Explanation Goals in Case-Based Reasoning

Frode Sørmo and Jörg Cassens

Norwegian University of Science and Technology (NTNU),
7491 Trondheim, Norway,
{frode.sormo|jorg.cassens}@idi.ntnu.no

Abstract. In this paper, we present a short overview of different theories of explanation. We argue that the goals of the user should be taken into account when deciding what is a good explanation for a given CBR system. Some general types relevant to many Case-Based Reasoning (CBR) systems are identified and we use these goals to identify some limitations in using the case as an explanation in CBR systems.

1 Introduction

Throughout our 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 more or less satisfactory explanations at different times. Explanation is one of those concepts that everyone has an intuitive understanding of, but which are very hard to explicitly define. In this paper, we will begin by looking at some attempts at formulating theories of explanation (section 2). We find that many of them recognize that the context of the explanation situation and the goals of the user influence what is and what is not a good explanation. In section 3 we will build on this and suggest some general explanation situations for CBR systems, before we in section 4 look at how the case as an explanation fulfill these explanation goals.

2 Theories of Explanation

Most people 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 [1]). 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", [2]). This is further complicated because 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 [3]. This makes it very hard to form a complete theory of explanation. In this paper we will focus on attempts to identify invariants in different kinds of explanations studied by the philosophical

community. Theories from the cognitive psychology community tend to focus on particular kinds of explanations (e.g. attribution theory or excuse theory [4]). A complete survey is beyond the scope of this paper, but we recommend the far wider survey of both empirical results and theories of explanations in use by knowledge-based systems given by Gregor and Bebasat [5].

2.1 Naïve Explanation

In Knowledge-Based Systems, the approach to explanation started out in a pragmatic manner. When a user asks why a conlusion has been reached, the obvious approach to explanation is to present the reasoning trace of the system. We will call this naïve explanation. This is for example done in rule-based systems like the early MYCIN. Its explanation system can give the user information about *how* a conclusion was reached and *why* a question was asked. A similar approach is used in Case-Based Reasoning, where the best matched case is displayed to the user as a form of justification for the conclusion reached by the system. This approach assumes that the reasoning method of the system is comprehensible to the user. For "black-box systems" such as Neural Nets, this assumption often does not hold, but methods such as Rule-Based Systems and (to an even greater degree) Case-Based Reasoning lend themselves to this approach.

2.2 Constructive Empiricism

There has been much debate about what constitutes an explanation within the Philosophy of Science community since what constitutes a good scientific explanation of a phenomena is important as a norm to identify what is and what is not good science. 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* [6]. Here, van Fraassen 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 S_0 relative to $S_1 \ldots S_n$. He suggests that this can be calculated by applying Bayes' Rule to each candidate answer. This may still leave us with quite a few candidate explanations, but as long as each satisfies the previous criteria 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 criteria an explanation must fulfill (it must favor the observed state) as well as a framework for understanding explanations (as answers to contrastive why-questions).

2.3 Natural Language Philosophy

The field of natural language studies focuses on explanations as a process of communication between people. Here, the goal of an explanation is to impart some piece of knowledge from the sender to a recipient. Achinstein [7] characterizes a request for explanation as a request for understanding of something. He believes this request can take many forms, not just the why-questions of 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 of 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 always full because of the high population density in an area is useful for a passenger, but for the train scheduling department a more useful explanation is that too few trains are scheduled to this part of the city.

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 to facilitate 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 or she 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 one hand, this emphasizes the value of knowing the recipient quite 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, we do not need to explain the inner workings of a Neural Network comparing two pictures. Presenting the pictures to the user so she can validate the similarity for herself can serve as explanation enough.

2.4 Goal-Based Explanation

David Leake [3] directs attention to the different goals the receiver of an explanation may have. Leake bases his theory of explanation on the work in cognitive psychology where explanations have two roles – either as a support of a claim or an argument against it. The work of Lalljee and Abelsen [8] suggests that explanations can be either 'constructive' or 'contrastive'.

Schank [9] is in the same tradition and further specifies that an explanation is required first and foremost in anomalous situations where a person is faced with a situation that does not fit her internalized model of the world. Although

this sounds like quite a specific explanation situation, Leake illustrates that the goals and needs of actors vary widely. As an example, he focuses on a situation where a person gets to know some unexpected news:

"Company X was beleaguered by high taxes, foreign competition, and outdated equipment, despite low labor costs [... The] managers announced their decision not to have layoffs. The next week, it was rumored that they would lay off 20% of their work force" [3, p. 258]

Further on, he gives examples of nine different actors which all have a different interest in the news. For all of them the news is surprising, which means that their model of the situation is flawed, and they all are interested in more information about the news in order to repair their models. These different interests range from pure disbelief in the news (the additional information sought would be an explanation where these rumours came from) to local politicians who want to know whether they could prevent this type of situations from occurring (information about the role the tax level played) to the managers themselves (who might want to avoid negative publicity by finding external reasons).

