MIS Quarterly
RESEARCH ARTICLE

EXPLANATIONS FROM INTELLIGENT SYSTEMS: THEORETICAL FOUNDATIONS AND IMPLICATIONS FOR PRACTICE[1]

By: Shirley Gregor
School of Computing and Information Systems
Faculty of Informatics and Communication
Central Queensland University
Rockhampton Queensland 4702
AUSTRALIA
s.gregor@cqu.edu.au

Izak Benbasat
CANFOR Professor of Management Information Systems
Faculty of Commerce and Business Administration
University of British Columbia
Vancouver, British Columbia V6T 1Z2
CANADA
izak@unixg.ubc.ca

Abstract

Information systems with an "intelligent" or "knowledge" component are now prevalent and include knowledge-based systems, decision support systems, intelligent agents, and knowledge management systems. These systems are in principle capable of explaining their reasoning or justifying their behavior. There appears to be a lack of understanding, however, of the benefits that can flow from explanation use, and how an explanation function should be constructed. Work with newer types of intelligent systems and help functions for everyday systems, such as word-processors, appears in many cases to neglect lessons learned in the past. This paper attempts to rectify this situation by drawing together the considerable body of work on the nature and use of explanations. Empirical studies, mainly with knowledge-based systems, are reviewed and linked to a sound theoretical base. The theoretical base combines a cognitive effort perspective, cognitive learning theory, and Toulmin's model of argumentation. Conclusions drawn from the review have both practical and theoretical significance. Explanations are important to users in a number of circumstances—when the user perceives an anomaly, when they want to learn, or when they need a specific piece of knowledge to participate properly in problem solving. Explanations, when suitably designed, have been shown to improve performance and learning and result in more positive user perceptions of a system. The design is important, however, because it appears that explanations will not be used if the user has to exert "too much" effort to get them. Explanations should be provided automatically if this can be done relatively unobtrusively, or by hypertext links, and should be context-specific rather than generic. Explanations that conform to Toulmin's model of argumentation, in that they provide adequate justification for the knowledge offered, should be more persuasive and lead to greater trust, agreement, satisfaction, and acceptance—of the explanation and possibly also of the system as a whole.

[1]Sirkka Jarvenpaa was the accepting senior editor for this paper.

Keywords: Explanation use, explanations, intelligent systems, knowledge-based systems, expert systems, intelligent agents, decision support systems, cognitive effort, cognitive learning

ISRL Categories: HA03, HA04, HC0101, HD01, ALO 305

Introduction

Knowledge-based (expert) systems (KBS), and intelligent systems in general, are important components of an organization's information systems portfolio (Hayes-Roth 1997; Hayes-Roth and Jacobstein 1994). While some of the initial claims about the contributions of such systems have been overstated and failures have taken place, senior managers still believe intelligent systems can contribute to organizational effectiveness and some organizations are strategically dependent on them (Gill 1995). For example, one organization has observed that an online advice-giving system that assists customers to configure their orders has improved order accuracy from 80% to over 95%, improved customer satisfaction, and reduced expenses substantially (Wanninger 1998). In the era of the Internet, a vital role is seen for systems with attributes similar to KBS as intelligent search engines and browsers for the Web. "For electronic commerce, a need exists to apply AI [Artificial Intelligence] technology for intelligent customer and vendor agents, interagent communication methods as they relate to AI in electronic commerce, and the like" (Liebowitz 1997).

This paper discusses the use of *explanations* in what we will label generically "intelligent systems" to indicate a broader focus than that of traditional KBS. The distinguishing feature of intelligent systems is that they commonly contain a knowledge component—a computerized version of human tacit and explicit knowledge. Such systems are based on the basic elements of artificial intelligence: knowledge representation, inference and control (Hayes-Roth 1997). Because of this basis, such systems are in principle capable of *explaining* to their human users both the knowledge they contain and the reasoning processes they go through.

Explanations serve to clarify and make something understandable, or are a "declaration of the meaning of words spoken, actions, motives, etc., with a view to adjusting a misunderstanding or reconciling differences" (*Macquarie Dictionary* 1981, p. 628). Two different aspects of explanations can be perceived even in this short definition. First, explanations can be *initiated by a provider* of information, with an aim of clarifying, justifying, or convincing. An explanation used in this sense may be viewed in terms of rhetoric or argumentation (Toulmin et al. 1984). Second, explanations can be *initiated by a receiver* of information to resolve misunderstanding or disagreement (Gilbert 1989; Ortony and Partridge 1987; Schank 1986).

Since the advent of advice-giving intelligent computer systems, explanation facilities have been one of their important and valued features (Berry and Broadbent 1987a; Shortliffe 1976; Stylianou et al. 1992). Explanations, by virtue of making the performance of a system transparent to its users, are influential for user acceptance of intelligent systems and for improving users' trust in the advice provided (Hayes-Roth and Jacobstein 1994). An explanatory capability is thought necessary to imitate behavior that has been found to be a characteristic of consultations with human experts (Goguen et al. 1983; Kidd 1985a, 1985b). Explanations provide information such as why certain questions were asked by the system, what some terms mean, how conclusions were reached, and why other conclusions were not reached.

Some examples illustrate the generality and currency of the topic. The developers of a patient advocate "intelligent assistant" to be delivered via the Internet found that an explanation facility was necessary to give patients information about their health conditions (Miksch et al. 1997). For an intelligent "information retrieval" tool that supported access to information resources, a form of explanation (terminological) support was found to improve the quality of the system (Brajnik et al. 1996). An "aiding function" (an explanatory capability) was found to help users achieve better performance with a statistical program (De Greef and Neerincx 1995). KBS now in wide use for government administrative and legislation-based purposes have a split screen

with explanations permanently in the right-hand side of the screen (Dayal et al. 1994).

Another group of knowledge systems for which explanation provision is potentially useful, but appear to be little used or studied, is referred to as "intelligent agents." A problem associated with these systems is trust in delegating tasks to an agent (Maes 1994). Possibly one would feel more comfortable and trusting of an agent if it is able to *explain* what it is doing and why. In addition, it is thought that artificial agents may need to communicate with each other about their knowledge and goals (Genesereth and Ketchpel 1994)—that is, give explanations. This aspect of agent behavior could be a promising area for further research on explanations.

Explanation facilities are also part of software in common daily use. Intelligent features of word processors, such as the grammar checker, have relatively primitive explanation facilities that could possibly be improved if attention was paid to some of the lessons learned from work summarized in this paper. Anecdotal evidence suggests that users of a word processor will discuss with each other at some length the reasons for a problem a grammar checker has identified ("which" instead of "that"), but do not know that the relevant stylistic convention can be obtained from the help facility, or indeed how to access the help facility to get this information. The grammar checker is a "computer program that critiques human-generated solutions" and it is thought that such systems should "provide feedback, criticism and *explanation* to the user, so the user may improve his or her solution or performance" (Silverman 1992, p. 107, emphasis added).

A further area where explanations can play a part, but one that is perhaps not yet recognized, is that of knowledge management. Knowledge management refers to a "systemic and organizationally specified process for acquiring, organizing, and communicating both tacit and explicit knowledge of employees so that other employees may make more use of it to be more effective and productive in their work" (Alavi and Leidner 1997, p. 7). The technologies associated with knowledge management include KBS. Since an aim of knowledge management systems is to communicate knowledge, the part that explanations can play in such communication, training, and learning should be recognized and investigated.

Although this paper has application in general to the kinds of intelligent systems discussed above, the majority of the relevant empirical work has been done with KBS or expert systems. The more recent types of systems, such as intelligent agents, have a basis in this earlier technology that is important to recognize so that lessons learned earlier are not lost and continuity in theoretical and empirical development is encouraged. With this goal in mind, the paper aims to provide answers, to the extent possible, to three *primary questions* concerning the importance of explanations:

1. Do users of intelligent systems want explanations? Why are explanations needed?
2. Do benefits arise from the use of explanations? What kinds of benefits?
3. What types of explanations should be provided?

In addition, a number of *subsidiary* questions are also of interest:

- When and how are explanations likely to be used in the course of advice-giving sessions? At the beginning? Throughout? After conclusions are presented?
- Are some tasks more likely to require explanations than others? Will different tasks need different types of explanations?
- Who is most likely to use explanations? Novices? Experts? Are there any other individual differences likely to affect explanation use?

Designers are unlikely to find answers in texts to these questions. At most, texts say that explanations are necessary and mention the most common forms and a few variants (for example, see Klein and Methlie 1990; Turban 1995; Zahedi 1993). The answers are most likely to be found in empirical work of which there is a reasonable inventory to provide guidelines for designers and researchers. However, we also need theories concerning KBS and other intelligent systems explanations to (1) assist in the design of explanation facilities, (2) understand users' behaviors when using explanations, and (3) identify the fac-

tors that are most promising to investigate in future empirical studies.

Our preferred approach, therefore, is to make progress toward having a top-down, theory-based approach to generate answers, instead of what we currently have: mostly a bottom-up one based on empirical studies. However, a problem in this field is that work has followed two separate streams, neither of which is theoretically based. In the "design" stream, desirable architectures and features for explanation facilities are proposed and sometimes used to construct prototype systems, but there are few or no theoretical bases and little empirical evaluation to support these proposals (Brady and Berwick 1983; Cawsey et al. 1992; Churchland 1990; Horacek 1992; Miller and Larson 1992; Paris 1987). In the "empirical" stream, studies to test and evaluate alternative explanation facilities are carried out with designs mostly based on considered opinion and wisdom, rather than on the basis of theory. These designs are primarily based on the *how* and *why* explanations inherited from a potentially promising KBS in the domain of medicine, MYCIN, with some additional features (Wick and Slagle 1989b).

Therefore, this paper first presents a synthesis of the current knowledge on explanations provided by intelligent systems, and then attempts to develop, based on the theories described in the third section, a unifying theoretical framework as the basis for the integration of empirical work. Evidence from empirical studies that have tested the predictions of the proposed theories is used to show the extent to which the theories chosen have support.

Other overviews of explanations in KBS can be found in Chandrasekaran, Tanner and Josephson (1989), Gilbert (1989), and Dhaliwal and Benbasat (1996). A special issue of *Expert Systems with Applications* (1995) dealt with the topic. Other useful collections of papers can be found in the proceedings of workshops on explanations (Brézillon 1992; Wognum 1991). None of these previous reviews, however, has integrated theoretical work and the body of empirical work that exists.

Space precludes a detailed examination of the large volume of work that relates to explanations in the human-human context. Themes in this work include (1) the nature of explanations from a philosophical point of view (Churchland 1990; Craik 1943), (2) the study and analysis of how explanations occur in conversations and consultation between people (Kidd 1985a, 1985b), (3) types of queries allowed (Gregor 1991; Hughes 1987; Lehnart 1978), (4) explanation as a social process, in particular contexts and with particular people (Goguen et al. 1983), (5) explanations prompted by expectation failures or anomalies (Schank 1986), and (6) the role of examples in explanations, particularly in the legal field (Rissland 1985).

The paper proceeds by first examining the constructs studied in empirical work. Theory proposed to account for the use of explanations in intelligent systems and associated phenomena is then analysed and evaluated in terms of evidence provided in the empirical studies. In the final section of the paper, the results of this evaluation are drawn on to answer the questions that motivated the paper and directions for future work are suggested.

