To appear in Cox, M. T., & Raja, A. (in press). *Metareasoning: Thinking about thinking*. Cambridge, MA: MIT Press.

Chapter 1 Metareasoning: An introduction

Michael T. Cox and Anita Raja

Philosophers and cognitive scientists of many persuasions have long wondered what is unique to human intelligence. Although a number of ideas have been proposed, a common differentiator appears to be a pervasive capacity for thinking about ourselves in terms of who we are, how others see us, and in terms of where we have been and where we want to go. As humans, we continually think about ourselves and our strengths and weaknesses in order to manage both the private and public worlds within which we exist. But the Artificial Intelligence community has not only wondered about these phenomena; it has attempted to implement actual machines that mimic, simulate, and perhaps even replicate this same type of reasoning called metareasoning.

The term is an overloaded one, and no single consensus exists as to its definition. Some have described metareasoning computationally in terms of specific programs and algorithms; whereas others have analyzed metacognition and focused on data from human experience and behavior. Indeed Ann Brown (1987) described research into metacognition as a "many-headed monster of obscure parentage." Many of the technical terms used in research on metareasoning and

related areas are quite confusing. Often, authors use different terms for the same concept (e.g., introspection and reflection), and sometimes the same terms are used in different ways (e.g., metareasoning has been cast as both process and object). The literature contains many related topics such as metaknowledge, metamemory, self-adaptation, and self-awareness. The index in the back of this book demonstrates the complexity of the subject by its length. So the main goal of this book is to assemble some measure of consistency and soundness in the topic.

To attempt to achieve progress toward this goal we have written a very brief summary of some existing research and put forth a simple, abstract model of metareasoning. We then asked numerous scientific researchers on the subject to address our "manifesto" by describing the relationship between their research and this model. The task is to compare and contrast separate theories and implementations to this sketch of what lies at the core of metareasoning. This model certainly has some weaknesses. The method of abstraction leaves out various details that may prove critical to a more in-depth understanding of the mechanisms behind the process. We also recognize that metareasoning is a much larger umbrella under which many related topics such as metaknowledge lie. Yet by going through this exercise, we hope that the reader and the researcher will both gain a deeper insight into the knowledge structures and computation involved.

Metareasoning: A manifesto

The 21st century is experiencing a renewed interest in an old idea within artificial intelligence that goes to the heart of what it means to be both human and intelligent. This idea is that much can be gained by thinking about one's own thinking. Traditionally within cognitive science and artificial intelligence, thinking or *reasoning* has been cast as a decision cycle within an action-perception loop similar to that shown in

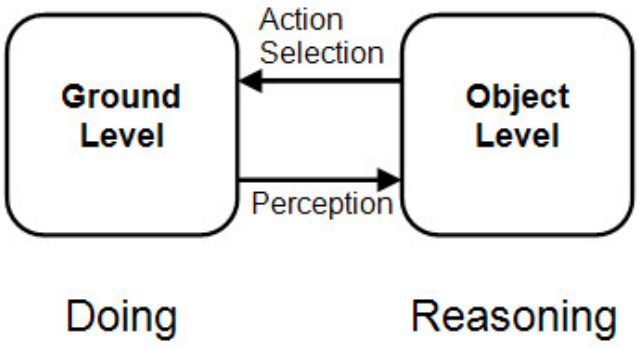

Figure 1.1. An intelligent agent perceives some stimuli from the environment and behaves rationally to achieve its goals by selecting some action from its set of competencies. The result of these actions at the ground level is subsequently perceived at the object level and the cycle continues. *Metareasoning* is the process of reasoning about this reasoning cycle. It consists of both the meta-level control of computational activities and the introspective monitoring of reasoning (see Figure 1.2). This cyclical arrangement represents a higher-level reflection of

the standard action-perception cycle, and as such, it represents the perception of reasoning and its control.

