Artificial Intelligence Review (2005) 24: 109–143
DOI 10.1007/s10462-005-4607-7

Explanation in Case-Based Reasoning–Perspectives and Goals

FRODE SØRMO, JÖRG CASSENS & AGNAR AAMODT
Department of Computer and Information Science (IDI), Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway
(E-mails: {frode.sormo, jorg.cassens, agnar.aamodt}@idi.ntnu.no)

Abstract. We present an overview of different theories of explanation from the philosophy and cognitive science communities. Based on these theories, as well as models of explanation from the knowledge-based systems area, we present a framework for explanation in case-based reasoning (CBR) based on explanation goals. We propose ways that the goals of the user and system designer should be taken into account when deciding what is a good explanation for a given CBR system. Some general types of goals relevant to many CBR systems are identified, and used to survey existing methods of explanation in CBR. Finally, we identify some future challenges.

Keywords: case-based reasoning, explanation

1. Introduction

The term explanation can be interpreted in two different ways in AI (Aamodt, 1991, p. 59). One interpretation deals with explanation as part of the reasoning process itself, for example used in the search for a diagnostic result in order to support a particular hypothesis. The other interpretation deals with usage aspects: attempting to make the reasoning process, its result, or the usage of the result understandable to the user. This paper primarily deals with the latter interpretation, but explanation as part of the reasoning process is also addressed where appropiate.

In our daily lives we experience explanations every day, and they seem to exist in an unlimited number of forms. Everything from "I didn't wash the dishes because there was no detergent", to "I hate shopping", and even "Because I said so!" can serve as satisfactory explanations in certain circumstances. Explanation is one of those concepts that everyone has a good intuition of, but which are very hard to explicitly define.

In this paper, we will attempt to characterize important aspects of an explanation, and relate them to explanations in and from case-based reasoning (CBR) systems.

When reviewing the literature we find that many accounts of explanation explicitly recognize that the context of an explanation situation, and the goal of the user in that situation, influence what is and what is not a good explanation. While goal situations may vary a lot among domains, systems, and users, some goal situations are common. We will present a framework of explanation based on important explanation goals, and discuss how they place limitations on each other and how different kinds of systems may be better suited to fulfill different goals.

We will begin by looking at foundational and theoretical issues of explanation, as developed within philosophy and cognitive science (Section 2). This is followed, in Section 3, by views and models of explanation from within the expert systems and intelligent tutoring communities. In Section 4 we review current accounts of explanation in CBR and present a set of explanation goals for CBR systems. A brief survey of explanation in different CBR systems follows in Section 5. In Section 6 we highlight some challenges for the future before concluding with Section 7.

2. Philosophical and Cognitive Accounts of Explanation

People tend to think of explanation as something identifying the cause for a particular event or state, as for example in the sentence "the train is late because of a faulty stop light". This is also the case in many philosophical theories of explanation (see for instance Salmon, 1984). However, in daily life we also use explanations that are functional ("there is rubber on the end of the pencil so you can erase mistakes") and intentional ("I turned off the light because I want to sleep" (Brewer et al., 1998)). This is further complicated by the fact that both the sender and recipient of an explanation have goals in the exchange, and their goals influence what candidate explanations are and are not acceptable (Leake, 1995b). Thus it may be very hard to form a complete theory of explanation. We will characterize some accounts of explanation discussed in the philosophical and cognitive science communities.

2.1. *Basic philosophical accounts*

The nature of explanation has been studied extensively by philosophers, particularly by researchers in the philosophy of science. Here

the targets for explanation are specific observations, predicted outcomes, or scientific theories themselves. Explanations are sought based on observations and existing knowledge. Two different approaches, or rather classes of approaches, emerged throughout the 1950s and 1960s. The logical deductive approach, suggested by Hempel and Oppenheim (1948) and Hempel (1965) was linked with a positivistic view on science. This approach was severely criticized by several people, resulting in several suggestions of different, and more pragmatic, approaches to explanation. Important early contributions were made by Harman (1965), Bromberger (1965), and Salmon (1971).

The positivist approach takes a scientific theory to be an axiomatic formalization of a set of sentences in a logic system. Hempel and Oppenheim refer to it as a "deductive-nomological" (deduction from laws) explanation, also referred to as the "covering law model", reflecting that the theory subsumes or covers the things that are explained. This work was subsequently extended with a formal model of probabilistic inference as well, the "inductive-statistical" model (Hempel, 1965). In order to analyze explanations formally, an explanation structure in both these models is defined to consist of two parts; the part that is to be explained, called the *explanandum*, and the explanatory expression, called the *explanans*. For example: The patient died (explanandum); The patient had cancer (explanans); The patient died because he had cancer (explanation).

While the pragmatic aspects of explanation are acknowledged by all philosophers of science (including Hempel), a characterization of the non-positivist tradition is that the pragmatics of an explanation situation, in terms of context, purpose, etc., is at the very basis of the nature of explanation. Pragmatics becomes the starting point for the understanding of explanation, rather than an additional challenge for axiomatic formalization.

Early advocates of pragmatic approaches criticized the deductive-nomological account for being too syntax-oriented, in that semantical interpretations (i.e. the content of theories) started out from the interpretation of the logical syntax of expressions, rather than from the needs of the real world. Pragmatic approaches attempt to offer a semantic that starts out from the real world, with the necessary or suitable syntax following from pragmatic needs. While deductive inference certainly is an important inference type, several philosophers have shown the importance of abductive inference – and particularly the form referred to as "inference to the best explanation" – as a frequently occurring inference type in hypothesis formation and

evaluation (Josephson and Josephson, 1994). The strict requirement of truth-preserving inference underlying logical deduction is relaxed here. Out of a set of hypotheses, the hypothesis that can best explain the facts is chosen. Originating from Charles Sanders Peirce, an early account in philosophy of science was suggested by Harman (1965). While other researchers have proposed abductive models of scientific discovery, Harman's model concentrated on justification. The basic idea behind his model was to argue that an inference from some data to the best explanation is a justified mode of inference and leads to true hypotheses.

CBR is concerned with problems that are open-ended, and often changing, and uncertainty as well as incompleteness of theories and input descriptions are typically assumed. Viewing explanations as deductive proofs will be too severe a limitation for our purpose, and hence less relevant for the type of explanations CBR systems need to generate. A pragmatic view of explanation will therefore be accounted for in the following, while the Hempel–Oppenheim account sometimes will be used for comparison.

Philosophers who study linguistics and everyday speech have also made significant contributions to the nature of explanations. An early influential example is Bromberger (1965), who in particular criticized two weaknesses of Hempel and Oppenheim's theory. Through a series of examples he showed that perfectly valid deductive-nomological explanations can be made with true but irrelevant premises.

The second problem was related to the symmetrical properties of logical inference, particularly when the explanatory law has a functional form. The equations can be rewritten so that any of the variables becomes the value to explain, i.e. the explanandum. One of his famous examples is the flagpole example. When the line of sight of the sun across the top of a flagpole is at a given angle with the ground, the height of the flagpole and the length of the shadow it casts are related. Under the deductive-nomological model, it can be explained why the length of the shadow takes a given value by citing this law and the height of the pole. So far so good. But the equation and the length of the shadow can equally well be used to explain the height of the flagpole, i.e. to explain why the flagpole has the height it has, which seems entirely inappropriate in all but very peculiar situations.

Bromberger analysed explanation triggering questions in the form of why-questions, and suggested that an important type of question arises "when one believes that the presupposition is true, views it as a departure from a general rule, and thinks that the conditions under

which departures from the general rule occur can be generalized" (Bromberger, 1966, p. 100). Asking this type of why-question would then imply that the person asking is in some way surprised about the fact implied in the why-question (the presupposition) while still believing its truth.

An early and influential approach to the treatment of causality in explanations was presented by Salmon (1971). Salmon characterizes explanation as the pursuit of understanding, and to explain as to attribute a cause. As opposed to Hempel's experimentalist position, Salmon worked in the realist tradition. Salmon's "causal realism" theory of explanation started out from Bayesian probability, viewing an explanation basically as a set of statistically relevant factors, but he later found that theory inadequate in accounting for how explanations produce scientific understanding (see the following subsection).

2.2. *Later philosophical accounts*

Later accounts include continued work on scientific explanation by van Fraassen (1980), Salmon (1984), and Thagard (1988), explanation in natural language by Achinstein (1983), as well as cognitive models of explanation, by Schank (1986), Keil and Wilson (2000), and Leake (1995a). Some of these theories are also applicable to everyday explanations.