Constructs in Empirical Work Relating To Explanations

Figure 1 presents an overview of the constructs used in KBS studies on explanation use and the way in which they have generally been perceived as being associated with explanation use. Appendix A provides an overview of the empirical studies reviewed in historical order. A wide search of literature in information systems, accounting information systems, computing, and artificial intelligence was undertaken to locate these studies. The criteria as to whether a study was "empirical" was that the study had to involve actual use of an intelligent system of some type, whether prototype or operational, by human users. For each study, the theoretical foundations of the study, the context, constructs studied, and results are described. A number of constructs recur in the different studies, either as determinants of explanation use (*triggers of explanation use, user characteristics*), as aspects of the explanation-use process, or as outcomes of explanation use (*performance, learning,* and *per-*

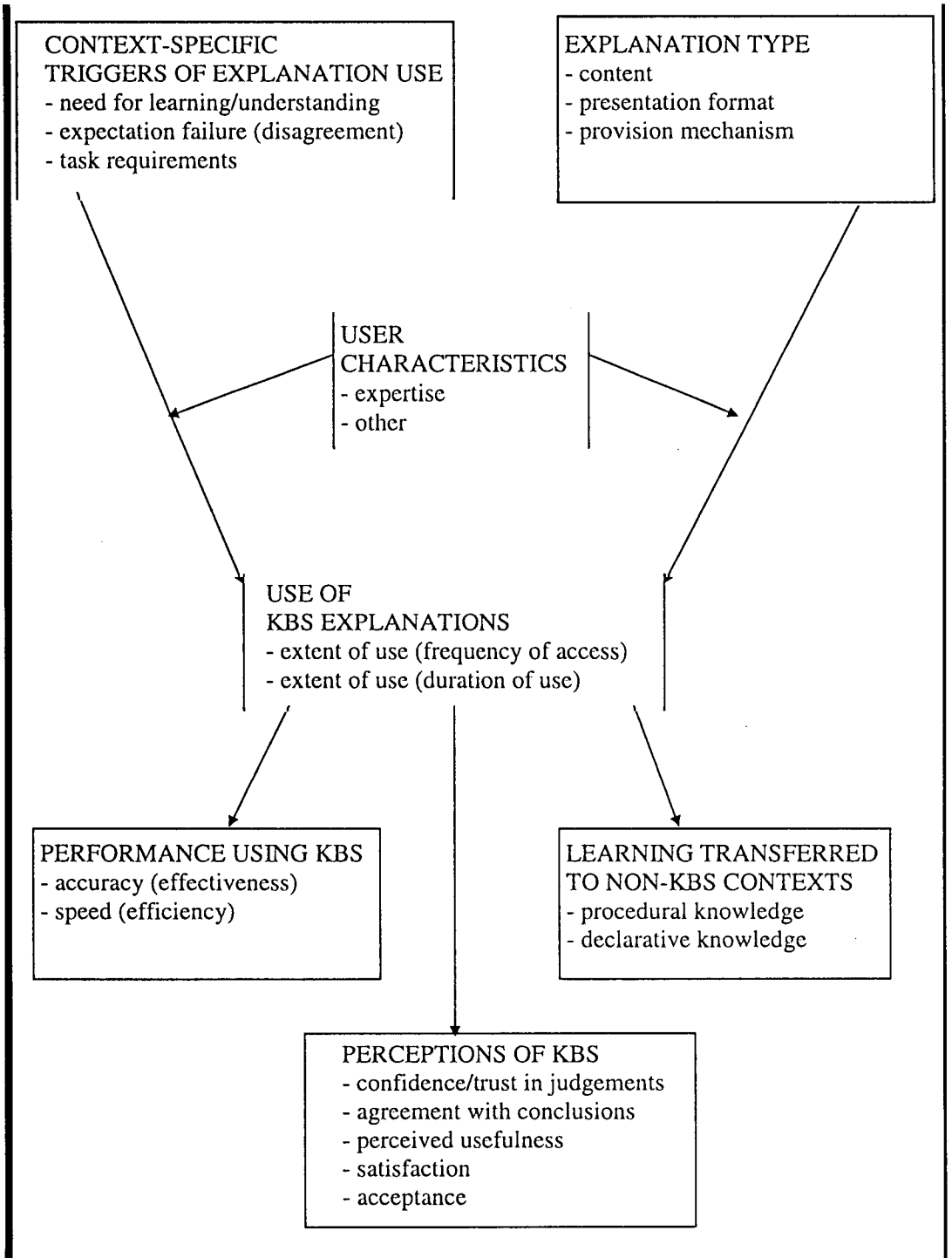

Figure 1. Constructs Associated With the Use of Explanations in KBS

ceptions). Figure 1 shows a general picture of the constructs and the relationships between them. All potential interaction effects between these constructs are not considered at this point.

It is necessary to arrive at some understanding of the constructs studied in empirical work in order for results from a number of separate empirical studies to be compared and synthesized. Although different terms and methods of operationalizing the constructs have been used, some common understanding of the constructs appears possible.

Chandrasekaran et al. (1989) provide a taxonomy for *explanation type*: basic *content*, human-computer *interface*, and *responsiveness*. The three dimensions represented in Tables 1 to 3 are similar to Chandrasekaran et al.'s classification. Table 1 shows examples of typical explanations based on the *content* of explanations. Tables 2 and 3 show how further variation in explanations can be achieved by changing the *presentation format* of explanations or the *provision mechanism*. Note the wide variety in explanations that is possible, in principle, by combining the different variants.[2] Appendix B gives a more detailed discussion of the types of explanation shown in Tables 1, 2, and 3 and their relationship to empirical work.

Context-specific triggers of explanation use identify the need for an explanation in a specific context or situation. A *learning* trigger can arise if the KBS is being used with a goal of learning or if learning is needed in the short-term so the user can contribute to problem solving. For example, the user may need a terminological explanation because he or she cannot understand a term occurring in a request for data (Gregor 1996a; Mao 1995). Learning may be more necessary with complex or uncertain tasks or when a KBS is being used in a supportive rather than a prescriptive role. In these situations, the user needs to contribute more to the problem solving process and so may need to learn more about how the KBS works. Another trigger of explanation use appears to be *lack of agreement* with a conclusion or an *expectation failure* (Dhaliwal 1993; Gilbert 1989; Mao 1995; Ye 1990).

User characteristics, expertise of the user in particular, have been investigated as an influence on explanation use. In Figure 1, we have shown user characteristics as a moderator type variable for the following reasons. There is evidence that expertise (Dhaliwal 1993; Mao 1995) and other individual characteristics, such as cognitive styles (Hsu 1993), interact with explanation provision; for example, expertise influences the content type (reasoning trace, justification), and amount, of explanations utilized. Similarly, there is evidence (see propositions P7 and P8) showing that experts react differently from novices in terms of explanation requests when they disagree with recommendations provided by a KBS and in their use of explanations for learning.

Bédard (1989) noted the difficulties in finding both a generally accepted definition for expertise and a method for operationalizing the concept. He suggests that more than one measure should be used to measure. Some researchers have operationalized it on the basis of the professional qualifications and years of experience of the users (Dhaliwal 1993; Mao 1995). Moffitt (1994) and Gregor (1996a) measured expertise by performance on a pretest that contained tasks similar to those to be later undertaken with a KBS. Lamberti and Wallace (1990) measured expertise as degree of proficiency in computer systems tasks, assessed by a questionnaire.

Use of KBS explanations is influenced by the triggers of use and by explanation provision characteristics, namely, content, format, and provision strategy. Note that there is some difficulty with the construct "use of explanations." In some studies, authors have noted that although they measure explanation access, they do not measure whether the explanation is actually read or utilized (Dhaliwal 1993; Gregor 1996a). Use is described in terms of *extent*, including the number of times explanations were accessed and the time users spent in reading explanations.

Constructs studied that relate to *outcomes of explanation* use include performance with the

[2]Other methods for classifying explanations have also been proposed. Gilbert (1989), for example, distinguished 12 different types of explanation, with a cross-classification scheme involving four different kinds of knowledge and three different levels of knowledge. Maybury (1992) provided a classification of explanatory utterances based on their content and communicative function.

Table 1. Classification of Explanations by Content Type

Type I. Trace or line of reasoning
Chandrasekaran et al.'s (1989) Type I explanations, which explain why certain decisions were or were not made by reference to the data and rules used in a particular case. How and why in MYCIN are this type.

Question: Why do you conclude that a tax cut is appropriate?
Explanation: Because a tax cut's preconditions are high inflation and trade deficits, and current conditions include those factors.

Type II. Justification or support
Chandrasekaran et al.'s (1989) Type II explanations, which justify part of a reasoning process by linking it to the deep knowledge from which it was derived.
These explanations were introduce in the Xplain system (Swartout 1983).

These explanations can be formed by attaching "deep" domain knowledge to portions of a procedure—for example, by attaching "see textbook, p. 36" to the preceding Type I explanation or a direct hypertext link to such text.

Type III. Control or strategic
Chandrasekaran et al.'s (1989) Type III explanations, which explain the system's control behavior and problem solving strategy. These explanation were introduced in NEOMYCIN (Clancy 1983).

Question: Why aren't you suggesting increased tariffs as a way of decreasing trade deficits?
Explanation: Because that plan involved political costs. My strategy is to consider politically easier plans first.

Type IV. Terminological
These explanations supply definitional or terminological information. They were distinguished by Swartout and Smoliar (1987).

Question: What is drug sensitivity?
Explanation: A drug sensitivity is an observable deviation that causes something dangerous that is also caused by the drug.

Note: Adapted from Swartout and Smoliar (1987) and Chandrasekaran et al. (1989).

KBS, learning, and perceptions of the KBS. Performance with the KBS is usually assessed by measures of accuracy, or time to complete tasks, or both. In the majority of studies, accuracy is measured in terms of degree of conformance to a prespecifed criterion. Eining and Dorr (1991), for example, compared audit evaluations made by participants with an evaluation made beforehand by an expert auditor.

The *learning* construct causes some difficulty because of the close link between learning and performance, and the fact that performance is often used as a measure of learning in non-KBS contexts (Anderson 1990). In this review, *learning* refers to long-term learning—a gain in knowledge that can be demonstrated in a context where the user solves the problem on his/her own. Studies have measured this type of learning with a posttest after a KBS is used (Eining and Dorr 1991; Gault 1994; Hsu 1993) or as the difference between posttest and pretest with intervening KBS use (Eining and Dorr 1991; Gregor 1996a; Moffitt 1994; Murphy 1990). A distinction is also drawn in some studies between the *learning of declarative knowledge* and the *learn-*

Table 2. Classification of Explanations by Presentation Format

Format	Description
Text-based	These explanations include: (1) a "rule" of the KBS in programming language, (2) a "canned text" equivalent of the rule, in a more readable form, (3) natural language.
Multimedia	Explanations can be enhanced by graphics, images, or animation. The *Expert Antenna Critic* (Silverman and Mezher 1992) showed an image of antenna placement on ships. *The Pilot's Associate* (Aretz et al. 1987) offered real-time advice to fighter pilots by voice synthesis.

Table 3. Classification of Explanations by Provision Mechanism

Provision Mechanism	Description
User-invoked	These explanations are provided at the request of the user. Selection methods include menu options, command, or hypertext links. They are also referred to as on-demand, optional, or voluntary.
Automatic	These explanations are not under the control of the user and are "always" presented. They are also referred to as "embedded" (Moffitt 1989) or "*omnipresent*" (Everett 1994).
Intelligent	These explanations are not fully under the control of the user. The KBS monitors the user in some way, perhaps building a model of the user. Such modeling allows explanations to be tailored to the user, either individually or as a member of a group (for example, novice or expert) (Clancey 1987) and explanations are provided when the KBS considers they are needed depending on the specific needs of a user at a certain point in a dialogue (Gilbert 1989). The KBS may detect user errors or omissions and provide explanations which assist with correction, or may even provide automatic correction (Carroll and McKendree 1987).

ing of procedural knowledge, to correspond with the stages of skill acquisition in cognitive learning theory (Fitts 1964).

Constructs studied that relate to *perceptions* of a KBS include confidence, trust, usefulness, satisfaction, and acceptance (Dhaliwal and Benbasat 1996). The notions of *confidence, trust,* and *belief* in a KBS, *user agreement* with the KBS conclusions, and *acceptance* of a KBS appear to be closely related, and no generally accepted, reliable scale appears to be available for any of these constructs. Lerch et al. (1997) suggest that trust in advice generated by a machine is a complex and multidimensional concept. Trust, following Rotter (1980, p. 2) is seen as "a generalized expectancy held by an individual that the word, promise, oral or written statement of another individual or group can be relied on." Rempel et al. (1985) identified three distinct and coherent dimensions of trust: predictability, dependability (confidence), and faith.

A number of studies have measured the constructs relating to perceptions of KBS with a single question answered on a Lickert-type scale. For example, Ye's (1990) measure of belief was the answer to the question "I believe the system's conclusion is true or reasonable." Gault (1994) asked, "How much confidence did you have in your answers?" Dhaliwal (1993) asked users to specify their level of agreement with each con-

clusion of the KBS. Mao (1995) and Mao and Benbasat (1998), however, provide a 10-item scale used to assess trust.

Perceived usefulness of a system is a construct investigated in a number of fields (Davis 1989; Davis et al. 1989; Moore and Benbasat 1991). Davis (1989, p. 320) defines this construct as "the degree to which a person believes that using a particular system would advance his or her job performance." Perceived usefulness is seen as a fundamental determinant of user *acceptance* of a system. A scale used for this construct has been adapted to assess the usefulness of explanations (Mao 1995; Mao and Benbasat 1998).

Theoretical Foundations: A Unifying View

In the introduction to this paper, we expressed the view that work in the field of explanations needs to be integrated and based on theory. We then discussed the constructs investigated in studies of KBS explanation use in the second section. In this section, theory is proposed to account for phenomena concerning explanation use. The theory is then used to generate propositions that are evaluated in terms of available empirical evidence (Appendix A provides an overview of empirical work that has been performed). The unified view that is presented provides the basis for answers to the questions that were the primary motivation for this paper: (1) Do users of intelligent systems want explanations? (2) Do benefits arise from using explanations? (3) What types of explanations should be provided?

