Proceedings of the AAAI Spring Symposium on Representing Mental States and Mechanisms, Stanford, CA, March 1995.

Representing Self-knowledge for Introspection about Memory Search

David B. Leake
Computer Science Department
Indiana University
Bloomington, Indiana 47405
leake@cs.indiana.edu

Abstract

This position paper sketches a framework for modeling introspective reasoning and discusses the relevance of that framework for modeling introspective reasoning about memory search. It argues that effective and flexible memory processing in rich memories should be built on five types of explicitly represented self-knowledge: knowledge about information needs, relationships between different types of information, expectations for the actual behavior of the information search process, desires for its ideal behavior, and representations of how those expectations and desires relate to its actual performance. This approach to modeling memory search is both an illustration of general principles for modeling introspective reasoning and a step towards addressing the problem of how a reasoner—human or machine—can acquire knowledge about the properties of its own knowledge base.

Introduction

When agents perform everyday tasks, they must frequently draw conclusions from uncertain or incomplete information. In addition, their conclusions must be formed using limited processing resources. Consequently, a fundamental question in developing intelligent agents is how to reason effectively under these constraints. The answer to that question depends on a number of pragmatic considerations: the goals driving processing and their precedences, the types of reasoning results that are useful for furthering those goals, the reasoning methods available for achieving those results, and the relative merits of alternative reasoning methods. In addition, because both tasks and environment may change in unpredictable ways, the long-term effectiveness of the reasoning process depends on being able to refine processing strategies. Performing that refinement depends on the capability to represent and reason about, and to learn from, the course of processing.

This position paper sketches a framework for characterizing the introspective reasoning process and discusses some of the types of knowledge needed for introspective reasoning. It illustrates its principles with points from research on applying introspective reasoning to an internal processing task seldom modeled as involving introspection or even deliberation: the information retrieval or "memory search" process. I argue that effective and flexible memory processing in complex domains should be built on introspectively accessible self-knowledge representing information needs, relationships between different types of information, expectations for the *actual* behavior of the information search process, desires for its *ideal* behavior, and representations of how those expectations and desires relate to its actual performance. The framework I describe is both an illustration of general principles for modeling introspective reasoning and a step towards addressing the problem of how a reasoner—human or machine—can acquire knowledge about the properties of its own knowledge base.

A Planful Framework for Internal Reasoning

The approach I take to the problem of modeling mental processes is one that has proven productive in a number of other investigations of introspective reasoning and learning: treating internal processing as a rational, deliberative process driven by explicit goals, subgoals, and plans for achieving them (e.g., Cox & Freed (1994), Freed & Collins (1994), Hunter (1990), Leake & Ram (1993), Oehlmann & Edwards (1995), Ram & Cox (1994)). In traditional research on planning, this type of framework has been used as a starting point for reasoning about tasks, means, and expectations for results of external actions in the physical world. In the context of introspective reasoning, it can be used as the basis for analogous reasoning about tasks, means, and expectations for the effects of "mental" operations. The motivation for applying a planning-based model to the reasoning process itself is twofold: to increase the flexibility of reasoning processes to solve new types of problems and to serve as a basis for analyzing and refining the conduct of internal reasoning.

In order to support such a framework, it is necessary to represent knowledge about possible internal reasoning tasks, the internal reasoning plans they engender, the expectations that are relevant to the reasoning process, and the relationships of these processes. The following sections both illustrate the approach and substantiate its usefulness for the task of memory search.

The Memory Search Problem

One crucial influence on the outcome of any reasoning process is the knowledge available to that process. When an agent with limited computational resources retrieves information from a very large knowledge base, it is infeasible for it to consider the relevance of *all* its stored beliefs. Consequently, the organization of beliefs and how they are accessed play a key role in processing: the effects of the agent's knowledge on its behavior depend on which of its beliefs are actually examined during a particular reasoning episode.