One of these is formulated by Bas van Fraassen in his book *The Scientific Image* (van Fraassen, 1980). Van Fraassen takes a strictly empiricist approach (often referred to as "constructive empiricism"), and claims that an explanation is always an answer to an implicit or explicit contrastive why-question. By "contrastive", he means a question of the form "Why S_0 rather than $S_1, \ldots, S_n$?" where one state or event is preferred over a set of alternatives. For example, the explanation "The train is late because of a faulty stop light" is an answer to the question "Why is the train somewhere else rather than here?" According to van Fraassen, an acceptable explanation must favor the observed state S_0 over the other states. By this, he means that the answer or explanation must increase the probability of the observed state S_0 relative to $S_1, \ldots, S_n$. He suggests that this can be calculated by applying Bayes' Rule to each candidate answer. As long as each candidate satisfies the previous criterion of favoring the observed state, van Fraassen claims there are no objective criteria for preferring one over another, but that the context of the question implicitly contains information about which answer the receiver would prefer. Perhaps

the most useful feature of van Fraassen's theory for application in knowledge based systems is that it suggests a minimum criterion an explanation must fulfill (it must favor the observed state) as well as a framework for understanding explanations (as answers to contrastive why-questions).

Salmon's later account of causal explanation was triggered by problems of causal relevance and causal asymmetry in his early account, and by the distinction between true causal processes and pseudoprocesses. An example illustrating the latter difference is the beam of a torch as the torch is moved by hand so the light describes an arc through the sky. The movement of the beam is a pseudoprocess, since later stages of the beam are not caused by earlier stages, while the hand movement of the torch itself is a true causal process – as is the electrical production of light within the torch. A central idea in his "causal mechanical" model of explanation is that a causal process is a physical process that is characterized by being able to transmit a "mark" in a continuous manner. A mark is a local modification to the physical structure involved, such as a scratch in its surface. True causal processes have marks, pseudoprocesses not. A second element in his theory is the notion of causal interaction, through which marks are transmitted between causal processes. According to the causal-mechanical model, an explanation of some phenomenon will trace the causal processes – including interactions – which lead up to the phenomenon, and describe the processes and interactions of the phenomenon itself. If successful, the explanation will show how the phenomenon to be explained fits into a causal structure. Salmon developed a detailed and complex theory, resulting in a set of instructions for how to produce an explanation by creating a causal model for a given phenomenon.

An influential follower of Bromberger in the philosophy of natural language is Achinstein (1983). He follows the tradition that a request for explanation is a request for understanding of something. He addresses questions such as: Why have the standard models of scientific explanation been unsuccessful? What is causal explanation, and must explanation in the sciences be causal? What is a functional explanation? He emphasizes the role of the explanation process – the explaining act in which someone writes or utters something to someone else. What is written or uttered in this process is called the explanation product. Achinstein's view is that an explanation (product) can not be understood or evaluated without reference to the explaining act, which leads to his "illocutionary" theory of explanation. The

explaining act defines some aspect of the context and purpose behind the explanation, which is needed for a correct and meaningful interpretation of the explanation product.

He believes this request can take many forms, not just the why-questions of Bromberger and van Fraassen but any number of questions (why, what, where, how, etc.). Achinstein says that an explanation is the intention of giving someone the knowledge to understand some phenomena from some frame of reference. Like van Fraassen, Achinstein suggests that there is further preference for some explanations over others, and that this preference is defined by the context of the conversation and ultimately in the control of the individual requesting the explanation. For example, an explanation that a train is full because it is the rush hour may be useful for a passenger, but for the train scheduling department a more useful explanation is that too few trains are scheduled at this time of the day.

This view of explanations suggests that a very wide variety of statements can serve as explanations. An explanation need not, for example, be a causal chain of events leading up to the matter to be explained. The explanation may have as a goal facilitating the formation of such a causal chain by the recipient, but it need not contain it explicitly. It is enough to supply the recipient with the knowledge that he needs in order to infer it. This is a case of observing one of the "rules of communication" often seen in human conversation: Only information that is not obvious should be communicated. If someone asks "Why is Peter not here?" a perfectly good explanation can be "Anne is sick" if the explainer is aware that the recipient knows that Peter has a daughter called Anne and that he has to stay at home and take care of her when she is sick.

On the one hand, this emphasizes the value of knowing the recipient well and it suggests that to form efficient explanations, accurate user models may be necessary. On the other hand, it alleviates the requirement of the explainer to put forward a complete explanation if the system can make reasonable assumptions about what the recipient knows and is capable of. For instance, an Artificial Neural Network that is trained to compare two pictures of a certain type can give a similarity measure, e.g. from 0 to 1, but it is difficult to explain how it came up with this score in a way most people can understand. However, presenting the pictures to the user so he can validate the similarity for himself can itself serve as an explanation. For many types of pictures, it is a reasonable assumption for the system to believe that the user is able to compare the pictures quite well on his own. Note

that this is only the case if the goal of the receiver is to gain understanding of how good an answer the system has supplied. If the goal is to gain understanding of how the system arrived at the conclusion, the above explanation is far from sufficient.

Like Achinstein (1983), Thagard (1988) is concerned with the pragmatics of an explanation. He developed what he calls a "computational philosophy of science", based on computational metaphors of epistemology, and by implementing and testing his theories in computer programs. Thagard also views explanation as a process of providing understanding, and understanding is to a large extent achieved through locating and matching. This is a view of reasoning based on retrieval and adaptation of knowledge structures, functionally similar to the MOPS (Memory Organization Packets) in Schank's (1982) theories, see Section 2.3. In Thagard's early model, called PI, the knowledge structures – based on concrete or generalized situations or episodes – are supplemented with more general knowledge in the form of rules. In order for an explanation to be understood, it must activate this "mental model" in a meaningful way – that is in a way that enables the existing knowledge structure to confirm the explanation without seriously contradicting other parts of the knowledge structure. On this basis, Thagard developed a theory referred to as "explanatory coherence", based on the notion of a coherent body of knowledge.

The notion of knowledge coherence – as a relaxation of the formal notion of consistency – has been adopted by many people, including AI researchers (e.g., Lenat and Feigenbaum, 1987). Thagard (1989), however, takes this further into a theory of explanation. Coherence, in this theory, is basically a property over a set of propositions. It only makes sense to talk about coherence of a single proposition if viewed with respect to another set of propositions. The notion of acceptability is introduced to characterize this property of single propositions. Starting out from a model of abductive inference, in the sense of inference to the best explanation, he identifies three important criteria for selecting the best explanation: *conscilience* (favoring explanatory breadth), *simplicity* (favouring explanations with few propositions), and *analogy* (favouring explanations based on analogies). Thagard's work not only presents an approach to scientific explanation, but also defines the role of explanation within a wider theory of coherence-seeking abductive inference. His research has focused on analogy and case-based reasoning (Thagard and Holyok, 1989), as well as other computational models, which include connectionist networks and probabilistic network models.

Additional philosophical accounts of explanation include the "unificationist" accounts of Friedman (1974) and Kitcher (1976) and the information theoretic model of Hanna (1982). The basic idea of the former is that a scientific explanation should attempt to unify a range of different phenomena. A successful unification may reveal relationships between phenomena that were previously unknown - which seems to be something that good explanations are expected to do. Hanna proposed the notion of "transmitted information", coming from information theory, as the basis for evaluating the goodness of an explanation. Hanna responds to a crucial problem with Hempel's inductive-statistical model in that it does not adequately take relevance into account. Transmitted information, according to Hanna, reflects a relevance relation, which in turn is linked to explanatory power.

2.3. *Cognitive science accounts*

Thagard's research, as described above, also spans the philosophy of mind, and hence is positioned within the field of cognitive science as well as philosophy. Several other researchers in this community have also addressed the issue of explanation related to cognition.

Roger Schank and colleagues further developed Schank's "dynamic memory" (Schank, 1982) theory of reminding, problem solving, and learning, into a theory of explanation generation and evaluation. As one of the founders of CBR as we know it today, he proposed a case-based approach to explanation, based on storing, indexing, and retrieval of "explanation patterns" (Schank, 1986). Explanation patterns are specific or generalized cases of explanation events. A particular focus has been the exploration of case-based reasoning as a platform for creativity (Schank and Leake, 1989). In this model, creativity comes from retrieving explanations related to a situation, but using them in new ways – referred to as "tweaking" of explanations. Depending on the retrieval and adaptation processes used, CBR has the potential to provide solutions to a range of creativity tasks, from close to copying old solutions up to producing novel ideas. The following has been a focusing problem for studying various types of explanations:

> In 1984, Swale was the best 3-year-old racehorse, and he was winning all the most important races. A few days after a major victory, he returned from a light morning gallop and collapsed

outside his stable. The shocked racing community tried to figure out why. Many hypotheses appeared, but the actual cause was never determined.

The experimental system that implements several of the methods investigated, called SWALE, attempts to explain the anomaly in Swale's premature death (Kass, Leake and Owens, 1986; Leake, 1992). It generates explanations of why Swale died by retrieving and tweaking remindings of explanation patterns for other cases of death. The approach has demonstrated the generation of a variety of interesting possible explanations of its death, including a heart attack (the "Jim Fixx explanation pattern"), and a drug overdose (the "Janis Joplin explanation pattern").

