belief system is defined to be an introspective belief system. Furthermore Konolige defines self-beliefs answered by M as extrinsic; intrinsic self-beliefs are answered solely by IM. Although some self-questions such as "Is my brother's name John?" can be answered extrinsically, only by introspective deduction through the system IM can it answer questions such as "Can M deduce some consequent given a particular deduction structure?" Moreover by separating the two levels, some problems of the liar's paradox and self-reference are eliminated (Attardi and Simi 1991). Unfortunately the drawback is that non-paradoxical self-referential and mutually referential sentences cannot be represented (see Perlis 1985; 1988).

McCarthy (1993) further formalizes the idea of introspection by introducing context as a first-class object about which a system can reason. By encapsulating mental situations in formalized contexts, the reasoner can view the mental state as providing an outer context. Reasoning about one's own thoughts then involves transcending the outer context (McCarthy 1993). However, the realization of such an introspective mechanism has not been implemented. Furthermore, McCarthy (1995) notes that even though reason maintenance systems (e.g., Doyle 1979) record justifications for their beliefs and can retract beliefs in response to new information, they do not have the capability of inspecting the justification structures or making specific assertions about them, nor do they have the power to derive explanations from such structures.[11]

Knowledge-Based Systems, Metareasoning, and Control

The expert system community has also invested much effort into the formalization of metareasoning and metaknowledge. It was recognized in the late 1970's that differences exist between domain knowledge in the form of expert rules, and declarative control knowledge in the form of meta-rules (Davis 1976, 1979, 1980; see also Clancey and Bock 1985). Metarules encode knowledge about how rules should be executed, whereas ordinary rules encode domain-specific knowledge. Barr (1977 1979) noted, as I do here, the parallel relation between higher-order knowledge and reasoning by knowledge-based systems and human metacognition (see also Lenat, Davis, Doyle, Genesereth, Goldstein and Schrobe 1983). Especially when trying to automate the transfer of domain knowledge from human expert to machine expert, these and other researchers have attempted to give programs abstract knowledge of human reasoning and inference procedures, so that programs can understand human experts (see for example Clancey 1987). Additionally, when expert systems explain

10.The paper repeatedly illustrates Advice Taker examples with propositions that use the indexical "I."

11.McCarthy (1979; 1995) also outlines a number of additional issues concerning the mental domain that have received lesser attention by the logic community. He raises the issue of consciousness, language, intentions, free will, understanding and creativity, all of which have come to represent provocative focal aspects of intelligent reasoning. But of course see Minsky (1968a; 1985) for further analyses of free will.

a conclusion by providing to the user a list of rules through which the system chained to generate the conclusion, the system is said to introspect about its own reasoning. This view appears, however, to be an over-simplified example of both metacognition and explanation.

Davis and Buchanan (1977) claim that four types of meta-level knowledge exist: knowledge about object representations (encoded in schemata), knowledge about function representation (encoded in function templates), knowledge about inference rules (encoded in rule models), and knowledge about reasoning strategies (encoded in metarules). But much of this information is less akin to metacognitive knowledge than it is to ordinary abstract knowledge. For example, the schematic object knowledge above is equivalent to class definitions in an object-oriented language such as Java. Furthermore to claim that default inheritance and learning are inherently introspective processes (Maes 1987b) or that extrapolating from past experience is reflective thinking (Smith 1982/1985) is perhaps stretching the definitions of introspection and reflection respectively.

As another example, Batali (1983; also Maes 1988) considers the meta-level to be that which decides about the base-level (or actions in the world) and explicitly includes *planning* as a meta-level reasoning process. This unfortunately conflates metareasoning with reasoning (c.f., the confusion between metacognition and cognition[12]), because the system is not reasoning about the reasoning process itself. A procedural difference exists between reasoning about a solution or a problem and the metareasoning directed at the reasoning that produces such solutions or engages such problems. For instance, Carbonell (1986) notes that in order to transfer knowledge from programming a quicksort problem on a computer in Pascal to solving the same problem in LISP, a student cannot analogically map the Pascal solution to LISP code. The languages are too dissimilar in data structures and process control. Instead the reasoner must reason about how the original solution was derived and what decisions were made while solving the first problem, analogically mapping the derivation to LISP. Reasoning is at the algorithm level, rather than the code level.

Another popular research issue has been to develop systems that can reason about LISP functions and the actual code that represents a program's control (Batali 1983; Davis and Buchanan 1977; Maes 1987a, 1988; Smith 1982/1985). However, this form of metacognition is at a low-level as compared to other methods covered here. Programs need to reason about the functioning at the level of cognitive or logical processes, as well as at the level of program execution.[13] Nonetheless, this research has motivated an important DARPA thrust (Laddaga, 1998) into self-adaptive software systems that adjust their configurations in response to experience.

Some in the AI community have come to recognize some of the more subtle differences between the different families of metareasoning. For example, Clancey (1992) notes that many of the metarules employed by systems such as TEIRESIAS (Davis 1979), although dealing with control, are nonetheless domain specific. He claims that strategic knowledge is inherently procedural whereas domain specific knowledge is rule-based. Moreover, unlike his previous work (e.g., Clancey 1987), he currently eschews modeling the mental process that the expert uses when reasoning about the domain, and instead he emphasizes modeling the domain that the expert knows. This change of focus to cognitive engineering, however, seems to be as much a concession to the difficulty of representing metacognitive knowledge as it is a necessity dictated by representation itself.

Although many in the artificial intelligence community have recognized the necessity of reasoning about one's own beliefs, few have both modeled and represented the processes that *generates* beliefs, and made them available to the reasoner itself. In this category of metacognitive system, a categorical distinction exists between those systems that reason forward to decide what action to perform or what computation to execute, and those that reason backward to explain a failure or to learn. This is related to the distinction made in the psychological literature between forward strategic control and backward metacognitive monitoring (see again Figure 2). In AI researchers use the terms metareasoning (or meta-level control) and introspection respectively.

In the former category, systems attempt to choose a reasoning action based on some knowledge of the mental actions at the disposal of the system. Doyle (1980), as well as Russell and Wefald (1991a, 1991b; Tash and Russell 1994), use probabilistic estimations and decision theory to select a computation that has the most expected utility. Etzioni (1991) uses decision-analytic methods to weigh the trade-off between deliberation cost, execution cost and goal value when choosing a goal toward which to direct attention and when deciding which action to take in service of a chosen goal.[14] The latter category of systems represents feedback from the reasoning process. This feedback can further inform the forward metareasoning, or it can be used in learning in causal abductive tasks such as explanation and interpretive understanding. The subsequent subsection looks at the metareasoning issues of decision making under limited conditions. It examines both control and monitoring sides. Discussion of introspective explanation and learning waits until the section after next.

12.For example, Derry (1989) claims that metacognitive components are associated with, not only knowledge of the problem-solving process, but with the ability of a subject to orchestrate and monitor these same processes (see the second subsection of section 2). Yet the paper often combines discussion of domain-independent problem solving processes with that of the orchestration and monitoring processes. Problem solving itself is often discussed in terms of strategy, thus further blurring the delineation between cognition and metacognition.

13.In the terms of Newell (1982), the reasoning should be at the symbol level as well as at the register-transfer level of intelligent systems.

14.The consensus is that Good's (1971) research on Type II rationality (i.e.,taking into consideration of the expected utility of action that includes the cost of deliberation itself) provided the foundation from which all such research began.

Limited Rationality

One of the core problems of AI (and indeed of human intelligence) is that of deciding what to do in any given situation. In all but the most trivial of conditions, many actions exist from which to choose, and the outcomes of such actions are often unclear or involve considerable uncertainty. Decision theory states that the most rational behavior is the action that maximizes the expected utility under all possible conditions (von Neumann and Morgenstern, 1944). The expected utility is defined as

$$E[U|Pr, X] = \sum_{x \in X} Pr(x)U(x)$$

where X is the set of possible outcomes, $Pr(x)$ is the probability of a particular outcome $x \in X$ and $U: X \rightarrow \Re$ is a real-valued utility function over outcomes. The best action, a^*, then is the one across whose possible resultant states sums to the highest expected utility.

$$a^* = argmax_{a \in A} E[U|Pr, results(a)]$$

Here A is the set of possible actions and *results(a)* is the distribution of states that could result from performing a particular action, a. Seen another way and given that an agent can be considered a function that maps to actions observations of the environment (including outcomes of its actions), a rational agent is represented by the optimal function, f^*, such that

$$f^* = argmax_f V(f, \mathbf{E}, U)$$

where V returns the global value of the expected utility in environment **E**. The problem with this solution is that, even if an agent could calculate the values of all possible states reachable with all available actions, the world will change while the calculation is being made. That is rational choice is resource-bounded by time, and the search space is so large that perfect rationality is impossible. Thus as mentioned at the very beginning of this paper, the agent must reason about both the benefits and costs of the actions and the associated benefits and costs of the reasoning about the actions. As such metacognition includes both control and monitoring components parallel to that in Figure 2.

Russell (1997) has outlined a comprehensive theoretical approach to this trade off that turns the imprecise question of preferred behavior of an abstract agent into the design of the optimal program on a specific machine. Traditional AI has operationalized the task of producing good behavior by substituting for *perfect rationality* the idea of computing a choice with an agent program or *calculative rationality*. But for computationally complex problems (e.g., chess), the fact that a program will eventually reach the best decision because it has encoded sufficient knowledge to ascertain the solution (e.g., knows the rules of chess and has the goal of achieving checkmate) does little to guarantee that an actual solution will be computed in feasible time frames. Instead *rational metareasoning* seeks to include into the calculation the cost of the time it takes to select an action. *Bounded optimality* seeks further to analyze reasoning and metareasoning using the tools of complexity theory.

Russell and Wefald's (1991) research seeks to find an equilibrium between reasoning and action using metacognitive control of computations. In AI this meta-level control problem is equivalent to the control information flow that Figure 2 shows from the perspective of psychology. Such tradeoffs between thinking and doing arise in anytime systems (Dean and Boddy, 1988; Horvitz, 1987).[15] An anytime system has the property that a current best choice always exists as an approximation to the perfect choice. The decision then is whether to execute the chosen action at time t or to perform additional reasoning with the hope of possessing a better choice at t + i, where i is a time increment often equal to 1. According to Russell and Wefald, the construction of a system to make this kind of decision is based upon two principles. First computations are to be treated as actions (i.e., mental actions) and thus selected as to their expected utility in the joint physical/mental space of outcomes. Second this utility is based upon the cost associated with doing nothing due to intervening changes in the world and upon the possible computational improvement due to the choice of a better mental action.