It is widely recognized that reasoning and decision-making play a crucial role in gathering information from the external world when performing tasks such as diagnosis. Less acknowledged, however, is the role of such reasoning in gathering information in the internal or "mental" sphere. Instead, the process used to search a reasoner's memory is often treated as ancillary to the reasoning process itself. The sophistication of the memory models used ranges from simple databases to memories with refined memory organization and retrieval schemes (e.g., Burke (1993), Domeshek (1992), Leake (1991, 1992), Owens (1991), Schank (1982)), but regardless of the subtlety of the approach, the methods are almost always applied by opaque procedures. (Notable exceptions include Cox (1994), Kennedy (1995), Kolodner (1984), and Rissland, Skalak, & Friedman (1994)). Psychological studies, however, show that people have knowledge about the contents of their memories and are able to draw conclusions both about what they know and what they do not know (e.g., Gentner & Collins (1981)).

My position is that memory search should be treated *as a reasoning task* on equal footing with other reasoning tasks. In such a model, introspective reasoning about how to search memory guides retrieval and enables learning to refine the effectiveness of the memory search process.

Representational Requirements for Introspective Reasoning

Introspective reasoning requires information about the tasks, operators, and performance characteristics of the reasoning system. The following sections sketch how general constructs for representing these types of self-knowledge relate to reasoning about the memory search process.

Representing the memory search task: Characterizing knowledge goals

Being able to reason about memory search depends on representing the particular type of information sought in a form accessible to introspective reasoning processes. In the planful model of introspective reasoning, needs for information are characterized as explicit *knowledge goals* (Hunter, 1990; Ram, 1987; Leake & Ram, 1993).

A crucial question for such models is how to represent knowledge goals. In previous representations of knowledge goals, the sought-after information has been represented with a *concept specification* (Ram, 1987). Concept specifications reflect the type of information that is often represented in queries to memory, such as "find another episode of a similar car breakdown." This information is crucial to guiding memory search, but I propose that another constraint is needed as well in order to judge the performance of the memory search process and to guide learning. This constraint, which I will call a *comparative specification,* describes the desired relationship of the retrieved information to other alternative information in memory.

Some sort of comparative specification is usually implicit in the implementation of a memory search process. For example, in case-based reasoning systems the implicit comparative specification might be that the retrieved case should be the one sharing the most features with the current situation. However, explicitly representing comparative specifications and reasoning about how to achieve them, rather than building them into the retrieval process, has the advantage of making it possible to use the same basic memory search framework to support retrievals that require different comparative specifications. For example, Burstein (1994) observes that recency may take precedence over similarity in contexts such as retrieving cases for real-estate appraisal, and Bain (1989) suggests that recency may also take precedence when judges use cases to sentence criminals. Making comparative constraints such as recency explicit obviates the need to redesign the retrieval process for different retrieval tasks.

Representing knowledge to select reasoning operators: Capturing information relationships with contentful memory links

In order to reason about how to find information in memory, a memory search system must have self-knowledge about the organization of its knowledge. This includes information about how particular types of information are indexed and related in memory.

Such information is particularly important for retrieving information that does not correspond to links explicitly precoded in memory. To locate such information, a memory search procedure may need to follow multiple related links and to perform multiple retrievals based on information that it finds incrementally. For example, consider the memory search problem of finding the location where someone works, supposing that that information is not explicitly pointed to by a memory link such as "business-address." To find the address, it may be necessary to first follow links to find the employee's home address, in order to find the employee's community, and to follow other links to find information about the person's background or interests, in order to find a likely type of career for the person. Additional links can then be followed to find candidate corporations in the community and to accumulate evidence about which one applies. Finally, additional links must be followed from the selected corporation to its divisions and their locations. By combining that information, it is possible to suggest a likely business address.

Most memory models, however, do not provide the information to support the needed search process. Instead, links between concepts in memory are simply atoms such as "employer" that have no meaning to the memory search system itself. As a result, although that link can be followed when "employer" information is requested in a direct query, it is impossible to reason about how to follow related links to gather new types of information. On the other hand, when self-knowledge includes explicit information about the meaning of memory links, it enables memory search to satisfy knowledge goals that were not anticipated when a memory was originally organized.