Abductive inference also has a central position in cognitive accounts of explanations (including Thagard's work, see Section 2.2). Extending from his earlier research on explanation patterns, Leake (1995a), in his work about models for everyday abductive explanations, identifies a set of issues related to comparing abductive reasoning methods. One is the issue of *when to explain something*, which links to the central ability of the reasoner itself to decide when an explanation is merited. Leake considers both plausibility criteria and the role of goals. He divides traditional plausibility criteria into the three groups of *structural minimality criteria*, motivated by the principle of Occam's razor, *proof-based approaches*, which are based on an evaluation of the generated proof-like explanations, and *probabilistic and cost-based criteria*, that focus on the costs and probabilities related to the generated explanations.

In contrast to these syntactic-oriented criteria, a set of goal-based criteria are suggested (Leake, 1995b). Explanations are assumed to have two roles – either as a support of a claim or an argument against it. This work follows the tradition of Lalljee et al. (1983), who suggest that explanations can be either 'constructive' or 'contrastive', and Schank (1982), who specifies that an explanation is required first and foremost in anomalous situations that do not fit a person's internalized model of the world (cf. the "surprises" assumed by Bromberger, Section 2.1). Leake's view on explanation is related to the natural language philosophy view outlined before, in the sense that it takes the recipient's frame of reference into account. However, Leake has an operational view on explanations and not a purely descriptive one. While Achinstein deals with general communication issues, Leake focuses on the evaluation of given explanations for the actor. In this

sense, Leake's theory can be seen as an operationalization of certain aspects of a more general theory of communication.

In the book *Explanation and Cognition*, Keil and Wilson (2000) collect recent research on explanations from a cognitive science point of view. They set out to study a set of questions about explanation, such as: "Are there different kinds of explanation?", "Do explanations correspond to domains of knowledge?", and "How central are causes to explanation?".

These questions are examined by studying for example whether there are fundamental differences between explanations offered and requested by children and those used by scientists (Brewer et al., 1998). Keil and Wilson describe three broad types of explanation; the scientific, the narrative and the goal-based. The narrative explanation is what we use in daily life to chain together events. An example of this would be to explain that a window is broken because the children playing football in the back yard accidentally kicked the ball through the glass. This kind of explanation contrasts with explanations that explain events from generalized principles, which Keil and Wilson call scientific explanations. The last type, the goal-based explanation, are useful to explain actions in terms of the actors' goals. For instance, the workings of a car may well be described by mechanical laws, but the reasons for building it are better explained in terms of the goals of car manufacturers and consumers.

While the scientific explanation typically can be used to predict events from a set of observations, the narrative explanation can be formed after the fact and has little in the way of predictive power. Keil and Wilson claim that narrative explanations are more intuitive to people. They suggest that these explanations are useful in that they may narrow down the inductive space or help us gather information in a more efficient fashion. For instance, a spectator at a cricket match may ask questions about the rules so that he is better able to understand and gather information about the game in real time. In this scenario, prediction may not be very relevant to the spectator – he is simply trying to understand the game.

From an AI perspective, the difference between the narrative and scientific explanations is interesting. In expert systems, explanations initially focused on how the system made the prediction by showing how it followed from generalized rules. In essence, the system attempted to show how the conclusion must follow from the knowledge contained in the system. Although the process used to do this was not necessarily or typically deductive (at least in expert systems),

the explanations produced seem to be closer to the scientific explanations than the narrative. Applying this to case-based reasoning, it seems likely that using a similar case to justify a conclusion is closer to a narrative account than using a rule from a rule-based system. However, the case will not typically contain a narrative account of how a conclusion followed from the findings. Rather, the way a case is used is that it represents a very local "rule" for drawing the conclusion but if the case contains a causal account of how the solution followed from the findings, it is not typically used by the system for explanation. Keil and Wilson suggest that depending on the goals of the users, they may not seek to know how the system found the case, but rather how the case's solution is a product of its findings.

We round off this section with a final remark about the goodness of an explanation. We have seen that the truth, or correctness, of an explanation is generally not sufficient to make it good. The flagpole height explanation is one example. An overly general explanation is another. What about necessity? Is correctness – or truth – a necessary criterion for a good explanation? One of the counter-arguments is related to the notion of truth. McDermott (1987) argues that an explanation may be good merely by making the observed facts probable, not necessarily proving their truth. Another argument is related to pragmatics. Achinstein (1983, p. 108) expresses it as follows: "The goodness or worth of an explanation is multidimensional; correctness is only one dimension in an evaluation. An explanation is evaluated by considering whether, or to what extent, certain ends are served. The ends may be quite varied. They may concern what are regarded as universal ideals to be achieved, particularly in science, e.g. truth, simplicity, unification, precision. Other ends are more 'pragmatic'."

3. Explanations in Expert Systems

In early rule-based expert systems like MYCIN the user could ask *how* the system reached the conclusion presented, and an explanation in the form of a reasoning trace from the system would be presented. This would offer the user a degree of transparency into how the system reached its conclusions. The user could also choose a *why* explanation that would provide a more local explanation that justified why a question was asked.

It was soon found that this capability was insufficient for answering many of the explanation requests from users. For instance, the problem solving strategy of a rule-based expert system is implicitly defined in the system, but was not explicitly encoded in such a way that it was accessible or easily explained to an end user. NEOMYCIN extended MYCIN's capabilities in this respect by explicitly encoding strategic information (Clancey, 1983).

Another notable extension was the XPLAIN system (Swartout, 1983). This system would record additional domain knowledge associated with each rule, so that the system could produce explanations that gave background information for the rule, and pointers to literature.

The focus of these early extensions was usually to extend the explanation capabilities by adding the type of knowledge required by the user. These explanations could be divided into four types (Swartout and Smoliar, 1987; Chandrasekaran et al., 1989; Gregor and Benbasat, 1999):

- **Reasoning Trace:** Producing an explanation from the trace of the reasoning process used by the system to find the solution. Examples are MYCIN's *how* and *why* explanations (Clancey, 1983).
- **Justification:** Providing justification for a reasoning step by referring to deeper background knowledge. This type of explanation was first offered by the XPLAIN system (Swartout, 1983).
- **Strategic:** Explaining the reasoning strategy of the system. The NEOMYCIN system first provided this kind of explanation (Clancey, 1983).
- **Terminological:** Defining and explaining terms and concepts in the domain. This type of explanation was identified in (Swartout and Smoliar, 1987).

Although it was found that expert system designers, and to some extent domain experts, appreciated the reasoning trace explanations, many end users did not understand or were not interested in the inner workings of the expert system. Later analysis of failed expert systems suggested that many of the attempts to provide explanations in early systems failed because they were incomprehensible to the user or failed to address the users' goals in demanding an explanation (Majchrzak and Gasser, 1991).

In response to this, further research went into how explanations could better be generated dynamically to fit the user's needs and goals. In (Swartout and Moore, 1993), five requirements for the explanation

capability of expert systems were put forth. The *fidelity* requirement says that the explanation given should mirror the knowledge used by the system in its reasoning. The explanation should also have *low construction overhead* or justify any increased resources spent on it. It must always remain *efficient* and not degrade runtime capability. The explanation produced must also be *understandable* to the user, and must be *sufficient* in that enough knowledge must be represented in the system to answer the question the user may have.

The fidelity criterion mentioned above appears more controversial than the other four. Wick and Thompson (1992) argue that explanation should be viewed as a problem-solving process separate from the process used to determine the conclusion in the first place. They contend that while expert system designers need explanations that accurately represent the reasoning done by the system, this may be inappropriate for an end user. They suggest three major explanation goals. *Verification* is the goal of the knowledge engineer in verifying that the system works as it should. A successful verification explanation would accurately and precisely convey the knowledge of the system on the knowledge level. *Duplication* is to help the domain expert examine the knowledge of the system. The system should not only expose its own knowledge, but help the user learn the methods and knowledge used in the problem solving process. Finally, the goal of *ratification* is to increase the end user's confidence in the system's conclusion.

Wick and Thompson suggest that each of these goals has different audience and focus. As the goal moves away from verification toward ratification, the explanation process should increasingly be decoupled from the reasoning process in order to provide explanations that focus on the solution. This allows the system to convey tailored information about the domain to the user. The higher degree of decoupling from the original reasoning processes will decrease the fidelity of the explanation as defined by Swartout and Moore (1993), but Wick and Thompson point out that explanations provided by human experts also tend to lack fidelity, although they are nevertheless perceived as useful.

