Anthropomorphic Self-Models for Metareasoning Agents

Andrew S. Gordon[1], Jerry R. Hobbs[2], and Michael T. Cox[3]

[1]Institute for Creative Technologies, University of Southern California, gordon@ict.usc.edu
[2]Information Sciences Institute, University of Southern California, hobbs@isi.edu
[3] Intelligent Computing, BBN Technologies, mcox@bbn.com

Abstract

Representations of an AI agent's mental states and processes are necessary to enable metareasoning, i.e. thinking about thinking. However, the formulation of suitable representations remains an outstanding AI research challenge, with no clear consensus on how to proceed. This paper outlines an approach involving the formulation of anthropomorphic self-models, where the representations that are used for metareasoning are based on formalizations of commonsense psychology. We describe two research activities that support this approach, the formalization of broad-coverage commonsense psychology theories and use of representations in the monitoring and control of object-level reasoning. We focus specifically on metareasoning about memory, but argue that anthropomorphic self-models support the development of integrated, reusable, broad-coverage representations for use in metareasoning systems.

Self-models in Metareasoning

Cox and Raja (2007) define reasoning as a decision cycle within an action-perception loop between the ground level (doing) and the object level (reasoning). Metareasoning is further defined as a second loop, where this reasoning is itself monitored and controlled in order to improve the quality of the reasoning decisions that are made (Figure 1).

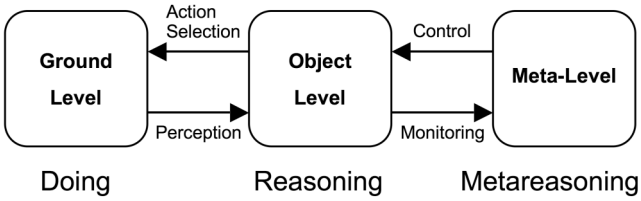

Figure 1. Multi-level model of reasoning

It has long been recognized (e.g., McCarthy, 1958) that to better understand and act upon the environment, an agent should have an explicit, declarative representation of the states and actions occurring in that environment. Thus the task at the object level is to create a declarative model of the world and to use such a representation to facilitate the selection of actions at the ground level. It follows also that to reason about other agents in the world (e.g., to anticipate what they may do in the future), it helps to have a representation of the agents in the world, what they know, and how they think. Likewise an explicit representation of the self supports reasoning about oneself and hence facilitates metareasoning. Representations provide structure and enable inference. They package together related assertions so that knowledge is organized and brought to bear effectively and efficiently.

One of the central concerns in the model of metareasoning as shown in Figure 1 is the character of the information that is passed between the object level and the meta-level reasoning modules to enable monitoring and control. Cast as a representation problem, the question becomes: How should an agent's own reasoning be represented to itself as it monitors and controls this reasoning? Cox and Raja (2007) describe these representations as *models of self*, which serve to control an agent's reasoning choices, represent the product of monitoring, and coordinate the self in social contexts.

Self-models have been periodically explored in previous AI research since Minsky (1968), and explicit self-models have been articulated for a diverse set of reasoning processes that include threat detection (Birnbaum *et al.*, 1990), case retrieval (Fox & Leake, 1995), and expectation management (Cox, 1997). Typically built to demonstrate a limited metareasoning capacity, these self-models have lacked several qualities that should be sought in future research in this area, including:

1. *Broad coverage*: Self-models should allow an agent to reason about and control the full breadth of their object-level reasoning processes.
2. *Integrated*: Self-models of different reasoning processes should be compatible with one another, allowing an agent to reason about and control the interaction between different reasoning subsystems.
3. *Reusable*: The formulation of self-models across different agents and agent architectures should have some commonalities that allow developers to apply previous research findings when building new systems.

Despite continuing interest in metareasoning over the last two decades (see Anderson & Oates, 2007; Cox, 2005), there has been only modest progress toward the development of self-models that achieve these desirable qualities. We speculate that this is due, in part, to an

emphasis on process rather than representation in the development of metareasoning systems. That is, researchers have tended to make only the representational commitments necessary to demonstrate the algorithmic approach that they advocate. As a predictable result, the collective set of representations across this field of research is modular, narrowly scoped, and specifically tied to particular agent architectures.