In order to support reasoning about memory search, a reasoner needs access to a general description of basic domain-independent relationships in memory, such as relationships involving abstraction and inheritance (Wilensky, 1986). It also needs more domain-specific information about particular types of memory links, reflecting constraints on role-fillers and the relationships that define the meaning of a link itself. These representations are the subject of an ongoing investigation begun in Leake (1994).

Representing self-knowledge of system performance: Modeling expected and desired performance of memory search

Explicit representations of the memory search process itself can also enable a reasoner to introspectively analyze and refine its memory search process. In order to perform that reasoning and learning, a reasoner needs two distinct models of those reasoning processes. The first is a model of the *expected behavior* of the process, which can be used to predict the performance of the reasoning system itself. This model reflects known strengths and weaknesses of the reasoning process, making it possible for the agent to guide its processing by reasoning about the types of processing strategies likely to be effective in particular circumstances. (E.g., for memory search this could represent information about search paths useful for satisfying particular types of knowledge goals.) This model can be augmented by storing representations of individual search episodes and their outcomes in particular classes of situations, making it possible to apply case-based reasoning to predicting the effects of memory search and to learn new memory search strategies.

The second model is of the *ideal behavior* of the process, including not only the desired *outcome* but also aspects of desired performance of the reasoning *process* as well—for example, the allowable processing cost. Such a model provides a benchmark for evaluating the performance of the reasoning process (e.g., Birnbaum, Collins, Freed, & Krulwich (1990), Collins, Birnbaum, Krulwich, & Freed (1993), Fox & Leake (1994, 1995)).

Both the models of expected and ideal behavior must be compared to the *actual behavior* of the reasoning process in reasoning episodes. It is clear that in an agent with perfect self-understanding and perfect understanding of the external world, actual performance would always bear out predictions and mirror the ideal. In a less than ideal reasoner, however, the distinct types of models are useful because discrepancies reveal the need for learning:

- Discrepancies between expected behavior and actual behavior are *expectation failures* (e.g., Collins & Birnbaum (1988), Hammond (1989), Leake (1992), Ram (1991), Riesbeck (1981), Schank (1982, 1986)) that, in the memory search context, show the need for a reasoner to refine its understanding of its own memory search capabilities or of the contents of its memory.
- Discrepancies between actual behavior and ideal behavior, whether or not these discrepancies are expected, show the need for learning to refine the processing of the reasoning system (e.g., Collins et al. (1993), Freed & Collins (1994), Fox & Leake (1994), Krulwich (1991), Ram & Cox (1994)).

The success of this method for guiding learning depends on addressing two key problems. First, it is obvious that both the models of expected and ideal behavior will in general be incomplete or too abstract to apply to some sub-parts of the reasoning process. (In fact, if the model of ideal behavior were specified at a sufficiently fine-grained level to determine the ideal decision at each processing step, that model would itself prescribe the desired reasoning process, obviating the need for any other reasoning process.) However, the model itself can be refined with experience.

Second, the credit assignment problem for failures may be quite difficult. In general, the knowledge required to judge the problem-solving process will not be available while the problem is being solved. (If that knowledge *were* available, the reasoner could simply use it during initial problem-solving to avoid following an incorrect problem-solving path.) However, after the problem has been solved, additional information is available, and that information can be used to analyze the effectiveness of a solution that was generated based on more limited information (Fox & Leake, 1994).

Representing needs for learning: Describing conflicts between models of memory performance

Introspective reasoning about the effectiveness of a reasoning process is only useful if it enables the reasoning process to become more effective. The previous section discusses how conflicts between expectations, ideal behavior, and actual behavior show the need for learning. Learning based on those conflicts can be facilitated by representing another type of information: characterizations of the types of conflicts involved. Such representations are useful provided that conflicts grouped by similar descriptions are resolvable in similar ways. When they are, the choice of learning strategies can be based on the descriptions themselves, using the knowledge that learning strategies that applied to similar conflicts in the past are likely to be appropriate to apply to current conflicts as well.

