Extending Introspective Learning from Self-Models

David Leake and Mark Wilson
Computer Science Department
Indiana University
Lindley Hall 215, 150 S. Woodlawn Avenue
Bloomington, IN 47405, U.S.A.
leake@cs.indiana.edu, mw54@cs.indiana.edu

Abstract

This position paper presents open issues for using self-models to guide introspective learning, focusing on five key types of areas to explore: (1) broadening the range of learning focuses and the range of learning tools which may be brought to bear, (2) learning for self-understanding as well as self-repair, (3) making model-based approaches more sensitive to processing characteristics, instead of only outcomes, (4) making model application more flexible and robust, and (5) increasing support for self-explanation and user interaction with the meta-level.

Introduction

Research on introspective reasoning has a long history in AI, psychology, and cognitive science (for an overview, see (Cox 2005)). One of the intriguing focuses of this work has been on *using self-models for introspective learning*—using explicit representation of and reasoning about internal processes to guide refinement of the system itself. In introspective learning approaches, a system exploits explicit representations of its own organization and desired behavior to determine when, what, and how to learn in order to improve its own reasoning.

This position paper aims to encourage reflection by the introspective reasoning community on how model-based introspective learning has been pursued and can be advanced. The paper begins by summarizing some dimensions of introspective learning, with illustrations from sample systems. With this background, it presents directions for increasing the robustness, effectiveness and flexibility of introspective learners.

The Nature of Introspective Learning

Cox and Raja (2008) present a general characterization of metareasoning, summarized diagrammatically in Figure 1. In their view, metareasoning includes both monitoring of object level reasoning processes (the reasoning processes performed in standard reasoning systems) and meta-level control of those reasoning processes. Just as the object level system perceives objects and events at the ground level and performs actions to affect the ground level, the metareasoning process monitors the object level and exerts control by applying actions to the object level. Metareasoning has been extensively studied (see (Cox 2005) and (Anderson & Oates 2007) for recent overviews.

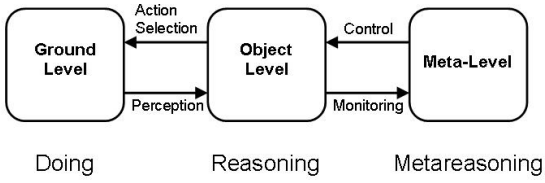

Figure 1: Duality in Reasoning and Acting (Cox & Raja 2008)

A rich current of metareasoning research addresses issues of bounded rationality, e.g., for anytime controllers (Raja & Lesser 2007; Hansen & Zilberstein 2001). Another current focuses on the use of general reasoning methods, applied to explicit models of desired system behavior, to support monitoring and learning by providing a standard against which to compare system performance to detect and diagnose failures and guide their repair (e.g., (Birnbaum *et al.* 1991; Fox & Leake 2001)).

Dimensions of Introspective Learning

Introspective reasoning systems may be characterized by a number of properties defining when, what, and how they learn. As background, we illustrate some of these properties. The following section considers opportunities for extending model-based introspective learning along these and other dimensions.

Focus of Learning

Introspective reasoning systems often learn in response to failures, focusing learning according to goals associated with particular failure types. For example:

- Meta-AQUA (Ram & Cox 1994), a story understander, learns from expectation failures. The system generates knowledge goals such as modifying indices, acquiring new knowledge, and re-organizing hierarchies in its knowledge base.

- ROBBIE (Fox & Leake 2001), a case-based route planner, includes a meta-reasoner which recognizes object reasoning failures, traces their root causes and, using declarative knowledge of the object-level domain, updates the case base's indexing scheme to repair the object reasoner.

Form of Meta-Knowledge

Model-based meta-reasoners frequently use static models of object-level reasoning or other pre-defined knowledge sources to guide introspective learning. For example:

- ROBBIE uses declarative structures to represent multiple levels of self-modeling, allowing it to describe ideal object-level reasoning and connect domain-level aspects for use in diagnosing and repairing expectation failures.
- Meta-AQUA uses two types of meta-level explanation patterns: Trace Meta-XPs, which package information about the system's goal-directed object reasoning, and Introspective Meta-XPs, which package information on why conclusions fail and how to learn from them by forming knowledge goals and choosing appropriate algorithms to satisfy them.
- DIAL (Leake, Kinley, & Wilson 1997a), a case-based planner, learns cases reflecting how to perform adaptations in its object-level system, which are also the basis for learning new similarity criteria.
- REM (Murdock & A.Goel 2008), uses adaptation of qualitative functional models to redesign case-based reasoning systems.