As expert systems have been deployed in production environments, empirical studies have been conducted to identify when different kinds of users ask for explanations, and what they expect to get from them. Results from this research include the observation that novices tend to ask for explanations to learn or clarify, thus preferring justification and terminological explanations (Mao and Benbasat, 2000). Experts

tend to require explanations to verify the reasoning of the system and explain away surprising results. As such, they tend to prefer strategic and reasoning trace explanations. A full survey of the empirical studies on explanations is beyond the scope of this paper, but we recommend (Gregor and Benbasat, 1999) for a more in-depth review.

A number of educational systems have also been built as extensions of expert systems. These systems have as their goal not only to help the user solve a problem, but also teach the user about the domain. One idea emerging from these systems is that it is often beneficial for learning if the user participates in the formation of explanations. The Cognitive Tutor (Aleven and Koedinger, 2002) system assists students in explaining solutions to geometry problems. They find that this helps the students learn the task better and helps them avoid bad generalization. Ford et al. (1993) use Concept Maps to help the student navigate an expert model to form explanations.

4. Explanation in CBR

We have reviewed several attempts to define criteria for explanations and categorizations of different kinds of explanations. Philosophical accounts focus on criteria for scientific explanations, while the cognitive accounts describe how humans use explanations in a wide range of contexts. However, many explanations may be produced that are not perceived as useful in a given context. This happens even if they fulfill criteria of what is considered a good explanation.

The research on explanation within expert systems provides a focus for a situational context that is similar to what we find with most case-based reasoning systems. Although the technology for generating and presenting advice is different from traditional rule-based expert systems, most CBR systems today are computer systems that give decision advice to human users. Because of this similarity in situational context, it is reasonable to believe that the typology of explanations useful in expert systems will be a good fit for CBR. In this section we introduce five explanation goals that are strongly influenced by expert systems.

Below the abstraction level of the explanation goals, we need to look at particular issues in applying these goals to CBR. For instance, traditional rule-based systems paraphrased the rules to form explanations. While CBR systems typically do not have rules, the basic unit of knowledge in CBR – the case – can also be used to produce expla-

nations. It has long been an article of faith in the CBR community that displaying an earlier solved case that represents a situation similar to the present problem situation can serve as a good explanation for adopting the solution of the previous case. After presenting the explanation goals, we will examine this approach further. In addition, we will discuss if cases are really the only source of knowledge that should contribute to explanations in a CBR system.

4.1. *Explanation goals*

We will designate our explanation categories based on a set of *explanation goals*. We do this in order to recognize that a single explanation technique can serve many of these goals at once, and that not all of these goals are of equal importance in all systems. The goals are based on the four content categories from Gregor and Benbasat (1999) as presented in the Section 3. In addition, we have a category that focuses on the learning perspective, similar to the Duplication goal of Wick and Thompson (1992). Our aim is not to provide an exhaustive list of goals – the rationale for introducing them is to discuss how some current explanation criteria, and methods, hold up in the light of these goals which have proved quite universal in expert systems.

4.1.1. *Explain How the System Reached the Answer (Transparency)*

"I had the same problem with my car yesterday, and charging the battery fixed it."

The goal of an explanation of this kind is to impart an understanding of how the system found an answer. This allows the users to check the system by examining the way it reasons and allows them to look for explanations for why the system has reached a surprising or anomalous result. If transparency is the primary goal, the system should not try to oversell a conclusion it is uncertain of. In other words, fidelity is the primary criterion, even though such explanations may place a heavy cognitive load on the user. The original *how* and *why* explanations of the MYCIN system would be good examples.

This goal is adapted from the reasoning trace type of explanations from Gregor and Benbasat (1999) and the verification goal of Wick and Thompson (1992). As they suggest, this goal is most important with knowledge engineers seeking to debug the system and possibly domain experts seeking to verify the reasoning process. It is also reasonable to think that in domains with a high cost of failure it can be

expected that the user wishes to examine the reasoning process more thoroughly.

4.1.2. *Explain Why the Answer is a Good Answer (Justification)*
"You should eat more fish – your heart needs it!"
"My predictions have been 80% correct up until now."

This is the goal of increasing confidence in the advice or solution offered by the system by giving some kind of support for the conclusion suggested by the system. This goal allows for a simplification of the explanation compared to the actual process the system goes through to find a solution. Potentially, this kind of explanation can be completely decoupled from the reasoning process such as advocated by the ratification goal of Wick and Thompson, but it may also be achieved by using additional background knowledge (as in XPLAIN) or reformulation and simplification of knowledge that is used in the reasoning process. As such, this goal also contains the category of justification explanations from Gregor and Benbasat (1999). Empirical research suggests that this goal is most prevalent in systems with novice users (Mao and Benbasat, 2000), in domains where the cost of failure is relatively low, and in domains where the system represents a party that has an interest in the user accepting the solution. Some e-commerce recommender systems fall into this category, although Herlocker et al. (2000) suggest that in high-cost domains (such as expensive vacation packages as opposed to relatively cheap books or music) users are unlikely to accept solutions without more in-depth explanations.

4.1.3. *Explain Why a Question Asked is Relevant (Relevance)*
"I ask about the more common failures first, and many users do forget to connect the power cable."

An explanation of this type would have to justify the strategy pursued by the system. This is in contrast to the previous two goals that focus on the solution. The reasoning trace type of explanations may display the strategy of the system implicitly, but it does not argue why it is a good strategy. In conversational systems, the user may wish to know why a question asked by the system is relevant to the task at hand. It can also be relevant in other kinds of systems where a user would like to verify that the approach used by the system is valid. In expert

systems, this kind of explanation was introduced by NEOMYCIN (and was one of the types of explanation discussed in Section 3).

4.1.4. *Clarify the Meaning of Concepts (Conceptualization)*
"By 'conceptualization' we mean the process of forming concepts and relations between concepts."

One of the lessons learned after the first wave of expert systems had been analyzed was that the users did not always understand the terms used by a system. This may be because the user is a novice in the domain, but also because different people can use terms differently or organize the knowledge in different ways. It may not be clear, even to an expert, what the system means when using a specific term, and he may want to get an explanation of what the system means when using it. This requirement for providing explanations for the vocabulary was first identified by Swartout and Smoliar (1987).

4.1.5. *Teach the User About the Domain (Learning)*
"When the headlights do not work, the battery may be flat as it is supposed to deliver power to the lights."

All the previous explanation goals involve learning – about the problem domain, about the system, about the reasoning process or the vocabulary of the system. Educational systems, however, have learning as the primary goal of the whole system. In these systems, we cannot assume that the user will understand even definitions of terms, and may need to provide explanations at different levels of expertise. The goal of the system is typically not only to find a good solution to a problem, but to explain the solution process to the user in a way that will increase his understanding of the domain. The goal can be to teach more general domain theory or to train the user in solving problems similar to those solved by the system. In other words, the explanation is often more important than the answer itself. Systems that fulfill the relevance and transparency goals may have some capabilities in this area, but a true tutoring system must take into account how humans solve problems. It should not attempt to teach the user a problem solving strategy that works well in a computer but that is very hard for people to reproduce.

This goal has similarities with the duplication goal of Wick and Thompson (1992), where the system should be able to explain itself on the knowledge level in order to transfer its knowledge to a user.

Although Wick and Thompson claim that this goal is primarily for the domain expert to gain an understanding of the system's capabilities, the name and description suggest that the goal is to transfer the knowledge contents and competence of the system to the user. Participatory explanation techniques (Ford et al., 1993; Aleven and Koedinger, 2002), where the system helps students form explanations, are good examples of techniques for achieving this goal.

4.2. *The case as explanation*

The case-based reasoning methodology seems quite transparent. It is fairly easy to understand the basic concept of searching for very similar, concrete cases to help solve the current problem. This understanding has supported the basic approach to explanation in CBR – displaying the case that is most similar to the current problem. In addition to the intuitive feeling and ad hoc reports that this works, there has been research showing that displaying cases along with the solution significantly improved user confidence in the solution compared to only showing the solution, or displaying a rule that was used in finding the solution (Cunningham et al., 2003).

There is also theoretical support for the case-as-explanation method fulfilling the justification goal by looking at it from the viewpoint of Achinstein's theory. It is likely that a previous example with a high degree of similarity would increase the relative probability of the solution from this case compared to other solutions (Faltings, 1997). However, the underlying assumption of this approach seems to be better represented by van Fraassen's (1980) framework. Displaying the retrieved case to the user is a kind of knowledge communication that allows the user to make his own judgment about the similarity of the old situation compared to the current one.

Both of these views depend on the user's ability to understand the case and to confirm the similarity assessment. In general, for the retrieved case to serve as an explanation to the user, the similarity between the retrieved case and current problem must be obvious to him. The difficulty for the user in comparing cases increases as the case structure becomes more complex and the similarity measures more convoluted. It also increases with the use of more complex adaptation techniques where the retrieved case may not be the most similar but one that fits the adaptation process, as suggested e.g. by Smyth and Keane (1998).

