AI Magazine Volume 10 Number 3 (1989) (© AAAI)

The 1988 AAAI Workshop on Explanation

Michael R. Wick

■ *This article is a summary of the Workshop on Explanation held during the 1988 National Conference on Artificial Intelligence in St. Paul, Minnesota. The purpose of the workshop was to identify key research issues in the rapidly emerging area of expert system explanation.*

Expert system explanation is the study of how to give an expert system the ability to provide an explanation of its actions and conclusions to a variety of users (including the domain expert, knowledge engineer, and end user). The 1988 AAAI Workshop on Explanation brought together many of the world's experts on expert system explanation in an attempt to highlight key research areas and questions that should be the focus of subsequent work. The one-day workshop was organized into five sessions of short presentations, each followed by panel-led open discussion among the 35 workshop participants. A proceedings of the workshop was compiled and is available through AAAI.

The first session, Text Planning I, focused on some of the issues involved in treating the process of explanation as a complex problem-solving task requiring knowledge beyond that used to solve the expert system's original problem. This session, led by Cecile Paris, Robert Schulman, Mike Wick, and Dan Suthers, brought up issues related to the generation of explanations, including the nature of the coupling between the processes and the knowledge involved in generating explanations and that involved in problem solving.

The second session, Explanation and Knowledge Acquisition, raised several issues regarding the connection between expert system explanation and the problem of knowledge acquisition. Led by Bill Mark, Alice Kidd, Sue Abu-Hakima, and Bruce Porter, the discussion focused on identifying how explanation can be used for knowledge acquisition as well as how knowledge acquisition can be used to acquire knowledge for explanation.

The third session, User Modeling, attempted to outline some of the major issues in employing user models to tailor expert system explanations to particular users or user types. Kathy McKeown, Robin Cohen, Ivan Rankin, and Robert Kass led the open discussion. The workshop participants outlined four major components central to a user model. In addition, a possible method for the automatic acquisition of some of these components was discussed.

The fourth session, Question Types, highlighted work designed to find useful categorizations of the queries that an explanation system can answer. The discussion, led by G. Nigel Gilbert, Mike Tanner, and Dave Schaffer, focused on two largely distinct approaches to finding such categorizations. First, an argument was presented that called for categorizing explanation responses rather than explanation queries. The second approach, focusing on query categorization, argued for the use of the domain model and explicit considerations of potential explanation queries during the design phase of the expert system. This session also raised the issue of canned text as an explanation paradigm and discussed its usefulness.

The fifth and final session, Text Planning II, centered on the issue of how to construct the information that is presented to the user. Johanna Moore, Dan Rochowiak, and Yee-Han Cheong led a discussion that focused on three main issues related to this problem: the ability to respond to follow-up questions by the user, the identification and use of implicit context in user queries, and the use of rhetorical devices to enhance explanations.

Text Planning I

This session raised the most hotly debated issue of the workshop. The debate centered around the question of whether explanations could or, in some cases, should be based on knowledge that is only loosely related to the knowledge used to solve a problem. This decoupling of the knowledge used for problem solving and the knowledge used for explanation raised a number of issues, including the faithfulness of the explanation system to the expert system's reasoning, the implications of subjective evidence in explanations, and the existence of goodwill between the explanation system and the user. In all, two main groups emerged from the discussion.

> *For an end-user audience, the often convoluted and opaque reasoning steps taken by the expert system are too complicated to provide the user with a basis for evaluation.*

The group in favor of decoupling (see Paris, Wick, and Thompson 1988, p. 4) argued that for an end-user audience, the often convoluted and opaque reasoning steps taken by the expert system are too complicated to provide the user with a basis for evaluation. They argued that an explanation system which can use a model of the domain specifically tailored to the knowledge understandable by the end user can present an argument, based on the same evidence as used by the expert system, that is more easily understood by the end user. In turn, this would give the end user the improved ability to evaluate the expert system's solution.

The majority of the workshop participants, however, argued that by totally decoupling the knowledge used by the expert system from the knowledge used by the explanation system, the faithfulness of the explanation to the

0738-4602/89/$3.50 ©1989 AAAI

reasoning of the expert system is jeopardized. Their concern was that an explanation system which is allowed to concentrate more on convincing the user the solution is correct than on presenting the user with an understandable version of the expert system reasoning might give rise to a situation in which the user is not presented with adequate information to evaluate the reasoning of the expert system. This group argued that the explanation must remain faithful to the execution trace to avoid potentially misleading the user. They acknowledged that the reasoning trace in its raw form is usually far too detailed and complex to be presented to users but that the proper way to deal with such complexity is to provide explanation routines which can select appropriate pieces of the trace to present to the user. They argued that this approach avoids overwhelming the user with complexity but still allows the explanations to be based on knowledge actually used to solve the problem.