The previous approach assumes that checking to ascertain the results of a computation is negligible. However such monitoring of computations may itself result in time and cost, so a more complete agent must reason about the monitoring of anytime calculations (Zilberstein, 1993). Consider that if the rate of decision improvement of reasoning is rather constant, then a contract can be made to specify the duration of running an anytime algorithm to achieve the maximum overall expected utility. However if a large amount of variability exists with the performance of the reasoning, the results must be periodically checked to ascertain the current progress and to determine whether or not to halt reasoning. Otherwise reasoning is wasted. Monitoring thus can serve two purposes. It can provide feedback as to the progress of the current reasoning and it can also be used to compile online (or offline) a profile of algorithm performance used to judge future reasoning.

Compiling performance profiles are especially important for complex algorithms that themselves may be composed of more primitive anytime algorithms (Zilberstein and Russell, 1996). The question then arises as to the allocation of computation resources to the individual pieces of the reasoning task. Algorithms may be arranged as a competing concurrent ensemble or in serial cascade such that the output of one provides the input to another. For example in a serial case, the meta-level control problem is how long to allow a vision computation before stopping it to run a path planning algorithm when the system must improve the overall robot trajectory. The vision component develops a terrain map that the planner uses. Whereas the planner initially creates an abstract general route and incrementally refines various path segments. Conditionalized performance profiles represent compiled introspective meta-knowledge (M** in Minsky's terms) used to estimate the

15.Note that this research focussed on one-step look-ahead local search rather than general anytime planning. We disregard the difference for the purpose of this discussion.

distribution of future run-times based upon input quality and past run-time.

Hansen and Zilberstein (1996) took this approach further by modeling the set of termination choices of the anytime process as a sequential Markov decision problem. The discrete states are levels of quality at a particular time, the action transitions between states are stop and continue, and the cost function is the expected value of the computation. The system then can use dynamic programming to determine an optimal stopping policy, $\pi^*(q,t)$. The important difference between this work and the previous is that the meta-level control information (i.e., the policy) is a statistical model of the reasoning process and its transitions rather than a statistical summary of the process behavior (i.e., the conditional performance profile).

Note that Hansen and Zilberstein's model is similar to the MDP developed for the algorithm selection task I used to motivate metareasoning in the beginning of this paper. This leads to the idea that reasoning is performed in a joint space of internal and external states and actions. The object level controls actions to be taken in the world and the meta-level controls the reasoning method to be taken in the mental world. Moreover just as an object level policy can be learned using reinforcement learning, so too can a meta-level policy be learned in the mental space. The advantage of using reinforcement learning is that it avoids myopic measurements that estimate the value of the computation solely based on local information. Harada and Russell (1999) has made some progress using this idea, and the concept has been implemented in the object domain of Tetris. Their approach uses semi-Markov decision processes (SMDPs) instead of MDPs, because SMDPs model the variable length of time between decisions.

One of the original theoretical goals of Russell and Wefald was to change the focus of finding the optimal agent, f^*, to the more concrete objective of designing the optimal agent program, l^*. written in a language, L, that runs on machine, M. with respect to the available computational resources in the environment E.

$$l^* = argmax_{l \in L_M} V(Agent(l, M), \mathbf{E}, U)$$

Russell and Subramanian (1995) have proved that this is possible for some small search tasks (e.g., automated mail sorting) and have argued for a more relaxed asymptotic version whose criterion for optimality depends upon a constant improvement in processor speed. This argument is not unlike the definition of optimality in complexity analysis.

Many other researchers have worked on problems of bounded rationality of course including Simon (1955, 1982) and Doyle (1980). See Horvitz, Cooper, and Heckerman (1989) for a emphasis on control of the decision making of bounded optimal agents and reasoning about the value of computations similar to that of Russell and Wefald. Note also that many researchers such as Fink (1998; 1999) use statistical methods to choose problem-solving strategies without ever framing the problem in terms of metacognitive processes.

Model-Based Reasoning, Case-Based Reasoning and Introspective Learning

Clearly people can and often do reason about their own reasoning and memory. Hayes (1979/1981) recounts a discussion he once had with a Texan about the number of scan lines in television screens in England. He thought it was one number whereas the Texan thought that it was another. At first Hayes was not sure about his answer. However if the number had changed at some point from his to the Texan's, it would have been an event that he would surely remember, but he did not. Thus after this realization in the dialogue, his confidence in the answer solidified. Hayes concludes that, more than simply not recalling the event, he had to realize that there was the lack of recall and actually use this fact as an argument in his reasoning.

The model-based reasoning and case-based reasoning communities have not missed such insights either. Like Minsky's insistence on a self-model and McCarthy's insistence on declarative knowledge, Collins, Birnbaum, Krulwich and Freed (1993) argue that to plan effectively a system must have an explicit model of its of planning and execution processes.[16] Given an explicit model of the causal and teleological properties of a standard physical device such as an electronic circuit (de Kleer 1984), a system can reason about future artifact design of similar electronics or can diagnose faults in specific circuits of that device class. Likewise researchers such as Stroulia (1994; Stroulia and Goel 1995) and Murdock (1998) treat the system itself as a device from whose model the system can generate a redesign or perform self-diagnosis.

Functional models are a particularly valuable form of knowledge for metacognitive reasoning. Whereas knowledge about the composition and behavior of reasoning strategies is important, such knowledge is more useful in supporting reflection and learning, if it is augmented by information about the functions of those strategies. Functional descriptions are particularly useful in metacognitive reasoning for three reasons: (a) functional descriptions can act as indices for retrieving relevant strategies to accomplish new requirements, (b) functional descriptions of required and retrieved strategies can be compared to compute differences to motivate adaptation, and (c) functional descriptions of the parts of a retrieved strategy can guide adaptation of the strategy to eliminate these differences (Murdock, personal communication).

At the heart of case-based reasoning (CBR) and case-based explanation (Kolodner 1993; Leake 1996a; Schank, Kass, and Riesbeck 1994) is the learning and use of episodic past experience in the form of a cases in a case memory. Given a new problem, a CBR system retrieves an older solution to a similar problem and then adapts it to fit the current problem-solving context. CBR systems have also been used to interpret actions and understand events in such

16.This contention concerning planning is also shared by Fox and Leake (1995a; Leake, 1996b) with respect to case-based planning and, moreover, was independently stated by Kuokka (1990) outside of the case-based reasoning community.

comprehension tasks as story understanding (natural language processing). Old explanation schemata or cases can be retrieved from memory and used to understand interesting or otherwise unusual events in the input. Finally learning has traditionally been central to CBR. It involves not only acquiring new case experience from success, but has focussed on repairing cases that fail and then learning to anticipate and avoid future performance failures by explaining what went wrong with executed actions in the world (e.g., Hammond 1990).

The theory presented in Cox (1996b; Cox and Ram 1999) is a computational model of introspection and failure-driven learning anchored firmly in the CBR tradition. In large part, the work represents a machine learning theory in the area of multistrategy systems that investigates the role of the planning metaphor as a vehicle for integrating multiple learning algorithms (Cox and Ram 1995; Ram and Cox 1994). To another extent, the research is a cognitive science treatise on a theory of introspective learning that specifies a mechanistic account of reasoning about reasoning failure. The central idea is to represent explicitly the reasoning of an intelligent system in specific knowledge structures[17] or cases called meta-explanation patterns (Meta-XPs) that explain *how* and *why* reasoning fails (Cox 1995; 1997a; Cox and Ram 1992). When failure occurs, the learner can then examine the trace structures (TMXPs; i.e., the how part), retrieve an introspective failure pattern (IMXP; i.e., the why part) from case memory, and unify the two to determine the proper learning methods. The overarching goal of the theory is to understand systems that turn inwards upon themselves in order to learn from their own mistakes.

The implementation of the theory is a case-based reasoning system called Meta-AQUA whose base performance task is story understanding (AQUA, Ram 1993; 1994). The idea is to have the system keep a trace of its explanation process, and when it generates an unsuccessful explanation of some event in the story, it needs to explain the explanation failure (hence meta-explanation). As Figure 1 shows, the AQUA component represents the story as a connect graph of action sequences that change the state of the environment in the story. When an unusual event occurs, AQUA will attempt to explain why the characters decided to perform the event. In a like manner, Meta-AQUA represents the actions and events in AQUA. Consider the following story (quasi-random generated).

```
Lynn was bored and asked Dad, Do you want
to play ball? Then Dad went to the garage
and picked up a baseball, and they went
outside to play with it. He hit the ball
hard so that it would reach her in left
field. Lynn threw the ball back. They
continued like this all afternoon.
Because of the game, Lynn was no longer
bored.
```

In the story Meta-AQUA finds it unusual for a person to strike a ball because its concept of “hit” constrains the object attribute to animate objects. It tries to explain the action by hypothesizing that Dad tried to hurt the ball (an abstract explanation pattern, or XP, retrieved from memory instantiates this explanation). However, the story specifies an alternate explanation (i.e., the hit action is intended to move the ball to the opposing person). This input causes an expectation failure (contradiction) because the system had expected one explanation to be true, but another proved true instead.

When Meta-AQUA detects an explanation failure, the performance module passes a trace of the reasoning (a TMXP) to the learning subsystem. The learner is composed of a CBR module for self-diagnosis and learning-goal generation and a non-linear planner for learning-strategy selection. At this time, the learner needs to explain why the failure occurred by applying an introspective explanation to the trace. An IMXP is retrieved using the failure symptom as a probe into memory. Meta-AQUA instantiates the retrieved meta-explanation and binds it to the trace of reasoning that preceded the failure. The resulting structure is then checked for applicability. If the explanation pattern does not apply correctly, then another probe is attempted. An accepted IMXP either provides a set of learning goals that are designed to modify the system’s memory or generates additional questions to be posed about the failure. Once a set of learning goals is posted, the goals are passed to the nonlinear planner for building a learning plan.

Table 1 lists the major state transitions that the learning processes produce. The learning plan is fully ordered to avoid interactions. For example, the abstraction step must precede the other steps. A knowledge dependency exists between the changes on the hit concept as a result of the abstraction and the use of this concept by both generalization and the indexing.[18] After the learning is executed and control returns to sentence processing, subsequent sen-

17. To support effective explanation of reasoning failure, and therefore to support learning, it is necessary to represent explicitly the thought processes and the conclusions that constitute the reasoning being explained. A large number of terms exist in the English language that concern mental activity. The earliest research to represent such content is Schank, Goldman, Rieger and Riesbeck (1972) who attempted to specify the primitive representations for all verbs of thought in support of natural language understanding. They wished to represent what people say about the mental world, rather than represent all facets of a complex memory and reasoning model. Schank’s conceptual dependency theory distinguishes between two sets of representations: primitive mental ACTs and mental CONCEPTUALIZATIONs upon which the ACTs operate. In addition, the theory proposes a number of causal links that connect members of one set with members of the other. They used only two mental ACTS, MTRANS (mental transfer of information from one location to another) and MBUILD (mental building of conceptualizations), and a few support structures such as MLOC (mental locations, e.g., working memory, central processor and long-term memory) to create a mental vocabulary. Schank’s theory has been corroborated by parts of the psychological literature, such as Schwanenflugel, Fabricius, Noyes, Bigler and Alexander’s (1994) analysis of folk theories of knowing. Subject responses during a similarity judgement task decomposed into memory, inference, and I/O clusters through factor analysis.