There is another problem. Displaying the case may serve as a window into the methodology of the reasoner. It does not, however, help the user to understand how the symptoms connect with the solution. End users may be less concerned about how the most similar case was found than why the solution in the presented case works. Based on Keil and Wilson's (2000) work (Section 2.3), we suggest that such an account would be required for the case to serve as a narrative episode and explanation for humans.

Schank's research suggests that people do use single cases to explain extraordinary situations where no more general theory covers the situation – they are a sort of index of situations where the general model failed. However, his theory also suggests that these single exceptions are perceived as very tentative in their predictive power compared to general knowledge that has been confirmed again and again. As we will see in Section 5, some recent research in CBR attempts to address these shortcomings.

4.3. *Knowledge containers*

The competence of a knowledge-based system depends on the knowledge sources available to it. Richter (1995) describes the knowledge sources used in problem solving as knowledge containers. Rule-based systems typically have *facts* and *rules* as knowledge containers, while Richter identifies four such containers for CBR systems – the *case base*, the *similarity measure*, the *adaptation knowledge* and the *vocabulary*.

The *vocabulary* provides the basis for the other knowledge containers by defining the terms and structure of the domain. The *case base* contains the concrete or prototypical problems previously solved by the system or otherwise provided to it. The *similarity measure* contains knowledge about how to compare cases and compute a similarity ranking of cases relative to a new problem, while the *adaptation knowledge* allows the reasoner to change the solution of a previous case to better fit a new problem.

Richter points out that given a complete case coverage of the problem domain, the similarity measure and adaptation problems become trivial since any problem can simply be looked up. Similarly, if we have perfect adaptation knowledge so that any previous solution can be adapted to a perfect new solution, the process only requires a starting position for the adaptation so that there is little need for cases or a similarity measure. Finally, if the system is always able to order the

cases so that the cases with the correct solution are ranked highest, a classification system only needs a case representing each solution class, and there is no need for adaptation knowledge. This means that CBR systems may put different weights on these containers depending on what is most convenient for the domain and system.

Roth-Berghofer (2004) points out that this insight by Richter places in doubt the idea that displaying the best case is a sufficient explanation – at least if the system places any weight on the other knowledge containers. If much of the competence of the problem solving emerges through adaptation, it will be hard to explain the reasoning of the system without using the adaptation knowledge. This is certainly true if the goal of the explanation is to provide transparency, but it can also become a problem in learning and justification if a solution is justified by a case that is not obviously similar and has a slightly different solution than that suggested by the system. In addition, conceptualization and relevance explanations cannot be provided by the case base. The vocabulary container seems perfect to provide explanations that serve to help conceptualization, but it is not clear which knowledge containers can support strategic explanations. Possibly this requires a fifth knowledge container in CBR in the same way that it required a different level of representation in rule-based expert systems.

5. Survey of Explanation in CBR

In this section, we review explanation techniques in different case-based reasoning systems, with an emphasis on the more recently developed techniques. Many early CBR systems also had explanation capabilities, extensive surveys of which have been published elsewhere, for instance (Kolodner, 1993).

5.1. *Displaying similar cases*

The most common form of explanation in CBR systems amounts to displaying the most similar case. This technique is used by many research systems, e.g. CARES (Ong et al., 1997), and in commercial CBR tools such as orenge (developed by empolis). In the previous section, we discussed limitations of this approach and recently CBR researchers have attempted to address some of these limitations.

Doyle et al. (2004) point out that the most similar case is not necessarily the most convincing case. When trying to convince his parents to let him see the latest Harry Potter movie, a child knows that friends that are younger than him are more convincing examples than his older best friend even if he is the closest match in terms of age. Doyle et al. suggest a method for selecting cases of the same solution class as the problem case that are closer to a class boundary than the problem case for explanation purposes. This has the effect of increasing the awareness of class borders in the user. However, it may also provide evidence that is atypical. Any parent knows that a child will choose his examples very carefully, avoiding those children that were not allowed to see the movie.

Recently, research on ensemble classifiers has shown that the aggregated output of a set of classifiers can be more accurate than a single classifier. Such an approach may make it harder to find a proper case to display as an explanation to the user. Zenobi and Cunningham (2002) have addressed this by introducing a meta-layer over the set of case-based classifiers that perform the aggregation step. Since this technique is also case-based, it also produces neighbour cases that can be used in explanation.

We have argued that when emphasis is placed on different knowledge sources than the cases, the nearest case may serve neither the justification nor the transparency goal. One way of dealing with this problem is to introduce explanations on multiple layers in the CBR process. The case may serve as a type of top-level explanation, with more detailed levels of explanations for each case feature. The feature weighting may be explained in probability terms and there may also be ways of illustrating the coverage of cases. In the CREEK system (Aamodt, 2004), the user may ask for explanations at the attribute level, and the generation of this explanation depends on the similarity measure. A simple example is when the similarity of attributes on an interval scale is explained, the range of all values for this attribute is shown to the user so he can more easily see how similar they are in the context of known cases.

This method may even be used to provide explanations for non-CBR systems, as demonstrated by Nugent and Cunningham (2005). They use this technique to justify solutions produced by black-box systems such as neural networks and support vector machines. This is done by extracting local feature weights for a given solution from the black-box system, and using these, the most similar case from the training data is retrieved and displayed to the user as a justification.

5.2. *Visualization*

Visualization can make it easier for a user to see whether a solution is correct. For example, McArdle and Wilson (2003) suggest a technique where the similarity of a set of cases is projected onto a two-dimensional surface in such a way that the distance between them roughly corresponds to the similarity. While this is a simplification of the similarity measure, it allows the user to get an overview of the case space.

Good visualization techniques may at the same time increase the understanding of the reasoning process and reduce the cognitive load for the user. As such, visualization techniques may at the same time serve the justification and transparency goals. One example of this is the way the FormuCaseViz system (Massie et al., 2004) visualizes how a number of cases differ on a number of attributes and how this leads to predictions. This is done by drawing a two-dimensional graph, where each attribute is represented by a vertical line and the values of the attributes are placed at intervals along that line. A case is then represented as a line along the horizontal axis that intersects the attribute lines at the points representing the value this case has for that attribute. This technique allows at-a-glance comparisons and makes it very easy for people to spot eventual attributes where the values of a problem case do not match those of the cases it is being compared to.

5.3. *Explanation models*

Knowledge-intensive systems may contain more generalized knowledge that can be of use to a human user in structuring his own internal model of the domain. This should allow knowledge-intensive systems to produce explanations that help in tying general domain knowledge and cases together. Examples of this are the IBP system (Brüninghaus and Ashley, 2003) and the CATO system (Aleven and Ashley, 1997) where model-based reasoning is combined with CBR to predict the outcome of legal cases. This is done by using both older cases and a weak domain model to produce legal arguments. In these systems the explanation is the solution, and the explanation (or argument) must be complete (fulfilling the transparency goal) in order to give justification to the prediction. This can make the argument complex, but as it uses the same problem-solving method as courts do in

solving these cases, the target users (lawyers) are able to make sense of them.

It is possible to use models that are built explicitly for explanation, e.g. models that are not used in the reasoning process and used only to generate explanations. In CREEK (Aamodt, 1991), the model-based reasoner can use a causal model to produce explanations of why observations in a case can cause or imply the solution suggested by the system. These explanations are produced purely through backward chaining of causal relations from a solution already given by the CBR component to find how it may be connected to the observed features. As such, the explanations produced tend to fulfill the justification goal. The downside is that these explanations are produced after the fact and are not an accurate representation of how the system found the solution. It also requires a knowledge acquisition effort in building the causal model, but this model can then be tailored to the typical user's level of expertise.

The Colibri environment (Díaz-Agudo and González-Calero, 2000; Bello-Thomás et al., 2004) assists the development of systems that utilize a task/method ontology to make an explicit model of the system structure. The user can see how the CBR reasoning tasks and problem solving methods are linked to the model of general domain knowledge. In this way, transparency of the reasoning process is achieved.

Bergmann et al. (1993) make use of general domain knowledge for explaining similarity. The mechanism is based on an abstraction method, involving the modeling of domain knowledge at several levels of abstraction. The explanation produced justifies the correctness of the solution, rather than reproducing its trace, and is used both for retrieval and adaptation purposes.

5.4. *Reasoning trace*

The reasoning trace method is feasible in systems that produce explanations as part of the reasoning process. The LID (Lazy Induction of Descriptions) system (Plaza et al., 2005) is an example of this. LID will attempt to find the categories that are maximally general while still as accurately as possible predicting the solution class of member cases. The induction process is similar to techniques used to induce decision trees, but is lazily applied at problem solving time. This process leaves a hierarchy of general-to-specific categories that may serve as an explanation as to the membership category of the problem case.