Although all these actors are referring to the same event, what makes up an explanation of the unexpected situation is very different for them. Leake's goal-driven model takes its starting point in these different interests. When the world is different from what is expected, existing goals and plans may have to be reconsidered. The kind of information that is needed for this purpose makes up a good explanation for the actor.

This 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. Achinstein deals with general communication issues whereas Leake focuses on the evaluation of given explanations for the actor: the user already has detected that his model of the world is faulty, he already has stated new goals, and the next step is to provide new information (explanations) and assess its usefulness. In this sense, Leake's theory can be seen as an operationalization of certain aspects of a more general theory of communication.

3 Explanation Goals in CBR

When discussing the use of explanations in CBR in the light of the theories presented in this paper, it is clear that the context of explaining is very important. This conclusion is also supported by research in Knowledge-Based Systems, which suggests that many of the attempts at providing explanations in earlier systems failed because they were incomprehensible to the user or failed to address the users' goals in demanding an explanation [10].

Often it is very hard for a system to obtain an accurate view of the users goals and knowledge. When asked to explain our reasoning, even we humans sometimes have to make assumptions about the asker. However, as real-life CBR systems are mostly made to perform a limited task for a limited audience, it should

be quite feasible to find reasonable assumptions about the users' goals and the explanation context.

In this section we will suggest some possible goals and contexts that may be reasonable assumptions in different kinds of case-based decision support and intelligent tutoring systems. Our aim is not to provide an exhaustive list – the rationale for introducing them is to discuss how some current explanation types hold up in light of these goals.

Explain Why the Answer is a Good Answer (Justification): This is perhaps the most obvious goal – for many it is the goal assumed when talking about explanations. This goal allows for a simplification of the explanation compared to the actual process the system goes through to find a solution. It will even allow a posteriori explanations formed after the solution is found, i.e. explanations that have nothing to do with how the reasoner came up with the answer. For example, a Neural Net can come up with the solution and a rule based reasoner can be applied to attempt to form justifications for it.

Explain How the System Reached the Answer (Transparency): This goal is subtly different from the previous goal in that it seeks to impart an understanding of how the system found the answer. This allows the user to control the system's quality by examining the way it reasons and allow them to look for explanations for why the system has reached a surprising anomalous result. This is one of the reasoning types (the how-explanations) given in Rule-Based Systems like MYCIN [11].

Explain Why a Question Asked is Relevant (Relevance): In conversational systems, the user may wish to know why a question asked by the system is relevant to the task at hand. An explanation of this type would have to justify the strategy pursued by the system and why a question is relevant in this strategy. This would normally require that the system is able to comprehensibly display its reasoning strategy. In the Mao and Benbasat [12] study, it was found that strategic explanations counted for about a fifth of the explanation requests in their system and this proportion was equally popular with novice and expert users.

Teach the User About the Domain (Learning): In Intelligent Tutoring Systems, it is often the goal not only to find a good solution to a problem, but to explain the solution 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 tutor the user in solving problems similar to those solved by the system. Systems that fulfill the Strategic and Transparancy goals may have some capabilitiy in this area, but a true tutoring system must also take into account that the users have different levels of understanding of the domain. This requires explanations that are at once simple enough for users to understand, and extensive enough to impart skill and knowledge in the domain.

4 Limitations of the Case as an Explanation

The Case-Based Reasoning methodology in itself is quite transparent since it is fairly easy for people to understand that the basic concept is to search for very similar, concrete cases and base the decision-making on them. The value of displaying the retrieved case as an explanation in support of the suggested solution is a truism repeated in CBR introductions everywhere. In addition to the intuitive feeling that this is true, there has been research showing that displaying cases along with the solution significantly improved a users confidence in the solution compared to only showing the solution, or displaying a rule that was used in finding the solution [13].

We can also find some support for the case-as-explanation method fulfilling the Justification goal by looking at it from the viewpoint of the theory of Achinstein. It seems 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.

