The Knowledge Engineering Review, Vol. 9:1, 1994, 78–82

Explanation as a Primary Task in Problem Solving

MICHAEL R. WICK
Computer Science Department, University of Wisconsin–Eau Claire, Eau Claire, WI 54702-4004, USA

Abstract

This article summarizes the author's perspective on the discussions that occurred at the Workshop on Explanation and Problem Solving held during the Thirteenth International Joint Conference on Artificial Intelligence*. Motivated by those discussions, the article argues for the promotion of expert system explanation from a secondary task, used mainly for communication, to a primary task that is tightly integrated with the domain problem solving of the expert system.

1 Introduction

Explanation is often cited as one of the most important advantages of the expert system methodology. Two reasons in particular appear to support this emphasis. First, end-user acceptance of the system is critically dependent on the expert system's ability to explain its decisions (Teach and Shortliffe, 1984). It was discovered early in the history of expert systems that it was not enough to be able to solve the problem, but one must also be able to explain the solution to that problem. Second, expert system construction is eased when the expert system is capable of explaining its reasoning to the system developer, i.e., knowledge engineer or domain expert (Neches et al., 1985). By giving an expert system the ability to present a trace of its execution as an explanation, system builders can better debug, analyse and maintain the knowledge contained in an expert system. These two uses, end-user acceptance and system maintainability, have been the driving force in expert system explanation.

The traditional method of generating an explanation in expert systems is to maintain, during execution, a trace of the expert system's line of reasoning (Paris et al., 1988). An explanation of the expert system's reasoning is then an augmented, pruned, or in some other way transformed presentation of this trace. This traditional high level of coupling between the knowledge used by the expert system and that presented in the explanation has caused knowledge engineers to carefully consider the explanatory implications of their design decisions. Researchers and practitioners alike have realized, often in retrospect, that the explanatory capabilities of a knowledge-based system are dependent on a combination of the problem being solved (e.g., diagnosis versus design), the model used to represent the artifact (e.g., functional versus structural), and the method used to direct problem solving (e.g., generate-and-test versus cover-and-differentiate). The result is an explanation matrix indexed by problem, model and method types.

The 1993 IJCAI Workshop on Explanation and Problem Solving brought together experts in an attempt to analyze explanation from this problem-solving perspective. The goal of the workshop was to highlight the potential relationship between each of these dimensions and the explanations that an expert system can produce and understand. This paper presents a brief summary of the highlights of that workshop, and points to an emerging area of expert system explanation that appears to have the potential of significantly affecting the development of knowledge-based systems.

2 Background

During the last 20 years, numerous efforts have been undertaken in an attempt to improve and use expert system explanations. For the most part, three core ideas form the foundation of nearly all research in this area. These three core ideas are commonly thought to represent the major research results in expert system explanation (Chandrasekaran et al., 1989).

*Reprints of the Workshop Working Notes are available by writing to the author at the address listed.

1. A trace of an expert system's execution can be used to provide an explanation of the expert system's problem solving. Mycin was one of the first systems to explain its actions (Shortliffe, 1976). Mycin provided two basic explanation queries: why and how. These two queries form the foundation of nearly all explanation facilities to date (Wick and Slagle, 1989).
2. An expert system's knowledge can be seen as being compiled from deeper justifying knowledge. Swartout (1983) introduced a system explicitly designed to attack the problem of explanation. Swartout used a *domain principle* and a *domain rationale* to record the designer's rule justification by using an automatic programmer to build the expert system. The system produced excellent explanations. However, the information appears to be best suited for the knowledge engineer (Wick and Thompson, 1989).
3. Explanations can be given at different problem-solving levels. Clancey has built an explanation system that augments the facility provided by Mycin (Hasling et al., 1984). Clancey's system shifts the focus from the domain knowledge to the strategic problem-solving knowledge, and is capable of generating why and how explanations about the strategy used to solve the problem.