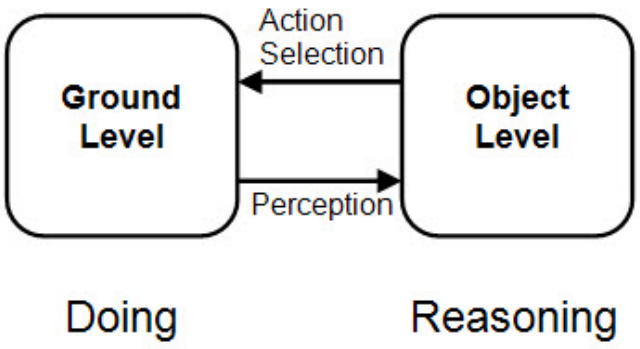

Figure 1.1. The action-perception cycle

The goal of *meta-level control* is to improve the quality of its decisions by spending some effort to decide what and how much reasoning to do as opposed to what actions to do. It balances resources between object level actions (computations) and ground level actions (behaviors). But while meta-level control allows agents to dynamically adapt their object level computation, it could interfere with ground level performance. Thus identifying the decision points that require meta-level control is of importance to the performance of agents operating in resource-bounded environments.

Introspective monitoring is necessary to gather sufficient information with which to make effective meta-level control decisions. Monitoring may involve the

gathering of computational performance data so as to build a profile of various decision algorithms. It could involve generating explanations for object-level choices and their effect on ground level performance. When reasoning fails at some task, it may involve the explanation of the causal contributions of failure and the diagnosis of the object-level reasoning process.

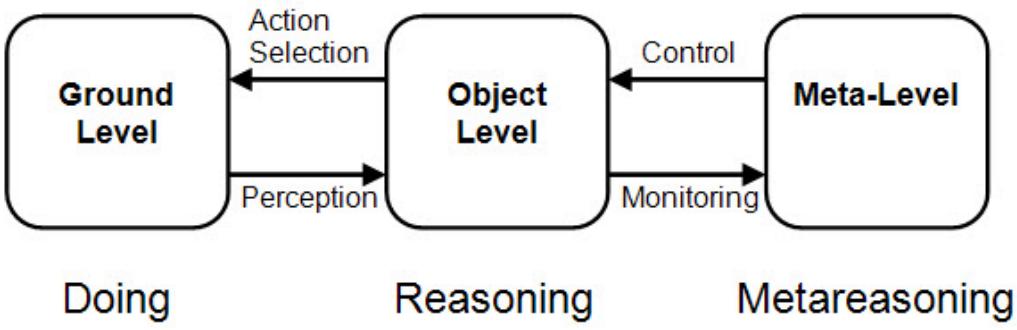

Figure 1.2. Duality in reasoning and acting

Under the banner of *distributed metareasoning*, significant research questions also exist concerning the extent to which meta-level control and monitoring affects multi-agent activity. In multi-agent systems, where the quality of joint decisions affects individual outcomes, the value obtained by an agent exploring some portion of its decision space can be dependent upon the degree to which other agents are exploring complementary parts of their spaces. The problem of coordinated meta-level control refers to this question of how agents should coordinate their strategies to maximize the value of their joint actions.

Finally any complete cognitive system that reasons about itself and its actions in the world will necessarily combine many aspects of metareasoning. A truly intelligent agent will have some conception of self that controls its reasoning choices, represents the products of monitoring, and coordinates the self in social contexts. Hence a comprehensive approach will include *models of self* in support of metareasoning and integrated cognition.

Meta-Level Control

A significant research history exists with respect to metareasoning (Anderson & Oates, 2007; Cox, 2005), and much of it is driven by the problems of limited rationality. That is because of the size of the problem space, the limitations on resources, and the amount of uncertainty in the environment, finite agents can often obtain only approximate solutions. So for example with an anytime algorithm that incrementally refines plans, an agent must choose between executing the current plan or further deliberation with the hope of improving the plan. When making this choice, the agent is reasoning about its own reasoning (i.e., planning) as well as its potential actions in the world (i.e., the plan). As such this represents the problem of explicit control of reasoning.