In van Fraassen's framework, displaying the retrieved case to the user would be a communication of knowledge required by 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 users ability to understand the case and to confirm the similarity assessment. In our own work, we are beginning to see that this may not always be the situation and we suggest that there are limits to the usefulness of cases as explanations.

4.1 Maintaining Transparancy in Complex Systems

In simple CBR system, displaying the retrieved case as a form of explanation provides complete transparancy into the reasoning process. When more advanced methods like feature weighting and complex similarity measures are introduced, however, it will be necessary to provide additional information in order to fulfill the transparancy goal. 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 adaption techniques where the retrieved case may not be the most similar but one which facilitates the adaption process (e.g. as in [14]).

In general, it can be argued that the use of other AI technologies in the CBR cycle (as suggested e.g. by Watson in [15]) 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 have to be explained, too. For example, consider a system where the solution of the retrieved case is altered by a rule-based system to fit the new problem. Then the adaptation steps of the rule-based system have to be explained alongside the presented case.

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. In the CREEK system [16], the user may ask for explanations on the attribute level, and the generation of this explanation depends on the similarity measure. A simple example is that when the similarity of attributes on a interval scale are 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 the known cases. The approach of different layers of explanation satisfies the Transparency goal, but 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 exceedingly hard problems, the value of the case as an explanation may go down.

4.2 Providing Justification to Novice Users

There is an implicit assumption in presenting the case to the user that she is able to do a similarity comparison herself. In general, for the retrieved case to serve as a justification explanation to the user, the similarity between the retrieved case and current problem must be obvious. 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 was deemed too complex to be useful by actual users [10].

For the novice users, a multi-level reasoning trace as suggested in the previous section will likely be too complex to understand. In a study performed with a financial advisor decision support system, it was found that while expert users preferred how-type transparancy-oriented explanations, novice users would more often ask for a simpler why-type justification explanation [12]. For novice users, it may be necessary to provide justifications that are based on simpler strategies than what the system uses internally.

This approach is taken by the ProCon system [17], where the system will identify 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 this attribute affects 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 it also helps the user to identify what attributes are important to the conclusion. The problem with this approach is that there may be domains where the simplification does not work well while the more advanced method does. Case-Based Reasoning is for instance usually quite good at capturing interaction effects between attributes, but the explanation system in ProCon would not be able to identify that e.g., two attributes in combination are a strong supporter of a conclusion while either of them in isolation is not.

A knowledge-intensive approach is used to similar effect in the CREEK system. In CREEK, the model-based reasoner can use a premade 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 of course 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 aquisition effort in building the causal model, but this model can then be tailored to the typical user's level of expertise.

It is also possible to sacrifice some accuracy or efficiency in order to choose a strategy that is easier to explain or seems more intutive to the user. This is the approach taken by the CBR Strategist system. This system is a mixed-initiative conversational diagnosis system where the user may enter a dialogue where she 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 asked of the user, the system would be able to give a good explanation for why this question was important to answer [18]. The extension of Strategist into a CBR system [19] does not form a decision tree in advance, but the question selection method is the same. As an example, the system would prefer questions that could confirm or eliminate possible outcome classes in the domain. This would allow it to form simple explanations of the relevance of questions that 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" [19, figure 7].

4.3 Connecting Cases to General Knowledge in Tutoring

Many cognitive theories (e.g., [9]) of learning assume that people start learning in a new domain by looking at concrete cases, or episodes. This seems to be a natural fit for Case-Based Reasoning – indeed one of the roots of the approach is in these theories. However, there is a separation between cognitive theory and CBR in that a basic principle in CBR is the value of lazy learning. The just-in-time approach to induction in CBR contrasts somewhat with cognitive theories in that most of them believe there is generalization going on as new situations are experienced. Simply presenting a high number of examples can be useful in learning about a domain, but it will not help the learner in generalizing these lessons.

As an example from our own work, we are currently doing experiments with a case-based tutoring system that assists first year students in solving programming exercises. Our tutoring system can assist a student by matching a half-finished program with another student's attempts at solving the same problem and displaying part of his solution. This may be of help, but only if the student is able to see why the suggested lines of program code work. If the system is required to provide an explanation for this, it must have a deeper understanding

of programming so it can for instance explain that when you want to repeat something a number of times, you can use a for-loop.

The above suggest that it may be necessary to have a knowledge-intensive system to fulfill the Learning goal, but generalization may also be done lazily by a number of machine-learning algorithms. The CBR Strategist [19] and Pro-Con [17] systems are examples of this as they do induction when presenting an explanation to the user, but they do so lazily. 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.