As in other areas of AI, a more balanced view of the relative contributions of process and representation may lead to new opportunities in this field that are currently obscured. Instead of avoiding representational commitments, we should encourage the development of systems that make these commitments in a principled manner. This paper describes an approach to representation in metareasoning, advocating principles to guide progress toward integrated, broad-coverage, reusable self-models.

Anthropomorphic Self-Models

One approach for achieving integrated, broad-coverage, reusable self-models for metareasoning is to try to mirror the sorts of representations that are used by people. To understand this approach, it is first necessary to recognize that people, too, have self-models that they employ to monitor and control their own reasoning. In the field of psychology, the model that people have of their own reasoning states and processes (as well as those of others) is commonly referred to as a Theory of Mind. The study of this model began in earnest with Fritz Heider (1958), and has received an enormous amount of theoretical attention over the last half century, cf. Smedslund (1997), particularly in the areas of developmental psychology (Wellman *et al*., 2001), cognitive anthropology (Lillard, 1998), and primate studies (Call & Tomasello, 1999). Although some philosophers have argued that a representational Theory of Mind is neither necessary (Goldman, 2006) nor beneficial (Churchland, 1986), process-oriented cognitive models of human first-person mind-reading (introspection) and third-person mind-reading (perspective-taking) have generally included a prominent role for explicit representations of mental state (e.g. Nichols & Stich, 2003).

The *anthropomorphism* approach to metareasoning representations is to formalize the self-model that people have of themselves and others, and utilize these representations to support monitoring and control in AI agents. In other words, rather than devising a new representational framework based on the functionality of the AI agents, we should identify and utilize a representational framework that is already successfully employed by billions of existing intelligent people.

Why Anthropomorphism?

The argument for pursuing an anthropomorphic approach to the representation of AI self-models is that people will be controlling, collaborating with, and designing these systems, and each of these activities will be facilitated if there are parallels that can be drawn between these AI self-models and the models that people use to think about themselves and others.

Parallelism between AI and human self-models is critical to enabling people to control these systems. As a strategy for managing the complexity of AI agents, the natural tendency of people will be to anthropomorphize these systems - seeing them as if they were people whose behavior is governed by a logic that parallels their own. Although unavoidable, adopting this intentional stance (Dennett, 1987) toward AI agents will only be fruitful if the constituents of this logic are grounded in the operation of the agent in some meaningful way. For example, consider the specific problem of interpreting natural language imperatives that a person might deliver to influence the metareasoning behavior of an AI agent, e.g. "Focus on what you are doing, Mr. Robot, and quit worrying about tomorrow's work!" People have specific expectations about how this directive should be executed by the AI agent, expectations that can only be met if there is something that parallels the concept of a *focus of attention*, among others, in the self-model of the AI agent.

The necessity of anthropomorphism in AI self-models is even more apparent when multi-agent systems consist of a heterogeneous mix of AI agents and people. Cox and Raja (2007) define *distributed metareasoning* as the coordination of problem solving contexts among agents in a multi-agent system, where the meta-control component of each agent should operate according to a multi-agent policy. Strategies for coordinating these problem-solving contexts are likely to be complex even if the agents (and their self-models) were homogenous. If these systems are a heterogeneous mix of people and AI agents, then each participant will need to be able to reason about the problem solving contexts of the others. Without a shared self-model or at least one that is compatible at a high level, the AI agents are faced with a much more difficult reasoning problem, and the humans are faced with an impossible one. If people in multi-agent collaborations are required to reason about the self-models of AI agents that are *different* than their own, then only the developers themselves will be able to participate.

For practical purposes, anthropomorphism also makes good engineering sense. Progress in the development of AI agents will continue to require the cooperative effort of large teams of researchers and developers. If these agents employ representational models of themselves that resemble those of their developers, then time and costs needed to understand and constructively contribute to these systems can be substantially reduced.