Figure 1.2 illustrates the control side of reasoning along its upper portion. Reasoning controls action at the ground level in the environment; whereas metareasoning controls the reasoning at the object level. For an anytime

controller, metareasoning decides when reasoning is sufficient and thus action can proceed. Although other themes exist within the metareasoning tradition (e.g., Leake, 1996), this characterization is a common one (e.g., Raja & Lesser, 2007; Hansen & Zilberstein, 2001; Russell & Welfald, 1991).

Now consider Figure 1.3. The most basic decision in classical metareasoning is whether an agent should act or continue to reason. For example the anytime planner always has a current best plan produced by the object level reasoning. Given that the passage of time itself has a cost, the metareasoner must decide whether the expected benefit gained by planning further outweighs the cost of doing nothing. If so it produces another plan; otherwise it executes the actions in the plan it already has. Note that this simple decision can be performed without reference to any perception of the ground level. Of course many more sophisticated meta-level control policies exist that include feedback.

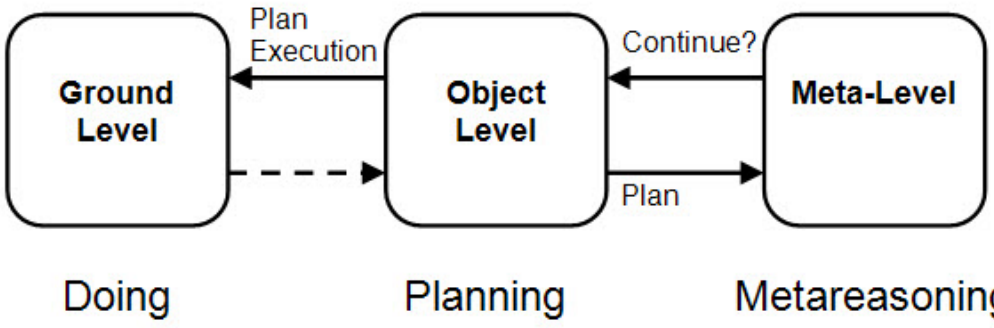

Figure 1.3. Classical metareasoning (from Russell & Wefald, 1991)

Introspective Monitoring

The complementary side of metareasoning is less well studied. The introspective monitoring of reasoning about performance requires an agent to maintain some kind of internal feedback in addition to perception, so that it can perform effectively and can evaluate the results of metareasoning. For instance Zilberstein (Zilberstein & Russell, 1996) maintains statistical profiles of past metareasoning choices and the associated performance and uses them to mediate the subsequent control and dynamic composition of reasoning processes.

But introspective monitoring can be even more explicit. If the reasoning that is performed at the object level (and not just its results) is represented in a declarative knowledge structure that captures the mental states and decision-making sequence, then these knowledge structures can themselves be passed to the meta-level for monitoring. For example the Meta-AQUA system (Cox & Ram, 1999) keeps a trace of its story understanding decisions in structures called a Trace Meta-eXplanation Pattern (TMXP). Here the object-level story understanding task is to explain anomalous or unusual events in a ground-level story perceived by the system (see Figure 1.4).[1] Then if this explanation process fails, Meta-AQUA passes the TMXP and the current story representation to a learning subsystem. The learner performs an introspection of the trace to obtain an explanation of the explanation failure called an Introspective Meta-

eXplanation Pattern (IMXP). The IMXPs are used to generate a set of learning goals that are passed back to control the object-level learning and hence improve subsequent understanding. TMXPs explain *how* reasoning occurs; IMXPs explain *why* reasoning fails.