18.During mutual re-indexing, the explanations are differentiated based on the object attribute-value of the hit. However, the abstraction repair changes this attribute. The generalization method applied to the new explanation also uses this attribute. See Cox (1996b) for a more complete analysis.

tences concerning the hit predicate causes no anomaly. Instead, Meta-AQUA predicts the proper explanation.

Table 1: Learning from explanation failure

Symptoms	Contradiction between input and memory. Contradiction between expected explanation and actual explanation.
Faults	Incorrect domain knowledge Novel situation Erroneous association
Learning Goals	Reconcile input with conceptual definition Differentiate two explanations
Learning Plan	Abstraction on concept of hit Generalization on hit explanation Index new explanation Mutually re-index two explanations
Plan Execution Results	Object of hit constrained to physical obj, not animate obj New case of movement explanation acquired and indexed Index of hurt-explan = animate obj; of move-explan = inanimate obj.

Several fundamental problems are addressed to create such learning plans or strategies. These problems are (1) determining the cause of a reasoning failure (introspective blame assignment, Ram and Cox 1994), (2) deciding what to learn (learning goal formulation, Cox 1997b; Cox and Ram 1995), and (3) selecting and ordering the best learning methods to pursue its learning goals (learning strategy construction, Cox and Ram 1991). The system can reason about both errors of inference as well as memory retrieval (e.g., forgetting, Cox 1994a; 1995). A large empirical evaluation of Meta-AQUA demonstrated the positive value of introspective reasoning for effective learning using a corpus of six runs that includes 166 stories and comprises a total of 4,884 sentences (Cox 1996a; Cox and Ram 1999).

In general, the orientation is similar to many approaches based on reasoning traces (e.g., Carbonell 1986; Minton 1988; Sussman 1975) or justification structures (e.g., Birnbaum, Collins, Freed, and Krulwich 1990; de Kleer, Doyle, Steele, and Sussman 1977; Doyle, 1979) to represent problem-solving performance and to other approaches that use characterizations of reasoning failures for blame assignment and multistrategy learning (e.g., Kass 1990; Mooney and Ourston 1994; Owens 1990; Stroulia and Goel 1995). Reasoning trace information has primarily been used for blame assignment during planning (e.g., Collins *et al*. 1993; Birnbaum *et al*. 1990; Veloso and Carbonell 1994) and for speedup learning (e.g., Mitchell, Keller, and Kedar-Cabelli 1986). In addition to Meta-AQUA, many other systems have used an analysis of reasoning failures to determine what needs to be learned. Some examples include Mooney and Ourston's (1994) EITHER system, the CASTLE system of Krulwich (1993; Collins *et al*. 1993), Fox's (1995; Fox and Leake 1995a, 1995b) ROBBIE path planning system, and Stroulia's (1994) Autognostic system.

The IULIAN system of Oehlmann, Edwards and Sleeman (1994; 1995) maintains metacognitive knowledge in declarative introspection plans. Freed's RAPTER system (Cox and Freed 1994; Freed and Collins 1994) uses three types of self-knowledge when learning. Records of variable bindings maintain an implicit trace of system performance, justification structures provide the knowledge of the kinds of cognitive states and events needed to explain the system's behavior, and transformation rules (Collins 1987; Hammond 1989) describe how the mostly implementation-independent knowledge in justification structures corresponds to a particular agent's implementation. In the Meta-AQUA system, however, TMXPs maintain reasoning traces explicitly, and most implementation-dependent knowledge is avoided.

Birnbaum *et al.* (1990) focuses on the process of blame assignment by backing up through justification structures but do not emphasize the declarative representation of failure types. They explicitly model, however, the planner. They also explicitly model and reason about the intentions of a planner in order to find and repair the faults that underlie a planning failure (see Freed, Krulwich, Birnbaum, and Collins 1992). Though much is shared between CASTLE and Meta-AQUA in terms of blame assignment (and to a great extent CASTLE is also concerned with deciding what to learn; see Krulwich 1991), CASTLE does not use failure characterizations to formulate explicit learning goals nor does it construct a learning strategy in a deliberate manner within a multistrategy framework. The only other system to introspectively deliberate about the choice of a learning method is the ISM system of Cheng (1995). ISM optimizes learning behavior dynamically and under reasoning failure or success, but the system chooses the best *single* learning algorithm, rather than composing a strategy from multiple algorithms. ISM does not therefore have to consider algorithm interactions. Regardless of the differences, all of the systems, representations, methods and theories described in this section have more in common than not with respect to metacognitive reasoning analyses.

Trends in Current Research

Perhaps one of the most influential research trends in artificial intelligence is that of control of anytime systems through metareasoning. Given that intelligent agents are necessarily resource bounded and that nontrivial problems tend to be computationally intractable, an agent must reason about the state of its reasoning process to make significant progress. However Conitzer and Sandholm (2003) recently proved that certain forms of the metareasoning problem are NP-hard whereas others are NP-complete. Recent research has made progress with respect to the problem nonetheless. A special issue of *Artificial Intelligence* (Horvitz and Zilberstein 2001) highlights this progress.[19]

One of the difficulties with earlier research such as Russell and Wefald's (1991a; 1991b) is that the estimate of the

19.For a brief informal introduction to the research, see Russell (1999).

utility of computation they use is myopic. That is, they base a decision to deliberate further, on whether the net expected utility of the solution after computation minus the cost of time is greater than the expected utility of the current solution. However the performance profiles of anytime algorithms are not always certain; indeed they can vary considerably. Therefore alternative algorithms, such as Hansen and Zilberstein's (2001) dynamic programming method that uses a more global utility estimate and that adds decisions about whether to monitor the state of the world or not at given points during the process, represent a more general and accurate treatment of the metareasoning problem.

Raja (2003; Raja and Lessor, 2004) also report progress related to the research of Harada and Russell using reinforcement learning techniques to generate a meta-level control policy that govern decisions in multiagent environments. They learn a meta-level MDP where the state consists of set of abstract qualitative features of a multiagent hierarchical task net (HTN) problem environment, the mental actions are processes such as scheduling a task and negotiating with another agent, and the reward function is the overall utility gained by the multiagent system as a result of the execution of the HTN plans. The system learns the MDP by making random decisions to collect state transition probabilities. Value iteration of Q-values computes an optimal meta-level policy. Finally, the system is reimplemented using the learned policy.

In contrast to advances such as those regarding metareasoning above, some recent research into introspective learning has strayed from its original formulation. Fox and Leake (1995a; 1995b) originally defined and continue to emphasize (Fox and Leake 2001) that introspective learning uses a model of the reasoning process to derive expectations concerning the behavior of the reasoning process. Thus by monitoring the reasoning process given these expectations, an introspective system can uncover failures that point to useful learning. As a result the system can adjust case indices to improve performance. However other researchers, such as Bonzano, Cunningham, and Smyth (1997), interpret introspective learning as monitoring the results of problem solving in relation to an objective function and adjusting memory indices as a function of the comparison. But to do so is to revert to a more standard machine learning perspective. They lack the emphasis on a declarative self-model of the reasoning that guides the detection of failure as opposed to an external objective function (specifically a training set of examples). This trend is continued by the research of Zhang and Yang (2001) and of Coyle and Cunningham (2004), still under the term of introspective learning where the learning is considered to be specific to learning index weights.

Others in the CBR community have successfully extended their previous research adding such constructs as meta-cases (Murdock and Goel 2001).[20] Research that makes computational use of meta-rules continues into the present as well. See for example the work of Cazenave (2003; Bouzy and Cazenave 2001) and the implemented Introspect system used to solve problems in the game of Go. Hudlicka (2005) also presents a novel implemented system that uses metacognition to mediate the effects of emotion on deliberation for action selection. Her research is inspired by the many new developments in the psychological metacognition literature.

One of the most encouraging trends has been the new research efforts that take a cross-disciplinary approach (e.g., Anderson 2003; Gordon 2004; Oehlmann 2003) where each integrates computational methods with psychological or philosophical approaches. A prominent example is the work of Gordon and Hobbs (2004; Hobbs and Gordon 2005). They have undertaken the first-order logical representation of 30 commonsense domains of mental activities and strategies such as memory, knowledge management, envisionment, planning, goals, and execution monitoring. But rather than using intuition to construct a competency formalism (c.f., McCarthy 1995; Schank *et al*. 1972), they have performed a large-scale analysis of human planning protocols (Gordon 2004), to obtain independent coverage first. That is, the representation of a content theory of logical terms depends upon a cognitive analysis of a natural corpus of mental terms. Note that this is in contrast to a third method whereupon the representation depends upon theoretical assumptions about metacognition (Cox 1995; Cox & Ram 1999).[21]

Anderson and Perlis (in press; 2005) also take a decidedly cognitive science direction. Anderson is a computationally-oriented philosopher by training who, from an embodied cognition perspective (Anderson in press; 2003), has studied technical problems associated with representation of the self. Countering the claims that the self is essentially an indexical, Anderson argues that self-representing mental tokens structurally organize self-knowledge, having a biological underpinning related to somatoception in the body. Furthermore Anderson and Perlis (2005) propose a computational theory of the "metacognitive loop" that accounts for improved performance in such behavioral components as reinforcement learning, robot navigation, and human-computer dialogue.

Most importantly many researchers have recently begun to work on significant architectures that specifically support metacognitive layers of monitoring and control of deliberation (i.e., cognition) of both inference and of memory. Examples include the work of Minsky, Sloman and colleagues (see McCarthy, Minsky, Sloman, Gong, Lau, Morgenstern, Meuller, Riecken, Singh, and Singh 2002 and Sloman 2001), Forbus and Hinrichs (2004), Anderson and

20.Notice the ambiguity of the term meta-case-based reasoning as used by Murdock and Goel. In their research meta-cases represent general cases that contain information about functional model cases; hence it is a case about a case. Whereas as used by Leake (1996b), the term can be construed as case-based reasoning about the case-based reasoning process itself.

21.The representational content theories of mental states and actions developed by Schank *et al*, by Gordon and Hobbs, and by Cox and Ram are all at the knowledge level. Newell (1982) used Schank's conceptual dependency representation as a specific example of a theory at the knowledge level, and both Gordon and Cox use the exact same approach as did Schank.

Perlis (2005), Schubert (2005), and Cox (2005). Minsky and Sloman have proposed a three-level architecture that mediates perception and action through reactive, deliberative, and reflective process layers. Forbus and Hinrichs (2004) propose a new architecture for "companion cognitive systems" that employ psychologically plausible models of analogical learning and reasoning and that maintain self-knowledge in the form of logs of activity. Cox (2005) proposes a preliminary architecture consisting of planning, understanding, and learning in which awareness is exhibited by an agent as it generates its own goals to change perceived anomalies in the world and in which self-awareness is exhibited as it generates explicit learning goals to change perceived anomalies in its own knowledge.