Formalizing Commonsense Psychology

Watt (1998) explored the relationship between anthropomorphism and Theory of Mind reasoning as it pertains to Artificial Intelligence systems, and argued for the central importance of commonsense psychology to understanding this type of reasoning. In artificial

intelligence, commonsense psychology has generally been approached in the same terms as commonsense (naïve) physics (Hayes, 1978), i.e. as a representational model (a set of logical axioms) that enables commonsense inference. However, instead of supporting commonsense inference about liquids, flow and physical force, this representational model attempts to reproduce the predictions and explanations that people make regarding the behavior of people based on their beliefs, goals, and plans. In this view, human anthropomorphic reasoning can be understood as the adoption of this representational model for predicting and explaining the behavior of non-human beings.

Commonsense reasoning about human psychology is *not* the same thing as metareasoning. Whereas commonsense reasoning about psychology serves the express purposes of prediction and explanation, metareasoning is specifically concerned with monitoring and control. The most successful optimization policies employed by an agent's metareasoning capabilities may be the product of calculations that are extremely different than those employed for logical inference. However, the anthropomorphism approach to metareasoning representations argues that the formalisms used for these two disparate functions should be roughly identical. That is, the *model of self* that is passed between reasoning and metareasoning modules in an agent should be expressed in the same vocabulary that drives commonsense psychological inference.

Hobbs and Gordon (2005) describe a large-scale effort to describe a model of commonsense psychology as a set of 30 content theories expressed in first-order logic. Aiming to achieve greater coverage over commonsense psychological concepts than in previous formalization projects, this work employed a novel methodology where the breadth of commonsense psychology concepts were first identified through a large-scale analysis of planning strategies, elaborated by an analysis of English words and phrases, and then formalized as axiomatic theories in which all of the identified concepts could be adequately defined.

Gordon and Hobbs (2003) presents one of the 30 content theories produced using this methodology, a commonsense theory of human memory. In the commonsense view, human memory concerns memories in the minds of people, which are operated upon by memory processes of storage, retrieval, memorization, reminding, and repression, among others. The formal theory of commonsense human memory presented by Gordon and Hobbs supports inferences about these processes with encodings of roughly three-dozen memory axioms in first-order logic. Key aspects of this theory can be characterized as follows:

1. Concepts in memory: People have minds with at least two parts, one where concepts are stored in memory and a second where concepts can be in the focus of one's attention. Storage and retrieval involve moving concepts from one part to the other.

2. Accessibility: Concepts that are in memory have varying degrees of accessibility, and there is some threshold of accessibility for concepts beyond which they cannot be retrieved into the focus of attention.

3. Associations: Concepts that are in memory may be associated with one another, and having a concept in the focus of attention increases the accessibility of the concepts with which it is associated.

4. Trying and succeeding: People can attempt mental actions (e.g. retrieving), but these actions may fail or be successful.

5. Remember and forget: Remembering can be defined as succeeding in retrieving a concept from memory, while forgetting is when a concept becomes inaccessible.

6. Remembering to do: A precondition for executing actions in a plan at a particular time is that a person remembers to do it, retrieving the action from memory before its execution.

7. Repressing: People repress concepts that they find unpleasant, causing these concepts to become inaccessible.

As an example of how these concepts are formalized, this theory defines memory retrieval as a change for a concept being in one's memory to being in one's focus of attention. This definition is reformulated below using the notation of eventualities (Hobbs, 1985) in the common logic interchange format.

```
(forall (a c)
   (iff (retrieve a c)
        (exists f m e1 e2)
           (and (agent a)
                (concept c)
                (focus f a)
                (memory m a)
                (change e1 e2)
                (inMemory' e1 c m)
                (inFocus' e2 c f)))))
```

Applied to the problem of representation in metareasoning, the formal commonsense theory of human memory provided by Gordon and Hobbs (2003) argues for representations of memory storage, retrieval, reminding, and repression, among other concepts. Although feasible, few argue that an agent's metareasoning functionality should be implemented as a logical theorem prover. Neither does the anthropomorphism approach to metareasoning representation take this route. Instead, the aim is to ensure that the representations used for monitoring and control of reasoning processes have some direct correspondence to the sorts of relations that appear in commonsense psychological theories. Specifically, the predicate relations that define mental states and processes in the theories should correspond to functionality that enable monitoring and control of reasoning processes.