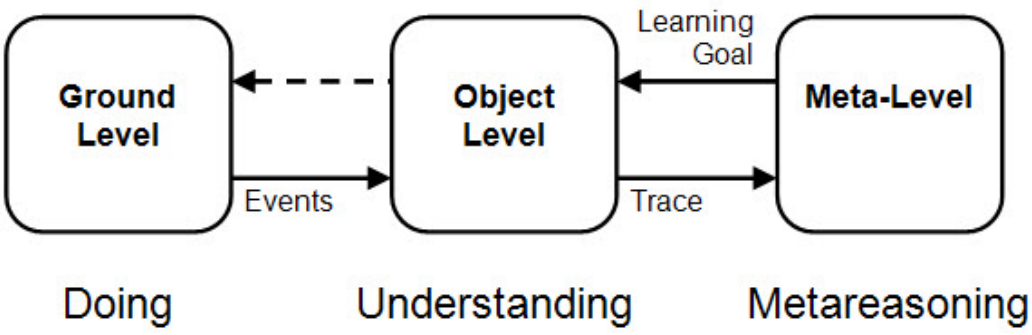

Figure 1.4. Introspective monitoring in Meta-AQUA

Note that the object-level process described above is a story understanding task without reference to the execution of personal actions at the ground level. The emphasis here is upon the perception and monitoring side of the model; that is, the understanding or comprehension processes in the model are equally as important as the action and control processes were in Figure 1.3, and indeed they can be treated independently. However most systems, especially agent-based systems, combine both in various fashions.

Distributed Metareasoning

In a multi-agent context, if two or more agents need to coordinate their actions, the agents' meta-control components must be on the same page. The agents must reason about the same problem and may need to be at the same stage of the problem-solving process. For example, suppose one agent decides to devote little time to communication/negotiation (Alexander, Raja, Durfee, & Musliner, 2007) before moving to other deliberative decisions while another agent sets aside a large portion of deliberation time for negotiation; the latter agent would waste time trying to negotiate with an unwilling partner.

We define an agent's problem solving context as the information required for deliberative-level decision making, including the agent's current goals, action choices, its past and current performance, resource usage, dependence on other agents, etc. Suppose the agent's context when it is in the midst of execution is called the current context, while a pending context is one where an agent deliberates about various what-if scenarios related to coordination with other agents. Distributed metareasoning can also be viewed as a coordination of problem solving contexts. One meta-level control issue would be to decide when to complete deliberation in a pending context and when to replace the current context with the pending context. Thus if an agent changes the problem solving context on which it is focused, it must notify other agents with which it may

interact. This suggests that the meta-control component of each agent should have a multi-agent policy where the content and timing of deliberations are choreographed carefully and include branches to account for what could happen as deliberation (and execution) plays out. Figure 1.5 describes the interaction among the meta-level control components of multiple agents.

Another meta-control question when there are multiple pending contexts is to determine which pending context should be allocated resources for deliberation. In all of these examples, the metareasoning issues are a superset of single agent cases.