We describe here the theory that we believe is most promising as a foundation for answering these questions and for further work with explanations in intelligent systems. This theory is based on aspects of cognitive psychology and human reasoning: (1) the cognitive effort perspective and the Production Paradox, (2) cognitive learning theory, and (3) Toulmin's model of argumentation. A cognitive or information-processing approach was chosen as it is reasonably well established (Best 1989) and has been used with some degree of success to predict outcomes in empirical work (see Appendix A). Cognitive learning theory and derivatives of the theory, especially those relating to expert-novice differences, have been used in a large number of studies. The cognitive effort perspective has been used in fewer studies. It appears, however, to be the only theory that offers an explanation for one of the more puzzling aspects of explanation use: the non-use or low use in some situations. Toulmin's model of argumentation—a theory of natural reasoning—has been used both in empirical studies (Gregor 1996a, 1997a, 1997b; Ye 1990, 1995) and as a basis for the design of explanation facilities (Wick and Slagle 1989b). Together these theory components offer explanations for the linkages between explanation use and the important constructs identified in Figure 1.

To answer our questions of interest and to evaluate the explanatory power and completeness of the theoretical foundation proposed, we derive a number of propositions from each theory. These propositions are stated using the constructs outlined in Figure 1 and presented in the second section. We then discuss the extent to which the propositions are supported based on the empirical evidence outlined in Appendix A. Any results counter to the propositions, or results which are not explained in terms of the theories, would suggest that the theoretical foundation is inadequate. Propositions that can be derived from the theory, but have not yet been tested, indicate areas in which further work should be considered.

A description of the three theory components follows, with the propositions derived from each and the degree of support that can be found for them in the literature. Note that the propositions are at a very general level. Constructs are operationalized differently in different studies. In some cases, the construct in a particular study may not match exactly the way in which a construct has been defined in this paper. For precise details, the original studies should be consulted.

The Cognitive Effort Perspective and the Production Paradox

The cognitive effort perspective or the Production Paradox helps us answer two of the primary questions we posed in the introduction. Propositions 1, 2, and 3 derived from these theo-

retical perspectives provide predictions associated with the following question: Do users of intelligent systems want explanations and why are explanations needed? Proposition 4 partially addresses the question: What types of explanations should be provided? These perspectives allow us to predict that there are certain circumstances in which users of intelligent systems *want* explanations and to predict the types they need; thus, they relate to the determinants of explanation use—more specifically, the independent variables "context-specific triggers" and "explanation type" in Figure 1. Propositions1, 2, and 3 also address to some extent the subsidiary question: When are explanations used? In addition, Propositions 2 and 3 address the subsidiary question: What tasks (or contexts) are more likely to lead to explanation use?

The cognitive effort perspective (Payne et al. 1993) and the Production Paradox (Carroll and Rosson, 1987) are theories that relate to limitations in human cognitive capacities. These theories are part of a tradition that includes Zipf's early work on the *Principle of Least Effort* (1949) and Simon's ideas of *bounded rationality* and *satisficing* (1955, 1956).

The cognitive effort perspective or cost-benefit principle was developed in the behavioral decision-making field, where the literature indicates that effort is an important factor in strategy selection in the decision-making process. This view is based on numerous empirical studies that are summarized in perspective by Payne et al. (1993). The cognitive effort perspective has been applied primarily to the choice of strategies in decision-making contexts, not to requirements for explanations (Todd and Benbasat 1991). The implication is, however, that users will not expend effort to access and read explanations unless the (expected) benefit of doing so is perceived to outweigh the cost of the mental effort.

A somewhat similar view is expressed in the Production Paradox or "learning versus working" argument (Carroll and McKendree 1987; Carroll and Rosson 1987). The Production Paradox refers to the conflicts between learning and working, constantly present in work settings: *learning is inhibited by lack of time and working is inhibited by lack of knowledge.* Whether requests for explanations will result in savings in cognitive resources and improvements in judgement may depend upon the usefulness and ease of use of the explanations. The motivational "cost" of learning may be reduced through the design of better explanation facilities and interfaces (Carroll and Rosson 1987). More learning may occur with the same amount of time and effort if learning is encouraged and made convenient and easy.

The general principle suggested by the cognitive effort perspective is that people in general won't use explanations without some specific reason, and anticipated benefit, as a consequence. In fact, the Production Paradox indicates that people often will not use explanations if access to explanations interfers with the goal of completing the task.

A number of propositions follow from this view. Users will tend not to use explanations unless they have a specific reason for doing so: when there is an expectation failure or anomaly, when they have an aim of long-term learning, or when they require a piece of information needed to get a task accomplished (P1, P2, P3). (These specific reasons for using explanations are consistent with aspects of cognitive psychology.) Seeking explanations because of curiosity alone could be seen as a hindrance to task accomplishment. Thus, in many situations, use of explanations will be low. Explanations will be used more, however, if they are easy to access—that is, the cognitive effort required is low (P4).

The specific propositions P1 to P4 derived from the cognitive effort perspective and the support for them follow.

P1: Explanations will be used when the user experiences an expectation failure or perceives an anomaly

Expectation failures and perceptions of anomalies have been identified as an occasion for explanations (Gilbert 1989; Schank 1986).

> Most explanations are triggered when users try, retrospectively, to account for or "understand" the system's output and find themselves either unable to do so, or able only using rules and concepts which conflict with their own beliefs (Gilbert 1989, p. 240).

From a cognitive learning perspective, Ausubel (1985) argues that when an individual cannot find a basis for reconciling apparently or genuinely contradictory ideas, he or she will (sometimes) attempt to resolve these differences so as to attempt synthesis and reorganization of his or her existing knowledge. Additional information could be sought by a request for an explanation. Ausubel suggested that the extent to which an individual attempts this reconciliation process depends on the individual's need for integrative meaning and the vigorousness of his or her self-critical faculty.

Dhaliwal (1993) found that explanations were used more when users had a moderate disagreement with the recommendations from a KBS. When disagreement was very low or very high, explanation use was less. Dhaliwal explained the result as follows. At high levels of agreement, explanations were not sought because there was no conflict between users and the KBS. On the other hand, when there was very little agreement between users and the KBS, users perceived their differences with the conclusion to be too large to reconcile, and therefore chose to ignore those conclusions without looking at the explanations. Dhaliwal noted that this inverted U-shaped relationship is common in other aspects of human information processing (Schroder et al. 1967).

Ye (1995, p. 553) concluded from a study of written comments gathered after use of a simulated KBS that

> experts were sometimes surprised by the system's conclusion. They could not recall the presence of data evidence on which the conclusion might be based, and they did not feel comfortable until they received explanations that provided the data needed.

A protocol analysis of the way in which explanations are used (Mao 1995; Mao and Benbasat 1996b) showed that generic and trace-type explanations were used for verification by experts of what they thought they already knew. An example illustrates this usage (Mao and Benbasat 1996b, p. 19). An expert read a recommendation and verbalized "*high level of debt*? . . . that's ridiculous, they've already been complaining about the fact that they are not investing enough." Apparently the recommendation was totally different from his expectation, therefore, he disagreed and requested a how explanation.

In conclusion, there appears to be support for proposition P1: explanations will be used when a user perceives an anomaly in the findings of the KBS (an expectation failure), or there is moderate disagreement between the user and the KBS. There is some evidence also that explanations are used in this way more by experts than novices.

P2: Explanations will be used more when the user has a goal of long-term learning (that is, learning that transfers to a non-KBS context).

Cognitive learning theory suggests that explanations are useful for learning, as described in the next section of this paper in connection with P6. Thus, an aim of learning is expected to be a trigger for use of explanations.

An experiment by Gregor (1996a) showed a difference in use of explanations depending on the goal of the user, whether learning or problem solving. Participants whose goal was learning used trace and justification-type explanations more than users whose goal was problem solving. Thus, there is evidence from one study to support proposition P2. No other relevant studies could be found.

P3: Explanations will be used when the user lacks knowledge needed so he or she can contribute to problem solving. The knowledge could be terminological knowledge or knowledge of a problem-solving procedure.

This further trigger of explanation use is deduced, on the basis that learning (at least short-term) is often needed when an intelligent system is used primarily for problem-solving. As for P2, an argument can be made from cognitive learning theory for the use of explanations as an aid to learning.

A user might be unable to contribute to problem solving properly, or unable to understand a KBS recommendation, if he or she cannot understand a term used by the KBS. In this case, a terminological explanation could be of assistance. Mao and Benbasat (1998) give a graphic illustration taken from a protocol analysis of the need for explanations when users encounter an unfamiliar term. Everett (1994) also found that subjects pre-

fer to invoke optional explanations only when they do not understand the KBS's question.

Users might also be unable to contribute to problem solving in cases where a KBS is being used in a supportive rather than prescriptive role, if they lack knowledge of the process behind the problem solving. Use of a system in a "supportive" role was envisaged by Luconi, Malone, and Scott Morton (1986) in their idea of "expert support systems." These systems allow a user to contribute on all four dimensions of a problem solving process: data, procedures, goals and constraints, and strategies. In contrast, a prescriptive system would allow the user less input into the problem solving process—perhaps contributing just the data. When a KBS is used in a supportive role, the user may need to choose between different constraints to enable alternative solutions to a problem to be generated, and then choose among alternatives. Gregor (1996a) showed that explanations were used more when a KBS was used in a supportive role, rather than a prescriptive role, by users who had made themselves familiar with the use of explanations in training activities. Thus, evidence supports the proposition that explanations will be used more when users lack knowledge needed for them to contribute properly to a problem solving process.

In the absence of the specific triggers for explanation use such as those specified in P1 through P3, it is possible that explanation use may be "low." It is known that there are some systems in use, apparently quite successfully, which have no explanation facilities. Thus, in at least some systems, usage is nil. Berry and Broadbent (1987a) studied the use of KBS in a number of organizations in the United Kingdom. They noted, "Despite a generally felt belief that explanations are fundamentally important, some systems are currently being developed without any explanation facility at all" (p. 22). In two cases—a route planning system and a manufacturing system—the clients had stipulated that explanations were unnecessary: they were "simply interested in systems which did the job" (p. 22). No further detail of these systems is given. It is possible that they are relatively prescriptive systems.

Explanation use has been measured in some studies. The figures given here are for explanations which are user-invoked rather than automatic. Dhaliwal (1993) found 25% to 30% of available explanations were requested. Mao (1995) observed that about 28% of the available reasoning-trace explanations and only 8% of the available deep explanations were requested. In Everett's (1994) study, only about half of the participants chose to view explanations. Only 21% of participants chose to view more than one explanation. Gregor (1996a) found with a relatively prescriptive KBS an explanation was requested, on average, 5.30 times in a 50 minute session. In a second study, an explanation was requested, on average, 1.45 times in a one hour session. Thus, explanation use is likely to be contingent upon specific triggers.

P4: Explanations that require less cognitive effort to access and assimilate will be used more and will be more effective with respect to performance, learning or user perceptions. The types of explanation for which this effect is expected include: (1) automatic (always present) explanations, (2) hypertext accessible explanations, (3) intelligent explanations (given to user automatically when system judges necessary), and (4) case-specific rather than generic explanations.

Cognitive effort is the number of elementary information processes (EIPs) that are needed to perform a task (Huber 1980; Johnson and Payne 1985; Newell and Simon 1972). An automatic explanation requires less cognitive effort to access because the user has only to read information that is already supplied on the screen. In contrast, a non-automatic explanation requires extra effort to bring the operator that is required to access explanations into short term memory (STM). A similar argument applies to an intelligent explanation that is automatically provided to the user. The user does not have to exert any effort to make it appear. In addition, an intelligent explanation can be tailored to a particular context or a particular user. In these cases, it should require even less cognitive effort because there will be less extraneous information to read. Case-specific and hypertext explanations should also require less cognitive effort because they allow the user to access needed information in the course of a consultation—information that applies directly to the data that is in STM. In contrast some generic explanations, perhaps

accessed before a consultation begins, will require the user to store extraneous information in long-term memory, resulting in additional effort for storing and fetching. Note that there are other aspects of the design of explanations that will affect the degree of cognitive effort expended in accessing and assimilating them. For example, the use of unfamiliar terms in a message may mean more effort to retrieve a meaning from long-term memory. These considerations are common to other areas of computer interface design and are too numerous to include in full in this paper.

The increased effectiveness of automatic explanations (P4a) has been demonstrated by Everett (1994) and Moffitt (1994). Everett found that subjects who always received explanations indicated lower perceived frustration with explanations. Moffitt (1994) found that learning was greater with automatic explanations compared with non-automatic explanations.

The increased effectiveness of hypertext accessible explanations (P4b) has been shown by Mao (1995). In this study, when hypertext was used to access explanations in the context of KBS output (recommendations and other explanations), deep explanations were used more and were more effective than other types of explanations in enhancing knowledge transfer from the KBS to the users. In contrast, in the study by Gault (1994), hypertext explanations were not found to be superior to rule-trace or fixed-text explanations.

No empirical studies can be found concerning the use and relative effectiveness of intelligent-type explanations (P4c).