Metareasoning About Memory in Agents

One of the hallmarks of human memory is that it is fallible; few among us can consistently remember everyone's

birthday, where we parked the car, or how many meters are in a mile. This fallibility is the reason that it is useful for people to engage in metareasoning about memory, which leads us to tie strings around our fingers, leave notes for ourselves, and schedule appointments using datebooks, among a suite of other memory-supporting strategies. It would be unfortunate if the hardware memory chips inside our computers were as problematic. However, when considering the utility of metareasoning about memory in software agents, these memory chips are not the central concern. Instead, it is useful to consider the broad set of agent functionality that can be viewed as analogous to human memory; anthropomorphically, if an agent had a set of memory functions, what would they be?

Standard database actions are perhaps the most straightforward analogs to human memory functions, in the commonsense view. Memory storage and retrieval can easily be viewed as the insertion and querying of records in database tables. Other commonsense memory concepts are analogous to the functionality of full-text search engines, where memory storage is accomplished through text indexing and remindings are analogous to ranking of documents based on their similarity to a text query. Conceivably, monitoring and control of these functions through metareasoning may optimize the performance of these software systems. However, the utility of anthropomorphic self-models in agent systems is most evident when these software systems employ the sorts of artificial intelligence reasoning techniques that are inspired by human cognitive function, e.g. planning, scheduling, prediction, explanation, monitoring, and execution.

From this perspective, the AI techniques that most directly align with commonsense models of human memory are those used to support Case-Based Reasoning (Aamodt & Plaza, 1994; Kolodner, 1993; Lopez de Mántaras *et al*, 2006). In this view, the case base is itself the agent's memory, the cases are its memories, case indexing is memory storage, and case retrieval is reminding. The use of commonsense concepts about human memory is one of the notable characteristics of the Meta-AQUA system (Cox & Ram, 1999), an implementation of a metareasoning agent that performs case-based explanation in a story-understanding task. Within the context of this research effort, explicit representations have been formulated for the majority of commonsense memory concepts identified that appear in Gordon and Hobbs's (2003) theory. In the following section, we describe how these representations are implemented in support of metareasoning within this type of case-based reasoning architecture.

Expectation-based Metareasoning

One of the most basic mental functions is to compare an agent's expectations with environmental feedback (or, alternatively, a "mental check" of the conclusions) to detect when the potential for improvement exists. The reasoner calculates some expected outcome (E) and compares it with the actual outcome (A) that constitutes the feedback. Figure 2 shows a graph structure where the comparison produces some relational state that links A and E, and Table 1 shows a declarative representation of the graph. When reasoning is successful, E is equal to A. When expectation failure occurs, R is equal to <>.

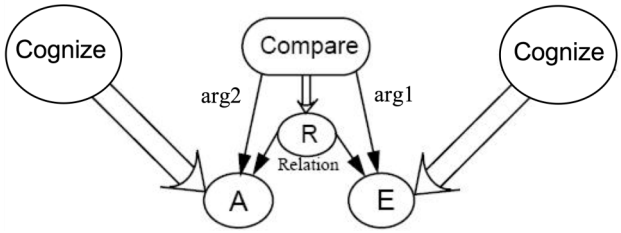

Figure 2. Basic comparison model.
A=actual; E=expected; R=relation

Table 1. Frame definition for comparison model

```
(define-frame basic-model
    (a (entity))
    (e (entity))
    (r (relation (domain =a)‡
                 (co-domain =e)))
    (mental-event (compare
                    (arg1 =e)(arg2 =a))
    (link1  (mentally-results
            (domain (cognize))
            (co-domain =e)))
    (link2  (mentally-results
            (domain (cognize))
            (co-domain =a)))
    (link3  (mentally-results
            (domain =mental-event)
            (co-domain =r))))
```

‡ *=X is a variable binding to property of name X.*
A relation r with domain d and co-domain c is equivalent to an RDF triple [subject predicate object] = [d r c]

