

## A Review of *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*

W. Lewis Johnson

*Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*, Johanna D. Moore, MIT Press, Cambridge, Massachusetts, 1995, 352 pp., \$29.95, ISBN 0-262-13301-6.

Johanna Moore's work in the area of computer-generated explanation has been highly influential. Her thesis work, as well as the subsequent work of her and her students, has helped to change the way we think about the problem of generating explanations. The crux of the explanation problem, according to Moore, is not how to present information as such but how to impart an understanding on the user. The explanation system should be flexible enough that if an initial explanation fails to convey the understanding, it can try explaining the concept in a different way. The system should be aware of what it previously said to the user and what its communicative goals were at the time. Both the success and the failure of previous explanation attempts can influence subsequent explanations. Consequently, the activity of explaining becomes an interactive process and must be viewed as a dialog rather than a series of answers to individual questions.

This book describes the explanation system that Moore developed for the Explainable Expert System (EES) Project and is based on her dissertation. Moore's system provides interactive explanations by the following means: When the user asks a question about a decision being made by an expert system, it extracts the

information necessary to generate the answer from the expert system's knowledge base. It assumes that the expert system was constructed using the EES development tool, which generates expert systems automatically from declarative descriptions of domain terminology and problem-solving methods. The declarative descriptions provide the design rationales and background information that are needed to answer many questions. The explainer then constructs a plan to achieve the goal of getting the user to understand the answer to the question. The plan consists of a series of speech acts and is constructed using a library of plan operators. The plans are then used to generate English-text phrases to present to the user.

Once an explanation is generated, the user can use the mouse to select some portion of the generated text and ask a follow-on question about it. The text plans generated by the explainer enable it to respond effectively to such questions. It checks the communicative goal that the selected text was intended to achieve and refers to that goal when answering the follow-on question. Answering the question can involve constructing an alternative plan to achieve the same goal. The beauty of this approach is that it does not require the user to articulate the follow-on question precisely; by referring back to the original explanation plan, the system is able to form reasonable conjectures about what the user needs clarified.

The techniques that Moore developed have found their way into a number of explanation systems and, no doubt, will continue to do so. Thus, this book should be of potential interest to anyone concerned with the explanation problem. The interest should also extend to related fields, including intelligent advisory systems, natural language processing and dialog modeling, and intelligent learning environments. The book is well written and edited and easy to read.

The main shortcomings of this book stem from the timing of its publication. Several years have passed since Moore's thesis work was completed, and research in the field has continued apace. A book based too closely on the original thesis risks being obsolete as soon as it is published. Moore is clearly aware of this problem, as evidenced by the book's rather odd concluding paragraph:

The prototype PEA system should be viewed as one possible implementation of the ideas presented in this book. Although I believe that some aspects of this implementation are fundamentally correct, others need further work and are already being extended and modified by myself and other researchers.

The following points are made in the book and require reexamination in view of more recent work: Moore cites a number of requirements that intelligent explanation systems should exhibit. The generated explanations should sound natural and abide by rules of discourse structure, particularly in multisentential explanations. They should accurately reflect the system's knowledge and reasoning. The explanation system should be capable of answering the range of questions a user might want to ask. The explanation system should be easily extensible. These points seem unarguable, but some of them in fact raise tricky issues.

First, let us consider the naturalness issue. Moore assumes that to be natural, the explanation system should have the same characteristics as human natural language dialog. However, computers have a range of communication media at their dis-

posal, including graphics and various types of structured presentation. Other researchers are looking at combining media and allocating presentations to various media (Arens and Hovy 1990; Feiner and McKeown 1990). These approaches limit the amount of text that must be generated, particularly multisentential text; this limitation is important because people tend to be less inclined to read multisentential descriptions on a computer screen than on a printed page. Multimodal explanations, although less "natural," are, in fact, more effective than text-based dialogs. Although it is also useful to track discourse context in the presence of these other media, the detailed mechanisms involved can differ, owing to the unique properties of each medium. For example, when a system constructs a picture to present to a user, the user can extract information from the picture that the system did not intend to convey. Moore appears to be more interested in experimenting with alternative ways of posing questions to explanation systems than with alternative ways of generating answers.