Singh (2005) has recently created an architecture called EM-ONE, that supports layers of metacognitive activities that monitor reasoning in physical, social, and mental domains. These layers range from the reactive, deliberative, reflective, self-reflective, and self-conscious to the self-ideals layer. Mental critics are represented as a case base of commonsense narratives that associate specific situations with a method of debugging the situation. Critics themselves are selected and retrieved by an executive set of meta-level critics that detect and classify problems in the environment or within the system.

The metacognition community in psychology has recently started a novel line of research on metacogntion and vision (see Levin 2004 in particular). Although some consider metacognition specifically related to higher order thought, this new research examines how people think about their own visual perception. Levin and Beck (2004) demonstrate that not only do people overestimate their visual capabilities, but most interesting, given feedback on their errors, they refuse to believe the evidence "before their eyes." For example humans will fail to perceive changes in clothing (e.g., a scarf that disappears) if the change occurs during video tape cuts or scene shifts. This robust effect is called change blindness. As Keil, Rosenblit, and Mills (2004) notes, this effect may be related to the illusion of explanatory depth, because human subjects do not fully understand the mechanisms behind their own visual perception, although they believe that they do.

Thus again I emphasize that metacognition in its many forms has limitations. As noted above in the general case metareasoning is intractable. But at the same time, it has the potential to provide a level of decision making that can make an intelligent system robust and tolerant of errors and of dynamic changing environments. As the twenty-first century opened, Bruce Buchanan in his AAAI Presidential Address (Buchanan 2000) claimed that the meta-level of computation provides a principled basis for genuine creativity. Surveying the literature in creativity, he argued that the feature that is most characteristic of creativity is the ability to bring something novel into existence. The argument was that search at the meta-level enables the identification of choices that are most effective for successfully completing particular tasks. This search allows the reasoner to modify the ontological vocabulary, the criteria, and the methods used at the object level to make decisions. Finally it allows an intelligent agent to define new problems for itself. Given these kinds of attributes, agents might have the capacity to go beyond the limitations of intelligent systems of the past. Yet at the current time, this is still a distant dream. Or is it?

Summary and Discussion

This paper outlined some of the research related to metacognition both from the artificial intelligence perspective and from the cognitive psychology point of view. This paper first examined psychological research into metacognition and human problems solving. It then described the genesis of interest in computational theories of introspection and metacognition during the formative years of AI. The logic community has a large part to play in this early research, because they established a formalism (and a legitimacy) for the representation of mental states and belief, including beliefs about a system's own beliefs. I also examined the research of the expert system community and others that have developed introspective systems, but take a different approach. I also discussed systems that combine metacognitive theories with model-based reasoning, case-based reasoning, and theories of learning. Finally I examined a set of more recent papers on all of these subjects that have been published since the turn of the century.

The computational community should take note of the results from other disciplines concerning metacognition. For example it is enticing to design an organized memory or knowledge base so that it is "indexed" to answer queries concerning the contents of memory. Indeed Nilsson (1980) begins his section on Meta-Knowledge with "We would like to be able to build systems that know or can deduce whether or not they know facts and rules about certain subjects without having to scan their large knowledge bases searching for these items." After all humans exhibit tip-of-the-tongue behavior, so this sounds reasonable.

However Reder and Ritter (1992) argue that such behavior (e.g., game-show events where people have to quickly hit a buzzer, if they think they can answer a question), is tied to familiarity with the questions rather than with the answers. This has important ramifications for those researchers like Nilsson wishing to build systems with metaknowledge. It indicates that direct access to memory content may not be fruitful and that inferential measures such as cue familiarity or current access to related concepts may provide a better measure (see the discussion in Dunlosky, 2004, for why these alternatives work with humans). In any case knowing the metacognitive literature and the human data can point computer scientists toward new possibilities and warn them about potential pitfalls.

Conversely Ghetti (2003) provides recent evidence to support Hayes' proposed metareasoning strategy about television scan lines discussed at the beginning of the section on model-based and case-based reasoning.[22] That is Ghetti showed that humans infer event nonoccurrences from the premise that, if they did occur, then the event would be memorable. Because they do not immediately retrieve the fact, they therefore must not know it. Regardless, computer scientists should have a working knowledge of the psychological literature on metacognition to provide evidence for or against their intuitions concerning the mental abilities of humans.

Yet many ambiguities and conflicting evidence exist within all of the disciplines enumerated here. Often, authors use different terms for the same concept (e.g., introspection and reflection[23]), and sometimes the same terms are used in different ways (e.g., metacognition is a multiple overloaded term). Indeed, Brown (1987) has described research into metacognition as a "many-headed monster of obscure parentage." This characterization applies equally as well to the many AI approaches that deal with metacognition, metareasoning, and metaknowledge and the relationships between each of them.

Finally, both metacognition theory and computational theories address the issue concerning a person's ability to assess the veracity of their own responses. In addition, because a person has a feeling of knowing, even when recall is blocked, the agent can make efficient use of search. Search and elaboration is pursued when an item is on the "tip of the tongue" and abandoned when an item is judged unfamiliar. This search heuristic provides some control of memory and helps to alleviate the combinatorial explosion of inferences (Lachman, Lachman and Thronesbery 1979; Miner and Reder 1994). Although people sometimes make spurious and biased inferences when assessing their own memories and reasoning, these inferences nonetheless affect people's decisions and thus are important components when understanding human decision-making.

By some measures, few people are working on metacognition, but in another sense used by some in the AI community, everyone in AI must be working on introspection and metareasoning. Most intelligent programs deliberate to some extent about the types of actions that are optimal given their goals. For example, Soar (Newell 1990; Rosenbloom, Laird, and Newell 1989; 1993), Theo (Mitchell, Allen, Chalasani, Cheng, Etzioni, Ringuette and Schlimmer 1991), and PRODIGY (Carbonell, Knoblock, and Minton 1991; Veloso, Carbonell, Perez, Borrajo, Fink, and Blythe 1995) are all programs that make deliberate control decisions as to the best action available in their domains. Moreover, if metaknowledge were taken to be any abstract knowledge (e.g., default knowledge), and metareasoning is any of the higher cognitive functions (e.g., planning), then virtually all AI programs would be metacognitive. Instead I echo Maes' assessment that an introspective system is one whose domain is itself (Maes 1987b). But in essence a metacognitive reasoner is a system that *reasons* specifically about itself (its knowledge, beliefs, and its reasoning process), not one that simply *uses* such knowledge.[24]

It needs to be better appreciated just how extensive the research is on metacognitive aspects of intelligent behavior. Indeed I have been forced to omit much important research such as the work on metacognitive monitoring in high-level perception and analogy (e.g., Marshall, 1999; Marshall and Hofstadter, 1998), active logic (Elgot-Drapkin and Perlis 1990) and more generally logic programming (but see Costantini, 2002), models of introspective distributed agents (e.g., Mason, 1994), self-adaptive software (e.g., Robertson, 2003) and BDI agents that reconsider intentions using decision-theoretic metareasoning (Schut, Wooldridge, and Parsons 2004; Schut and Wooldridge, 2001). But much of the past research covered in this paper contains valuable lessons to teach us and provides firm foundations with which to make progress in our individual fields of expertise. In any case and as is with all careful research, we should be aware of the work that has preceded us, if for nothing else than to prevent ourselves from reinventing the wheel or repeating past failures.

Acknowledgments

I thank David Leake and Mike Anderson for recent comments on the content of this article. I also thank the anonymous reviewers for their insights and suggestions. This paper started with a literature search for a paper (Cox 1994b) that was originally written for a graduate school seminar in metacognition and cognitive aging taught by Chris Hertzog at Georgia Tech. Over the years many people have provided me with pointers into the various literatures and feedback on portions of the material contained here. For a rather large set of acknowledgments see Cox (1996b). Because the material I cover is so broad, I necessarily have both sins of omission as well as commission in the historical references. See also the Introspection home page that I maintain on an irregular basis: `hcs.bbn.com/cox/Introspect/`

References

Anderson, J. R. 1983. *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., and Thompson, R. 1989. Use of Analogy in a Production System Architecture. In S. Vosniadou and A. Ortony eds. *Similarity and Analogical Reasoning*, 267-297. Cambridge: Cambridge University Press.

Anderson, M. L. in press. How to Study the Mind: An Introduction to Embodied Cognition. In F. Santoianni and C. Sabatano, eds. *Brain Development in Learning Environments: Embodied and Perceptual Advancements*. New York: Cambridge University Press.

Anderson, M. L. 2003. Embodied Cognition: A Field

22.Moore (1985) uses the same logic in his example of autoepistemic reasoning whereby one concludes the lack of an older brother given that the experience of having such a brother would be saliently represented and given the lack of an assertion concerning a brother in the knowledge base.

23.Although note that I have differentiated these two terms when discussing Minsky's use of M* and M**.

24.Thus systems that use metaknowledge are not necessarily metacognitive. For example metaknowledge concerning the properties of constraints may assist CSP solvers to be more efficient in terms of reducing the number of arc consistency checks (Bessiere, Freuder and Regin 1999), but I assert that such algorithms in isolation should not be included in metacognition in computing activities.

Guide. *Artificial Intelligence* 149(1): 91-130.

Anderson, M. L., and Oates, T. (Eds.) 2005. *Metacognition in Computation: Papers from 2005 AAAI Spring Symposium*. AAAI Technical Report SS-05-04. Menlo Park, CA: AAAI Press.

Anderson, M. L., and Perlis, D. R. 2005. Logic, Self-Awareness and Self-Improvement: The Metacognitive Loop and the Problem of Brittleness. *Journal of Logic and Computation*, 15(1): 21-40.

Anderson, M. L., and Perlis, D. R. in press. The Roots of Self-Awareness. *Phenomenology and the Cognitive Sciences*.

Antaki, C., and Lewis, A. eds. 1986. *Mental Mirrors: Metacognition in Social Knowledge and Communication*. London: Sage Publications.

Attardi, G., and Simi, M. 1991. Reflections about Reflection. In J. Allen, R. Fikes, and E. Sandewall eds. *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference* (KR91). San Mateo, CA: Morgan Kaufmann.

Augustine 1955. De Trinitate. In J. Burnaby, trans. and ed. *Augustine: Later works* (Vol. 8, Library of Christian Classics, Bk. 10, Sec. 7), 80. SCM Press. (Original work published around 1600)

Barr, A. 1977. *Meta-Knowledge and Memory*, Technical Report, HPP-77-37. Stanford University, Department of Computer Science, Stanford, CA.

Barr, A. 1979. Meta-Knowledge and Cognition. In *Proceedings of the Sixth International Joint Conference on Artificial Intelligence* 31-33. Los Altos, CA: Morgan Kaufmann.