A node labeled Cognize may be instantiated by any mental mechanism including inferential and memory processes. Figure 3 shows a representation for a successful memory retrieval where the value of the right Cognize node is a memory process. This representation captures the distinctions between an incidental reminding, a deliberate recall, and recognition. The structural differences depend on the nodes C and G, and the temporal order of the causal links resulting in nodes E and A (see Table 2). If there is no *knowledge goal* (Ram, 1991; Cox & Ram, 1999) to retrieve some memory item, only cues in the environment, and if E is retrieved before A is produced, then the structure is a reminding. On the other hand, if there is a deliberate attempt to retrieve a memory item that is later compared to some feedback, A, then the structure represents recall. Finally, if A is presented followed by a memory probe, then the structure represents recognition, whether or not a retrieval goal exists. It is also significant to note that memory "elaboration" can be represented as a feedback loop in Figure 3 from E to C such that each new

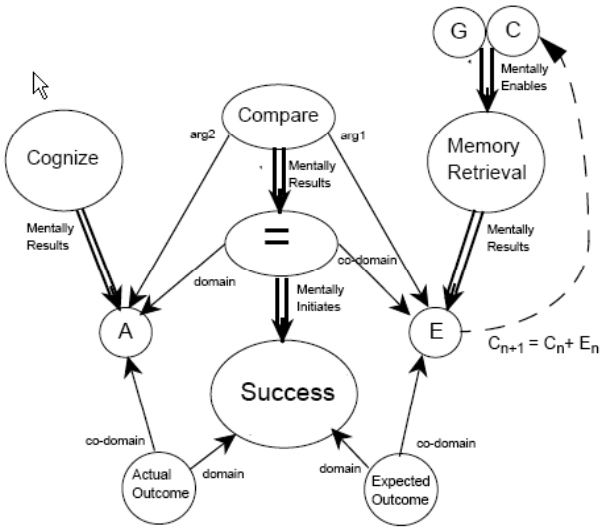

Figure 3. Representation of various remembering events. A=actual; E=expected; R=relation

item retrieved adds to the context that enables further memory retrieval.

The kinds of representations above are useful when reasoning (or memory) fails. It is at this point that a trace of the prior reasoning in a declarative form can be passed to a metareasoner for inspection. If the system is to learn from its mistakes and improve its performance, it needs to consider what happens or what does not happen at the object level. The object level can pass to the meta-level a representation of reasoning (i.e., introspective monitoring), the metareasoner can infer what went wrong (i.e., metareasoning), and then it can pass some direction back to an object level learner to change its knowledge or the reasoning mechanism (i.e., meta-level control). Consider what happens if memory fails.

The structure of Figure 4 represents a memory retrieval attempt enabled by goal, G, and cues, C, that tried to retrieve some memory object, M, given an index, I, that did not result in an expectation (or interpretation), E, that should have been equal to some actual item, A. The fact that E is out of the set of beliefs with respect to the reasoner's foreground knowledge (FK), that is, is not present in working memory, initiates the knowledge that a

Table 2. Structural differences between remembering events in Figure 3

Memory Term	Structural Features	Description
Reminding	Has only Cues; E before A	Incidental; No Knowledge Goal
Recall	Cues and Goal; E before A	Deliberate; Has Knowledge Goal
Recognition	May or may not have Goal; A before E	Borderline between items above; Has judgment

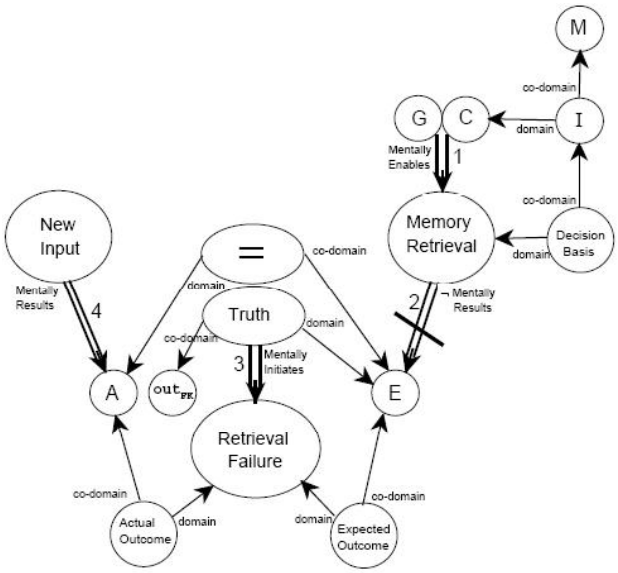

Figure 4. Representation of forgetting. A=actual; E=expected; G=goal; M=memory item; I=memory index.

retrieval failure had occurred.