The relevance goal can also be fulfilled by offering explanations to the user that increase the understanding of the reasoning process. The Top Case mixed-initiative recommender system pursues a strategy where it selects questions that potentially strengthen the match for its currently selected best hypothesis case (McSherry, 2005). This strategy is explained to the user by showing how an answer to this question could affect the recommendation. Top Case can for instance ask what region the user would like to take a holiday in. If the user would like to know why this is relevant in recommending a trip, the system can offer an explanation like "Because if the region = Tyrol this will increase the similarity of Case 510 from 0.28 to 0.44 and eliminate 866 cases, including Case 574". Because Top Case always displays the best matching cases found so far, the user can relate to these case labels and see how his answer affects the recommendation process.

5.5. *Case space awareness*

In case-based reasoning it is important that the transparency goal is not only applied to the reasoning process but also to the case base itself as much of the competence of the system lies in its collection of cases. The visualization techniques discussed above can help to achieve this, as can displaying similar cases, both opposing and supporting the conclusion.

The Stamping Advisor (Leake et al., 2001a) is a system to support feasibility analysis for the production of sheet metal parts in the automotive industry. For the feasibility analysis, it is important to understand the potential problems of a new design. The Stamping Advisor therefore displays two so called "bracketing cases", one where an identified problem exists and the most similar one without the problem. The user can thus more easily identify the limits of the design.

Reilly et al. (2005) suggest that their system's compound critiques can play a similar role in recommender systems. The compound critiques generated by their system identify sets of attribute values that are correlated so that the user can see what kind of trade-offs he must make when deciding on a product. An example is that "higher price" and "bigger screen" may correlate when browsing for a TV. While this may not be an example of explanation in the usual sense, it illustrates that quite a wide range of techniques may have explanatory properties as long as they impart knowledge that increase the user's awareness of the problem domain.

5.6. *Contrasting evidence*

The goal of transparency demands that the system does not try to hide conflicting evidence to its recommendation. In CBR systems this can be achieved by displaying the most similar case(s) that are not of the proposed solution class to the user. The Stamping Advisor (Leake et al., 2001a) mentioned in Section 5.5 displays cases which are close to each other but with different findings. Compound confidence measures can also be calculated as for instance in (Cheetham and Prince, 2004).

McSherry's ProCon system (McSherry, 2004) identifies which attributes of the input case support the suggested solution and which attributes oppose it. The attributes are identified as opposers or supporters of a solution based on how the attributes affect the probability of the solution. This allows the system to present justifications that are not only simpler to understand than possibly complex case similarity measures, but also help the user to identify what attributes are important to the conclusion.

The AHEAD system (Murdock et al., 2003) is an interpretative CBR system (Kolodner and Leake, 1996) designed to detect potential asymmetric threat situations (such as a terrorist attack). A situational interpretation is formed by constructing a trace of events, attempting to match it to prototypical threat situations. When matching this trace, AHEAD attempts to justify its conclusion by forming an argument that lists factors for and against the hypothesis based on what matches and does not match the expected findings in the prototypical threat situation. This allows the user to see evidence both for and against the conclusion. The difference between how AHEAD and ProCon identify contrastive evidence is that AHEAD is a knowledge-rich system where expectations about threat situations are modeled in advance by an expert, while ProCon uses machine learning techniques to generalize such knowledge from the case base.

5.7. *Simplified problem solving strategy*

The conversational CBR community has developed methods that are particular to the relevance explanation goal. One such method is used by the Strategist system (McSherry, 1998), a mixed-initiative conversational diagnosis system where the user may enter a dialog where he is asked a single question at a time. The original Strategist induced a decision tree from a set of instances with the explicit goal that for

each question the user is asked, the system would be able to give a good explanation for why this question was important to answer. The extension of Strategist into a CBR system (McSherry, 2001) does not form a decision tree in advance, but the question selection method is the same. As an example, the system prefers questions that could confirm or eliminate possible outcome classes in the domain. This allows it to form simple explanations of the relevance of questions the user is asked. In the computer fault domain, for example, the relevance of the question "Can you hear the fan?" might be explained, in the context of other reported evidence, by telling the user "Because if the fan cannot be heard this will confirm faulty power cord as a possible cause" (McSherry, 2001, Figure 7).

5.8. *Concept maps*

Semantic network representation of knowledge such as in the CREEK system (Aamodt, 1991; Sørmo and Aamodt, 2002) may provide some explanatory support showing the part of the network around the concept the user is interested in. In particular this method may further the conceptualization goal by showing how the system views concepts in relation to other concepts and thus helps the user understand the system's conceptualization. Methods for sharing conceptualizations through two-dimensional visual-based representations are often referred to as topic or concept maps. There has been some work using these in CBR (e.g. Leake et al., 2001b), although the focus on this work has so far not been on its use for explanation.

5.9. *Machine learning induction*

The learning goal seems to have a strong preference for knowledge-intensive methods, but generalization may also be done lazily by a number of machine-learning algorithms. The CBR Strategist (McSherry, 2001) and ProCon (McSherry, 2003, 2004) systems mentioned earlier are examples of this as they do induction when presenting an explanation to the user, but they do so lazily. In the example in Section 5.7, CBR Strategist observed that all surviving cases with "fan cannot be heard" have the same solution ("faulty power cord") and can inform the user that this feature is enough to confirm the solution. The CBR Strategist system may be fairly effective in training

users in the skill of identifying computer faults. A limitation of this approach is that the system cannot introduce higher-order concepts or relate to how generalized concepts are used in the environment outside the system.

6. Challenges

Recently, there has been a renewed focus on explanation in case-based reasoning. There are, however, still many challenges that remain to be addressed. In this section we identify four such challenges for the future of explanation research in CBR.

6.1. *Maintaining transparency in complex systems*

Displaying the closest case is quite near the actual reasoning process in simple case-based reasoning systems, but when more advanced methods like feature weighting and complex similarity measures are introduced, it will be necessary to provide additional information in order to fulfill the transparency goal. For example, in a k-nearest neighbour system, the transparency goal is no longer fulfilled by only displaying the best case if $k > 1$. The difficulty for the user in comparing cases increases as the case structure becomes more complex and the similarity measures more convoluted. It also increases with the use of more complex adaptation techniques where the retrieved case may not be the most similar but one which fits the adaptation process (e.g. Smyth and Keane, 1998).

In general, it can be argued that the use of other AI technologies in the CBR cycle as suggested by Watson (1999) increases the difficulty for the user to see the explanative character of the case since it is necessary to have an at least intuitive understanding of the different techniques used in order to understand why the case presented offers a solution to the problem. If we cannot expect such an understanding, the steps taken by the different components also have to be explained. For example, consider a system where the cases contain image data, and the similarity of two images is assessed by a neural network. Then the similarity measured through the neural network will have to be explained alongside the presented case – at least if complete transparency is the goal.

One way of dealing with this problem, as suggested in Section 5.1, is to introduce explanations on multiple layers in the CBR process.

The case may serve as a type of top-level explanation, with more detailed levels of explanations for each case feature. One problem with this approach is that although it satisfies the transparency goal, the cognitive load of the user increases as similarity measures increase in complexity. This has the interesting effect that as case-based systems grow more complex and are more able to help with extremely hard problems, the value of the case as an explanation may go down.

6.2. *Providing justification to novice users*

As we have mentioned before there is an implicit assumption in presenting the case to the user that he is able to do a similarity comparison himself. Just as an explanation may not be required when the solution offered by a system matches the beliefs of the user, an explanation may not be necessary when the similarity between two cases is obvious. No new knowledge is required from the system in these cases in order for the conclusion to be accepted.

In complex domains with complex similarity measures, the similarity may not be so clear, especially to novice users. This has been seen in other kinds of knowledge-based systems, where explanation methods based on showing in detail how the problem-solver found the answer were deemed too complex to be useful by actual users (Majchrzak and Gasser, 1991). For the novice users, a multi-level reasoning trace places a high cognitive load on the user and may be too complex or too time consuming to understand. In Section 5, we have reviewed methods for simplifying this explanation as a means to achieve the justification goal, but many of these come at a cost to the fidelity of the explanation. While this may be acceptable in some domains, it is usually a goal to find simplification methods that preserve as much of the fidelity as possible. If a system uses justification explanations to overstate its confidence in the conclusion, it is likely that the user's confidence in the system will decrease over time.

However, research in the cognitive science and expert systems communities suggest that the goal of the user is not necessarily to gain an understanding of how the system solved the problem. When presented with a similar case, it may not be obvious to the user why the solution of the retrieved case was good even for the retrieved case itself. For these situations, providing justification explanations that do not stem from the reasoning process is not misleading the user but is addressing a different explanation goal.