The degree of decoupling that is acceptable between the expert system and the explanation system remains an open question.

Another concern about decoupling was the relative subjectivity of the evidence used by the expert system and the user. Most workshop participants argued that in some cases, the user rejects the conclusion of the expert system even though the reasoning used is accepted. Such a rejection might result when the subjective value of the evidence used by the expert system is different than the subjective value of this evidence to the user. Therefore, it was argued that the explanation system must make it apparent what value it places on the evidence used by the expert system, thus forcing the explanation to be coupled to the execution trace. This conclusion was contested by those advocating decoupling. They argued that although it is certainly possible the user might reject the solution of a correct reasoning process based on different value judgments, this possibility did not force the explanation to be coupled to the execution trace. They argued that an explanation system which uses a model of the domain largely distinct from the expert system's model could use a model that reflects the user's (assumed) values, allowing the explanation system to construct an explanation that would be the most convincing to the user.

Another issue raised during the discussion of decoupling was the idea of goodwill between the explanation system and the user. Many of the workshop participants raised the concern that an explanation system which is decoupled from the expert system and which has the goal of convincing the user that the solution is correct could intentionally mislead the user by concealing evidence which is strongly against the conclusion. Both those in favor of decoupling and those against it agreed this problem is possible.

An overwhelming majority of the workshop participants were against any large degree of decoupling. However, those in favor of decoupling (including this author!) were able to present substantive counterarguments to the points raised. Thus, the degree of decoupling that is acceptable between the expert system and the explanation system remains an open question.

Although the coupling issues constituted the majority of the open discussion during this session, other work was presented. This work included a method for producing strategic explanations from a problem-solving trace (Schulman, Hayes-Roth, and Johnson Jr. 1988, p. 8), a categorization of expert system knowledge according to three epistemological dimensions (Suthers 1988, p. 12), and a discussion of the rhetorical justification of an expert system's operation (Maybury 1988, p. 16).

Explanation and Knowledge Acquisition

The session on the relation between explanation and knowledge acquisition gave raise to many interesting issues. For the most part, there was agreement that to be useful, explanation must be considered from the beginning of the expert system production process. It must be involved during the knowledge-acquisition process. Also agreed on was the notion that the explanation knowledge (the supporting knowledge) must be acquired just as the problem-solving knowledge is acquired. Explanation knowledge does not come free with the problem-solving knowledge. Thus, two knowledge-acquisition problems must be solved, namely, the acquisition of problem-solving knowledge and the acquisition of the explanation or support knowledge.

The issue arises of to what extent we can rely on the domain expert to explicitly give us explanations and, thus, explanation knowledge. Two significantly different approaches were advocated. One approach treats the problem-solving process as the process of building an explanation. In this approach, the expert enters an explanation for each element of the problem-solving knowledge (Abu-Hakima 1988, p. 26). Because the problem-solving knowledge is constrained to be in the form of an explicit explanation, a solution to the problem is an explanation by definition. The explanations for each concept in the solution are then presented to the user. This approach can be seen as putting a large amount of work on the expert.

Another approach advocates less interference in the acquisition of problem-solving knowledge from the domain expert. In this approach, the system uses a relatively small set of explanation primitives (rationales) to actively construct plausible explanations of the expert's problem solving (Mark 1988, p. 22). Periodically, the system presents these plausible explanations to the expert and asks for their modification (if necessary) so that the true rationale for the problem solving is represented. This approach puts added burden on the system because it is responsible for generating possible accounts of the expert's problem solving that are only verified by the expert on occasion.

Explanation was presented as a possible tool in knowledge acquisition with the use of explained examples as a method of acquiring domain knowledge (Porter, Branting, and Murray 1988, p. 30). In this approach, a system is presented with examples of solutions with explanations. Using the explanation, the system attempts to learn rules that allow the system to obtain the same solutions. Explanation was also shown to aid learning during knowledge acquisition. Here, new information that is presented to the expert system is explained within the context of the information already in the system. Thus, the new information is integrated with the old by explicit explanations. Another potential benefit of explanation to knowledge acquisition was shown by considering problem solving as a cooperative task between the expert system and the user (Kidd 1988, p. 34). In this context, the explanations given by the system and the user allow both to acquire information relevant for current problem solving. For example, specific constraints that the user needs to impose on the expert system, such as solution, must be quick.