The increased effectiveness of case-specific explanations (P4d) has been shown by Berry and Broadbent (1987b) and Dhaliwal (1993). Berry and Broadbent found that multiple case-specific explanations led to better performance than a single general explanation given at the beginning of a session with an advice-giving system. Dhaliwal found that feedback (case-specific) explanations were used more than feedforward (generic) and that feedback explanations improved the accuracy of decision making. Mao (1995) found that deep explanations provided from within the context of case specific recommendations and reasoning traces led to higher learning on the part of novice subjects.

To summarize, there is support for a cognitive effort perspective on the use of explanations. There is support for all four propositions derived from this perspective and no substantive evidence counter to any of the propositions. It appears that explanations are not necessarily accessed as a matter of course or general curiosity: a specific trigger is needed. In addition, the amount of cognitive effort required to access a particular type of explanation will affect how likely it is to be used and be useful.

Cognitive Learning Theory

Four propositions are derived from cognitive learning theory. Propositions 5 and 6 provide a basis for answering the question: Do benefits arise from the use of explanations and what kinds of benefits? Propositions 6 and 7 allow predictions to be made about benefits and how they are moderated by the expertise level of the users, thus addressing one of our subsidiary questions concerning the effect of user differences on explanation use. The four propositions, P5 to P8, relate primarily to the outcomes of explanation use—more specifically, the dependent variables of learning and performance identified in Figure 1.

Theories of learning have been used by many KBS researchers because of expected relationships between explanations and learning (Eining 1988; Eining and Dorr 1991; Gregor 1996a, 1997a; Hsu 1993; Mao 1995; Mao and Benbasat 1996b; Moffitt 1994, 1984; Murphy 1990). The cognitive approach to learning (Anderson 1983; Ausubel 1968; Schuell 1986) is one of several approaches in the educational field. Others are the earlier behavioral approaches (Skinner 1985) and the more recent constructivist approach (Cooper 1993). The different theories of learning emphasize different aspects of learning and to some extent are complementary.

The cognitive (information processing) approach to learning is based on the conception of short-term and long-term memory and the way knowledge is organized in memory (Ausubel 1968). This approach emphasizes the distinction between declarative and procedural knowledge

and shows stages in skill acquisition in these terms (Anderson 1990). Declarative knowledge is knowledge of facts: "knowing that." Procedural knowledge is knowledge of a skill: "knowing how." Fitts (1964) describes how skill learning occurs in stages. In the initial "cognitive" stage, a set of facts is learned, which may include a description of a procedure. In a second "associative" stage, the declarative information is transformed into a procedural form. Errors in initial understanding are eliminated and the elements necessary to perform the task become more strongly connected. In a third "autonomous" stage, the procedure becomes more automated and rapid and the ability to verbalize knowledge of the skill may be lost. Anderson (1982, 1983) built on these ideas to form a theory known as adaptive control of thought (ACT). ACT is based on the idea that elements of permanent memory are stored in propositional networks. Among several later developments is ACT*, a system that simulates language and skill acquisition.

The specific propositions P5 to P8 derived from cognitive learning theory and the support for them follow.

P5: Use of explanations improves the performance achieved with a KBS as an aid.

Explanations are expected to aid performance primarily because they can assist users with the understanding of unfamiliar terms and requests during data input, and thus lead to greater accuracy of input. This assistance is more likely to be required by novices than experts, since the cognitive theory of skill acquisition shows that in early stages of knowledge acquisition, declarative knowledge of terms and procedures are incomplete. In addition, with less prescriptive systems, explanations can help the user better understand what the KBS is doing, so that the "collaboration" between user and system is more effective. In short, explanations aid learning (at least short-term), which is reflected in improved performance.

Several studies support proposition P5. Wognum (1990) found that explanations improved decision making in a pencil-and-paper study. Dhaliwal (1993) found that the use of feedback (reasoning trace) explanations improved the accuracy of decision making. Gregor (1996a) found the use of terminological explanations was related to improved problem solving performance with a relatively prescriptive KBS. In her second study, she found that use of explanations of all types was related to improved problem solving performance. Mao (1995) found that increased use of deep explanations led novice subjects to make judgements that were similar to those of the experts who contributed their knowledge to the development of the KBS. De Greef and Neerincx (1995) found that an "aiding" interface improved performance with a statistical program.

Some studies appear contrary to the proposition. Gault (1994) found no difference in performance (accuracy) between groups with explanations and a control group without explanations. Note that in Gault's study, there was no measure of how many times explanations were accessed. Similarly, in Gregor's (1996a) second study, there was no difference between groups with and without explanations in terms of performance. Differences were only observed in relation to the number of times explanations were accessed in the treatments where explanations were available. Thus, it appears there is support for the proposition as it refers to a relationship between performance and the amount of *use* of explanations. There does not appear to be evidence for a similar relationship between the *availability* of explanations and improved performance.

P6: Use of explanations aids learning (transfer of knowledge to non-KBS contexts).

Explanation use should contribute to long-term learning that transfers to non-KBS contexts, as well as contributing to short-term learning for improved problem solving (P6). Explanations can supply the declarative knowledge that is needed in the first cognitive stage of skill acquisition. By explaining unfamiliar terms and procedures, explanations can allow new knowledge to be better assimilated with existing knowledge structures (Ausubel 1985). This process is particularly important when there are discrepancies or anomalies between the KBS and the user's prior knowledge or expectations.

The evidence for proposition P6 is inconclusive. There is evidence for a learning-by-doing effect,

where repeated problem solving with a KBS leads to improved problem solving when an individual has to solve a problem without the aid of the KBS (Eining and Dorr 1991; Fedorowicz et al. 1992; Gregor 1996a). It is not clear, however, that explanations add appreciably to this learning-by-doing effect.

It may be that this expected relationship, if it does exist, is subtle and difficult to detect. Careful examination of previous studies tends to support this view. A number of studies have failed to detect any difference in learning between KBS groups with and without explanations (Eining and Dorr 1991; Murphy 1990). Gregor (1996a) found no relationship between the use of explanations and the amount learned. Studies which have found a link between the availability of explanations and learning did so under rather extreme circumstances. Moffitt (1994) found that a group with one type of explanation learned more than a control group. The control group had no learning treatment at all (no KBS and no other aid), so this result is hardly surprising. Gault (1994) also found that groups with particular types of explanations learned more than a control group without explanations. In this case, the control group had a KBS but no explanation facility. There was no way for them to learn the declarative knowledge that was tested in the posttest. Procedural learning was not tested. The study by Mao (1995) is not directly comparable. Mao found that use of "deep" (domain knowledge) explanations improved knowledge transfer, but here the emphasis was on the degree to which the KBS led the user to adopt the KBS advice in one particular case. The KBS used was a simulation and the users did not enter data or perform procedures themselves, so did not in fact "practice" the problem with the KBS.

The study by Everett (1994) is difficult to evaluate in the terms used in this review. His experiment showed learning did occur in many groups. It is difficult to distinguish, however, what could be attributed to explanations and what to learning-by-doing. Everett found that subjects who invoked explanations in optional-explanation conditions showed more declarative learning. He also found that procedural knowledge gain was greater in an automatic-explanation condition than in an optional-explanation condition. The study by De Greef and Neerincx (1995), with a statistical program and two versions of an interface, an "aiding" interface and a "plain" interface, showed that use of the aiding interface led to greater knowledge gains with a group of students with some prior statistical knowledge.

Thus, the evidence for proposition P6 must be regarded as inconclusive at this point.

P7: Novices will use explanations more for learning (short- and long-term) than experts.

P8: Experts will use explanations more for resolving anomalies (disagreement) and for verification than novices.

The view of the development of expertise in cognitive learning theory allows a number of predictions to be made about expert-novice differences with respect to the types of explanations used and the reasons for explanation use. It is difficult to make general statements about whether novices will use *more* explanations overall than experts. Arguments from the cognitive effort perspective showed how explanation use is highly context-dependent. The context is expected to be more important overall than expert-novice differences. If there are many occasions for perceptions by experts of anomalous output then they may use more explanations than novices. If novices are required to use the KBS for learning, they may use explanations more than experts. Nevertheless, some general predictions concerning expert-novice differences can be made. As novices have more to learn, we would expect them to make greater use of explanations for learning than experts (P7).

Johnson (1983) defines an expert as a person "who, because of training and experience, is able to do the things that the rest of us cannot, experts are not only proficient but also smooth and efficient in the actions they take" (p. 78). Anderson (1990) describes a number of dimensions in the development of expertise: the conversion of declarative, factual knowledge into more efficient procedural representations, tactical and strategic learning that leads to faster and better problem solving, abstract rather than surface-level representations, and improved memory and memory structures in the domain of expertise.

Experts should be more able to understand terms used by the KBS and also, in many cases, per-

form the tasks performed by the KBS themselves. They may use the KBS to assist them with tedious calculations and avoid errors, but are capable of forming some judgement as to the accuracy of the KBS conclusion. Thus, experts will use explanations more for resolving perceived anomalies in the KBS output. The explanations they will use to do this are reasoning trace, justification, and control type explanations. Wognum (1990, p. 122) noted that

> More experienced users may use the system in an informative role. In this role users are allowed to reject a conclusion. The users are considered to be superior to the system. They are familiar with the knowledge in the system and know how to use it.

There is support for propositions P7 and P8. Mao and Benbasat (1996b) found that novices used more deep, but not reasoning-trace, explanations compared to experts. Experts mainly used deep knowledge and reasoning-trace for verifying conclusions against their own knowledge. Experts were more likely to identify potential inconsistencies in the KBS output and to resort to explanations to resolve the differences in judgement. Novices were more likely to request explanations for learning. Ye (1995) also noted that experts were more likely to look at explanations because they were surprised by conclusions.

In summary, there appears to be support for a number of propositions taken from cognitive learning theory. Explanations aid performance and in some cases learning. Experts will use explanations more for resolving anomalies and novices more for learning.

Toulmin's Model of Argumentation

Toulmin's model of argumentation (Toulmin 1958; Toulmin et al. 1979), a model of human reasoning, provides a basis for answering one of our primary questions—What types of explanations should be provided?—and links it to another primary question concerning the types of benefits that arise from the use of explanations. Proposition 9 relates to the design of explanation facilities and outcomes of explanation use—more specifically, the independent variable "explanation type" and the dependent variable "perceptions of KBS" in Figure 1. The latter construct includes user perceptions of confidence, trust, agreement with conclusions, perceived usefulness, satisfaction, and acceptance.

Toulmin's model has been used as a basis for constructing explanation capabilities (Miller and Larson 1992; Wick and Slagle 1989a) and in empirical work (Gregor 1996a; Ye 1990, 1995; Ye and Johnson 1995). The model provides a basis for the examination of practical reasoning and argumentation, as distinct from formal logic. The model distinguishes the following different parts of an argument:

- *claims:* the assertions or conclusions that are put forward for acceptance,
- *grounds:* the factual data that is the foundation for the argument,
- *warrants:* the justification for moving from the grounds to the claims (examples are rules of thumb and laws of nature),
- *backing:* the authorization for the warrant (an example is a legal statute),
- *qualifiers:* phrases expressing the degree of certainty placed on a claim,
- *possible rebuttals:* the extraordinary or exceptional circumstances that might undermine the force of the argument.

Arguments that are strong and well-founded are thought to be convincing, while others that are weak or baseless are unconvincing. The model can be applied to explanations in KBS. A rule-trace explanation, which has a rule with data premise, certainty factor, and conclusion, corresponds to the grounds, qualifier, and claim in Toulmin's model. In justification-type explanations, a warranty and possibly a backing will also be distinguished.

Explanations that conform to Toulmin's model should be more persuasive because they contain the elements that are present in convincing human-human arguments. Thus, they should lead to greater trust, agreement, satisfaction, and acceptance.

P9: Use of explanations conforming to Toulmin's model (justification explanations) will give rise to more positive user perceptions of a KBS than other explanations (trace and strategic explanations).

Some evidence for this proposition may be discerned in the perceived inadequacy of early sys-

tems, such as MYCIN, that had only rule-trace explanations (Buchanan and Shortliffe 1984). Considerable support is also found in the work by Everett (1994) and Ye (1990).

Everett (1994) found that confidence, satisfaction, and perceptions of usefulness, effectiveness, and ease-of-use were all strongly affected by the presence of justification-type explanations. Frustration with the KBS also decreased when automatic justification-type explanations were given.

> The effect of including justification explanations was so strong that it affected perceptions which should not have differed depending on their presence or absence, notably system ease of use. Although there was absolutely no difference in any aspect of ease of use, subjects receiving justification explanations rated the system significantly easier to use than did those not receiving justifications. All of the justification effects were significant regardless of invocation mode or explanation content treatment (Everett 1994, p. 307).

Ye (1995, p. 553) found "justification to be more effective than trace and strategy in getting the system's conclusions accepted, as evidenced by its higher perceptual value (usefulness and preference) and, more importantly, its higher usage rates (choice of explanation). Participants' informal comments also provided support for their discrete usage patterns, as a number of them suggested that they would always want to see the justification for a conclusion."