Batali, J. 1983. *Computational Introspection*, Technical Report, 701. Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. 1995. Metacognition and Problem-Solving: A Process-Oriented Approach. J*ournal of Experimental Psychology: Learning, Memory, and Cognition 21*(1): 205-223.

Bessiere, C., Freuder, E. C., and Regin, J-C. 1999. Using Constraint Metaknowledge to Reduce Arc Consistency Computation. *Artificial Intelligence* 107: 125-148.

Bhatta, S. 1995. Model-Based Analogy in Innovative Device Design. Ph.D. diss., College of Computing, Georgia Institute of Technology, Atlanta.

Bhatta, S., and Goel, A. 1992. Use of Mental Models for Constraining Index Learning in Experience-Based Design. In *Proceedings of the AAAI-92 Workshop on Constraining Learning with Prior Knowledge*.

Birnbaum, L., Collins, G., Freed, M., and Krulwich, B. 1990. Model-Based Diagnosis of Planning Failures. In *Proceedings of the Eighth National Conference on Artificial Intelligence,* 318-323. Menlo Park, CA: AAAI Press.

Bonzano, A., Cunningham, P., and Smyth, B. 1997. Using Introspective Learning to Improve Retrieval in CBR: A Case Study in Air Traffic Control. In D. B. Leake and E. Plaza eds. *Case-Based Reasoning Research and Development: Second International Conference on Case-Based Reasoning* 291-301. Berlin: Springer.

Boring, E. G. 1953. A History of Introspection. *Psychological Bulletin* 50(3): 169-189.

Bouzy, B., and Cazenave, T. 2001. Computer Go: An AI Oriented Survey. *Artificial Intelligence* 132: 39-103.

Brachman, R. J. 2002, Nov/Dec. Systems That Know What They Are Doing. *IEEE Intelligent Systems* 67-71.

Brazdil, P. B., and Konolige, K. eds. 1990. *Machine learning, meta-reasoning and logics.* Norwell, MA: Kluwer Academic.

Brown, A. 1987. Metacognition, Executive Control, Self-regulation, and Other More Mysterious Mechanisms. In F. E. Weinert and R. H. Kluwe eds. *Metacognition, Motivation, and Understanding* 65-116. Hillsdale, NJ: LEA.

Buchanan, B. G. (2000). Creativity at the Meta-Level: AAAI-2000 Presidential Address. *AI Magazine* 22(3): 13-28.

Carbonell, J. G. 1986. Derivational Analogy: A Theory of Reconstructive Problem Solving and Expertise Acquisition. In R. Michalski, J. Carbonell and T. Mitchell eds. *Machine learning: An artificial intelligence approach, Vol. 2,* 371-392. San Mateo, CA: Morgan Kaufmann Publishers.

Carbonell, J. G., Knoblock, C. A., and Minton, S. 1991. PRODIGY: An Integrated Architecture for Planning and Learning. In K. VanLehn ed. *Architectures of Cognition: The 22nd Carnegie Mellon Symposium on Cognition* 241-278. Hillsdale, NJ: LEA.

Cazenave, T. 2003. Metarules to Improve Tactical Go Knowledge, *Information Sciences*, 154(3-4):173-188.

Ceci, S. J. ed. 1987. *Handbook of Cognitive, Social, and Neuropsychological Aspects of Learning Disabilities* (Vol. 2). Hillsdale, NJ: LEA.

Cheng, J. 1995. Management of Speedup Mechanisms in Learning Architectures, Technical Report, CMU-CS-95-112. Ph.D. diss., School of Computer Science, Carnegie Mellon University, Pittsburgh.

Chi, M. T. H. 1987. Representing Knowledge and Metaknowledge: Implications for Interpreting Metamemory Research. In F. E. Weinert and R. H. Kluwe eds. *Metacognition, Motivation, and Understanding* 239-266. Hillsdale, NJ: LEA.

Chi, M. T. H. 1995. *Revising the Mental Model As One Learns*. Plenary address to the Seventeenth Annual Conference of the Cognitive Science Society. Pittsburgh (July 23).

Chi, M. T. H., Bassok, M., Lewis, M., Reimann, P., and Glasser, R. 1989. Self-Explanations: How Students Study and Use Examples in Learning to Solve Problems. *Cogni-*

tive Science 13: 145-182.

Clancey, W. J. 1987. *The Knowledge Engineer As Student: Metacognitive Bases for Asking Good Questions*, Technical Report, STAN-CS-87-1183, Department of Computer Science, Stanford University, Stanford, CA.

Clancey, W. J. 1992. Model Construction Operators. *Artificial Intelligence* 53: 1-115.

Clancey, W. J., and Bock, C. 1985. *Representing Control Knowledge as Abstract Task and Metarules., Technical Report,* STAN-CS-87-1168, Department of Computer Science, Stanford University, Stanford, CA.

Cointe, P. ed. 1999. *Meta-Level Architectures and Reflection: Second International Conference, Reflection '99*. Berlin: Springer.

Collins, G. 1987. Plan Creation: Using Strategies as Blueprints, Technical Report, 599. Ph.D. diss., Department of Computer Science, Yale University, New Haven, CT.

Collins, G., Birnbaum, L., Krulwich, B., and Freed, M. 1993. The Role of Self-models in Learning to Plan. In A. Meyrowitz ed. *Machine Learning: Induction, analogy and discovery*. Boston: Kluwer Academic Publishers.

Conitzer, V, and Sandholm, T. 2003. Definition and Complexity of Some Basic Metareasoning Problems. In G. Gottlob and T. Walsh eds., Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), 1099-1106. San Francisco: Morgan Kaufmann.

Costantini, S. 2002. Meta-Reasoning: A Survey. In A. Kakas and F. Sadri (Eds.), *Computational Logic: From Logic Programming into the Future*. Special volume in honour of Bob Kowalski. Berlin: Springer.

Cox, M. T. 1994a. Machines That Forget: Learning from Retrieval Failure of Mis-indexed Explanations. In A. Ram and K. Eiselt eds. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society,* 225-230. Hillsdale, NJ: LEA.

Cox, M. T. 1994b. Metacognition, Problem Solving and Aging (Cognitive Science Tech. Rep. No. 15). Atlanta: Georgia Institute of Technology, College of Computing.

Cox, M. T. 1995. Representing Mental Events (or the lack thereof). In M. T. Cox and M. Freed eds. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms*, 22-30. Menlo Park, CA: AAAI Press. (Available as Technical Report, SS-95-08)

Cox, M. T. 1996a. An Empirical Study of Computational Introspection: Evaluating Introspective Multistrategy Learning in the Meta-AQUA System. In R. S. Michalski and J. Wnek, eds. *Proceedings of the Third International Workshop on Multistrategy Learning* 135-146. Menlo Park, CA: AAAI Press / The MIT Press.

Cox, M. T. 1996b. Introspective Multistrategy Learning: Constructing a Learning Strategy under Reasoning Failure. Technical Report, GIT-CC-96-06. Ph.D. diss., College of Computing, Georgia Institute of Technology, Atlanta. (`hcs.bbn.com/cox/thesis/`)

Cox, M. T. 1997a. An Explicit Representation of Reasoning Failures. In D. B. Leake and E. Plaza eds. *Case-Based Reasoning Research and Development: Second International Conference on Case-Based Reasoning* 211-222. Berlin: Springer.

Cox, M. T. 1997b. Loose Coupling of Failure Explanation and Repair: Using Learning Goals to Sequence Learning Methods. In D. B. Leake and E. Plaza eds. *Case-Based Reasoning Research and Development: Second International Conference on Case-Based Reasoning* 425-434. Berlin: Springer-Verlag.

Cox, M. T. 2005. Perpetual Self-Aware Cognitive Agents. In M. Anderson and T. Oates eds., *Metacognition in Computation: Papers from 2005 AAAI Spring Symposium*, 42-48. Technical Report SS-05-04. Menlo Park, CA: AAAI Press.

Cox, M. T. and Freed, M. 1994. Using Knowledge from Cognitive Behavior to Learn from Failure. In J. W. Brahan and G. E. Lasker eds. *Proceedings of the Seventh International Conference on Systems Research, Informatics and Cybernetics: Vol. 2. Advances in Artificial Intelligence - Theory and Application II,* 142-147. Windsor, Ontario, Canada: The International Institute for Advanced Studies in Systems Research and Cybernetics.

Cox, M. T. and Freed, M. eds. 1995. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms*, Technical Report, SS-95-08. Menlo Park, CA: AAAI Press.

Cox, M. T., and Ram, A. 1991. Using introspective reasoning to select learning strategies. In R. S. Michalski and G. Tecuci eds. *Proceedings of the First International Workshop on Multistrategy Learning*, 217-230. Washington, DC: George Mason University, Artificial Intelligence Center.

Cox, M. T., and Ram, A. 1992. Multistrategy learning with introspective meta-explanations. In D. Sleeman and P. Edwards eds. *Machine Learning: Ninth International Conference*, 123-128. San Mateo, CA: Morgan Kaufmann.

Cox, M. T., and Ram, A. 1995. Interacting Learning-goals: Treating Learning as a Planning Task. In J.-P. Haton, M. Keane and M. Manago eds. *Advances in Case-Based Reasoning,* 60-74. Berlin: Springer-Verlag.

Cox, M. T., and Ram, A. 1999. Introspective Multistrategy Learning: On the Construction of Learning Strategies. *Artificial Intelligence* 112: 1-55.

Coyle, L., and Cunningham, P. 2004. Improving Recommendation Rankings by Learning Personal Feature Weights. Technical Report TCD-CS-2004-21. The University of Dublib, Trinity College, Ireland.

Davidson, J.E., Deuser, R., and Sternberg, R.J. 1994. The role of metacognition in problem solving. In J. Metcalfe and A. P. Shimamura eds. *Metacognition* 207-226. Cambridge, MA: The MIT Press.

Davis, D. N. in press. *Visions of Mind: Architectures for*

Cognition and Affect. Hershey, PA: Idea Group Inc.

Davis, R. 1976. Applications of Meta-Level Knowledge to the Construction, Maintenance, and Use of Large Knowledge Bases. Stanford HPP Memo 76-7. Stanford University.

Davis, R. 1979. Interactive Transfer of Expertise: Acquisition of New Inference Rules. *Artificial Intelligence* 12: 121-157.

Davis, R. 1980. Meta-Rules: Reasoning about Control. *Artificial Intelligence* 15: 179-222.

Davis, R., and Buchanan, B. 1977. Meta-Level Knowledge: Overview and Applications. In *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* 920-927. Los Altos, CA: Morgan Kaufmann.

Dean, T, and Boddy, M. 1988. An Analysis of Time-Dependent Planning. In T. M. Mitchell, and R. G. Smith eds., *Proceedings of the Seventh National Conference on Artificial Intelligence*, 49-54. Menlo Park, CA: AAAI Press.

de Kleer, J., and Brown, J. S. 1988. Assumptions and Ambiguities in Mechanistic Mental Models. In A. Collins and E. E. Smith eds. *Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence*, 270-287. San Mateo, CA: Morgan Kaufmann. (Original work published 1983)

de Kleer, J. 1984. How Circuits Work. *Artificial Intelligence* 24: 205-280.

de Kleer, J., Doyle, J., Steele, G. L., and Sussman, G. J. 1977. Explicit Control of Reasoning. *SIGPLAN Notices, 12*.