In previous research on explaining anomalies during story understanding, I developed a representation for expectation failures and belief conflicts that includes four types of information (Leake, 1991, 1992). The first is the *expected behavior or state;* the second is the *actual behavior or state*. These two components describe the surface conflict. The third and fourth components are the *source of the reasoning that failed*—the theory or model that needs to be revised—and a description of how what actually occurred deviated from previous expectations. This four-part structure applies to a wide range of conflicts, and can serve as a basis for organizing specific problems and response information. I am investigating the application of such a framework to characterizing the conflicts that prompt learning about memory search.

Managing the Reasoning Process

The previous sections consider the knowledge needed for introspective reasoning about memory search. A final crucial issue is how this knowledge should be applied. Two points to address are the computational cost of generating memory search plans and the need for the memory search process to respond appropriately to idiosyncrasies in the contents and organization of the memory being searched. A promising approach for addressing these problems is case-based reasoning (Kolodner, 1993; Riesbeck & Schank, 1989). In case-based reasoning, new problems are solved by retrieving and applying relevant prior solutions. Case-based reasoning to reuse previous introspective reasoning is fundamental to a number of models of introspective reasoning (e.g., Cox (1994), Leake (1994), Ram & Cox (1994), Oehlmann & Edwards (1995)), and is promising for guiding memory search as well. In a similar spirit, Kennedy (1995) has proposed the *internal analogy* process, which compiles solutions to previous memory search problems and re-applies them to new search problems. Case-based approaches to memory search in turn raise new questions about the representation and organization of the memory search cases stored in memory, and those issues are another focus of ongoing research.

Perspective

As I observed in the previous sections, fundamental notions such as knowledge goals, models of expected and ideal performance, and learning from expectation failures have wide applicability for introspective reasoning and are now being investigated by a number of researchers. I have argued that those notions can also be applied productively to reasoning about and refining the memory search process. In the context of the memory search task, the specific representational requirements for those fundamental constructs include knowledge goals with both concept specifications and comparative specifications; contentful representations of memory links; models of expected and desired memory search behavior; and a representation for the ways that actual memory search behavior can conflict with those models.

Applying introspective reasoning to memory search appears promising both as a way of achieving more effective retrieval and as a way of building up an introspective model of the contents of memory. As learning refines the reasoner's model of the types of information it can find in memory, that model could form a starting point for drawing conclusions about the nature of its knowledge as a whole, rather than simply about isolated facts in its memory.

Acknowledgment

This paper is based on research supported by the National Science Foundation under Grant No. IRI-9409348. I would like to thank Michael Cox, Susan Fox, and David Wilson for their helpful comments on a draft of this paper.

References

Bain, W. (1989). Judge. In Riesbeck, C. & Schank, R. (Eds.), *Inside Case-Based Reasoning*, chap. 4, pp. 93–140. Lawrence Erlbaum Associates.

Birnbaum, L., Collins, G., Freed, M., & Krulwich, B. (1990). Model-based diagnosis of planning failures. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pp. 318–323 Boston, MA. AAAI.

Burke, R. (1993). Retrieval strategies for tutorial stories. In Leake, D. (Ed.), *Proceedings of the AAAI-93 Workshop on Case-Based Reasoning*, pp. 118–124 Washington, DC. AAAI. AAAI Press technical report WS-93-01.

Burstein, M. (1994). Case age: selecting the best exemplars for plausible reasoning using distance in time or space. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 106–111 Atlanta, GA. Cognitive Science Society.

Collins, G. & Birnbaum, L. (1988). An explanation-based approach to the transfer of planning knowledge across domains. In *Proceedings of the 1988 AAAI Spring Symposium on Explanation-based Learning* Stanford, CA. AAAI.

Collins, G., Birnbaum, L., Krulwich, B., & Freed, M. (1993). The role of self-models in learning to plan. In *Foundations of Knowledge Aquisition: Machine Learning*, pp. 83–116. Kluwer Academic Publishers.

Cox, M. (1994). Machines that forget: learning from retrieval failure of mis-indexed explanations. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 225–230 Atlanta, GA.