Monitoring Methods

Introspective learners commonly build a solution, or a partial solution, and then allow a meta-reasoner to examine the results and make corrections if necessary:

- Meta-AQUA attempts to explain events in stories and build representations for them. It triggers meta-reasoning if this task fails.
- ROBBIE checks a trace of its object-level case-based process against assertions characterizing an ideal process. If these fail, the meta-reasoner traces through the model for failures which identify possible faults.
- SPARK (Morley & Myers 2004), an agent framework based on a belief-desire-intention model, provides meta-predicates which may be used to monitor task execution, and to trigger meta-procedures, which can override certain of the system's default behaviors. The system also provides predicates and actions which provide introspective access to the current task-execution structure.
- Autognostic (Stroulia & Goel 1995) uses structure-behavior-function patterns to model processes with a "road map" of object-level problem solving. Each subtask can be checked for correct output, and the meta-reasoner can assign blame as it notices incorrect results, or after performing diagnosis on a complete reasoning trace.

Opportunities and Challenges

The previous examples illustrate valuable steps but also suggest a set of fundamental issues for further exploration, which we sketch in the following sections.

Reasoning About Failure Detection and Response

In model-based introspective reasoners, the model provides a gold standard against which performance can be compared, to detect failures. However, applying this apparently simple process may not be straightforward, and provides a new target for introspective learning. For example, to detect failure to retrieve the proper case, ROBBIE relies on the built-in strategy of performing an additional retrieval *after* a problem is successfully solved, using information about its solution to provide additional indices. This works well in ROBBIE's domain, but other strategies might be more appropriate in other situations. Likewise, once a failure has been detected, it may not be obvious whether to learn from it immediately, whether to wait to gather more information, or whether to forgo action (e.g., for an isolated exception). Consequently, managing failure detection and response may require introspective reasoning and learning in its own right.

Flexible Learning Focus

A large body of introspective learning research focuses on the use of introspection with a narrow focus, to monitor and repair a specific portion of the system. However, in introspective reasoners for rich object-level systems, what to repair to address a problem may not be straightforward. For example, case-based reasoning systems typically solve problems by generating indices from new problems, using those indices to guide retrieval of candidate stored cases for similar prior problem-solving, performing similarity assessment to select the most relevant prior case, and adapting that case's solution to fit new needs (Mantaras *et al.* 2005).

A poor solution might be ascribed to defects in many aspects of the process or to the knowledge it involves. For example, the problem might be repairable by adding a case, revising index generation or similarity assessment, or refining case adaptation knowledge. How to address this is still an open question. Focusing on a single aspect (e.g., indexing in ROBBIE) may miss opportunities.

The alternative strategy of simply attempting to perform all possible repairs is problematic as well. In addition to the potential costs, (Leake, Kinley, & Wilson 1997b) show that independent uncoordinated repairs of multiple knowledge sources may degrade performance. Consequently, related repairs will need to be coordinated. This coordination process may be seen as related to the problem of coordinated distributed metareasoning (Cox & Raja 2008), in that it may require individual introspective learning processes to coordinate their models and processes. One possible approach to such coordination would be case-based, in the spirit of meta-XPs (Cox & Ram 1999), using introspective reasoning cases to package coordinated combinations of learning actions.

Enabling Multistrategy Introspective Learning

Just as many introspective learning systems focus on a particular class of repair, introspective systems often rely on a

single learning method. However, the usefulness of multistrategy learning methods for object-level learning (e.g., (Michalski & Tecuci 1994)) suggests the value of exploring the spectrum of methods which could be applied to introspective learning. For example, changes to general indexing knowledge in a case-based reasoning system might be augmented with learned information about specific exception cases for which standard indices fail.

Learning for Self-Understanding in Addition to Self-Repair

When a self-model of desired behavior is used to guide learning, it is expected that the system's behavior may conflict with the desired behavior: The conflicts guide learning to improve the performance system.