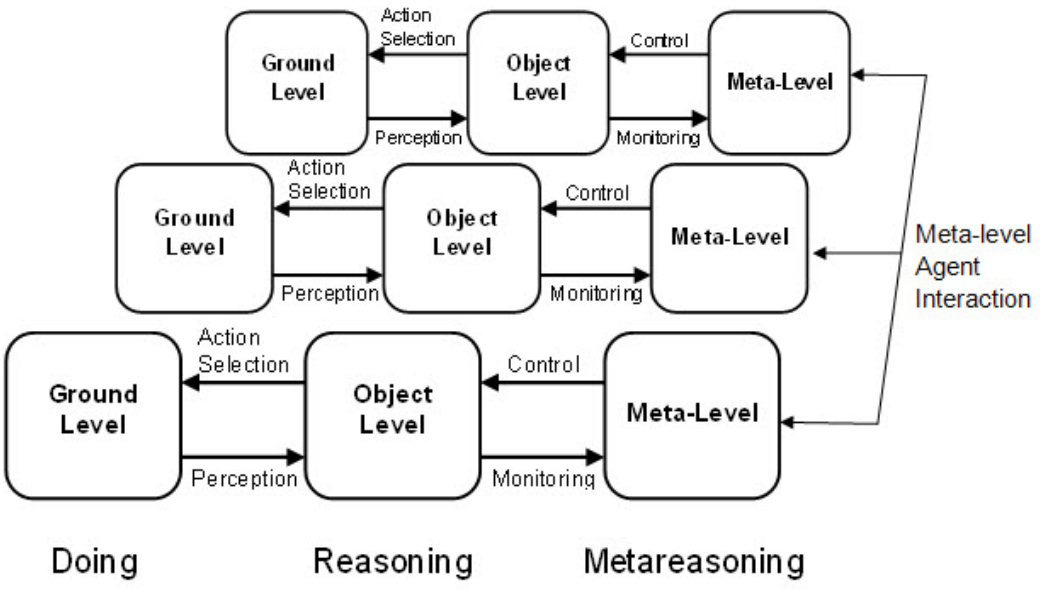

Figure 1.5. Meta-level reasoning among multiple agents

Models of Self

For a cognitive agent to behave intelligently in a physical and social environment with complex, dynamic interactions, many if not all of the features necessary for an integrated human-level model of intelligence are required. For it to succeed in such an environment, an agent must perceive and interpret events in the world including actions of other agents, and it must perform complex actions and interact in a social context. These constitute the minimal object level requirements. At the meta-level, an agent must have a model of itself to represent the products of experience and to mediate the choices effectively at the object level. Facing novel situations the successful agent must learn from experience and create new strategies based upon its self-perceived strengths and weaknesses. Consider Figure 1.6.

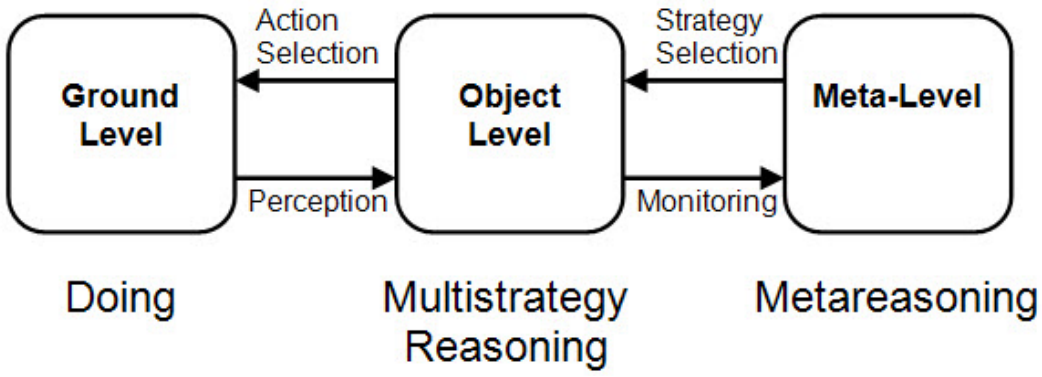

Figure 1.6. Example integrated model of self

Monitoring at the meta-level can determine the kinds of mental actions at which the agent excels and those it fails. Using such introspective information allows the agent to choose reasoning strategies that best fit future intellectual demands like the agent that selects actions based on past task performance. In more complicated approaches, the agent may actually construct a complex reasoning strategy rather than simply choose an atomic one. In either case, the basis for such metareasoning comes from a picture of itself, its capacities (both physical and mental), and its relationships to other agents with which it must interact to recognize and solve problems.

Many theorists have speculated as to the interactions between levels of representation and process (i.e., the architecture), but few researchers have attempted to implement the full spectrum of computation implied in a comprehensive model of self (see Singh, 2005, for one such attempt). However we challenge the AI community to consider seriously the problems of metareasoning in this larger context. How would an agent best understand itself and use such insight to construct a deliberate knowledge-level reasoning policy? Can an agent know enough about itself and its colleagues' self-knowledge to communicate its meta-level needs for coordination? Can it estimate the time it might take to negotiate a coordination policy with its fellow agents and hence negotiate the time and length of a negotiation session? Finally could an intelligent soccer agent decide that it is good at planning but getting weak at passing and so

aspire to becoming a coach? We claim that the model of acting, reasoning, and metareasoning put forth in this document can help maintain clarity if this challenge is to be embraced and answering questions like these pursued.