Cox, M. & Freed, M. (1994). Using knowledge of cognitive behavior to learn from failure. In *Proceedings of the Seventh International Conference on Systems Research, Informatics and Cybernetics*, pp. 142–147 Baden-Baden, Germany.

Domeshek, E. (1992). *Do the Right Things: A Component Theory for Indexing Stories as Social Advice*. Ph.D. thesis, The Institute for the Learning Sciences, Northwestern University.

Fox, S. & Leake, D. (1994). Using introspective reasoning to guide index refinement in case-based reasoning. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 324–329 Atlanta, GA.

Fox, S. & Leake, D. (1995). Planning for repairing reasoning failures. In *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms* Stanford, CA. AAAI. In press.

Freed, M. & Collins, G. (1994). Learning to prevent task interactions. In Ram, A. & desJardins, M. (Eds.), *Proceedings of the 1994 AAAI Spring Symposium on Goal-Driven Learning*, pp. 28–35 Stanford, CA. AAAI.

Gentner, D. & Collins, A. (1981). Studies of inference from lack of knowledge. *Memory and Cognition*, *9*(4), 434–443.

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

Hunter, L. (1990). Planning to learn. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, pp. 261–268 Cambridge, MA. Cognitive Science Society.

Kennedy, A. (1995). Using a domain-independent introspection mechanism to improve memory search. In *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms* Stanford, CA. AAAI. In press.

Kolodner, J. (1984). *Retrieval and Organizational Strategies in Conceptual Memory*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Kolodner, J. (1993). *Case-Based Reasoning*. Morgan Kaufmann, San Mateo, CA.

Krulwich, B. (1991). Determining what to learn in a multicomponent planning system. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, pp. 102–107 Chicago, IL. Cognitive Science Society.

Leake, D. (1991). An indexing vocabulary for case-based explanation. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pp. 10–15 Anaheim, CA. AAAI.

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

Leake, D. (1994). Towards a computer model of memory search strategy learning. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 549–554 Atlanta, GA.

Leake, D. & Ram, A. (1993). Goal-driven learning: fundamental issues (a symposium report). *The AI Magazine*, *14*(4), 67–72.

Oehlmann, R. & Edwards, P. (1995). Introspection planning: representing metacognitive experience. In *Proceedings of the 1995 AAAI Spring Symposium on Representing Mental States and Mechanisms* Stanford, CA. AAAI. In press.

Owens, C. (1991). *Indexing and retrieving abstract planning knowledge*. Ph.D. thesis, Yale University.

Ram, A. (1991). A theory of questions and question asking. *The Journal of the Learning Sciences*, *1*(3 & 4), 273–318.

Ram, A. & Cox, M. (1994). Introspective reasoning using meta-explanations for multistrategy learning. In Michalski, R. & Tecuci, G. (Eds.), *Machine Learning: A Multistrategy Approach*. Morgan Kaufmann.

Ram, A. (1987). Aqua: asking questions and understanding answers. In *Proceedings of the Sixth Annual National Conference on Artificial Intelligence*, pp. 312–316 Seattle, WA. American Association for Artificial Intelligence, Morgan Kaufmann Publishers, Inc.

Riesbeck, C. (1981). Failure-driven reminding for incremental learning. In *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 115–120 Vancouver, B.C. IJCAI.

Riesbeck, C. & Schank, R. (1989). *Inside Case-Based Reasoning*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Rissland, E., Skalak, D., & Friedman, M. (1994). Heuristic harvesting of information for case-based argument. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, pp. 36–43 Seattle, WA. AAAI.

Schank, R. (1982). *Dynamic Memory: A Theory of Learning in Computers and People*. Cambridge University Press, Cambridge, England.

Schank, R. (1986). *Explanation Patterns: Understanding Mechanically and Creatively*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wilensky, R. (1986). Knowledge representation—a critique and a proposal. In Kolodner, J. & Riesbeck, C. (Eds.), *Experience, Memory and Reasoning*, chap. 2, pp. 15–28. Lawrence Erlbaum Associates.