Delclos, V. R., and Harrington, C. 1991. Effects of Strategy Monitoring and Proactive Instruction on Children's Problem-Solving Performance. *Journal of Educational Psychology* 83(1): 35-42.

Dennett, D. C. 1978. *Brainstorms: Philosophical Essays on Mind and Psychology.* Cambridge, MA: MIT Press/Bradford Books.

Derry, S. J. 1989. Strategy and Expertise in Solving Word Problems. In C. B. McCormick, G. E. Miller, and M. Pressley eds. *Cognitive Strategy Research: From Basic Research to Educational Applications*, 269-302. Berlin: Springer-Verlag.

Dominowski, R. L. 1998. Verbalization and problem solving. In D. L. Hacker, J. Dunlosky, and A Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 25-45). Mahwah, NJ: Lawrence Erlbaum Associates.

Dörner, D. 1979. Self-Reflection and Problem-solving. In F. Klix ed. *Human and Artificial Intelligence*, 101-107. Amsterdam: North Holland.

Doyle, J. 1979. A Truth Maintenance System. *Artificial Intelligence* 12: 231-272.

Doyle, J. 1980. A Model for Deliberation, Action, and Introspection, Technical Report, TR-581. Ph.D. diss., Department of Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Dunlosky 2004. Metacognition. In Hunt and Ellis (Eds.), *Fundamentals of Cognitive Psychology*, 7th Ed. (pp. 232-262). New York: McGraw Hill.

Elgot-Drapkin, J., and Perlis, D. 1990. Reasoning Situated in Time I: Basic Concepts. *Journal of Experimental and Theoretical Artificial Intelligence* 2: 75-98.

Epstein, R. L., and Carnielli, W. A. 1989. *Computability: Computable Functions, Logic, and the Foundations of Mathematics.* Pacific Grove, CA: Wadsworth and Brooks.

Etzioni, O. 1991. Embedding Decision-Analytic Control in a Learning Architecture. *Artificial Intelligence* 49: 129-159.

Fink, E. 1999. Automatic Representation Changes in Problem Solving. Technical Report, CMU-CS-99-150, Ph.D. Thesis, Computer Science Department, Carnegie Mellon University.

Fink, E. 1998. How to Solve It Automatically: Selection among Problem-Solving Methods. In *Proceedings of the Fourth International Conference on Artificial Intelligence Planning Systems*, 128-136.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3*(4), 552-564.

Flavell, J. H. 1971. First Discussant's Comments: What Is Memory Development the Development of? *Human Development* 14: 272-278.

Flavell, J. H. 1976. Metacognitive Aspects of Problem Solving. In Resnick ed. *The Nature of Intelligence*, 231-235. Hillsdale, NJ: LEA.

Flavell, J. H., and Wellman, H. M. 1977. Metamemory. In R. V. Kail, Jr., and J. W. Hagen eds. *Perspectives on the Development of Memory and Cognition,* 3-33. Hillsdale, NJ: LEA.

Forbus, K. and Hinrichs, T. 2004. Companion Cognitive Systems: A step towards human-level AI. In *AAAI Fall Symposium on Achieving Human-level Intelligence through Integrated Systems and Research*, October, Washington, DC.

Forrest-Pressley, D. L., MacKinnon, G. E., and Waller, T. G. eds. 1985. *Metacognition, Cognition and Human Performance* (Vol. 2, Instructional Practices). New York: Academic Press.

Fox, S. 1995. Introspective Learning for Case-Based Planning. Unpublished, Ph.D. diss., Department of Computer Science, Indiana University, Bloomington, IN.

Fox, S., and Leake, D. 1995a. Modeling Case-Based Planning for Repairing Reasoning Failures. In M. T. Cox and M. Freed eds. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms,* 31-38. Menlo Park, CA: AAAI Press. (Available as Technical Report, SS-95-08)

Fox, S., and Leake, D. 1995b. Using Introspective Reasoning to Refine Indexing. In C. S. Mellish ed. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 391-397. San Mateo, CA: Morgan Kaufmann.

Fox, S., and Leake, D. 2001. Introspective Reasoning for Index Refinement in Case-Based Reasoning. *Journal of Experimental and Theoretical Artificial Intelligence* 13: 63-88.

Freed, M., and Collins, G. 1994. Learning to Cope with Task Interactions. In A. Ram and M. desJardins eds. *Proceedings of the 1994 AAAI Spring Symposium on Goal-Driven Learning* 28-35. Menlo Park, CA: AAAI Press.

Freed, M., Krulwich, B., Birnbaum, L., and Collins, G. 1992. Reasoning about Performance Intentions. In *Proceedings of Fourteenth Annual Conference of the Cognitive Science Society* 7-12. Hillsdale, NJ: LEA.

Garner, R. 1987. *Metacognition and reading comprehension*. Norwood, NJ: Ablex Publishing Corporation.

Gavelek, J. R., and Raphael, T. E. 1985. Metacognition, Instruction, and the Role of Questioning Activities. In D. L. Forrest-Pressley, G. E. MacKinnon, and T. G. Waller eds. *Metacognition, Cognition and Human Performance*. Vol. 2 (Instructional Practices), Academic Press, Inc., New York, 103-136.

Genesereth, M. R. 1983. An Overview of Meta-Level Architecture. In *Proceedings of the Third National Conference on Artificial Intelligenc*e, 119-123. Los Altos, CA: William Kaufmann.

Ghetti, S. 2003. Memory for Nonoccurrences: The Role of Metacognition. *Journal of Memory and Language* 48(4): 722-739.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1992). The illusion of knowing: Failure in the self-assessment of comprehension. In T. O. Nelson (Ed.), *Metacognition: Core readings* (pp. 185-195). Boston: Allyn and Bacon. (Original work published 1982)

Gombert, J. E. 1992. *Metalinguistic Development*. Chicago: University of Chicago Press.

Good, I. J. 1971. Twenty-Seven Principles of Rationality. In V. P. Godambe and D. A. Sprott eds. *Foundations of Statistical Inference*. Toronto: Hold, Rinehart, Winston.

Gordon, A. S. 2004. *Strategy Representation: An Analysis of Planning Knowledge*. Mahwah, NJ: LEA.

Gordon, A. S., and Hobbs, J. R. 2004. Formalizations of Commonsense Psychology. *AI Magazine* 25(4): 49-62.

Hammond, K. J. 1989. *Case-Based Planning: Viewing Planning as a Memory Task*. *Vol. 1.* of *Perspectives in Artificial Intelligence*. San Diego, CA: Academic Press.

Hammond, K. J. 1990. Explaining and Repairing Plans That Fail. *Artificial Intelligence* 45: 173-228.

Hansen, E. A., and Zilberstein, S. 1996. Monitoring Anytime Algorithms. *ACM SIGART Bulletin 7*(2): 28-33.

Hansen, E. A., and Zilberstein, S. 2001. Monitoring and Control of Anytime Algorithms: A Dynamic Programming Approach. *Artificial Intelligence* 126(1-2): 139-157.

Harada, D., and Russell, S. J. 1999. Extended Abstract: Learning search strategies. In W. Zhang and S. Koenig (Eds.), *Search Techniques for Problem Solving under Uncertainty and Incomplete Information: Papers from the AAAI Spring Symposium* (pp. 48-52). AAAI Technical Report SS-99-07. Menlo Park, CA: AAAI Press.

Haugeland, J. ed. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Hayes, P. J. 1981. The Logic of Frames. In B. L. Webber and N. J. Nilsson eds. *Readings in Artificial Intelligence,* 451-458. Los Altos, CA: Morgan Kaufmann. (Original work published 1979)

Hayes-Roth, F., Waterman, D. A., and Lenat, D. B. eds. 1983. *Building Expert Systems*. London: Addison-Wesley Publishing.

Hobbs, J. R., and Gordon, A. S. 2005. Toward a large-scale formal theory of commonsense psychology for metacognition. In M. Anderson and T. Oates eds., *Metacognition in Computation: Papers from 2005 AAAI Spring Symposium*, 49-54. Technical Report SS-05-04. Menlo Park, CA: AAAI Press.

Hofstadter, D. R. 1989. *Gödel, Escher, Bach: An Eternal Golden Braid*. New York: Vintage Books. (Original work published in 1979)

Horty, J., and Shoham, Y. eds. 1993. *Proceedings of the 1993 AAAI Spring Symposium on Reasoning about Mental States: Formal Theories and Applications.* Menlo Park, CA: AAAI Press.

Horvitz, E. 1987. Reasoning about Beliefs and Actions under Computational Resource Constraints. In *Proceedings of the Third Workshop on Uncertainty in Artificial Intelligence*, 429-444. Seattle, Washington. Also in L. Kanal, T. Levitt, and J. Lemmer, ed., 1990. *Uncertainty in Artificial Intelligence 3*, 301-324. Elsevier.

Horvitz, E. J., Cooper, G., and Heckerman, D. 1989. Reflection and Action under Scarce Resources: Theoretical Principles and Empirical Study. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann.

Horvitz, E., and Zilberstein, S. 2001. Special Issue on Computational Tradeoffs under Bounded Resources. *Artificial Intelligence* 126(1-2).

Hudlicka, E. 2005. Modeling Interaction between Metacognition and Emotion in Cognitive Architectures. In M. Anderson and T. Oates eds., *Metacognition in Computation: Papers from 2005 AAAI Spring Symposium*, 55-61. Technical Report SS-05-04. Menlo Park, CA: AAAI Press.

Johnson-Laird, P. N. 1983. *Mental models: Toward a Cognitive Science of Language, Inference, and Consciousness*.

Cambridge: Cambridge University Press.

Johnson-Laird, P. N. 1988. *The Computer and the Mind: An Introduction to Cognitive Science*. Cambridge, MA: Harvard University Press.

Kass, A. 1990. Developing Creative Hypotheses by Adapting Explanations. Ph.D. diss., The Institute for the Learning Sciences, Northwestern University, Evanston, IL.

Keil, F., Rosenblit, L., and Mills, C. M. 2004. What Lies Beneath? Understanding the Limits of Understanding. In D. T. Levin ed. *Thinking and Seeing*, 227-249. Cambridge, MA: The MIT Press.

Kolodner, J. L. 1993. *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann Publishers.

Konolige, K. 1985. A Computational Theory of Belief Introspection. *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, 502-508. Los Altos, CA: Morgan Kaufmann.

Konolige, K. 1986. *A Deduction Model of Belief*. Los Altos, CA: Morgan Kaufmann.