In summary, there appears to be a considerable degree of support for proposition P9.

To conclude this section we need to evaluate the success of the three theories we have suggested as a unifying foundation for work with explanations in KBS. Table 4 gives a summary of the propositions discussed and associated empirical studies for each. The propositions drawn from this theory appear to be largely supported by empirical evidence with the exception of proposition P6, which stated that the use of explanations is expected to aid long-term learning that transfers to non-KBS contexts. The evidence for this proposition is equivocal and it was concluded that this effect, if it does exist, is probably subtle and difficult to detect.

We need also to consider whether there have been empirical findings outside of the theoretical framework we have proposed. Examination of Figure 1 and the studies in Appendix A shows that there is one area in which the theoretical foundation could possibly be deficient. It is possible that individual differences, apart from expert-novice differences, could influence explanation use. In general, however, attempts to find other individual differences related to the use of explanations have been unsuccessful. Gault (1994) studied the user's attributional style as a possible determinant of explanation use, but found no significant effect.

The only study that shows a relationship between individual differences and explanation use is that of Hsu (1993). He found that cognitive style was related to knowledge transfer (learning) with a KBS. In addition, *field-independents,* as measured by the GEFT scale of Oltman, Raskin, and Witkin (1971), were more affected by different explanation types than *field-dependents.* Field-independents learned better with flexible (user-invoked) and justification explanations than they did with rule-trace explanations. Hsu related cognitive style to cognitive restructuring skills, and thus to cognitive learning theory (Anderson 1983). Further work in this area appears warranted.

Various other individual differences have been included in studies as covariates but none have been found to be significant when the influence of other constructs is taken into account. The covariates examined include age, gender, computer experience (Eining 1988; Everett 1994; Gregor 1996a; Murphy 1989) and need-for-cognition (Gregor 1996a).

We conclude that the theoretical foundation proposed is reasonably adequate as an aid for understanding and predicting phenomena relating to the use of explanations in intelligent systems.

Discussion And Conclusions

In the preceding section, we proposed a combination of the cognitive effort perspective, cognitive learning theory, and Toulmin's model of argumentation as a unifying foundation for work with explanations in intelligent systems. Analysis

Table 4. Propositions Derived From Theory and Relevant Empirical Studies

	Proposition	Studies[a]
From the cognitive effort perspective		
P1	Explanations will be used when the user experiences an expectation failure, or perceives an anomaly.	Dhaliwal (1993), Mao and Benbasat (1996b), Ye (1995).
P2	Explanations will be used more when the user has a goal of long-term learning (learning that transfers to a non-KBS context).	Gregor (1996a).
P3	Explanations will be used when the user lacks knowledge needed (terminological knowledge or problem-solving procedures) so he or she can contribute to problem solving.	Everett (1994), Gregor (1996a), Mao (1995).
P4	Explanations that require less cognitive effort to access and assimilate will be used more and will be more effective with respect to performance, learning, or user perceptions. The types of explanation for which this effect is expected include:	
a	automatic (always present) explanations,	Everett (1994), Moffitt (1989).
b	hypertext accessible explanations,	Gault (1994), Mao (1995).
c	intelligent explanations (given to user automatically when system judges necessary),	No empirical tests found.
d	case-specific rather than generic explanations.	Berry and Broadbent (1987b), Dhaliwal (1993).
From cognitive learning theory		
P5	Use of explanations improves the performance achieved with a KBS as an aid.	De Greef and Neerincx (1995), Dhaliwal (1993), Gregor (1996a) Mao (1995), Wognum (1990).
P6	Use of explanations helps in learning (transfer of knowledge to non-KBS contexts).	Differing results: De Greef and Neerincx (1995), Eining (1988), Everett (1994), Gault (1994), Gregor (1996a), Moffitt (1989), Murphy (1990).
P7	Novices will use explanations more for learning (short- and long-term) than experts.	Mao (1995).
P8	Experts will use explanations more for resolving anomalies (disagreement) and for verification than experts.	Mao (1995), Ye (1990).
From Toulmin's model		
P9	Explanations conforming to Toulmin's model (justification explanations) will give rise to more positive user perceptions of a KBS than other explanations (trace and strategic explanations).	Everett (1994), Ye (1990).

[a]Only primary references are given for supporting studies.

Note: All propositions refer to a context in which a KBS is used, unless stated otherwise. That is, propositions compare KBS use with explanations to KBS use without (or with fewer or different) explanations. The propositions do not compare a KBS situation with a non-KBS situation.

of empirical work showed that this theoretical foundation appears to be reasonably adequate and complete. In this concluding section we use the theoretical foundation to answer the questions that motivated the paper. The answers to these questions should be of benefit to designers of intelligent systems.

1. *Do users of intelligent systems want explanations? Why are explanations needed?*

It appears that explanations should be provided in intelligent systems, despite the low use observed in some situations. This low use may be, at least in part, occasioned by the desire to avoid expending cognitive effort and the Production Paradox. The occasions on which users want explanations are likely to be highly context-specific. These occasions include the need to resolve perceived anomalies, the desire to learn on the part of the user, and a lack of knowledge of the terms or procedures used by the intelligent system. Particular tasks such as report production or debugging may also necessitate the use of an intelligent-system explanation.

2. *Do benefits arise from the use of explanations? What kinds of benefits?*

Explanation use has been shown to have positive outcomes—better performance, higher user perceptions of the system, and, in some cases, improved learning.

3. *What types of explanation should be provided?*

The explanations that are needed in intelligent systems should be considered in terms of the three classification methods shown in Tables 1 to 3: content, presentation format, and provision mechanism.

With respect to the content of explanations, it appears, congruent with Toulmin's model of argumentation, that justification-type explanations are particularly efficacious. This effect has been shown empirically by Everett (1994). In addition, terminological explanations appear to be generally useful. Gregor (1996a) found with a prescriptive system terminological-type explanations were the type significantly related to performance. Wognum (1990) found in her study of operational systems that users had actually demanded a terminological-type explanation function in one system and use of this facility reduced the use of other explanation facilities. Note that a terminological-type explanation function may not be provided for in KBS shells. Mao (1995) found deep (terminological) explanations to be more useful than reasoning-trace explanations for novice users to acquire the knowledge contained in the KBS.

In terms of presentation format, there is little empirical evidence as to the relative worth of different methods available (text, graphics, sound). It may be that general rules for interface design are at present the best guide available for choice of presentation method.

With respect to the provision mechanism, methods that reduce the cognitive effort needed to access the explanations are desirable. There is compelling evidence for the advantages of automatic provision of explanations compared with user-invoked explanations. In Everett's words (1994, p. 308),

> being provided an explanation and being provided the opportunity to invoke an explanation are not identical in the perception of an expert system user. It is possible that the perceived effort of requesting an explanation, even through a single keystroke, might be sufficient to discourage explanation requests.

Everett concluded that any important explanations during the course of the expert system consultation should *always* be presented to the subject.

Hypertext appears also to offer advantages, such as reducing the cost of accessing explanations, thus increasing their use (Mao 1995). There is little empirical evidence for the worth of intelligent explanations, although it is expected that they would also offer advantages.

The work reviewed also allows some subsidiary questions to be addressed.

When and how are explanations likely to be used in the course of advice-giving sessions? At the beginning? Throughout? After conclusions are presented? When there is an "expectation failure"?

Terminological-type explanations are likely to be used throughout a problem solving session for assistance with data input. Reasoning trace, justification, and control explanations, particularly

justification-type explanations, are likely to be used at the end of a consultation, to resolve expectation failures and perceptions of anomalies, or for learning.

Are some tasks more likely to require explanations than others? Will different tasks need different types of explanations?

Reasoning trace, justification, and control explanations are likely to be used more if the "task" is learning rather than problem solving. Use of KBS in a supportive rather than a prescriptive role may also mean greater use of explanations (Gregor 1996a). Specific task requirements, such as report production or debugging, can also necessitate explanations. Wognum's (1990) study of use of operational systems noted that explanations were found useful for report production.

Who is most likely to use explanations? Novices? Experts? Are there any other individual differences likely to affect explanation use?

There appear to be expert-novice differences in the use of explanations (Mao 1995; Ye 1990). Novices use explanations more for learning and understanding, experts more for verification. Many other individual differences have been studied in relation to explanations without significant results. The only relationship found was between cognitive style and learning: field-independents showed increased learning with user-invoked and strategic explanations (Hsu 1993).

To conclude, there is now a body of empirical work relating to the use of explanations in KBS that shows a considerable degree of convergence. Explanations, when available, are not used to the degree that might be expected. If explanations are used, they can result in improved performance, more positive user perceptions, and in some cases, long-term learning. The degree to which explanations are used appears to be related to the effort that needs to be expended in accessing them. Thus, explanations that require less effort to access, particularly automatic and hypertext explanations, are likely to be most efficacious. Terminological and justification explanations also appear to be particularly effective. The results observed appear to be congruent with the theoretical base proposed, which combines a cognitive effort perspective, cognitive learning theory, and Toulmin's model of argumentation.

Further work is suggested in areas where few empirical investigations have been undertaken. The first is the investigation of the usefulness of intelligent-type explanations, which are given to the user automatically when the KBS judges they are necessary, and may involve user modeling. Second, there appears to have been little work on the relative merits of comparatively novel presentation formats such as multimedia. Third, work could be extended to some of the newer types of systems, such as intelligent agents and knowledge management systems. Fourth, the relationship between the use of explanations and long-term learning that can be transferred to other contexts (Proposition 6) has not been clearly established due to equivocal results from empirical studies. This important issue needs to be explored further. Finally, we suggest that there is a greater payoff in comparing explanation facilities with different features (for example, with and without intelligent explanations) within the same study, rather than making gross comparisons between intelligent systems that do or do not have explanation facilities.

In addition, further work could investigate the relevance of broader theoretical perspectives to the use of explanations. This paper has proposed a unifying theoretical foundation based on a cognitive psychology approach, which focuses on the use of explanations on a individual basis. It would be possible also to consider whether social, ethical, or organizational theories provide additional insight into the use of explanations. Certainly in some contexts an explanation is required to fulfil a legal or reporting requirement. For example, *ExperTAX*, a system for audit and tax planning, has simple, brief explanations that satisfy a reporting requirement (Shpilberg et al. 1986). Only one empirical study of the use of explanations in an organizational setting was identified (Wognum 1990). Apart from use of explanations for report generation, Wognum found that explanations were used as a basis for negotiation with clients. Further work in a broader context appears to be warranted, as Goguen et al. (1983) concluded from a study of naturally occurring explanations that:

> Explanation is a social process, in the sense that explanation actually occurs in particular social contexts involving particular people having their own particular assumptions and dispositions, which in turn significantly influence how the explanation is actually presented and understood (p. 553).

Overall, the conclusion for the practical manager or developer of information systems is that much greater attention should be paid to the inclusion of explanations in any system that has an "intelligent" component. These are systems which contain knowledge in some form, whether it be the meaning of a term, the reasons for advising a particular course of action, or the justification for a particular piece of knowledge. Today these systems could be referred to as knowledge systems, intelligent agents, intelligent assistants, or critiquing systems, as well as the more familiar decision support, expert, or knowledge-based systems. Attention should be paid to the inclusion of justifications or backing for knowledge fragments when they are added to organizational knowledge management systems.

It should also be realized that what some people regard as "help" can have an explanatory capability. For example, the help function in a grammar checker may provide a grammatical rule as justification for its advice. The grammar checker in a word processor could automatically offer an explanation of a grammatical rule in a position of the screen where it does not interfere with current work. The work reviewed in this paper suggests that such a feature would lead to a greater knowledge of grammatical rules and better use of these rules. Our observations are that users in general either do not know this help feature exists, or find it too much trouble to access, and will as a consequence continue in ignorance of why the systems behaves as it does. They may even disable the grammar checker because of dissatisfaction and the lack of understanding of its actions. Thus, a deficiency in the design of an explanation facility can lead to a number of undesirable outcomes.

The reasons for including explanations in intelligent systems are that they have been clearly shown, when suitably designed, to improve performance, learning and result in more positive user perceptions of a system. The design is important, however, because it appears that explanations will not be used if the user has to exert "too much" effort to get them. Explanations could be provided automatically if this can be done relatively unobtrusively, or by hypertext links. Designers of explanation facilities and help functions should heed the results reported here.

Acknowledgements

The assistance of Professor Ron Weber in foundation work for this review is acknowledged as is support from the Natural Sciences and Engineering Research Council of Canada.

References

Adelson, B. "Problem Solving the Development of Abstract Categories in Programming Languages," *Memory and Cognition* (9:4), 1981, pp. 422–433.

Alavi, M., and Leidner, D. "Knowledge Management Systems: Emerging Views and Practices from the Field," INSEAD Working Paper Series, Fontainebleau, France, 1997.