However, there is no guarantee that the system will be able to repair all problems. Thus the ability of a system to predict its own performance—its limitations—may be crucially important. If the introspective system can choose between alternative methods, it needs criteria for understanding when those methods may fail, given its own characteristics, in order to anticipate and avoid failures. Efforts to achieve this can draw on numerous methods developed at the object level, such as the explanation-based generation of anticipation rules (Hammond 1989) (if the causes of the failures can be explained) or case-based methods if they cannot.

Monitoring Processing Characteristics in Addition to Outcomes

Despite the attention to resource issues in much metareasoning research, there has been little overlap between work focusing on model-based system repair and work focusing on metamanagement of processing resources. Consequently, an interesting question is how to integrate the two approaches. Such considerations may be especially important in distributed scenarios, for which object-level and introspective processes may run on different hardware and methods are needed to reason about—and handle—network and process interruptions. Such issues have begun to receive attention in the context of self-managing grid applications (Parashar *et al.* 2006), but primarily through the study of specific architectures, rather than of reasoning models.

Enabling More Flexible Use of Self-Models

Adjustable Modeling Levels Self-models often capture high-level descriptions of idealized processing, in order to increase generality by capturing domain-independent characteristics of the reasoning process. This approach provides generality, but removes the ability to apply model-based reasoning to lower-level subprocesses. Likewise, it may be challenging to connect the model to domain- and implementation-specific details. Consequently, an interesting question is how to flexibly support the operationalization of abstract models for specific systems and domains, and how a system can choose for itself the operationalization strategies and level of modeling, given particular goals.

Extending Models and Handling Imperfect Self-Models Approaches based on self-models often assume that the system's self-model is perfect. However, just as it may be difficult to pre-determine all the needs of the object-level system and to encode them perfectly—motivating the use of introspective learning—it may be difficult to do so at the meta-level. This raises questions for how a system might construct or extend a self-model and how it might refine a self-model which is partial or flawed. Blame assignment for focusing repair of flawed self-models might be guided by learned or externally-provided information on levels of trust for different components.

Likewise, when self-models or repair strategies may be imperfect, reasoners must predict whether the repairs suggested by introspection will really be beneficial. One approach is to monitor the introspective learning process itself, to determine whether candidate changes should be retained (Arcos, Mulayim, & Leake 2008).

Supporting Self-Explanation

The Cognitive Science literature supports the value of self-explanation in human learning (VanLehn, Jones, & Chi 1992), and some evidence suggests that human experts have greater awareness of their own problem-solving process than those with less expertise (e.g., (Chi *et al.* 1989)). However, the functional role of self-explanation in humans is poorly understood, as is its potential for improving the performance of AI systems. Cox (2007) argues for self-explanation as a path to self-awareness; this is an open area which may have profound ramifications for introspective reasoning. A particular challenge is to characterize what makes a good self-explanation: to understand the system's information needs for learning, how to transform them into *explanation goals* (Leake 1992), and how the system can generate the right explanations to satisfy those goals.

Exploiting Interaction at the Meta Level

A final opportunity concerns broadening the scope of introspective systems. Standard views of introspective systems, as exemplified by Figure 1, model the introspective process in terms of an agent whose interaction with the external world is entirely at the ground level. However, if an introspective system can explain its behaviors—behaviors either at the object or meta levels—and can communicate them, it becomes possible for them to accept guidance from a human as an external source. In the short term, such a strategy enables fielding introspective systems which can profit from the strengths of both the human and automated system, as well as providing an additional source of feedback (and possibly interactive guidance on new procedures), assisting in the process of diagnosis, in building a shared model, and in user acceptance (Cox 2007). We have begun to explore such an approach in a system which refines its reasoning in light of user feedback concerning confidence judgments by its monitoring component (Arcos, Mulayim, & Leake 2008). However, fully addressing such issues raises a host of issues for explanation generation and understanding.

Likewise, addressing such issues requires examining how interacting agents can best build and refine models of each

other's internal reasoning processes.