Konolige, K. 1988. Reasoning by Introspection. In P. Maes and D. Nardi eds. *Meta-Level Architectures and Reflection,* 61-74. Amsterdam: North-Holland.

Krulwich, B. 1991. Determining What to Learn in a Multi-Component Planning System. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society.* Chicago, IL, (August 7-10), 102-107.

Krulwich, B. 1993. Flexible Learning in a Multicomponent Planning System, Technical Report, 46. Ph.D. diss., The Institute for the Learning Sciences, Northwestern University, Evanston, IL.

Kuokka, D. R. 1990. The Deliberative Integration of Planning, Execution, and Learning, Technical Report, CMU-CS-90-135. Ph.D. diss., Computer Science Dept., Carnegie Mellon University, Pittsburgh.

Lachman, J. L., Lachman, R., and Thronesbery, C. 1979. Metamemory through the Adult Life Span. *Developmental Psychology* 15(5): 543-551.

Laddaga, R. 1998. Self-Adaptive Software. DARPA Solicitation BAA 98-12.

Lagoudakis, M. G., Littman, M. L., and Parr, R. 2001. Selecting the right algorithm. In Carla Gomes and TobyWalsh, eds., *Proceedings of the 2001 AAAI Fall Symposium Series: Using Uncertainty within Computation*, Menlo Park, CA: AAAI Press.

Lagoudakis, M. G., Parr, R., and Littman, M. L. 2002. Least-Squares Methods in Reinforcement Learning for Control. In *Proceedings of the 2nd Hellenic Conference on Artificial Intelligence* (pp. 249-260) Lecture Notes on Artificial Intelligence, Vol. 2308. Berlin: Springer.

Leake, D. B. ed. 1996a *Case-based Reasoning: Experiences, Lessons, & Future Directions*. Menlo Park, CA: AAAI Press / The MIT Press.

Leake, D. B. 1996b. Experience, Introspection, and Expertise: Learning to Refine the Case-based Reasoning Process. *Journal of Experimental and Theoretical Artificial Intelligence* 8(3-4): 319-339.

Lenat, D. B., Davis, R., Doyle, J., Genesereth, M., Goldstein, I., and Schrobe, H. 1983. Reasoning about Reasoning. In F. Hayes-Roth, D. A. Waterman, and D. B. Lenat eds. *Building Expert Systems,* 219-239. London: Addison-Wesley Publishing.

Levin, D. T. ed. 2004. *Thinking and Seeing*. Cambridge, MA: The MIT Press.

Levin, D. T., and Beck, M. R. 2004. Thinking about Seeing: Spanning the Difference between Metacognitive Failure and Success. In D. T. Levin ed. *Thinking and Seeing*, 121-143. Cambridge, MA: The MIT Press.

Lyons, W. 1986. *The Disappearance of Introspectio*n. Cambridge, MA: Bradford Books/MIT Press.

Maes, P. 1987a. Computational Reflection, Technical Report, 87-2. Ph.D. diss., Artificial Intelligence Laboratory, Vrije Universiteit Brussels, Belgium.

Maes, P. 1987b. Introspection in Knowledge Representation. In Du Boulay, B., Hogg, D., and Steels, L. eds. *Advances in Artificial Intelligence - II*, 249-262. Amsterdam: North-Holland.

Maes, P. 1988. Issues in Computational Reflection. In P. Maes and D. Nardi eds. *Meta-level Architectures and Reflection,* 21-35. Amsterdam: North Holland.

Maes, P., and Nardi, D. eds. 1988. *Meta-Level Architectures and Reflection.* Amsterdam: North-Holland.

Marshall, J. 1999. Metacat: A Self-Watching Cognitive Architecture for Analogy-Making and High-Level Perception, Ph.D. Dissertation, Indiana University, Bloomington.

Marshall, J. and Hofstadter, D. 1998. Making sense of analogies in Metacat. In K. Holyoak, D. Gentner, and B. Kokinov eds., *Advances in Analogy Research: Integration of Theory and Data from the Cognitive, Computational, and Neural Sciences*, Berlin: Springer.

Mason, C. 1994. Introspection as Control in Result-Sharing Assumption-Based Reasoning Agents. In *International Workshop on Distributed Artificial Intelligence*. Lake Quinalt, WA.

McCarthy, J. 1959. Programs with Common Sense. In *Symposium Proceedings on Mechanisation of Thought Processes* (Vol. 1), 77-84. London: Her Majesty's Stationary Office.

McCarthy, J. 1968. Programs with Common Sense. In M. L. Minsky ed. *Semantic Information Processing* 403-418. Cambridge, MA: MIT Press.

McCarthy, J. 1979. Ascribing Mental Qualities to Machines. In M. Ringle ed. *Philosophical Perspectives in Artificial Intelligence* 161-195. Atlantic Highlands, NJ: Humanities Press.

McCarthy, J. 1993. Notes on Formalizing Context. In R. Bajcsy ed. *Proceedings of the Thirteenth International*

Joint Conference on Artificial Intelligence (Vol. 1) 555-560. San Mateo, CA: Morgan Kaufmann.

McCarthy, J. 1995. Making Robots Conscious of Their Mental States. In M. T. Cox and M. Freed eds. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms,* 89-96. Menlo Park, CA: AAAI Press. (Available as Technical Report, SS-95-08)

McCarthy, J. (chair) and Chaudri, V. (co-chair). 2004. *DARPA Workshop on Self Aware Computer Systems*. SRI Headquarters, Arlington, VA, Apr. 27-28.

McCarthy, J., and Hayes, P. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence* 4: 463-502.

McCarthy, J., Minsky, M., Sloman, A., Gong, L., Lau, T., Morgenstern, L., Meuller, E., Riecken, D., Singh, M., and Singh, P. 2002. An Architecture of Diversity for Commonsense Reasoning. *IBM Systems Journal* 41(3): 530-539.

McNamara, T. P., Miller, D. L., and Bransford, J. D. 1991. Mental Models and Reading Comprehension. In R. Barr, M. L. Kamil, P. Mosenthal, and P. D. Pearson eds. *Handbook of Reading Research* (Vol. 2), 490- 511. New York: Longman.

Metcalfe, J. 2000. Metamemory: Theory and Data. In E. Tulving and F. I. M. Craik (Eds.), *The Oxford Handbook of Memory* (pp. 197-211). New York: Oxford University Press.

Metcalfe, J. 1998a. Cognitive Optimism: Self-Deception or Memory-Based Processing Heuristics? *Personality and Social Psychology Review* (Special Issue: Metacognition, J. Metcalfe ed.) *2*(2): 100-110.

Metcalfe, J. ed. 1998b. Special Issue: Metacognition. *Personality and Social Psychology Review*, 2(2).

Metzinger, T., and Chalmers, D. J. 1995. Selected Bibliography, Consciousness in Philosophy, Cognitive Science and Neuroscience: 1970-1995. Appendix I in T. Metzinger ed. *Conscious Experience*. Schoning, UK: Imprint Academic.

Miner, A. C., and Reder, L. M. 1994. A New Look at Feeling of Knowing: Its Metacognitive Role in Regulating Question Answering. In J. Metcalfe and A. P. Shimamura eds. *Metacognition: Knowing about knowing*, 47-70. Cambridge, MA: MIT Press/Bradford Books.

Minsky, M. L. 1963. Steps towards Artificial Intelligence. In E. A. Feigenbaum and J. Feldman eds. *Computers and Thought,* 406-450. New York: McGraw Hill. (Original work published 1961)

Minsky, M. L. 1965. Matter, Mind, and Models. In *Proceedings of the International Federation of Information Processing Congress 1965* (Vol. 1) 45-49.

Minsky, M. L. 1968a. Matter, Mind, and Models. In M. L. Minsky ed. *Semantic Information Processing*, 425-432. Cambridge, MA: MIT Press.

Minsky, M. L. ed. 1968b. *Semantic Information Processing*. Cambridge, MA: MIT Press.

Minsky, M. L. 1985. *The Society of Mind*. New York: Simon and Schuster.

Minsky, M., Singh, P., and Sloman, A. 2004. The St. Thomas Common Sense Symposium: Designing Architectures for Human-Level Intelligence. *AI Magazine*, Summer: 113-124.

Minton, S. 1988. *Learning Search Control Knowledge: A Explanation-Based Approach.* Boston: Kluwer Academic.

Mitchell, T. M., Allen, J., Chalasani, P., Cheng, J., Etzioni, O., Ringuette, M., and Schlimmer, J. C. 1991. Theo: A Framework for Self-Improving Systems. In K. VanLehn ed. *Architectures of Cognition: The 22nd Carnegie Mellon Symposium on Cognition* 323-355. Hillsdale, NJ: LEA.

Mitchell, T. M., Keller, R., and Kedar-Cabelli, S. 1986. Explanation-Based Generalization: A Unifying View, *Machine Learning* 1(1): 47-80.

Mooney, R., & Ourston, D. (1994). A multistrategy approach to theory refinement. In R. S. Michalski & G. Tecuci (Eds.), *Machine learning IV: A multistrategy approach* (pp. 141-164). San Francisco: Morgan Kaufmann.

Moore, R. C. 1985. Semantical Considerations on Nonmonotonic Logic. *Artificial Intelligence* 25(1): 75-94.

Moore, R. C. 1995. *Logic and Representation*. Stanford, CA: CSLI Publications.

Murdock, J. W. 1998. A Theory of Reflective Agent Evolution. Technical Report, GIT-CC-98-27. Ph.D. proposal, College of Computing, Georgia Institute of Technology, Atlanta.

Murdock, J. W., and Goel, A. 2001. Meta-Case-Based Reasoning: Using Functional Models to Adapt Case-Based Agents. In D. W. Aha, I. Watson, and Q. Yang eds., *Case-Based Reasoning Research and Development: Proceedings of the 4th. International Conference on Case-Based Reasoning, ICCBR-2001*, 407-421. Berlin: Springer

Nelson, T. O., and Narens, L. 1992. Metamemory: A Theoretical Framework and New Findings. In T. O. Nelson ed. *Metacognition: Core Readings*, 9-24. Boston: Allyn and Bacon. (Originally published in 1990.)

Newell, A. 1982. The Knowledge Level. *Artificial Intelligence* 18: 87-127.

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

Nilsson, N. 1980. *Principles of Artificial Intelligence*. Los Altos, CA: Morgan Kaufmann.

Nisbett, R. E. and Wilson, T. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84(3): 231-259.

Oehlmann, R. 2003. Metacognitive and Computational Aspects of Chance Discovery. *New Generation Computing*

21(1): 3-12.

Oehlmann, R., Edwards, P., and Sleeman, D. 1994. Changing the Viewpoint: Re-indexing by Introspective Questioning. In A. Ram and K. Eiselt eds. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* 675-680. Hillsdale, NJ: LEA.