Recently, a new development has emerged in expert system explanation. In this new research, explanation is no longer viewed as an add-on to the expert system's reasoning, but as a problem-solving activity in its own right (David & Krivine, 1989; Moore & Swartout, 1989; Ryan & Bridges, 1988; Tanner & Josephson, 1988). This new research appears to be headed towards adding a fourth core idea to expert system explanation.

4. Explanation can be viewed as a complex problem solving process largely distinct from the expert system's original problem-solving process. Wick and Thompson (1992) have demonstrated that improved expert system explanations can be generated within a *reconstructive explanation* paradigm in which the explanation is reconstructed by a separate problem-solving process.

It is this most recent finding, combined with the outcome of discussion from the workshop, that highlight a new view of explanation that has particular implications in the design and application of expert systems.

3 Principle themes of the workshop

The workshop participants reached several interesting conclusions by analysing explanation from the perspective of the three-dimensional explanation matrix (see Wick & Dieng, 1993, for a more detailed summary):

- With traditional expert system explanation techniques (i.e., trace-based explanation), the method that is used to solve the problem is primarily useful in generating comparative explanations, such as what-if and why-not explanations.
- Reconstructive explanation techniques, while raising a number of difficult issues, provide a means of eliminating the need to access the method of the expert system for comparative explanations for some classes of user.
- The original three-dimensional explanation matrix lacked two important dimensions that impact explanation and problem solving: context and user. The context and the user were found to interact with all three of the other matrix dimensions.
- Explanation is a primary task in expert systems at the same level of the domain problem solving conducted by the expert system.
- Explanation is sufficiently supported by six expert system models: the artifact, solution, process, domain, control and user models.
- The models used by an expert system must support multiple viewpoints so that the corresponding explanations can be tailored to particular users.

Although several of these conclusions are evolutionary in that they have directly evolved from previous research, the conclusion that "explanation is a primary task in expert systems" is much more revolutionary. The workshops on explanation held during the 1988 and 1990 National Conferences on Artificial Intelligence involved heated discussion over this very topic. In contrast, most of the participants of the 1993 workshop were supportive of this view. This represents a significant change in mind-set that warrants particular attention.

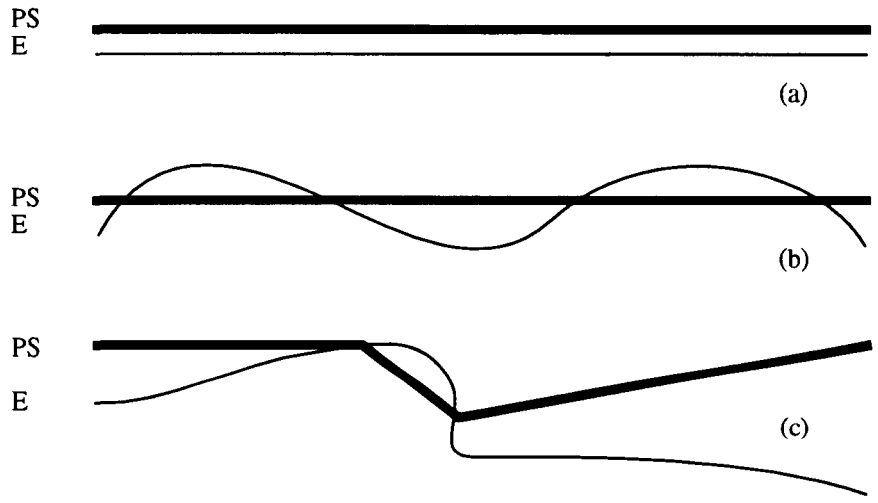

Figure 1 The role of explanation in problem solving

4 Explanation as a primary task

Expert system explanation has traditionally been viewed as involving the generation of explanations based on the reasoning of an expert system. However, as was discussed during the workshop, this view is unnecessarily restrictive. Expert system explanation also involves the consumption of explanations into the reasoning of an expert system. In some domains, the explanation system and the reasoning system are almost completely the same. For example, in the diagnosis of a traffic accident scene, explanation plays a paramount role in attempting to understand how and why the accident occurred. Further, the common notion that an explanation system only produces one class of explanation is also misleading. The explanations that an expert system must produce and consume fall into two general classes; *world explanation* and *decision explanation* (Chandrasekaran et al., 1989). World explanations involve explaining objects or interactions in the world around us. For example, an explanation of how the heart operates. Decision explanations, however, involve the explanation of the steps behind problem solving. The workshop participants concluded that the role of explanation in problem solving includes the generation and consumption of both world and decision explanations. Figure 1 illustrates three general roles that explanation can play in problem solving, as proposed by the workshop participants.