Overall, the session determined that explanation does have several implications for knowledge acquisition and vice versa. Explanation was shown to be both an additional burden on knowledge acquisition and an additional aid.

User Modeling

The workshop participants agreed about the importance of the ability to tailor an explanation to the specific audience that requested it. This session highlighted several ways in which a user model can be used to help tailor the explanation to meet the user's specific needs.

Implicit in most of the discussion on user models was the notion of a user's expertise. This information classifies a user's knowledge of a given subject (or concept) according to a spectrum from novice to expert. The user's knowledge about the domain can also be represented explicitly, pointing to a subset of the expert system's knowledge base. This information can influence both the amount of information presented as well as the actual content or form of the information. By varying the explanation to the user's level, the system can avoid both boring advanced users with too much information and handicapping naive users with too little information.

Central to the discussion of user models was the idea of goals. By understanding the user's goals in asking a question, the system can better tailor its response to meet these goals. Thus, goals allow the explanation system to identify what information the user is after in the query. In this manner, goals were highlighted as a major element of a user model. In addition to pointing to information to include in the explanation, it was also shown that goals can be used to help prune information from the explanation (McKeown and Weida 1988, p. 38). Goals can be used to help a system infer the need for including certain elements of an explanation. For example, if an inferential step in the problem-solving trace does not affect the goal of the end user's query, it can be left out of the explanation. Thus, goals can be used to decrease the amount of information presented to the user and, in this way, highlight or draw attention to the significant aspects of the explanation.

Background knowledge also emerged as a critical element of a user model (Cohen 1988, p. 44). By having a representation of the user's knowledge of the domain, the explanation system can provide responses that better match the user's understanding. In most cases discussed during the workshop, the user's knowledge was assumed to be a subset of the expert system's knowledge. Therefore, background knowledge provides the explanation system with additional information beyond the user's level and goal.

Cognitive preferences were also introduced as playing a key role in user modeling (Rankin, Hagglund, and Waern 1988, p. 48). *Cognitive preferences* represent the method that the user perceives is being used to solve the problem. The goal of many explanation systems is to tutor the user so that the perceived method is the actual method used by the expert system. In other words, the explanation system attempts to correct any misconceptions in the user's cognitive model of the expert system's problem solving. Cognitive preferences are more general than the goals of the user's query because they can provide information concerning possible pointers to misconceptions in the user's knowledge. This information can then be used to focus the explanation on revealing the misconceptions and correcting them.

During the session, four major components of a user model were highlighted as important in tailoring an explanation to a specific user: (1) the user's expertise (be it level or explicit knowledge), (2) goals, (3) background knowledge, and (4) cognitive preferences. In addition to producing these features of a user model, a possible approach to the automatic acquisition of user models was discussed (Kass and Finin 1988, p. 51). This approach focuses on inferring user goals from a series of interactions between a user and the expert system. Sequences of user queries are used to determine implicit goals in the interaction. These goals can then be used to determine appropriate responses by the expert system. Overall, this session provided both a partial set of useful features for a user model as well as potential methods to automatically acquire some of these features.

Question Types

This session was aimed at presenting and discussing useful categorizations of explanation. It turned out to be the second most hotly debated session. The session opened with a discussion of the hypothesis that categorizing according to explanation response instead of explanation request leads to a more useful categorization (Gilbert 1988, p. 72). In this approach, explanations are grouped according to the type of knowledge they require. With this scheme, 12 categories were presented that handle a wide variety of naturally occurring explanations. With these categories intact, the problem reduces to finding methods to produce each explanation category. Most of the workshop participants agreed that this categorization provides a useful structure for explanation.

Following the more traditional method of classifying explanation queries, a second approach was discussed that uses an explicit model of the problem-solving process to clearly define the set of reasonable queries (Tanner and Josephson 1988, p. 76). For the problem of diagnosis, viewing the diagnostic process as an abductive process provides a context for the interpretation of explanation queries. The diagnostic model is used to explicitly list the potential error types that could occur during problem solving. Requests for justifying the expert system's problem solving can then be interpreted as requests for assuring that none of these error types occurred. The major concern expressed during the discussion was that such a categorization of explanation in terms of the diagnostic model does not give any information about how to answer the requests.