There is also some issue about the extent to which the explanation should be faithful to the knowledge and reasoning of the expert system. Wick and Thompson (1989) argued that explanation is fundamentally a *reconstructive process*: People do not trace their reasoning process when

explaining their conclusions and neither should expert systems. The structure of explanations can be radically different from the structure of the original problem solving, and the knowledge used in explanations can also be in a radically different form. Moore's approach permits the expert system's reasoning process and the rhetorical structure of the explanation to deviate; however, the desire for fidelity in practice places limits on this deviation. Explanations are generated based on the design history of the expert system and on the knowledge base used in designing the expert system. Basing the explanation on the design history can limit the system's ability to deviate from the structure of the design history when necessary. Also, using a design knowledge base for explanation can be problematic if there is knowledge that is needed for explanation but not for design. Thus, Moore's system does not fully address Wick and Thompson's concerns, although it appears to be adequate for generating cogent explanations in the context where it was tested. Part of the reason that the explanations work is that Moore's users typically interrupt the system's problem solving to ask questions, requiring the system to explain its reasoning at that moment in time. Explanations, therefore, tend to focus on the rationales for the current reasoning step, which are obtained readily from the design history. If,



*It should be emphasized that these criticisms are at the margin and do not detract from the overall significance of the contributions of this book.*

instead, a user were to request an explanation after the problem solving is completed, a more thorough reorganization of the design history might be required to produce an effective explanation. Moore's approach to explanation also places heavy demands on the knowledge engineer: The deep knowledge underlying the expert system must be represented explicitly for the explainer to use it, raising questions of where the deep knowledge comes from and how it is obtained. At the least, extensive acquisition effort is required to obtain the deep knowledge, which can get in the way of the practical concerns of developing and fielding the knowledge-based system. At worst, as Clancey (1991) argued, the representations of deep knowledge are abstractions that are invented by knowledge engineers and are unfamiliar to users. One approach to this problem, described in Johnson (1994), is to employ induction techniques to construct the abstract knowledge representations automatically by observing the knowledge-based system's behavior in various hypothetical situations. This approach does not eliminate the knowledge-acquisition bottleneck engendered by explanation, but it can make it less severe. It also provides a mechanism for restructuring knowledge along the lines suggested by Wick and Thompson.

Moore's technique employs a conventional hierarchical linear-planning paradigm for generating text. It is an important technical advance over schema-based explanation methods but still might not be suited ideally to conversational interaction. Suchman (1987) argued that the classical planning model is not a realistic model of human behavior in the world. A number of developers of intelligent agents, for example, Tambe et al. (1995), find it inadequate for dynamic environments. Given Moore's objective of a dynamic explanation facility that can respond to user interruptions, it might be appropriate to consider approaches that do not rely as heavily on declarative plan structures. Moore argues that a rich plan representation would

also be necessary in an interactive explanation approach; however, this claim remains to be demonstrated. It might be possible for an explainer to do less planning during the normal course of explanation and reconstruct a more detailed plan only when it appears that the user does not understand the explanation.

It should be emphasized that these criticisms are at the margin and do not detract from the overall significance of the contributions of this book. Anyone contemplating work in the area of automated explanation is well advised to study Moore's work thoroughly and should therefore find this book to be valuable.

### References

- Arens, Y., and Hovy, E. H. 1990. How to Describe What? Towards a Theory of Modality Utilization. In Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, 487-494. Cambridge, Mass.: Cognitive Science Society.
- Clancey, W. 1991. Situated Cognition: Stepping Out of Representational Flatland. *AI Communications—The European Journal on Artificial Intelligence* 4(1): 4-10.
- Feiner, S. K., and McKeown, K. R. 1990. Coordinating Text and Graphics in Explanation Generation. In Proceedings of the Eighth National Conference on Artificial Intelligence, 442-449. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Johnson, W. L. 1994. Agents That Learn to Explain Themselves. In Proceedings of the Twelfth National Conference on Artificial Intelligence, 1257-1263. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Suchman, L. A. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. New York: Cambridge University Press.
- Tambe, M.; Johnson, W. L.; Jones, R. M.; Koss, F.; Laird, J. E.; Rosenbloom, P. S.; Schwamb, K. 1995. Intelligent Agents for Interactive Simulation Environments. *AI Magazine* 16(2): 15-39.
- Wick, M. R., and Thompson, W. B. 1989. Reconstructive Explanation: Explanation as Complex Problem Solving. In Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, 135-140. San Mateo, Calif.: International Joint Conferences on Artificial Intelligence.