Conclusion

This manifesto has tried to present in plain language and simple diagrams a brief description of a model of metareasoning that mirrors the action-selection and perception cycle in first-order reasoning. Many theories and implementations are covered by this model including those concerning meta-level control, introspective monitoring, distributed metareasoning, and models of self. We claim that it is flexible enough to include all of these metacognitive activities, yet simple enough to be quite parsimonious. Figures 1.3 through 1.6 and their accompanying examples suggest some variations on the potential implementations rather than dictate an agenda. We offer the model as a framework to which the community can compare and contrast individual theories, but most of all, we hope that this model can clarify our thinking about thinking about thinking.

Overview

Now each chapter considers this model at some level of detail. Starting with this chapter the first section sets the stage by providing some of the fundamental themes within this book. Perlis (Chapter 2) notes the ubiquity of self-reference

within the metareasoning literature (e.g., the previous sentence) and argues that reference in general has at its core a concept that is at the heart of what it means for an object to refer to itself. Zilberstein (Chapter 3) examines several approaches to building rational agents and the extent to which they rely on metareasoning. He demonstrates the application of an optimal metareasoning approach using anytime algorithms and discusses its relationships with the other approaches to bounded rationality. The rest of the book follows the structure of the manifesto and is divided into four parts: Part II on Meta-level Control; Part III on Introspective Monitoring; Part IV on Distributed Metareasoning and Part V on Models of Self.

In examining Meta-level Control in Part II, Epstein and Petrovic (Chapter 4) employ metareasoning to manage large bodies of heuristics and to learn to make decisions more effectively. Their approach gauges the program's skill within a class of problems and determines when learning for a class is complete and whether it has to be restarted. Alexander, Raja and Musliner (Chapter 5) discuss their efforts to add metalevel control to a Markov Decision Process-based deliberative agent. The agent uses heuristic guidance to incrementally expand its considered state space and solve the resulting MDP. Kim, Meyers, Gervasio and Gil (Chapter 6) describe a metalevel framework for coordinating different agents using explicit learning goals. By supporting both top-down and bottom-up control strategies, the framework enables flexible interaction among learners and is

shown to be effective for coordinating learning agents to acquire complex process knowledge for a medical logistics domain. Robertson and Laddaga (Chapter 7) discuss metareasoning in an image interpretation architecture called GRAVA where the goal is to produce good image interpretations under a wide range of environmental conditions. The section concludes with Conitzer's (Chapter 8) discussion on how to formulate variants of the metareasoning problem as formal computational problems. He also presents the implications of the computational complexity of these problems.

In exploring Introspective Monitoring in Part III, Cox (Chapter 9) examines the role of self-modifying code, self-knowledge, self-understanding, and self-explanation as aspects of self from a computational stance. Goel and Jones (Chapter 10) describe the use of meta-knowledge for structural credit assignment in a classification hierarchy when the classifier makes an incorrect prediction. They present a scheme in which the semantics of the intermediate abstractions in the classification hierarchy are grounded in percepts in the world and show that this scheme enables self-diagnosis and self-repair of knowledge contents at intermediate nodes in the hierarchy. Arcos, Mulayim and Leake (Chapter 11) present an introspective model for autonomously improving the performance of CBR systems. To achieve this goal, the model reasons about problem solving failures by monitoring the reasoning process, determining the causes of the failures, and performing actions that will improve future reasoning

processes. Schmill, et al. (Chapter 12) describe the Meta-Cognitive Loop (MCL), a human-inspired meta-cognitive approach to dealing with failures in automated systems behavior. MCL attempts to improve robustness in cognitive systems in a domain-general way by offering a plug-in reasoning component that will help decrease the brittleness of AI systems.