Figure 1(a) illustrates the situation where problem solving and explanation are two relatively independent processes. This is representative of the traditional approach to explanation where the expert system first solves the problem and then an explanation system is invoked to communicate this solution to the user. The explanations that are produced may focus on either the process that led to the solution (i.e., a decision explanation) or on the solution itself (i.e., a world explanation). Figure 1(b) illustrates a more integrated role of explanation in problem solving. In this scenario, problem solving occasionally produces or consumes explanation as the reasoning is in progress. These explanations, however, do not alter the overall reasoning strategy that the expert system is using. For example, an expert system may use an explanation routine to simulate the effects of alternative decisions. The results of those explanations are used as support for the various alternatives. These explanations are typically world explanations that model the artifact under consideration. This use of explanation as one of the processes invoked by the reasoning strategy of the expert system is typical of the type of multi-process/multi-representation organization found in second generation expert system (David et al., 1993). Figure 1(c) illustrates the most general and active use of explanation in the reasoning process. In this situation, not only does the reasoning strategy consume explanations during its execution, but those explanations may alter the basic strategy that is being followed. For example, in a cooperative environment where the user and the expert system are working together to solve a problem, information provided in a user's explanation may cause the system to abandon its current strategy in favor of another that more accurately matches the emerging problem. The explanations in this scenario may be either world

explanations or decision explanations, but are usually generated by the user and consumed by the expert system.

The potential impact of this active role of explanation in problem solving is significantly increased when one considers a reconstructive approach to explanation (Wick & Thompson, 1992). In the traditional approach to expert system explanation, the reasoning trace of the expert system is completely known by the explanation system. Although only a subset of this information is included in the final explanation, along with other auxiliary information, the basis for all information in the explanation is the line of reasoning. However, consider a reconstructive approach to expert system explanation. With reconstruction, a distinction arises between the line of reasoning as followed by the expert system and the *line of explanation* postulated by the explanation process. Expert system explanation becomes a complex problem-solving process of reconstructing plausible connections between observed elements of the expert system's line of reasoning. This reconstructive approach appears to model more closely the processes involved in human explanation. In expert problem solving, many of the details of how and why things happened are not available from a memory of the problem solving (Ericsson & Simon, 1984). When asked to report this information, the expert will reconstruct an explanation that integrates the elements of a partial memory trace with the memory of other related entities (for example, textbook information). This freedom to reconstruct an explanation based on information in addition to the information and processes used during problem solving may be in part responsible for the high quality of human explanations. Reconstructive expert system explanation is the study of how to give an expert system this reconstructive ability. Although a complete discussion of the issues involved in reconstructive explanations is beyond the scope of this paper (see Wick & Thompson, 1992), the use of reconstructive techniques strongly interacts with each of the three views illustrated in Figure 1.