Nearly all the workshop participants were in agreement that categorization of both explanation queries and responses yields valuable structure to the problem of explanation generation. Representatives from

Four major components of a user model were highlighted as important in tailoring an explanation to a specific user.

industry introduced the hypothesis that the categorization of explanation queries was useful for another reason (Wexelblat 1988, p. 80). It was claimed that having an explicit list of the potential explanation queries gives the expert building the system a priori knowledge of what kinds of help might be needed by the user during problem solving. With this information, the expert is able to provide canned text that can be used in response to any of the legal-explanation queries. It was also argued that the use of canned text is cheaper because it bypasses the need to explicitly represent and encode explanation knowledge. The main objection brought up during the discussion was the inconsistency between the explanations and the operation of the expert system. The fear is that as the expert system develops and matures, the canned text will need to be updated to reflect the changes and, therefore, will not be cheaper in the long run. After much debate, it was more or less agreed that the usefulness of canned text as an explanation paradigm depends on several features of the expert system as well as the explanation being sought. Although inappropriate for dynamic explanations, canned text is at least partially acceptable for conveying "help" information to the user on the operation of the expert system or on ways to perform the instructions given by the expert system.

Other issues presented during this session were the need for explanations to be viewed as logical proofs (Bruffaerts and Henin 1988, p. 83), and the use of high-level design constraints to influence explanation (Josephson 1988, p. 87).

Text Planning II

The last session of the day was designed to address some of the issues involved in generating the text presented to the user. This session focused on how to construct the information that is presented in the final English response. Three novel approaches to text planning were discussed, each designed to help an explanation system overcome some particular shortcoming. First, an approach was described that is designed to address the problem of follow-up questions (Moore and Swartout 1988, p. 91). It was pointed out that for most explanation systems, the process of explanation is a one-shot attempt. This one-shot process is in direct contrast to human interaction in which the process of explanation is a highly interactive refinement of a final response. To address this problem, a model was proposed that builds an explicit plan of the text to be presented to the user. This plan not only includes the information which is actually presented but also information which is helpful in the analysis of follow-up questions. For example, the intent of each explanation statement is explicitly encoded so that when confusion arises, other methods which achieve the same intent might be used to replace the failing method. This model also advocates the categorization of follow-up questions in much the same way as earlier work advocates the categorization of initial explanation questions based on the query presented to the system. This categorization is then used to structure methods of updating an explanation plan to improve the answer given to the user.

A second approach presented a model designed to precisely define the intent of the user's explanation query (Rochowiak 1988, p. 95). This work hypothesizes that for every question of the form Why P? there is a contrast class of the form Why P rather than Q? Q is the contrast class in this case. It was argued that knowing the contrast class is essential to determining the type of response which can clarify the user's confusion. In this sense, the process of querying the system becomes the process of interactively defining the contrast class of the initial explanation question. Once the contrast class is established, the proposed approach uses an argumentation model to structure the text of the explanation.

The third approach advocated the use of rhetorical devices to help explanations overcome user misconceptions (Cheong and Zukerman 1988, p. 99). Here, the explanation system simulates the effects of the explanation on a model of the user using commonsense inferential rules. Once this effect is known, misconceptions in the user's new knowledge state are identified and reversed using specifically tailored rhetorical devices, such as analogy or contrast. This work focuses on the categorization of misconceptions in user knowledge to organize rhetorical devices that can be used to overcome such impairments.

This session proved to be insightful within the context of the earlier text-generation session in the workshop. Common themes emerged, such as the value of categorizing explanation (either responses or questions) and the importance of an explanation context that allows the intent and meaning of the user interaction to be better defined and used. One research area, namely, application architectures, included in the proceedings was not used as a panel session during the workshop (Garzotto et al. 1988, p. 56)

Acknowledgments

I would like to thank the organizing committee, Cecile Paris, Bill Swartout, and Bill Thompson, for its generous help with the preparation of both the workshop and this summary. Also, I am indebted to the workshop participants for their insightful discussions and probing questions. It is they who deserve the credit for the success of this workshop. Thanks to you all.

References

All citations are from the Proceedings of the 1988 AAAI Workshop on Explanation, eds. M. R. Wick, C. L. Paris, W. B. Swartout, and W. B. Thompson. Menlo Park, Calif.: American Association for Artificial Intelligence.

Michael R. Wick recently received his Ph.D. in computer science from the University of Minnesota. He has now joined the faculty at Washington State University, Computer Science Department, Pullman, WA 99164. Wick received his B.S. in mathematics and computer science from the University of Wisconsin–Eau Claire in 1984 and his M.S. in computer science from the University of Minnesota in 1986. His research interests include AI, expert systems, expert problem solving, and explanation.