In Part IV on Distributed Metareasoning, Raja et al. (Chapter 13) present a generalized meta-level control framework for multi-agent systems and discuss the issues involved in extending single-agent meta-level control to a team of cooperative agents requiring coordination. They present a methodology for constructing a class of MDPs that can model the interactions necessary for coordinating meta-level control among multiple agents. Rubinstein, Smith and Zimmerman (Chapter 14) consider the role of metareasoning in achieving effective coordination among multiple agents that maintain and execute joint plans in an uncertain environment. They identify several degrees of freedom in configuring the agent's core computational components, each of which affects the proportion of computational cycles given to local scheduling and inter-agent coordination processes. They also motivate the need for on-line reasoning by considering how aspects of the current control state impact the utility of different configurations. Kennedy (Chapter 15) presents a distributed metareasoning architecture for a single cognitive agent where the meta-level and object-level components form a non-hierarchical network in which the meta-levels mutually

monitor and protect each other. She argues that coordination among meta-levels can also allow the agent to explain itself in a coherent way. Borghetti and Gini (Chapter 16) present a metareasoning system that relies on a prediction performance measurement and propose a novel model performance measurement called Weighted Prediction Divergence that fulfills this need.

In Part V, several approaches to building models of self are presented. Morbini and Schubert (Chapter 17) highlight the importance of meta-reasoning for self-aware agents and discuss some key requirements of human-like self-awareness including using a highly expressive representation language for the formalization of meta-level axioms. Hart and Scassellati (Chapter 18) discuss an approach to building rich models of the sensory and kinematic structure of robots and examine tasks to which such models may be applied. Here the task is for a robot to recognize itself in a mirror. Gordon et al. (Chapter 19) describe anthropomorphic self-models as an alternative approach to current approaches. They argue that developing integrated, broad-coverage, reusable self-models for metareasoning can be achieved by formalizing the commonsense theories that people have about their own human psychology.

In the concluding chapter, Sloman (Chapter 20) surveys varieties of meta-cognition and draws attention to some types that appear to play a role in

intelligent biological individuals (e.g. humans) and which could also help with practical engineering goals.

Conclusion

The goal of this book is to present a comprehensive narrative that incorporates an integrated set of chapters on various themes pertaining to metareasoning from both artificial intelligence and cognitive science perspectives. It includes concepts from research on multiagent systems, planning and scheduling technology, learning, case-based reasoning, control theory, logic programming, autonomic computing, self-adaptive systems, and cognitive psychology. We hope the reader will find that the model described in the manifesto operates as a central theme that supports a larger narrative. The manifesto is intended to be a shared organizational framework to which each author compares and contrasts their theory, results, and implementational details. For the most part, the authors have found this to be a useful abstraction. In the end, we hope that the reader will as well.

Acknowledgments

The views, opinions, and findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the

Department of Defense. This document has been approved for public release by DARPA for unlimited distribution.

References

Alexander, G., Raja, A., Durfee E., & Musliner, D. (2007). Design paradigms for meta-control in multi-agent systems In A. Raja & M. T. Cox (Eds.), *Proceedings of the First International Workshop on Metareasoning in Agent-based Systems* (pp. 92-103). AAMAS-07.

Anderson, M. L., & Oates, T. (2007). A review of recent research in metareasoning and metalearning. *AI Magazine, 28*(1), 7-16.

Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65-116). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cox, M. T. (2005). Metacognition in computation: A selected research review. *Artificial Intelligence, 169*(2), 104-141.

Cox, M. T., & Ram, A. (1999). Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence, 112*, 1-55.

Hansen, E., & Zilberstein, S. (2001). Monitoring and control of anytime algorithms: A dynamic programming approach. *Artificial Intelligence, 126*(1-2), 139-157.

Leake, D. B. (1996). Experience, introspection, and expertise: Learning to refine the case-based reasoning process. *Journal of Experimental and Theoretical Artificial Intelligence, 8*(3), 319-339.

Raja, A., & Lesser, V. (2007). A framework for meta-level control in multi-agent systems. *Autonomous Agents and Multi-Agent Systems, 15*(2), 147-196.

Russell, S. J., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence, 49*, 361-395.

Singh, P. (2005). *EM-ONE: An architecture for reflective commonsense thinking*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, Boston.

Zilberstein, S., & Russell, S. J. (1996). Optimal composition of real-time systems. *Artificial Intelligence, 82*(1-2), 181-213.

Endnotes

[1] Meta-AQUA does no action at the ground level. Rather it perceives events representing characters in the story doing actions.