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. One example of this is Brüninghaus and Ashley's [20] IBP system in which 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 users (lawyers) are able to make sense of them. While systems like IBP are highly specialized, this specialization allows them to focus very well on the explanation goals of the users, and the system's combination of models and cases allows them to communicate using both concrete examples and generalized structure.

5 Conclusions

One of the things the different theories of explanation have in common is that context and goals influence to a great deal what is and what is not a good explanation. This is also an experience we make in our own work on intelligent tutoring systems and when trying to understand how systems function in a work context. In this work we have also seen that there are situations where the explanation goals of the users are not fulfilled by simply displaying the best matched case. While the value of case-as-explanation should not be dismissed, neither should it be overestimated.

When designing a CBR system, we suggest an approach where the explanation goals of the potential users should be considered in the design of the explanation mechanism. It is hard to model an individual user's goals and intentions completely, but the designer of the system can often make assumptions on the goals and capabilities of prototypical users of the system. We also believe that explicitly formulating such explanation goals facilitate the discussion of possible conflicts between goals and makes clear how different approaches tend to favor different types of goals.

References

1. Salmon, W.: Scientific Explanation and the Causal Structure of the World. Princeton University Press, Princeton (1984)
2. Brewer, W.F., Chinn, C.A., Samarapungavan, A.: Explanations in Scientists and Children. Minds and Machines **8** (1998) 119–136
3. Leake, D.B.: Goal-Based Explanation Evaluation. In: Goal-Driven Learning. MIT Press, Cambridge (1995) 251–285
4. Mehlman, R., Snyder, C.: Excuse Theory: A Test of the Self-protective Role of Attributions. Journal of Personality and Social Psychology **49** (1983) 994–1001
5. Gregor, S., Benbasat, I.: Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Quarterly **23** (1999) 497–530
6. van Fraassen, B., ed.: The Scientific Image. Clarendon Press, Oxford (1980)
7. Achinstein, P.: The Nature of Explanation. Oxford University Press, Oxford (1983)
8. Lalljee, M., Watson, M., White, P.: Attribution Theory: Social and Functional Extensions. In: The Organization of Explanations. Blackwell, Oxford (1983)
9. Schank, R.C.: Dynamic Memory: A Theory of Reminding and Learning in Computers and People. Cambridge University Press, Cambridge (1983)
10. Majchrzak, A., Gasser, L.: On using Artificial Intelligence to Integrate the Design of Organizational and Process Change in US Manufacturing. AI and Society **5** (1991) 321–338
11. Buchanan, B.G., Shortliffe, E.H.: Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison Wesley, Reading (1984)
12. Mao, J.Y., Benbasat, I.: The Use of Explanations in Knowledge-Based System: Cognitive Perspectives and a Process-Tracing Analysis. Journal of Managment Information Systems **17** (2000) 153–179
13. Cunningham, P., Doyle, D., Loughrey, J.: An Evaluation of the Usefulness oof Case-Based Reasoning Explanation. In: Case-Based Reasoning Research and Development: Proceedings ICCBR 2003. Number 2689 in LNAI, Trondheim, Springer (2003) 122–130
14. Smyth, B., Keane, M.T.: Adaptation-Guided Retrieval: Questioning the Similarity Assumption in Reasoning. Artificial Intelligence **102** (1998) 249–293
15. Watson, I.: Case-Based Reasoning is a Methodology, not a Technology. Knowledge-Based Systems (1999) 303–308
16. Aamodt, A.: Explanation-driven Case-Based Reasoning. In: Topics in Case-Based Reasoning: Proceedigs EWCBR 1993. LNAI, Springer (1994) 274–288
17. McSherry, D.: Explanation in Case-Based Reasoning: an Evidential Approach. In Lees, B., ed.: Proceedings of the 8th UK Workshop on Case-Based Reasoning, Cambridge (2003) 47–55
18. McSherry, D.: Strategic Induction of Decision Trees. Proceedings of ES98 (1998) 15–26
19. McSherry, D.: Interactive Case-Based Reasoning in Sequential Diagnosis. Applied Intelligence **14** (2001) 65–76
20. Brüninghaus, S., Ashley, K.D.: Combining Case-Based and Model-Based Reasoning for Predicting the Outcome of Legal Cases. In: Case-Based Reasoning Research and Development: Proceedings ICCBR 2003. Number 2689 in LNAI, Trondheim, Springer (2003) 65–79