Conclusions

This position paper has considered some fundamental aspects of introspective reasoning and opportunities for extending approaches in which a self-model guides the learning process. It has presented a range of challenges falling into five main categories. The first is to broaden the range of learning focuses and methods, to examine how systems can simultaneously refine multiple aspects of their behavior by multistrategy methods. The second is to relax the assumption that the goal of learning is always to achieve perfect performance, instead—when perfect performance is impossible—focusing on self-understanding of system capabilities, to enable the system to make the best use of its abilities. The third is to make model-based approaches more sensitive to processing characteristics, instead of only outcomes, helping to combine the flexibility of the model-based approach with the performance advantages of methods aimed at tuning performance. The fourth is making model application more flexible and robust, relaxing assumptions for a perfect model or a single level of description. The fifth is to increase support for self-explanation and external explanation, for richer user interaction with the meta-level, helping a user to understand and help facilitate the system's introspective reasoning tasks. Fully addressing these challenges will require broad effort across the introspective reasoning community, to leverage the results along many dimensions into a cohesive whole.

Acknowledgment

This work is supported in part by the National Science Foundation under grant OCI-0721674.

References

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

Arcos, J.-L.; Mulayim, O.; and Leake, D. 2008. Using introspective reasoning to improve CBR system performance. In *Proceedings of the AAAI 2008 Workshop on Metareasoning: Thinking About Thinking*. In press.

Birnbaum, L.; Collins, G.; Brand, M.; Freed, M.; Krulwich, B.; and Pryor, L. 1991. A model-based approach to the construction of adaptive case-based planning systems. In *Proceedings of the DARPA Case-Based Reasoning Workshop*, 215–224. San Mateo: Morgan Kaufmann.

Chi, M.; Bassok, M.; Lewis, M.; Reimann, P.; and Glaser, R. 1989. Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science* 13:145–182.

Cox, M., and Raja, A. 2008. Metareasoning: A manifesto. Technical Report, BBN TM-2028, BBN Technologies. www.mcox.org/Metareasoning/Manifesto.

Cox, M., and Ram, A. 1999. Introspective multistrategy learning: On the construction of learning strategies. *Artificial Intelligence* 112(1-2):1–55.

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

Cox, M. 2007. Metareasoning, monitoring, and self-explanation. In *Proceedings of the First International Workshop on Metareasoning in Agent-based Systems, AAMAS-07*, 46–60.

Fox, S., and Leake, D. 2001. Introspective reasoning for index refinement in case-based reasoning. *JETAI* 13(1):63–88.

Hammond, K. 1989. *Case-Based Planning: Viewing Planning as a Memory Task*. San Diego: Academic Press.

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

Leake, D.; Kinley, A.; and Wilson, D. 1997a. A case study of case-based CBR. In *Proceedings of ICCBR 1997*, 371–382. Berlin: Springer Verlag.

Leake, D.; Kinley, A.; and Wilson, D. 1997b. Learning to integrate multiple knowledge sources for case-based reasoning. In *Proceedings of IJCAI 1997*, 246–251. Morgan Kaufmann.

Leake, D. 1992. *Evaluating Explanations: A Content Theory*. Hillsdale, NJ: Lawrence Erlbaum.

Mantaras, R.; McSherry, D.; Bridge, D.; Leake, D.; Smyth, B.; Craw, S.; Faltings, B.; Maher, M.; Cox, M.; Forbus, K.; Keane, M.; Aamodt, A.; and Watson, I. 2005. Retrieval, reuse, revision, and retention in CBR. *Knowledge Engineering Review* 20(3).

Michalski, R., and Tecuci, G. 1994. *Machine Learning: A Multistrategy Approach*. Morgan Kaufmann.

Morley, D., and Myers, K. 2004. The SPARK agent framework. In *Proceedings of AAMAS-04*, 712–719.

Murdock, J., and A.Goel. 2008. Meta-case-based reasoning: self-improvement through self-understanding. *JETAI* 20(1):1–36.

Parashar, M.; Liu, H.; Li, Z.; Matossian, V.; Schmidt, C.; Zhang, G.; and Hariri, S. 2006. Automate: Enabling autonomic grid applications. *Cluster Computing: The Journal of Networks, Software Tools, and Applications* 9(2):161–174.

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

Ram, A., and Cox, M. 1994. Introspective reasoning using meta-explanations for multistrategy learning. In Michalski, R., and Tecuci, G., eds., *Machine Learning: A Multistrategy Approach*. Morgan Kaufmann. 349–377.

Stroulia, E., and Goel, A. 1995. Functional representation and reasoning in reflective systems. *Applied Artificial Intelligence* 9(1):101–124.

VanLehn, K.; Jones, R.; and Chi, M. 1992. A model of the self-explanation effect. *The Journal of the Learning Sciences* 2(1):1–59.