6.3. *Connecting cases to general knowledge in tutoring*

Cognitive theories of learning (e.g. Schank, 1982) assume that people start learning in a new domain by looking at concrete cases, or episodes. At some point, however, humans start to generalize the concrete episodes. This is in contrast to those approaches to CBR that rely on pure just-in-time induction. These lazy learners are well equipped to provide the student with example cases, and although this can be useful, they are ill equipped to assist the learner in generalizing from these examples.

Today, most systems that attempt to tutor rely on generalized knowledge in addition to cases. Kolodner's (1997) more recent work takes this approach, as does our own (Sørmo and Aamodt, 2002; Sørmo, 2005). As mentioned in Section 5, there are knowledge-light techniques that do produce generalizations that may be useful for learning in humans e.g. CBR Strategist (McSherry, 2001) and ProCon (McSherry, 2004), but these techniques are currently not applied to tutoring.

6.4. *Scope of explanation efforts*

In Section 4, we noted that the case-as-explanation method uses only one of the four knowledge containers Richter identified in CBR (Richter, 1995) – the case base. Competence arising from the three other containers (similarity measure, adaptation knowledge and vocabulary) is not used for explanation. We have surveyed several innovative systems, e.g. FormuCaseViz (Massie et al., 2004) and ProCon (McSherry, 2004) that explain and visualize the similarity measure, but after CASEY (Koton, 1988), we have seen few efforts at explaining adaptation or vocabulary.

In our own research, we are working on combining different views on explanation. The goal is to integrate them into the CBR system design process in order to be able to make better use of the explanatory potential of the different knowledge containers (Roth-Berghofer et al., 2005; Roth-Berghofer and Cassens, 2005).

A parallel to the above is seen if we look at explanation efforts in the light of the CBR Cycle (Aamodt and Plaza, 1994). Explanation efforts seem to focus on the *retrieve* step with little effort used to explain the other three steps (*revise*, *reuse* and *retain*). This is perhaps a natural consequence of the greater focus *retrieve* receives in problem-solving, but a CBR system that does not, for instance, retain all cases should be able to explain why a case is dropped or merged into another.

7. Conclusions

We have surveyed theories of explanation from the philosophy of science, linguistic and cognitive science communities, and also attempted to draw on the experiences with explanations from the expert-systems community in AI. From these theories and experiences, we believe it is useful to analyze the explanation requirements in the form of explanation goals. The goals that an explanation is required to achieve vary with the domain, system, and user. It can be hard to model these dynamically for the system itself, but the designer of the system can often make assumptions about the goals and capabilities of prototypical users of the system. We also believe that explicitly formulating such explanation goals facilitates the discussion of possible conflicts between goals and makes clear how different approaches tend to favor different types of goals. Although the goals discussed in this paper are abstract goals made to fit a wide range of CBR systems, they are not an attempt at completeness. There will be some CBR systems that fall outside the situational context we have defined for our explanation goals, and individual systems will also benefit from formulating more specific explanation goals that are tailored to their context.

In knowledge-intensive systems there has been continuous work on explanation, but recently this topic has received wider interest as exemplified by many of the methods we survey in Section 5 of this paper. However, these have mainly been focused on the *retrieve* step in the CBR cycle. Although we are starting to see explanation methods that address competence arising from the similarity-measure knowledge container in addition to the case base, methods explaining vocabulary and adaptation are still rare.

Acknowledgements

We would like to thank the anonymous reviewers and the participants and organizers of the ECCBR-04 Workshop on Explanation in CBR for their helpful and constructive feedback.

References

Aamodt, A. (1991). A Knowledge-Intensive, Integrated Approach to Problem Solving and Sustained Learning. Ph.D. thesis, Norwegian Institute of Technology, Department of Computer Science, Trondheim (http://www.idi.ntnu.no/grupper/su/publ/phd/aamodt-thesis.pdf).

Aamodt, A. (2004). Knowledge-Intensive Case-Based Reasoning in CREEK. In Funk and González–Calero, (eds.) 1–15, Springer: Berlin Heidelberg.

Aamodt, A. & Plaza, E. (1994). Case-based Reasoning; Foundational Issues, Methodological Variations and System Approaches. *AI Communications*. **7**(1): 39–59.

Achinstein, P. (1983). *The Nature of Explanation*. Oxford University Press: Oxford.

Aleven, V. & Ashley, K. D. (1997). Teaching Case-Based Argumentation through a Model and Examples: Empirical Evaluation of an Intelligent Learning Environment. In du Boulay, B. and Mizoguchi, R. (eds.) *Artificial Intelligence in Education, Proceedings of AI-ED 97 World Conference*. 87–94. IOS Press. Amsterdam:

Aleven, V. & Koedinger, K. R. (2002). An Effective Metacognitive Strategy: Learning by Doing and Explaining With a Computer Based Cognitive Tutor. *Cognitive Science* **26**: 147–179.

Ashley, K.D. & Bridge, D.G. (eds.) (2003). Case-Based Reasoning Research and Development: Proceedings ICCBR 2003, No. 2689 in LNAI. Springer: Berlin Heidelberg.

Bello-Thomás, J. J., González-Calero, P. & Díaz-Agudo, B. (2004). JColibri: an Object-Oriented Framework for Building CBR Systems. In Funk and González–Calero (eds.) 32–46, Springer: Berlin Heidelberg.

Bergmann, R. and Pews, G. & Wilke, W. (1993). Explanation-Based Similarity: A Unifying Approach for Integrating Domain Knowledge into Case-Based Reasoning for Diagnosis and Planning Tasks. In *Topics in Case-Based Reasoning: Proceedings EWCBR 1993*. 182–196.

Brewer, W.F. & Chinn, C.A. & Samarapungavan, A. (1998). Explanations in Scientists and Children. *Minds and Machines* **8**: 119–136.

Bromberger, S. (1965). An Approach to Explanation. In R. J. Butler (ed.) *Analytical Philosophy*, vol. 2.: 72–105. Basil Blackwell: Oxford.

Bromberger, S. (1966). Why Questions. In Colodny, R. G. (ed.) *Mind and Cosmos*. 86–111. Pittsburgh University Press: Pittsburgh.

Brüninghaus, S. & Ashley, K.D. (2003). Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases. In Ashley & Bridge (eds.) 65–79. Springer: Berlin Heidelberg.

Chandrasekaran, B., Tanner, M.C. & Josephson, J. R. (1989). Explaining Control Strategies in Problem Solving. *IEEE Expert*, **4**(1): 9–15.

Cheetham, W. & Price, J. (2004). Measures of Solution Accuracy in Case-Based Reasoning. In Funk & González–Calero (eds.) 106–118. Springer: Berlin Heidelberg.

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

Cunningham, P., Doyle, D. & Loughrey, J. (2003). An Evaluation of the Usefulness of Case-Based Reasoning Explanation. In Ashley & Bridge (eds.) 122–130. Springer: Berlin Heidelberg.

Díaz-Agudo, B. & González-Calero, P. (2000). An Architecture for Knowledge-Intensive CBR systems. In *Advances in Case-Based Reasoning: Proceedings EWCBR 2000*. Springer: Berlin.

Doyle, D., Cunningham, P. Bridge, D. & Rahman, Y. (2004). Explanation Oriented Retrieval. In Funk & González–Calero (eds.), 157–168. Springer: Berlin Heidelberg.

Faltings, B. (1997). Probabilistic Indexing for Case-Based Prediction. In Leake, D. B. & Plaza, E. (eds.) *Case-Based Reasoning Research and Development: Proceedings ICCBR 1997*, vol. 1266 of *Lecture Notes in Artificial Intelligence*. 611–622. Springer: Berlin Heidelberg.

Ford, K. M., Cañas, A. J. & Coffey, J. (1993). Participatory Explanation. In *Proceedings FLAIRS*. 111–115.

Friedman, M. (1974). Explanation and Scientific Understanding. *Journal of Philosophy* **71**: 5–19.

Funk, P. & González–Calero, P. A. G. (eds.) (2004). Advances in Case-Based Reasoning: Proceedings ECCBR 2004. No. 3155 in LNAI. Springer: Berlin Heidelberg.

Gregor, S. & Benbasat, I. (1999). Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly* **23**(4): 497–530.

Hanna, J. (1982). Probabilistic Explanation and Probabilistic Causality. *Philosophy Socity American*, vol. 2.

Harman, G. (1965). The Inference to the Best Explanation. *The Philosophical Review* **74**(1): 88–95.

Hempel, C. G. (1965). *Aspects of Scientific Explanation*. Free Press: New York.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science* **15**: 135–175.

Herlocker, J. L., Konstan, J. A. & Riedl, J. (2000). Explaining Collaborative Filtering Recommendations. In *Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work* 241–250.

Josephson, J. R. & Josephson, S. G. (eds.) (1994). *Abductive Inference Computation: Philosophy, Technology*. Cambridge University Press: New York.