Anderson, J. R. "Acquisition of Cognitive Style," *Psychological Review* (89:4), 1982, pp. 369–406.

Anderson, J. R. *The Architecture of Cognition,* Harvard University Press, Cambridge, MA, 1983.

Anderson, J. R. *Cognitive Psychology and Its Implications,* 2nd ed., W. H. Freeman, New York, 1985.

Anderson, J. R. *Cognitive Psychology and Its Implications,* 3rd ed., W. H. Freeman, New York, 1990.

Aretz, A. J., Hickox, J. C., and Kesler, S. R. "Dynamic Function Allocation in Fighter Cockpits," *Proceedings of the Human Factors Society Thirty-first Annual Meeting,* 1987, pp. 414–418.

Ausubel, D. "Learning as Constructing Meaning," in *New Directions in Educational Psychology,* N. Entwhistle (ed.), Falmer Press, London, 1985, pp. 71–82.

Balzer, W., Doherty, M. E., and O'Connor, R. "Effects of Cognitive Feedback on Performance," *Psychological Bulletin* (106:3), 1989, pp. 410–433.

Bandura, A. "Human Agency in Social Cognitive Theory," *American Psychologist* (44:9), 1989, pp. 1175–1184.

Bédard, J. "Expertise in Auditing: Myth or Reality?" *Accounting Organizations and Society* (14:1/2), 1989, pp. 113–131.

Berry, D. C., and Broadbent, D. E. "Expert Systems and the Man-machine Interface. Part Two: The User Interface," *Expert Systems* (4:1), 1987a, pp. 18–27.

Berry, D. C., and Broadbent, D. E. "Explanation and Verbalization in a Computer-assisted Search Task," *Quarterly Journal of Experimental Psychology* (39A), 1987b, pp. 585–609.

Best, J. B. *Cognitive Psychology,* West Publishing Co., St. Paul, MN, 1989.

Brady, M., and Berwick, R. C. (eds.). *Computational Models of Discourse,* MIT Press, Cambridge, MA, 1983.

Brajnik, G., Mizzaro, S., and Tasso, C. "Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support, *SIGIR'96,* Zurich, 1996, pp. 128–136.

Brézillon, P. (ed.). *Proceedings of the ECAI-92 Workshop W15 "Improving the Use of Knowledge-based Systems with Explanations,"* Institut Blaise Pascal, Paris, June 1992.

Bruner, J. S., Goodnow, J. L., and Austin, G. A. *A Study of Thinking,* Wiley, New York, 1956.

Brunswik, E. *The Conceptual Foundation of Psychology,* University of Chicago Press, Chicago, 1952.

Brunswik, E. *Perception and the Representative Design of Experiments,* University of California Press, Berkeley, CA, 1956.

Buchanan, B. G., and Shortliffe, E. H. *Expert Systems,* Addison-Wesley, Reading, MA, 1984.

Carroll, J. M., and McKendree, J. "Interface Advice Issues for Advice-giving Expert Systems," *Communications of the ACM* (30), January 1987, pp. 14–31.

Carroll, J. M., and Rosson, M. B. "Paradox of the Active User," in *Interfacing Thought,* J. M. Carroll (ed.), 1987, pp. 81–111.

Cawsey, A., Galliers, J., Reece, S., and Sparck Jones, K. "The Role of Explanations in Collaborative Problem Solving," in *Proceedings of the ECAI-92 Workshop W15 "Improving the Use of Knowledge-based Systems with Explanations,"* P. Brézillon (ed.), Institut Blaise Pascal, Paris, June 1992, pp. 97–106.

Chandrasekaran, B., and Mittal, S. "Deep versus Compiled Knowledge Approaches to Diagnostic Problem-solving," *International Journal of Man-Machine Studies* (19), 1983, pp. 425–436.

Chandrasekaran, B., Tanner, M. C., and Josephson, J. R. "Explaining Control Strategies in Problem Solving," *IEEE Expert,* Spring 1989.

Chi, M. T. H., Feltovich, P. J., and Glaser, R. "Categorization and Representation of Physics Problems by Experts and Novices," *Cognitive Science* (5:1), January–March 1981, pp. 121–152.

Churchland, P. M. "On the Nature of Explanation: A PDP Approach," *Physica D* (42), 1990, pp. 281–292.

Clancey, W. J. "The Epistemology of a Rule-based Expert System: A Framework for Explanation," *Artificial Intelligence* (20:3), 1983, pp. 215–251.

Clancey, W. J. "Heuristic Cassification," *Artificial Intelligence* (27), 1985, pp. 289–350.

Clancey, W. J. *Knowledge-based Tutoring: The GUIDON Program,* MIT Press, Cambridge, MA, 1987.

Cooper, P. "Paradigm Shifts in Designed Instruction: From Behaviourism to Cognitivism to Constructivism," *Educational Technology,* May 1993, pp. 12–19.

Craik, K. *The Nature of Explanation,* Cambridge University Press, Cambridge, England, 1943.

Davis, F. "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly* (13:3), 1989, pp. 319–340.

Davis, F., Bagozzi, R. P., and Warshaw, R. P. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models," *Management Science* (35), 1989, pp. 982–1003.

Dayal, S., Johnson, P., and Mead, D. "Natural Language: An Appropriate Knowledge Representation Scheme for the Administrative Domain," in *Proceedings of the Second World Congress on Expert Systems,* J. Liebowitz (ed.), Cognizant Communication Corp., Elmsford, NY 1994.

De Greef, H. P., and Neerincx, M. A. "Cognitive Support: Designing Aiding to Supplement

Human Knowledge," *International Journal of Human-Computer Studies* (42), 1995, pp. 531–571.

Dhaliwal, J. S. *An Experimental Investigation of the Use of Explanations Provided by Knowledge-based Systems,* Unpublished Doctoral Dissertation, University of British Columbia, 1996.

Dhaliwal, J. S., and Benbasat, I. "The Use and Effects of Knowledge-based System Explanations, Theoretical Foundations and a Framework for Empirical Evaluation," *Information Systems Research* (7:3), 1996, pp. 342–362.

Einhorn, H. J. "Expert Judgement: Some Necessary Conditions and an Example," *Journal of Applied Psychology* (59), 1974, pp. 562–571.

Eining, M. *The Impact of an Expert System as a Decision Aid on Learning During the Audit Process: An Empirical Test,* Unpublished Doctoral Dissertation, Oklahoma State University, 1988.

Eining, M., and Dorr, P. B. "The Impact of Expert System Usage on Experiential Learning in an Auditing Setting," *Journal of Information Systems* (5:1), 1991, pp. 1–16.

Everett, A. M. *An Empirical Investigation of the Effect of Variations in Expert System Explanation Presentation on Users' Acquisition of Expertise and Perceptions of the System,* Unpublished Doctoral Dissertation, University of Nebraska, 1994.

Expert Systems with Applications (8:4), 1995.

Fedorowicz, J., Oz, E., and Berger, P. D. "A Learning Curve Analysis of Expert System Use," *Decision Sciences* (23), 1992, pp. 797–818.

Fitts, P. M. "Perceptual-motor Skill Learning," in *Categories of Human Learning,* A. W. Melton (ed.), Academic Press, New York, 1964, pp. 243–284.

Fitts, P. M., and Posner, M. I. *Human Performance,* Brooks Cole, Belmont, CA, 1976.

Flowler, C. .J., Macauley, L. A., and Flowler, J. F. "The Relationship Between Cognitive Style and Dialogue Style," in *People and Computers: Designing the Interface,* P. Johnson and S. Cook (eds.), Cambridge University Press, Cambridge, England, 1985.

Gagne, E. D. *The Cognitive Psychology of School Learning,* Little Brown, Boston, 1985.

Galbraith, J. R. *Organization Design,* Addison-Wesley, Reading, MA, 1977.

Gault, R. W. *Learning and Explanation Type in a Knowledge-based Arms Control Inspection Assistant: An Empirical Evaluation (Strategic Arms Reduction Treaty),* Unpublished Doctoral Dissertation, George Washington University, 1994.

Genesereth, M. R., and Ketchpel, S. P. "Software Agents," *Communications of the ACM* (37:7), 1994, pp. 48–53.

Gentner, D., and Stevens, A. L. *Mental Models,* Lawrence Erlbaum, Hillsdale, NJ, 1983.

Gilbert, N. "Explanation and Dialogue," *Knowledge Engineering Review* (4:3), 1989, pp. 205–231.

Gill, G. T. "Early Expert Systems: Where Are They Now," *MIS Quarterly* (19:1), 1995, pp. 51–81.

Goguen, J. A., Weiner, J/ L., and Linde, C. "Reasoning and Natural Explanation," *International Journal of Man-Machine Studies* (19), 1983, pp. 521–559.

Gregor, S. D. "Explanations in an Expert System for Capital Gains Tax," in *Proceedings of the IJCAI Workshop on Explanation Generation for Knowledge-based Systems,* N. Wognum (ed.), University of Twente, Enschede, The Netherlands, 1991, pp. 43–56.

Gregor, S. D. *Explanations from Knowledge-based Systems for Human Learning and Problem Solving,* Unpublished Doctoral Dissertation, University of Queensland, Brisbane, Australia, 1996a.

Gregor, S. D. "A Personal Financial Planning System for Everyday Use?" in *Advanced IT Tools IFIP World Conference on IT Tools,* N. Terashima and E. Altman (eds.), Chapman and Hall, London, 1996b, pp. 189–196.

Gregor, S. D. "Explanations from Knowledge-based Systems and the Learning versus Working Argument," Working Paper, Central Queensland University, 1997a.

Gregor, S. D. "Knowledge-based Systems in Prescriptive versus Supportive Roles and the Use of Explanations," Working Paper, Central Queensland University, 1997b.

Hayes-Roth, F. "Artificial Intelligence: What Works and What Doesn't," *AI Magazine,* Summer 1997, pp. 99–113.

Hayes-Roth, F., and Jacobstein, N. "The State of Knowledge-based Systems," *Communications of the ACM* (37:3), 1994, pp. 27–39.

Horacek, H. "The Role of Explanation in Interactive Problem Solving," in *Proceedings of the ECAI-92 Workshop W15 "Improving the Use of Knowledge-based Systems with Explanations,"* P. Brézillon (ed.), Institut Blaise Pascal, Paris, June 1992, pp. 125–134.

Huber, O. "The Influence of Some Task Variables on Cognitive Operations in an Information-Processing Decision Model," *Acta Psychologica* (45), 1980, pp. 187–196.

Hughes, S. "Question Classification in Rule-based Systems," in *Proceedings of Fourth Technical Conference of the British Computer Society Specialist Group on Expert Systems,* M. A. Bramer (ed.), Cambridge University Press, Cambridge, England, 1987.

Hsu, K. *The Effects of Cognitive Styles and Interface Designs on Expert Systems Usage: An Assessment of Knowledge Transfer,* Unpublished Doctoral Dissertation, Memphis State University, 1993.

Johnson, E. J., and Payne, J. W. "Effort and Accuracy in Choice," *Management Science* (31:4), April 1985, pp. 395–415.

Johnson, H., and Johnson, P. "Explanation Facilities and Interactive Systems," *Intelligent User Interfaces '93,* 1993, pp. 159–465.

Johnson, P. E. "What Kind of Expert Should a System Be?" *Journal of Medicine and Philosophy* (8), 1983, pp. 77–97.

Johnson-Laird, P. N. *Mental Models,* Harvard University Press, Cambridge, MA, 1983.

Kidd, A. L. "What Do Users Ask? Some Thoughts on Diagnostic Advice," in *Expert Systems '85,* M. Merry (ed.), Cambridge University Press, Cambridge, England, 1985a.

Kidd, A. L. "The Consultative Role of an Expert System," in *People and Computers: Designing the Interface,* P. Johnson and S. Cook (eds.), Cambridge University Press, Cambridge, 1985b, pp. 228–254.

Klein, M., and Methlie, L. B. *Expert Systems: A Decision Support Approach,* Addison-Wesley, Wokingham, England, 1990.

Kolodner, J. L. "Towards an Understanding of the Role of Experience in the Evolution from Novice to Expert," *International Journal of Man-Machine Studies* (19), November 1983, pp. 497–518.

Koonce, L. "Explanation and Counterexplanation During Audit Analytical Review," *The Accounting Review* (67), 1992, pp. 59–76.

Lamberti, D. M., and Wallace, W. A. "Intelligent Interface Design: An Empirical Assessment of Knowledge Presentation in Expert Systems," *MIS Quarterly,* September 1990, pp. 279–311.

Lehnart, W. G. *The Process of Question Answering: A Computer Simulation of Cognition,* Lawrence Earlbaum Associate, Hillsdale, NJ, 1978.

Lerch, F. J., Prietula, M. J., Kim, J., and Buzas, T. "Measuring Trust in Machine Advice," Unpublished Manuscript, Graduate School of Industrial Administration, Carnegie Mellon University, Pittsburgh, PA, 1993.