This representation captures an entire class of memory failures: failure due to a missing index, I; failure due to a missing object, M; failure because of a missing retrieval goal, G (the agent never attempted to remember); or failure due to not attending to the proper cues, C, in the environment. Such a representation allows the system to reason about these various causes of forgetting; it can inspect the structural representation for a memory failure and therefore analyze the reasons for the memory failure. Such an ability facilitates learning because it allows a learner to explain the reasoning failure and to use the result in determining what needs to be learned and so avoid the failure in the future (Cox & Ram, 1999).

Discussion

Research toward the development of an effective metareasoning component for agents and agent systems has been slow, producing a modest number of prototype systems designed to demonstrate the utility of a particular metareasoning approach. Much of this work has been successful by focusing on process rather than representation, by making only the representational commitments necessary to support metareasoning within the context of a given agent architecture and task. As a consequence, the collective set of representations used in these systems has none of the characteristics that are needed to enable this technology to move forward: representations that have broad coverage, are integrated across object-level reasoning subsystems, and are reusable across different agents and agent architectures. In this paper we have argued for making representational commitments in metareasoning systems in a principled

manner, i.e. through the development of anthropomorphic self-models.

The representational approach that we describe in this paper involves two distinct research activities. First, formal theories of commonsense psychology are developed using a combination of empirical, analytical, and traditional knowledge engineering techniques. Our efforts in this task (Hobbs & Gordon, 2003) have aimed to develop theories that have both broad coverage of commonsense psychology concepts (breadth), and have the competency to draw commonsense inferences in support of automated prediction and explanation (depth). Second, the commonsense concepts that appear in these theories are explicitly represented for use in metareasoning in agent systems. Our efforts in this task (Cox, 2007; 1997) have advanced a comparison-based approach, where the expected outcome of object-level reasoning behavior is compared with the actual reasoning outcomes that are observed.

In this paper, we describe how these two research activities relate to each other in support of metareasoning about memory. The formalization of commonsense concepts of memory help us identify the breadth of concepts that will be passed between object-level and meta-level reasoning components in support of monitoring and control. The comparison-based implementation of metareasoning demonstrates that representations at this level of abstraction can be effectively grounded for use in real agent systems. Although these two activities were pursued completely independently by the coauthors of this paper, we see that closer coordination of these efforts in the future offer a principled approach to developing integrated, reusable, broad-coverage representations for metareasoning systems.

First, anthropomorphic self-models can achieve broad coverage by hitting the right level of representational abstraction. Commonsense psychological concepts like *reminding* and *forgetting* are general enough that they can be easily aligned with a large number of disparate object-level software functions when agents are viewed from an anthropomorphic perspective. Conversely, these concepts are specific enough to provide the meta-level reasoning component of an agent enough information and control to diagnose and correct problems as they arise.

Second, anthropomorphic self-models can achieve the goal of integrated representations by working to mirror the integrated coherence of human commonsense psychological models. Although not without its inconsistencies, commonsense psychology is remarkable in that it allows people to predict and explain behavior by drawing coherent connections between a wide variety of mental states and processes. It allows us, for example, to tell a coherent story about how forgetting something can result in an incorrect prediction about a world state during the execution of a plan, and how the failure to achieve the goal of the plan subsequently leads to emotional feelings of guilt. The ease in which people effortlessly reason *about* memory, prediction, execution, goal management, and emotion in an integrated manner should serve as inspiration for the representations used in metareasoning agents. By deriving their representational commitments from commonsense psychology, anthropomorphic self-models aim to enable this level of integration as more and more of these object-level reasoning functions are included in AI-based agent systems in the future.