Kass, A., Leake, D. & Owens, C. (1986). SWALE: A Program that Explains. In Schank, R.C. (ed.) Explanation Patterns: Understanding Mechanically and Creatively, 232–254. Lawrence Erlbaum Associates: Hillsdale, NJ.

Keil, F. C. & Wilson, R. A. (eds.) (2000). Explanation and Cognition Bradford Books: Boston, MA.

Kitcher, P. (1976). Explanation, Conjunction, and Unification. *Journal of Philosophy* **73**: 207–12.

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

Kolodner, J. L. (1997). Educational Implications of Analogy: A View from Case-Based Reasoning. *American Psychologist* **52**(1): 57–66.

Kolodner, J. L. & Leake, D. B. (1996). A Tutorial Introduction to Case-Based Reasoning. In Leake, D. B. (ed.) *Case-Based Reasoning: Experiences, Lessons, & Future Directions*. MIT Press & AAAI Press, Cambridge, MA.

Koton, P. (1988). Reasoning about Evidence in Causal Explanations. In *Proceedings of AAAI-88* vol. 1. 256–261. AAAI Press/MIT Press: Cambridge, MA.

Lalljee, M., Watson, M. & White, P. (1983). Attribution Theory: Social and Functional Extensions. In *The Organization of Explanations*. Blackwell: Oxford.

Leake, D. B. (1992). *Evaluating Explanations: A Content Theory*. Lawrence Erlbaum Associates: New York.

Leake, D. B. (1995a). Abduction, Experience, and Goals: A Model of Everyday Abductive Explanation. *Journal of Experimental and Theoretical Artificial Intelligence* **7**: 407–428.

Leake, D. B. (1995b). Goal-Based Explanation Evaluation. In *Goal-Driven Learning*. 251–285 MIT Press: Cambridge.

Leake, D. B., Birnbaum, L. Hammond, K. Marlow, C. & Yang, H. (2001a). An Integrated Interface for Proactive, Experience-Based Design Support. In *Proceedings of the 6th International Conference on Intelligent User Interfaces*. Santa Fe. 101–108.

Leake, D. B., Maguitman, A. & Cañas, A. (2001b). Assessing Conceptual Similarity to Support Concept Mapping. In *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*. AAAI Press, Menlo Park. 172–186.

Lenat, D. & Feigenbaum, E. (1987). On the Thresholds of Knowledge. In *Proceedings IJCAI 1987*. 1173–1182.

Majchrzak, A. & Gasser, L. (1991). On using Artificial Intelligence to Integrate the Design of Organizational and Process Change in US Manufacturing. *AI and Society* **5**: 321–338.

Mao, J.-Y & Benbasat, I. (2000). The Use of Explanations in Knowledge-Based Systems: Cognitive Perspectives and a Process-Tracing Analysis. *Journal of Managment Information Systems* **17**(2): 153–179.

Massie, S., Craw, S. & Wiratunga, N. (2004). Visualisation of Case-Based Reasoning for Explanation. In Gervás, P. & Gupta, K. M. (eds.) *Proceedings of the ECCBR 2004 Workshops*. Madrid. 135–144.

McArdle, G. P. & Wilson, D. C. (2003). Visualising Case-Base Usage. In McGinty, L. (ed.) *Workshop Proceedings ICCBR 2003*. 105–114, Trondheim.

McDermott, D. (1987). A Critique of Pure Reason. *Journal of Computational Intelligence*. **3**(3): 151–160.

McSherry, D. (1998). Strategic Induction of Decision Trees. In Milne, R. & Bramer, M. A. (eds). 15–26. *Proceedings of ES98*.

McSherry, D. (2001). Interactive Case-Based Reasoning in Sequential Diagnosis. *Applied Intelligence* **14**: 65–76.

McSherry, D. (2003). Explanation in Case-Based Reasoning: an Evidential Approach. In Lees, B. (ed.) *Proceedings of the 8th UK Workshop on Case-Based Reasoning*. 47–55. Cambridge.

McSherry, D. (2004). Explaining the Pros and Cons of Conclusions in CBR. Funk, P. & González–Calero, P. (eds.) *Proceeding of ECCBR 2004*, 317–330. Springer: Berlin Heidelberg.

McSherry, D. (2005). Explanation in Recommender Systems. *Artificial Intelligence Review* (This issue).

Murdock, J., Aha, D. & Breslow, L. (2003). Assessing Elaborated Hypotheses: An Interpretive Case-Based Reasoning Approach. In Ashley & Bridge (eds.) 332–346. Springer: Berlin Heidelberg.

Nugent, C. & Cunningham, P. (2005). A Case-Based Explanation System for Black-Box Systems. *Artificial Intelligence Review* (This issue).

Ong, L., Shepard, B. Tong, L. Seow-Choen, F. Ho, Y. Tong, L. Ho, Y. & Tan, K. (1997). The Colorectal Cancer Recurrence Support (CARES) System. *Artificial Intelligence in Medicine* **11**(3): 175–188.

Plaza, E., Armengol, E. & Ontañón, S. (2005). The Explanatory Power of Symbolic Similarity in Case-Based Reasoning. *Artificial Intelligence Review* (This issue).

Reilly, J., McCarthy, K. McGinty, L. & Smyth, B. (2005). Explaining Compound Critiques. *Artificial Intelligence Review*. (This issue).

Richter, M. M. (1995). The Knowledge Contained in Similarity Measures. Invited Talk at the First International Conference on Case-Based Reasoning, ICCBR'95, Sesimbra, Portugal.

Roth-Berghofer, T. R. (2004). Explanations and Case-Based Reasoning: Foundational Issues, In Funk & González–Calero (eds.), 389–403. Springer: Berlin Heidelberg.

Roth-Berghofer, T. R. & Cassens, J. (2005). Mapping Goals and Kinds of Explanations to the Knowledge Containers of Case-Based Reasoning Systems. In Muñoz-Avila, H. & Ricci, F. (eds.) *Proceedings of ICCBR-05*, 451–464. Springer.

Roth-Berghofer, T. R., Cassens, J. & Sørmo, F. (2005). Goals and Kinds of Explanations in Case-Based Reasoning. In Althoff et al. (eds.) *Proceedings of WM 2005*, 264–268. DFKI: Kaiserslautern.

Salmon, W. (1971). Statistical explanation. In Colodny, R. G. (ed.) *The Nature and Function of Scientific Theories*. 173–231. Pittburgh University Press: Pittburgh.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press: Princeton.

Schank, R. & Leake, D. (1989). Creativity and Learning in a Case-Based Explainer. *Artificial Intelligence* **40**(1–3): 353–385.

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

Schank, R. C. (1986). *Explanation Patterns – Understanding Mechanically and Creatively*. Lawrence Erlbaum: New York.

Smyth, B. & Keane, M. T. (1998). Adaptation-Guided Retrieval: Questioning the Similarity Assumption in Reasoning. *Artificial Intelligence* **102**(2): 249–293.

Sørmo, F. (2005). Case-Based Student Modeling using Concept Maps. In Muñoz-Avila, H. & Ricci, F. (eds.) *Proceedings of ICCBR-05*, 492–506. Springer.

Sørmo, F. & Aamodt, A. (2002). Knowledge Communication and CBR. In González-Calero, P. (ed.) *Proceedings of the ECCBR-02 Workshop on Case-Based Reasoning for Education and Training*. 47–59, Aberdeen.

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

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

Swartout, W. R. & Moore, J. D. (1993). Explanation in Second Generation Expert Systems, In David, J. Krivine, J. & Simmons, R. (eds.) *Second Generation Expert Systems*. 543–585, Springer Verlag: Berlin.

Thagard, P. (1988). *Computational Philosophy of Science*. MIT Press/Bradford Books: Boston.

Thagard, P. (1989). *Explanatory Coherence*, Behavioral and Brain Sciences. **12**(3): 435–467.

Thagard, P. & Holyok, K. (1989). Why Indexing is the Wrong Way to Think About Analog Retrieval. In *Proceedings of the DARPA Workshop on Case-Based Reasoning*. 36–40. Morgan Kaufman: San Mateo.

van Fraassen, B. (ed.) (1980). *The Scientific Image*. Clarendon Press: Oxford.

Watson, I. (1999). Case-Based Reasoning is a Methodology, not a Technology. Knowledge-Based Systems. **12**(5–6): 303–308.

Wick, M. R. & Thompson, W. B. (1992). Reconstructive Expert System Explanation *Artificial Intelligence* **54**(1–2): 33–70.

Zenobi, G. & Cunningham, P. (2002). An Approach to Aggregating Ensembles of Lazy Learners That Supports Explanation. In Craw, S. & Preece, A. (eds.) *Advances in Case-Based Reasoning: Proceedings ECCBR 2002*. 436–447, Springer: Berlin Heidelberg.