Lerch, F. J., Prietula, M. J., and Kulik, C. T. "The Turing Effect: The Nature of Trust in Expert System Advice," in *Expertise in Context: Human and Machine,* P. J. Feltman, K. M. Ford, and R. R. Hoffman, (eds.), The MIT Press, Cambridge, MA, 1997.

Liebowitz, J. "Worldwide Perspectives and Trends in Expert Systems: An Analysis Based on the Three World Congresses on Expert Systems," *AI Magazine,* Summer 1997, pp. 115–119.

Luconi, F. L., Malone, T. W., and Scott Morton, M. S. "Expert Systems: The Next Challenge for Managers," *Sloan Management Review,* Summer 1986, pp. 3–14.

The Macquarie Dictionary, Macquarie Library, McMahons Point, Australia, 1981.

Maes, P. "Agents that Reduce Work and Information Overload," *Communications of the ACM* (37:7), 1994, pp. 30–40.

Mao, J. *An Experimental Study of the Use and Effects of Hypertext Based Explanations in Knowledge-based Systems,* Unpublished Doctoral Dissertation, University of British Columbia, 1995.

Mao, J., and Benbasat, I. "The Effects of Hypertext-based Explanations in Knowledge-based Systems on Explanation Use and Knowledge Transfer," Working Paper 96-MIS-001, Faculty of Commerce, University of British Columbia, 1996a.

Mao, J., and Benbasat, I. "Exploring the Use of Explanations in Knowledge-based Systems: A

Process Tracing Analysis," Working Paper 96-MIS-002, Faculty of Commerce, University of British Columbia, 1996b.

Mao, J., and Benbasat, I. "Contextualized Access to Knowledge in Knowledge-based Systems: A Process Tracing Study," *Information Systems Journal,* (8), 1998, pp. 217–239.

Mao, J., Benbasat, I., and Dhaliwal, J. S. "Enhancing Explanations in Knowledge-based Systems with Hypertext," *Journal of Organizational Computing and Electronic Commerce* (6:3), 1996, pp. 239–268.

Maybury, M. T. "Communicative Acts for Explanation Generation," *International Journal of Man-Machine Studies* (37), August 1992, pp. 135–172.

Mayer, R. E. "Elaboration Techniques that Increase the Meaningfulness of Technical Text: An Experimental Test of the Learning Strategy Hypothesis," *Journal of Educational Psychology* (72:6), 1980, pp. 770–784.

Mayer, R. E. "Structural Analysis of Science Prose: Can We Increase Problem-solving Performance?" in *Understanding Expository Text: A Theoretical and Practical Handbook for Analyzing Exploratory Text,* B. K. Britton and J. B. Black (eds.), Lawrence Erlbaum Associates, Hillsdale, NJ, 1985, pp. 65–86.

Miller C. A., and Larson, R. "An Explanatory and 'Argumentative' Interface for a Model-based Diagnostic System," *User Interface Software and Technology '92,* Monterey, CA, 1992.

Miksch, S., Cheng, K., and Hayes-Roth, B. "An Intelligent System for Patient Health Care," *Autonomous Agents 97,* ACM Press, Marina Del Ray, CA, 1997, pp. 458–465.

Miyake, N. "Constructive Interaction and the Iterative Process of Understanding," *Cognitive Science* (10), 1986, pp. 151–177.

Moffitt, K. E. *An Empirical Test of Expert System Explanation Effect on Incidental Learning and Decision-making,* Unpublished Doctoral Dissertation, Arizona State University, 1989.

Moffitt, K. "An Analysis of the Pedagogical Effects of Expert System Use in the Classroom," *Decision Sciences* (25:3), 1994, pp. pp. 445–460.

Moore, G., and Benbasat, I. "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation," *Information Systems Research* (2), 1991, pp. 192–222.

Murphy, D. S. "Expert System Use and the Development of Expertise in Auditing: A Preliminary Investigation," *Journal of Information Systems,* Fall 1990, pp. 18–35.

Neches, R., Swartout, W. R., and Moore, J. "Enhanced Maintenance and Explanation of Expert Systems Through Explicit Models of Their Development," *IEEE Workshop on Principles of Knowledge-based Systems,* IEEE Computer Society Press, Los Alamitos, CA, 1984, pp. 173–183.

Newell, A., and Simon, H. A. *Human Problem Solving,* Prentice Hall, Englewood Cliffs, NJ, 1972.

Oltman, P. K., Raskin, E., and Witkin, H. A. *Group Embedded Figures Test,* Consulting Psychologists Press, Inc., Palo Alto, CA, 1971.

Ortory, A. and Partridge, D. "Surprisingness and Expectation Failure: What's the Difference," in *Proceedings of the 10th International Joint Conference on Artificial Intelligence,* Vol. 1, J. McDermott (ed.), Interprint, San Francisco, August 23–28, 1987, pp. 106–116.

Paris, C. "Combining Discourse Strategies to Generate Descriptions to Users Along a Naive/Expert Spectrum," in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (Volume 2), J. McDermott (ed.), San Francisco, August 23–28, 1987, pp. 626–635.

Payne, J. W., Bettman, J. R., and Johnson, E. J. *The Adaptive Decision Maker,* Cambridge University Press, Cambridge, England, 1993.

Rempel, J. K., Holmes, J. G., and Zanna, M. P. "Trust in Close Relationships," *Journal of Personality and Social Psychology* (49:1), 1985, pp. 95–112.

Rissland, E. L. "The Ubiquitous Dialectic," in *Advances in Artficial Intelligence,* T. O'Shea, (ed.), Elsevier Science, Amsterdam, 1985, pp. 367–372.

Rook, F. W. *Human Cognition and the Expert System Interface: Mental Models and Inference Explanations,* Unpublished Doctoral Dissertation, Catholic University of America, 1990.

Ross, L., Lepper, M., and Hubbard, M. "Perseverance in Self-perception and Social Perception: Biased Attributional Processes in

the Debriefing Paradigm," *Journal of Personality and Social Psychology* (32), 1975, pp. 880–892.

Rotter, J. B. "Interpersonal Trust, Trustworthiness, and Gullibility," *American Psychologist* (35), 1980, pp. 1–7.

Schank, R. C. "Explanation: A First Pass," in *Experience, Memory, and Reasoning,* J. L. Kolodner and C. K. Riesbeck (eds.), Lawrence Erlbaum Associates, Hillsdale, NJ, 1986, pp. 139–165.

Schroder, H. M., Driver, M. J., and Streufert, S. *Human Information Processing,* Holt Rinehart and Winston, New York, 1967.

Schuell, T. "Cognitive Conceptions of Learning," *Review of Educational Research* (56:4), 1986, pp. 411–436.

Seligman, M. E. P. *Learned Optimism,* Pocket Books, New York, 1990.

Shortliffe, E. H. *Computer-based Medical Consultations: MYCIN,* Elsevier Computer Science Library, New York, 1976.

Shpilberg, D., Graham, L., and Schatz, H. "ExperTAX: An Expert System for Corporate Tax Planning," *Expert Systems* (3:3), July 1986.

Silverman, B. G. "Survey of Expert Critiquing Systems: Practical and Theoretical Frontiers," *Communications of the ACM* (35:4), 1992, pp. 107–127.

Silverman, B. G., and Mezher, T. M. "Expert Critics in Engineering Design: Lessons Learned and Research Needs," *AI Magazine* (13:1), Spring 1992, pp. 45–62.

Simon, H. A. "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics* (69), 1955, pp. 99–118.

Simon, H. A. "Rational Choice and the Structure of the Environment," *Psychological Review* (63), 1956, pp. 129–138.

Skinner, B. "Cognitive Science and Behaviourism," *British Journal of Psychology* (76:3), 1985, pp. 291–301.

Stylianou, A. C., Madey, G. R., and Smith, R/ D. "Selection Criteria for Expert System Shells: A Socio-technical Framework," *Communications of the ACM* (35:10), 1992, pp. 30–48.

Swartout, W. R. "What Kind of Expert Should a System Be? XPLAIN: A System for Creating and Explaining Expert Consulting Programs," *Artificial Intelligence* (21), 1983, pp. 285–325.

Swartout, W. R., and Smoliar, S. W. "On Making Expert Systems More Like Experts," *Expert Systems* (4:3), 1987, pp. 196–207.

Swinney, L. "The Explanation Facility and the Explanation Effect," *Expert Systems with Applications* (9:4), 1995, pp. 557–567.

Todd, P., and Benbasat, I. "An Experimental Investigation of the Impact of Computer Based Decision Aids on Decision Making Strategies," *Information Systems Research* (2:2), 1991, pp. 87–115.

Toulmin, S. *The Uses of Argument,* Cambridge University Press, Cambridge, England, 1958.

Toulmin, S., Rieke, R., and Janik, A. *An Introduction to Reasoning,* Macmillan, New York, 1984.

Turban, E. *Decision Support and Expert Systems: Management Support Systems,* 4th ed., Prentice Hall, Englewood Cliffs, NJ, 1995.

van Dijk, T.A., and Kintsch, W. *Strategies of Discourse Comprehension,* Academic Press, New York, 1983.

Wick, M. R., and Slagle, J. R. "The Partitioned Support Network for Expert System Justification," *IEEE Transactions on Systems, Man, and Cybernetics* (19:3), 1989a, pp. 528–535

Wick, M. R., and Slagle, J. R. "An Explanation Facility for Today's Expert Systems," *IEEE Expert,* Spring 1989b, pp. 26–35.

Wanninger, L. "Profitable Electronic Commerce: Frame Work, Examples, Trends," in *Eleventh International Bled Electronic Commerce Conference "Electronic Commerce in the Information Society" Proceedings* (Volume 2), G. J. Doukidis, J, Gricar, J. Novak (eds.), Bled, Slovenia, June 8–10 1998, pp. 3–27.

Wason, P. C., and Johnson-Laird, P. N. *Psychology of Reasoning: Structure and Content,* Batsford, London, 1972.

Wognum, P. M. *Explanation of Automated Reasoning: How and Why?* Unpublished Doctoral Dissertation, University of Twente, Enschede, The Netherlands, 1990.

Wognum, P. M. (ed.). *Proceedings of the IJCAI Workshop on Explanation Generation for Knowledge-based Systems,* University of Twente, Enschede, The Netherlands, 1991.

Ye, L. R. *User Requirements for Explanation in Expert Systems,* Unpublished Doctoral Dissertation, University of Minnesota, 1990.

Ye, L. R. "Value of Explanations in Expert Systems for Auditing: An Experimental Investigation," *Expert Systems with Applications* (9:4), 1995, pp. 543–556.

Ye, L. R., and Johnson, P. E. "The Impact of Explanation Facilities on User Acceptance of Expert System Advice," *MIS Quarterly,* June 1995, pp. 157–172.

Yourdon, E. *Modern Structure Analysis,* Prentice-Hall, Englewood Cliffs, NJ, 1989.

Zahedi, F. *Intelligent Systems for Business: Expert Systems with Neural Networks,* Wadsworth, Belmont, CA, 1993.

Zipf, G. K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology,* Addison-Wesley, Reading, MA, 1949.

About the Authors

Shirley Gregor is head of the School of Computing and Information Systems at Central Queensland University, Rockhampton, Australia. She holds a Ph.D. in information systems from the University of Queensland (1996). Shirley spent a number of years in the computing industry in Australia and the United Kingdom before beginning an academic career. Her current research interests include intelligent systems, human-computer interaction, the development of web information systems, and electronic commerce in agribusiness.

Izak Benbasat is CANFOR Professor of Management Information Systems and associate dean, Faculty Development, at the Faculty of Commerce and Business Administration, The University of British Columbia (UBC), Vancouver, Canada. He received his Ph.D. (1974) in management information systems from the University of Minnesota. His current research interests are in investigating the role of explanations in intelligent systems; measuring information systems competence in line managers and business competence in information systems professionals; and the impact of these competencies on the successful deployment of information technologies; evaluating human-computer interfaces; and comparing methods for conducting information systems research.