Third, anthropomorphic self-models can achieve the goal of reusable representations, where the content of these representations is not inextricably tied to one particular agent implementation. Representational commitments are made not at the level of software, but rather to the conceptual framework that is used to characterize an agent's reasoning functions. By standardizing the representations used in metareasoning systems around this framework, we can begin to conceptualize metareasoning systems that are interchangeable across agent architectures and tasks. This would, in turn, enable some form of comparison and competition between different approaches, and would allow developers to apply validated research findings when building new metareasoning systems.

Acknowledgments

The project or effort described here has been sponsored, in part, by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Aamodt, A., and Plaza, E. 1994. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications* 7(1): 39-59.

Anderson, M., and Oates, T. 2007. A Review of Recent Research in Metareasoning and Metalearning. *AI Magazine* 28(1):7-16.

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

Call, J., and Tomasello, M. 1999. A Nonverbal False Belief Task: The performance of children and great apes. *Child Development* 70(2):381-395.

Churchland, P. 1986. *Neurophilosophy*. Cambridge, Mass.: Bradford Books/MIT Press.

Cox, M. T. 1997. An Explicit Representation of Reasoning Failures. 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. 2005 Metacognition in Computation: A Selected Research Review. *Artificial Intelligence* 169(2):104-141.

Cox, M. T. 2007. Perpetual Self-Aware Cognitive Agents. *AI Magazine* 28(1): 32-45.

Cox, M. T., and Raja, A. 2007. Metareasoning: A Manifesto. BBN Technical Memo, BBN TM-2028. http://www.mcox.org/Metareasoning/Manifesto

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

Dennett, D. 1987. *The Intentional Stance*, Cambridge, Mass.: MIT Press

Fox, S., and Leake, D. 1995 Using Introspective Reasoning to Refine Indexing. Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

Goldman, A. 2006. *Simulating Minds: The Philosophy Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

Gordon, A., and Hobbs, J. 2003. Coverage and Competency in Formal Theories: A Commonsense Theory of Memory. *Proceedings of the 2003 AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, Stanford University, March 24-26, 2003.

Hayes, P. 1978. The Naive Physics Manifesto, *Expert Systems in the Microelectronic Age*, D. Michie, Ed. Edinburgh, Scotland: Edinburgh University Press, pp. 242-270.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. New York: Wiley.

Hobbs, J. 1985. Ontological Promiscuity, *Proceedings, 23rd Annual Meeting of the Association for Computational Linguistics*, 61-69. Chicago, Illinois.

Hobbs, J., and Gordon, A. 2005. Encoding Knowledge of Commonsense Psychology. 7th International Symposium on Logical Formalizations of Commonsense Reasoning. May 22-24, 2005, Corfu, Greece.

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

Lillard, A. 1998. Enthopsychologies: Cultural Variations in Theories of Mind. *Psychological Bulletin* 123(1): 3-32.

Lopez de Mántaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., Faltings, B., Maher, M. L., Cox, M. T., Forbus, K., Keane, M., Aamodt, A., & Watson, I. 2006. Retrieval, reuse and retention in case-based reasoning. *Knowledge Engineering Review* 20(3), 215-240.

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

Minsky, M. L. 1968. Matter, Mind, and Models. M. L. Minsky (Ed.), *Semantic information processing,* 425-432. Cambridge, MA: MIT Press.

Nichols, S., and Stich, S. 2003. Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding Other Minds. Oxford: Clarendon Press.

Ram, A 1991. A Theory of Questions and Question Asking. *Journal of the Learning Sciences* 1(3&4): 273-318.

Smedslund, J. 1997. *The Structure of Psychological Common Sense*. Mahwah, NJ: Lawrence Erlbaum.

Watt, S. 1998. Seeing this as people: Anthropomorphism and common-sense psychology. PhD Thesis, Open University.

Wellman, H.M., Cross, D., and Watson, J. 2001. Meta-analysis of theory of mind development: The truth about false-belief. *Child Development* 72: 655–684.