Oehlmann, R., Edwards, P., and Sleeman, D. 1995. Introspection Planning: Representing Metacognitive Experience. In M. T. Cox and M. Freed eds. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms,* 102-110. Menlo Park, CA: AAAI Press. (Available as Technical Report, SS-95-08)

Owens, C. 1990. Indexing and Retrieving Abstract Planning Knowledge. Ph.D. diss., Department of Computer Science, Yale University, New Haven.

Perlis, D. 1985. Languages with Self-Reference I: Foundations. *Artificial Intelligence* 25: 301-322.

Perlis, D. 1988. Languages with Self-Reference II: Knowledge, Belief and Modality. *Artificial Intelligence* 34(2): 179-212.

Perlis, D. in press. Theory and Application of Self-Reference: Logic and Beyond. *CSLI*. To appear as chapter.

Pirolli, P., and Recker, M. 1994. Learning Strategies and Transfer in the Domain of Programming. *Cognition and Instruction*, *12*(3), 235-275.

Pollock, J. L. 1989a. *How to Build a Person.* Cambridge, MA: MIT Press/Bradford Books.

Pollock, J. L. 1989b. OSCAR: A General Theory of Rationality. *Journal of Experimental and Theoretical Artificial Intelligence* 1: 209-226

Pressley, M., and Forrest-Pressley, D. 1985. Questions and Children's Cognitive Processing. In A. C. Graesser and J. B. Black eds. *The Psychology of Questions*, 277-296. Hillsdale, NJ: LEA.

Raja, A. 2003. *Meta-Level Control in Multi-Agent Systems*. Ph.D. dissertation. Department of Computer Science. University of Massachusetts, Amherst, MA.

Raja, A., and Lessor, V. 2004. Meta-Level Reasoning in Deliberative Agents. In *Proceedings of the International Conference on Intelligent Agent Technology* (pp. 141-147). Piscataway, NJ: IEEE Computer Society.

Ram, A. 1993. Indexing, Elaboration and Refinement: Incremental Learning of Explanatory Cases. *Machine Learning* 10: 201-248.

Ram, A. 1994. AQUA: Questions That Drive the Understanding Process. In R. C. Schank, A. Kass, and C. K. Riesbeck eds. *Inside Case-Based Explanation* 207-261. Hillsdale, NJ: LEA.

Ram, A., and Cox, M. T. 1994. Introspective Reasoning Using Meta-Explanations for Multistrategy Learning. In R. S. Michalski and G. Tecuci eds. *Machine Learning: A Multistrategy Approach IV*, 349-377. San Mateo, CA: Morgan Kaufmann.

Ram, A., and Leake, D. 1995. Learning, Goals, and Learning Goals. In A. Ram and D. Leake eds. *Goal-Driven Learning,* 1-37. Cambridge, MA: MIT Press/Bradford Books.

Recker, M., & Pirolli, P. (1995). Modeling individual differences in student's learning. *The Journal of the Learning Sciences, 4*(1), 1- 38.

Reder, L. M., and Ritter, F. 1992. What Determines Initial Feeling of Knowing? Familiarity with Question Terms, Not with the Answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 18(3): 435-451.

Reder, L. M., and Schunn, C. D. 1996. Metacognition Does Not Imply Awareness: Strategy Choice Is Governed by Implicit Learning and Memory. In L. Reder ed. *Implicit Memory and Metacognition*, 45-77. Mahwah, NJ: LEA.

Rice, J. R. 1976. The Algorithm Selection Problem. *Advances in Computers 15*: 65-118.

Robertson, P. 2003. Confidence from Self Knowledge and Domain Knowledge. In *Self-Adaptive Software: Applications*. Berlin: Springer-Verlag LNCS 2614.

Rosenbloom, P. S., Laird, J. E., and Newell, A. 1989. Metalevels in SOAR. In P. Maes and D. Nardi eds. *Meta-Level Architectures and Reflection,* 227-240. Amsterdam: North-Holland.

Rosenbloom, P. S., Laird, J. E., and Newell, A. eds. 1993. *The Soar Papers: Research on Integrated Intelligence*. Cambridge, MA: MIT Press.

Russell, S. J. 1997. Rationality and Intelligence. *Artificial Intelligence 94*, 57-77.

Russell, S. J. 1999. Metareasoning. In Wilson, R. A., and Keil, F. C. eds. *The MIT Encyclopedia of the Cognitive Sciences (MITECS).* Bradford Books. Cambridge, MA: MIT Press. (Also available at http://cognet.mit.edu/)

Russell, S. J., and Subramanian, D. 1995. Provably Bounded-Optimal Agents. *Journal of Artificial Intelligence Research 2*: 575-609

Russell, S. J., and Wefald, E. 1991a. *Do the Right Thing: Studies in Limited Rationality*. Cambridge, MA: MIT Press.

Russell, S. J., and Wefald, E. 1991b. Principles of Metareasoning. *Artificial Intelligence 49:* 361-395.

Schank, R. C., Goldman, N., Rieger, C., and Riesbeck, C. K. 1972. *Primitive Concepts Underlying Verbs of Thought* (Stanford Artificial Intelligence Project Memo No. 162. Stanford, CA: Stanford University, Computer Science Department. (NTIS No. AD744634)

Schank, R. C., Kass, A., and Riesbeck, C. K. 1994. *Inside Case-Based Explanation*. Hillsdale, NJ: LEA.

Schneider, W. 1985. Developmental Trends in the Metamemory-Memory Behavior Relationship: An Integrative Review. In D. L. Forrest-Pressley, G. E. MacKinnon, and T. G. Waller eds. *Metacognition, Cognition and Human Performance*, Vol. 1 (Theoretical Perspectives), 57-109.

New York: Academic Press.

Schubert, L. K. 2005. Some KR&R Requirements for Self-Awareness. In M. Anderson and T. Oates eds., *Metacognition in Computation: Papers from 2005 AAAI Spring Symposium*, 106-113. Technical Report SS-05-04. Menlo Park, CA: AAAI Press.

Schut, M., and Wooldridge, M. 2001. Principles of Intention Reconsideration. *Proceedings of the 5th International Conference on Autonomous Agents*, 340-347. Montreal, Quebec, Canada.

Schut M. C., Wooldridge M. J., and Parsons S. D., 2004. The Theory and Practice of Intention Reconsideration. *Journal of Experimental and Theoretical Artificial Intelligence* 6(4): 261-293.

Schwanenflugel, P. J., Fabricius, W. V., Noyes, C. R., Bigler, K., D., and Alexander, J. M. 1994. The Organization of Mental Verbs and Folk Theories of Knowing. *Journal of Memory and Language, 33,* 376-395.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics 69*:99-118.

Simon, H. A. (1982). *Models of Bounded Rationality: Behavioral Economics and Business Organization*, vol. 2. Cambridge, MA: MIT Press.

Singh, P. 2005. *EM-ONE: An Architecture for Reflective Commonsense Thinking*. Ph.D. dissertation. Department of Electrical Engineering and Computer Science. Massachusetts Institute of Technology. Boston, MA.

Skinner, B. F. 1950. Are Theories of Learning Necessary? *Psychological Review* 57: 193-216.

Skinner, B. F. 1956. What is Psychotic Behavior? In F. Gildea ed. *Theory and Treatment of the Psychoses: Some Newer Aspects*. St. Louis: Washington University Press.

Sloman, A. 2001. Beyond Shallow Models of Emotion. *Cognitive Processing* 1(1).

Smith, B. C. 1985. Prologue to "Reflection and semantics in a procedural language." In R. J. Brachman and H. J. Levesque eds. *Readings in Knowledge Representation* 31-40. San Mateo, CA: Morgan Kaufmann. (Original work published 1982)

Stein, G. and Barnden, J. A. 1995. Towards More Flexible and Common-Sensical Reasoning about Beliefs. In M. T. Cox and M. Freed eds. *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms,* 127-135. Menlo Park, CA: AAAI Press. (Available as Technical Report, SS-95-08)

Stroulia, E. 1994. Failure-Driven Learning as Model-Based Self-Redesign. Ph.D. diss., College of Computing, Georgia Institute of Technology, Atlanta.

Stroulia, E. and Goel, A. K. 1995 Functional Representation and Reasoning in Reflective Systems, *Applied Intelligence*, Special Issue on Functional Reasoning 9(1): 101-124.

Sussman, G. J. 1975. *A Computer Model of Skill Acquisition*. New York: American Elsevier.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology, 82*(2), 306-314.

Tash, J. and Russell, S. 1994. Control Strategies for a Stochastic Planner. In *Proceedings of the Twelfth National Conference on Artificial Intelligence, II,* 1079-1085. Cambridge, MA: MIT Press.

Titchener, E. B. 1912. The Schema of Introspection. *The American Journal of Psychology* 23(4): 485-508.

VanLehn, K., Ball, W., and Kowalski, B. 1990. Explanation-Based Learning of Correctness: Towards a Model of the Self-Explanation Effect. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: LEA.

VanLehn, K., Jones, R. M., and Chi, M. T. H. 1992. A Model of the Self-Explanation Effect. *Journal of the Learning Sciences* 2(1): 1-60.

Veloso, M., and Carbonell, J. G. 1994. Case-Based Reasoning in PRODIGY. In R. S. Michalski and G. Tecuci eds. *Machine Learning IV: A Multistrategy Approach* 523-548. San Francisco: Morgan Kaufmann.

Veloso, M.,Carbonell, J. G., Perez, A., Borrajo, D.,Fink, E., and Blythe, J. 1995. Integrating Planning and Learning: The PRODIGY Architecture. *Journal of Theoretical and Experimental Artificial Intelligence* 7(1): 81-120.

Von Neumann, J., and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. New York: John Wiley and Sons.

Watson, J. B. 1919. *Psychology from the Standpoint of the Behaviorist*. Philadelphia: J. B. Lippincott.

Wellman, H. M. 1983. Metamemory Revisited. In M. T. H. Chi ed. *Contributions to Human Development*. Vol. 9 (Trends in memory development research). Basel, Switzerland: S. Karger, AG.

Wellman, H. M. 1985. The Origins of Metacognition. In D. L. Forrest-Pressley, G. E. MacKinnon, and T. G. Waller eds. *Metacognition, Cognition and Human Performance*. Vol. 1 Theoretical perspectives 1-31. New York: Academic Press.

Wellman, H. M. 1992. *The Child's Theory of Mind*. Cambridge, MA: MIT Press.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*(2), 181-192.

Yussen, S. R., ed. (1985). *The Growth of Reflection in Children*. New York: Academic Press.

Zhang, Z., and Yang, Q. 2001. Feature Weight Maintenance in Case Bases Using Introspective Learning. *Journal of Intelligent Information Systems*, 16: 95-116.

Zilberstein, S. 1993. Operational Rationality through Compilation of Anytime Algorithms. Ph.D. Dissertation, University of California at Berkeley.

Zilberstein, S., and Russell, S. J. 1996. Optimal Composition of Real-time Systems. *Artificial Intelligence 82*(1-2):181-213.