The use of reconstructive techniques in the situation of Figure 1(a) may reduce the need for knowledge engineers to consider the end-user explainability of the reasoning strategy used by the expert system. For example, if the expert system's reasoning strategy is too complicated for the end-user to understand, a second more understandable approach could be substituted when the solution is explained to this particular end-user. In the situation of Figure 1(b), reconstructive explanation techniques may allow the problem-solving strategy of the expert system to incorporate decision explanations into its reasoning process. Again, the information consumed does not alter the overall strategy used by the expert system, but it may influence the status of some of its hypotheses. For example, an expert system could ask a reconstructive explanation routine to explain how certain hypotheses might have been generated. Those hypotheses that are supported by a variety of problem-solving methods could have their corresponding level of confidence increased significantly. Finally, the situation in Figure 1(c) appears to offer the most exciting interaction between the reconstructive approach and the role of explanation in problem solving. Nearly everyone has had the experience of explaining to someone a problem that you are struggling with at work when, during the process of explaining your approach, you see something you had missed before. The process of explaining your reasoning provides new insight into the problem that may alter your approach. This phenomenon is difficult to account for when the traditional trace-based explanation techniques are used. However, the use of a separate reconstructive technique could easily account for the inclusion of information or hypotheses not present in the original line of reasoning. In the context of Figure 1(c), this means that the expert system might alter its reasoning strategy based on information uncovered by the reconstructive explanation of its own reasoning. This feedback from explanation to problem solving truly creates an *active explanation* system in which the process of explanation plays a key role in the overall reasoning of the expert system. Such an active explanation module might be extremely useful in domains where the strategy of the expert system is not known to be perfectly sound. For example, the expert system may contain rules that it has learned through experience. A reconstructive explanation that attempts to justify the use of those rules may very well point to exceptions that were not noticed by the learning system.

5 Conclusion

Although serious questions remain open, the increasing realization that explanation is a primary task in problem solving, and the growing acceptance of reconstructive explanation techniques were

the most significant trends of the 1993 IJCAI Workshop on Explanation and Problem Solving. Over the next few years, one can expect to see many results on the implications of each of these trends, their interaction, and their relationship to other fields of research such as second generation expert systems and knowledge acquisition.

References

Chandrasekaran, B, Tanner, M and Josephson, J, 1989. "Explaining control strategies in problem solving". *IEEE Expert* **4** (1) 9–24.

David, JM and Krivine, JP, 1989. "Augmenting experience-based diagnosis with causal reasoning". *Applied AI Journal.*

David, JM, Krivine, JP and Simmons, R (eds.), 1993. *Second Generation Expert Systems.* Berlin: Springer-Verlag.

Hasling, D, Clancey, WJ and Rennels, G, 1984. "Strategic explanations for a diagnostic consultation system". *International Journal of Man–Machine Studies* **20** 3–19.

Moore, J and Swartout, WR, 1989. "A reactive approach to explanation". In: *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence.*

Neches, R, Swartout, WR and Moore, JD, 1985. "Enhanced maintenance and explanation of expert system through explicit models of their development". *IEEE Transactions on Software Engineering* **11** (11) 1337–1351.

Paris, CL, Wick, MR and Thompson, WB, 1988. "The line of reasoning versus the line of explanation". In: *Proceedings of the 1988 AAAI Workshop on Explanation*, pp 4–7.

Ryan, J and Bridges, S, 1988. "Constructing explanations from Conceptual graphs". In: *Proceedings of the Third Annual Workshop on Conceptual Graphs.*

Shortliffe, EH, 1976. *Computer Based Medical Consultations: MYCIN.* New York: Elsevier/North Holland.

Swartout, WR, 1983. "XPLAIN: A system for creating and explaining expert consulting programs". *Artificial Intelligence* **21** (3) 285–325.

Tanner, MC and Josephson, JR, 1988. "Justifying diagnostic conclusions". In: *Proceedings of the 1988 AAAI Workshop on Explanation*, pp 76–79.

Teach, RL and Shortliffe, EH, 1984. "An analysis of physician's attitudes". In: *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project.* Reading, MA: Addison-Wesley.

Wick, MR and Slagle, JR, 1989. "An explanation facility for today's expert systems". *IEEE Expert* **4** (1) 26–36.

Wick, MR and Thompson, WB, 1989. "Reconstructive explanation: Explanation as complex problem solving". In: *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence (IJCAI).*

Wick, MR and Thompson, WB, 1992. "Reconstructive expert system explanation". *Artificial Intelligence* **52** 33–70.

Wick, MR and Dieng, R, 1993. "The 1993 IJCAI Workshop on Explanation and Problem Solving". Submitted to *AI Magazine.*