APPENDIX A

Empirical Studies Related to Explanations in KBS

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
Berry and Broadbent 1987b	Human reasoning (Wason and Johnson -Laird 1972), concept-learning (Bruner et al. 1956).	Three experiments. Task was testing for river pollutants with aid of advice-giving system.	(1) Prior general explanation vs. multiple case-specific explanations (2) Prior general explanation with verbalization vs. verbalization alone (3) As for (b) but also a no-KBS condition.	(1) Learning (2) Performance	(1) Performance better with multiple case-specific explanations, rather than single general explanation. (2) Performance better with single general explanation and verbalization rather than verbalization alone. (3) Multiple case-specific explanations and general explanation/verbalization both contributed to learning.
Eining 1988; Eining and Dorr 1991	Cognitive learning theory and expert-novice differences (Einhorn 1974; Gagne 1985).	Task was evaluation of internal control over factory payroll. Five week laboratory study with 191 novice auditors and purpose-built KBS.	(1) Type of decision aid (none, questionnaire, KBS with explanations, KBS without explanations) (2) Level of feedback.	(1) Procedural learning (measured as decrease in time and increase in accuracy).	Use of expert system resulted in greater learning, but no difference between groups with and without explanations.
Moffitt 1989, 1994	Cognitive learning theory (Anderson 1982).	Task was two scheduling problems. Experiment with 362 student subjects and purpose-built KBS.	Type of decision aid (none, KBS without explanation, KBS with rule-trace, KBS with user-invoked canned-text, KBS with automatic. canned-text explanation).	(1) Declarative learning (2) Procedural learning (3) Perceptions of KBS.	Both types of learning greater in automatic-explanation condition compared with no-aid condition. Ratings for usefulness for learning highest for automatic-explanation condition, then canned-text, then rule-trace, then no-explanation condition.

APPENDIX A. CONTINUED

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
Lamberti and Wallace 1990	Expert-novice differences (Adelson 1981; Anderson 1982; Chi et al. 1981), task uncertainty in decision making (Galbraith 1977).	Quasi-experimental field study of operational KBS with 90 programmers and diagnostic problems in computer support center over two-year period.	(1) Expertise (2) Knowledge presentation format (procedural vs. declarative) (3) Question type (abstract vs. concrete) (4) Task uncertainty.	(1) Performance (speed and accuracy) (2) Confidence in KBS (3) Satisfaction with KBS.	Higher-expertise users performed better and showed greater confidence. KBS had more impact on performance of low-expertise users. Users with different levels of expertise needed different presentation formats. Other interaction effects.
Murphy 1990	Learning theory and development of expertise (Anderson 1985; Einhorn 1974; Fitts and Posner 1976; Kolodner, 1983).	Task was auditing problems. Experiment with 67 accounting students and production KBS.	(1) Type of decision aid (KBS with explanations, KBS without explanations, non-automated aid).	(1) Declarative learning (2) Procedural learning.	Declarative and procedural learning greater in non automated aid condition.
Rook 1990	Mental models (Craik 1943; Gentner and Stevens 1983; Johnson-Laird 1983).	Task was space-station fault diagnosis. Experiment with 30 students and purpose-built KBS.	(1) Explanation type (graphic vs. text) (2) Mental model (graphic, text, none).	(1) Ability to reconstruct KBS reasoning (performance) (2) Perceptions of KBS.	Performance higher in graphic mental model condition. Improved performance also when mental model matched explanation type.
Wognum, 1990	Explanation strategies and architectures (Buchanan and Shortliffe 1984; Neches, et al. 1984).	(1) Retrospective study of nine operational systems, (2) Paper test with eight social security workers and 40 problems.	(1) Not applicable. (2) Explanation type (legal rules vs. handbook).	(1) Not applicable. (2)(a) performance, (b) Helpfulness of explanation.	(1) In operational systems explanations not always needed. Importance of explanations depended on users' experience and role of the system. (2) Explanations improved decision making. Explanations based on handbook perceived as most useful. (not statistically significant).

APPENDIX A. CONTINUED

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
Ye 1990; Ye 1995; Ye and Johnson 1995	Mental models (Johnson-Laird 1983), task classification (Clancey 1985), model of argument (Toulmin 1958).	Task was audit analytical review problems. Experiment with 10 novice and 10 expert practicing auditors and simulated KBS.	(1) Expertise (2) Task type (data abstraction vs. heuristic match) (3) Type of explanation.	(1) Change in belief (before and after receipt of explanation) (2) Choice of explanation (3) Reading time for explanation (4) Perceived usefulness of explanation (5) Degree of preference for explanation type.	Belief in conclusions increased after explanations were given. Preferences were for (1) justification-type, (2) trace, then (3) strategic explanations. The order for reading times (greatest to least) was the same. Novices perceived justification to be more useful and preferable than other two types. Experts perceived justification to be more useful and preferable than strategy, but about as useful as trace.
Dhaliwal 1993.	Cognitive learning theory (Balzer et al. 1989); lens model (Brunswik 1952, 1956).	Task was financial statement analysis. Experiment with 40 students, 40 practicing accountants and simulated KBS.	(1) Explanation type (feedforward, feedback, feedforward and feedback, none) (2) Explanation use (3) Expertise (novice or expert) (4) Agreement with KBS conclusion.	(1) Explanation use (trace, justification, strategic) (2) Performance (accuracy) (3) Perception of usefulness of explanation.	Trace and justification-type used more than strategic explanations. Feedback were used more than feedforward explanations. Explanations were used least when agreement was very high or very low. Feedback explanations improved the accuracy of decision making. Experts used trace more than strategic explanations. Novices used more justification than strategic explanations.

APPENDIX A. CONTINUED

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
Everett 1994	Developed a framework based on content and purpose of explanations and prior empirical work, no theory.	Task was product/ process matching. Experiment with 260 students and simulated KBS.	(1) Explanation type (facts, rules, pseudo-rules, choice of other three types) (2) Justification- type presence (yes or no), (3) Invocation mode (user-invoked vs. automatic), (4) Expertise	(1) Declarative learning (clarification) (2) Procedural learning (duplication) (3) Agreement (ratification) (4) Perceptions of KBS.	Justification-explanations were of critical importance, affecting user acceptance. Procedural learning was greater in the automatic-explanation condition compared with the user-invoked condition.
Gault 1994	Learning theory (Anderson 1983), social learning theory (Bandura 1989), attributional style (Seligman 1990).	Task was application of a treaty for arms control. Experiment with 132 students at a military college and simulated KBS.	Explanation type (none, rule- trace, fixed-text, hypertext).	(1) Performance (accuracy, time, confidence) (2) Declarative learning(accuracy, time, confidence) (3) Satisfaction.	Learning (accuracy and confidence) was greater in fixed-text and rule-trace conditions compared with no-explanation condition. Confidence was greater in all explanation conditions compared with no-explanation condition.
Hsu 1993	ACT* learning theory (Anderson 1983), function-mechanism model of understanding (Miyake 1986), cognitive style (Flowler et al. 1985).	Task was financial statement analysis. Laboratory experiment over four weeks with 287 accounting students with purpose-built KBS.	(1) Explanation type (rule-trace, justification, user-invoked) (2) Cognitive style.	(1) Procedural learning (2) Explanation use (three types) (3) Perceived usefulness (4) Perceptions of KBS.	Cognitive style and interface design affected procedural learning. Procedural learning was greater with justifications rather than rule-based explanations alone.

APPENDIX A. CONTINUED

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
De Greef and Neerincx 1995	Model-based design of human-computer systems (Yourdon 1989).	Two experiments, each with 20 students using statistical program.	Explanation (aiding) availability	(1) Performance, (2) Learning.	Aiding interface improved performance and learning in second experiment where participants had some statistical knowledge.
Mao 1995; Mao and Benbasat 1996a, 1996b, 1998	Discourse comprehension (Mayer 1980, 1985; van Dijk and Kintsch 1983), Anderson's ACT* model (Anderson 1983), nature of explanation (Schank 1986).	Task was financial analysis. Experiment with 29 students, 26 professionals and simulated KBS. Included verbal protocol analysis.	(1) Expertise (novice, expert) (2) Explanation provision type (linear text vs. hypertext) (3) Frequency of use of explanations (generic, trace).	(1) Knowledge transfer (2) Usefulness of explanations (3) Trust in KBS. (4) Frequency of explanations use (generic, trace) (5) Timing of explanation use.	More general-domain explanations used in hypertext condition. Higher knowledge transfer from KBS to users with general-domain explanations than with trace explanations. Use of trace explanations was related to perceived usefulness of explanations. Novices used explanations more for learning and understanding, experts more for verification.
Swinney 1995	The explanation effect (Koonce 1992; Ross et al. 1975).	Audit assessment task. Experiment with 41 practicing auditors and paper output from operational KBS.	(1) Explanation source (self-generated, KBS-generated, none) (2) Explanation direction (positive vs. negative).	User's judgement.	The most influential explanations were those that were expert-system generated and negative (conservative) in direction.
Brajnik et al. (1996)	None	Information retrieval system. Experiment with 45 students.	(1) System type (with or without a query reformulation capability), (2) External support type.	(1) user satisfaction, (2) performance, (3) user behavior.	The majority of help requests were for specific terms to include in the query. Terminological help was requested in contextual form. Strategic help was not requested but was needed.

APPENDIX A. CONTINUED

Study	Theoretical foundation[a]	Task and context	Independent variables	Dependent variables	Results[b]
Gregor 1996a, 1996b, 1997a, 1997b	Cognitive learning theory (Anderson 1990; Ausubel 1985), Toulmin's model of argument (Toulmin 1958; Toulmin et al. 1979), cognitive-effort perspective (Payne et al. 1993).	(1) Experiment with operational tax system and 84 students. (2) Experiment with a purpose-built financial planning system and 91 members of the general public.	(1) (a) User's goal (learning vs. problem-solving), (b) explanation use (reasoning trace, justification, control, terminological) (2) (a) KBS role (prescriptive vs. supportive), (b) Availability of explanations, (c) Explanation use (general-domain, reasoning trace, justification, control).	(1) (a) Explanation use (b) Performance (c) Learning (d) Confidence. (2) (a) Explanation use (b) Performance (c) Confidence.	Explanations were used more when the user's goal was learning rather than problem solving. Under some conditions the role of the KBS affected explanation use and use improved performance. Confidence was related to use of explanations but the direction of the relationship was different in the two studies.
Lerch et al. (1997)	Interpersonal trust (Rotter 1980; Rempel et al. 1985)	Experiment with 67 students and financial decision problems.	Explanation type (none, canned text, rule).	(1) Agreement with advice, (2) Confidence in source, (3) Performance attributions.	Agreement with advice greater when explanations given.

Note: [a] Some selectivity was employed in choosing references. [b] Only primary, significant results reported.

APPENDIX B

Explanation Types

The classification of explanations into four different types by content (Table 1) reflects also to some extent the historical development of explanation facilities. It includes the types of explanations found in most operational systems and expert system shells (Chandrasekaran et al. 1989; Wick and Slagle 1989b; Wognum 1990). This classification also enables explanations to be discussed in terms of Toulmin's (1958) model of argumentation.

Explanations Types I, II, and III in Table 1 are the three types proposed by Chandrasekaran et al. (1989): trace, justification, and control. These explanation types have been used in a number of studies (Dhaliwal 1993; Hsu 1993; Mao 1995; Ye 1990). Most explanation facilities available in expert system shells (expert system building tools) are limited to the two reasoning trace queries (Type I): *How* and *Why* (Wick and Slagle 1989b). These queries were introduced in MYCIN, a system developed in the early 1970s for diagnosing infectious blood diseases (Clancey 1983; Shortliffe 1976).

Justification-type explanations (Type II) require "deep" domain knowledge, causal knowledge or generally accepted rules or principles in the relevant field. Deep explanations can incorporate many different types of knowledge: analogies, cases, textbook knowledge, and so on. The role of deep knowledge in explanations can be explicated further by considering the model of practical reasoning and argumentation provided by Toulmin (Toulmin 1958; Toulmin et al. 1983) and discussed in the paper. Toulmin's model shows how "warrants" and "backing" are elements in any explicit argument (explanation). These warrants and backing are drawn from the deep knowledge in a particular field. In science, a warrant may be a law of nature and the backing may be the degree to which the law has been investigated and confirmed. In law, a warrant may be a legal principle or statute and the backing the knowledge that the statute has been validly legislated. Further discussion of deep versus surface knowledge in expert systems can be found in Chandrasekaran and Mittal (1983).

In principle, deep knowledge could be included with any of the different explanation-content types (Table 1) except reasoning trace explanations. Reasoning trace is differentiated from justification in that the latter has deep knowledge and the former does not. Thus, an explanation of terminology could have a textbook reference attached to show the authority from which it was drawn. A control or strategic explanation could have deep knowledge attached in the form of evidence that the particular strategy was used successfully by experts in the field.

Terminological explanations are the "knowledge of the concepts and relationships of a domain that experts use to communicate with one another" (Swartout and Smoliar 1987, p. 198). Mao (1995) uses this category, but includes it with "deep" explanations. Terminological explanations are comparable to the *facts* referred to by Everett (1994), the *Answer Help* of Gregor (1996a), and the *which* facility of Wognum (1990).

A further distinction in explanation-content types can be made between *generic* explanations and *case-specific* explanations. An explanation could be couched in general terms, as in the description of a general problem solving method, and given at any time in the course of a consultation. Terminological explanations are generally of this type. In contrast, some explanations are case-specific, and are given in the context of solving a specific problem with reference to the data for that case. Reasoning trace, justification, and control explanations are mostly case-specific. Chandrasekaran et al. (1989), for example, applied the justification-type category to refer to explanations that support a link from specific data to a specific conclusion.