The Role of Explanations in Collaborative Problem Solving*

Alison Cawsey, Julia Galliers, Steven Reece and Karen Sparck Jones

Computer Laboratory, University of Cambridge,
New Museums Site, Pembroke St, Cambridge CB2 3QJ, England
Email: ac@cl.cam.ac.uk
Phone: +44-223-334622

May 5, 1992

Abstract

In this paper I discuss the role of explanations in collaborative problem solving. In particular, I will show how explanations may be used to achieve agreement, when each participant has their own beliefs and areas of expertise. True agreement may only be achieved if each participant autonomously decides that the proposed solution is to be preferred, and these preferences are influenced by the explanations or justifications given for the proposals. Explanations are thus an essential aspect of communication and collaborative problem solving, and influence how that problem solving proceeds. A prototype information retrieval expert based on this framework will be described.

Topic: Relation between explanation generation, task execution, and user-system interactions.
Key Words: Negotiation, Belief revision, Justification, Cooperation.

*This research is supported by grant 90/CS42 of the UK Tri-Council Initiative on Cognitive Science and Human-Computer Interaction.
1 Introduction

The majority of existing expert systems (and their associated explanation facilities) are based on the premise that the system is the authoritative expert, able to advise on the user's problem and suggest solutions. Explanations are normally only provided at the end of the problem solving activity, at the user's explicit request, to justify the system's already determined conclusions.

For most realistic problem situations this model of an authoritative expert advising a passive, receptive user is clearly invalid. The 'expert' and user each have their own beliefs and their own individual (and possibly overlapping) areas of expertise. A suggestion from either participant (made at any point in the problem solving process) may be consistent with the other's prior beliefs, or it may conflict with them. In the face of such conflict the participant must reason about what they prefer to believe — they should neither assume that the other participant is an 'absolute authority', to be believed automatically, nor assume that their own prior beliefs are automatically more reliable (Galliers, 1990).

When two participants disagree, either may offer explanations (or strictly, justifications)\(^1\) to try and get the other participant to change their beliefs, so that agreement may be achieved. An extended negotiation process may be involved, with each participant justifying their own position, until participants either agree or agree to disagree. The result of this negotiation cannot be determined in advance. One or both participants may have changed their beliefs, as a result of the negotiation. Explanations, in this view, are an intrinsic part of the process of collaborative problem solving, influencing how the problem solving proceeds as partial solutions are negotiated among participants. The role of an explanation is to persuade another participant that a proposed partial solution is correct, but the explainer remains open to changing its own beliefs, in the light of the response.

This view of collaborative problem solving applies between any number of participants, human or computer. In our example application (described below), the system 'expert' consists of a number of independent, communicating agents, collaborating among themselves and with the user. Explanations

\(^1\)In this paper we will use the term explanation to refer to a participant's justification of their beliefs or reasoning, used to achieve mutual acceptance of a proposal. We do not consider explanations used to achieve mutual understanding (cf. Karzenty and Falzon, this volume), but believe an analogous treatment of this type of explanation would be fruitful.
may be given (and reasoned about) by any participant, including the user(s).

The background of the work is a general theory of communication, proposed by Galliers (Galliers, 1990), where conflict is viewed as a positive force, motivating communication (and thus enabling further problem solving). Each participant reasons about what they prefer to believe through a process of autonomous belief revision. In this theory, utterances are strategically planned, to achieve predicted belief changes in others. Thus a suggestion may be accompanied by an unprompted justification if this is believed necessary for the suggestion to be accepted.

A model of (human) belief revision is crucial to this theory, influencing both the way agents reasons about incoming information, and the way they strategically plan utterances for others. The particular model of belief revision that has been developed (Galliers, in press) provides a partial preference ordering among alternative consistent sets of beliefs. This preference depends on the internal connectivity of the belief sets; on a heuristic description of the type of source (or endorsement) for unjustified (e.g., communicated) beliefs; and a principle of minimal change from previously preferred beliefs.

2 Example Human Information Retrieval Dialogue

We are testing these ideas by developing an 'information retrieval expert', able to collaborate with the user in order to decide on an appropriate literature search. This is a complex and difficult problem, discussed in (Belkin, Seeger & Wersig, 1983) for example, where both user and 'expert' each have complementary areas of expertise. The user knows about their information need and about their subject domain, while the expert knows about retrieval strategies and sources, for example. All these types of knowledge are vital if an appropriate search is to be made.

To illustrate the kind of negotiation (involving explanations) that goes on, consider the following fragment, taken from transcripts of human information retrieval dialogues between an 'expert' and a user (Brooks, 86):

1. **Expert:** Um, the only other possibility is Historical Abstracts but it it
2. **User:** No.
3. **Expert**: it is fairly, they can include some recent material...


5. **Expert**: We'll think about it we'll see we'll put a query by that one. Mm.

6. **User**: OK OK alright.

7. **Expert**: It's the only database which has really, obviously because it deals with history tried to, cope with this time limitation.

In this dialogue the user wants to get material about ‘Greek-Turkish relations after 1974’. The expert initially suggests that ‘Historical Abstracts’ may be a useful literature database to search, because it flags material explicitly by the time period to which it refers. However, the user rejects this, possibly believing that the database would not include recent material. The expert therefore provides an explanation in utterance 3, justifying their choice by showing that an argument against this choice is invalid. In utterance 7 further explanation is given, with a positive justification provided for the choice of database (i.e., it allows searches for material with particular date restrictions). The result of the negotiation is that the user is persuaded that the database may possibly be appropriate, and the expert is (possibly) less confident that the database is acceptable. The construction and selection of justifications by the expert appears to involve reasoning about the user’s hypothesised prior beliefs and justifications.

This fragment forms part of an extended dialogue, where expert and user negotiate different aspects of the problem to be solved. Justifications are not just given at the end of a whole questioning and problem solving process but as an integral part of collaborative problem solving. The may be provided unprompted, if one participant predicts that the other might otherwise have difficulty accepting a suggestion; in response to apparent conflict in beliefs; or prompted by explicit requests for explanations. In human collaborative dialogues it is unprompted or spontaneous explanations that occur most frequently, and requested explanations that occur least frequently (Belkin, 1988; Karzenty and Falzon, this volume; Draper, 1987).
3 A Prototype System

The prototype expert system we are developing consists of a number of communicating agents, but for the purpose of this paper we will focus on the communication and problem solving between the system as a whole and its user. Further details of the system are given in (Cawsey et al., 1992a).

We assume that neither user nor system will unequivocally accept the other’s assertions. In our domain, the user’s conception of their information need will be partial and possibly faulty. If an undergraduate, for example, claims that they want journal articles to complete their student project, the expert may believe that this is unlikely, and try to persuade them that they want an introductory text book. And if the system suggests a particular literature database or search strategy (as we saw above), the user may reject this suggestion, if they have their own reason to believe it is inappropriate. And when a participant accepts the other’s proposals, despite contradictory prior beliefs, they may have to revise a whole network of related beliefs, supporting or supported by the now rejected prior belief.

In order to decide what to believe in the face of such conflicts the system incorporates a belief revision system based on Gallier’s approach, as referred to above (Galliers, in press). This mechanism is used to decide how and whether to believe new information suggested by the user, and also as the basis for planning utterances and explanations for the user.

Utterance planning involves predicting the effect of possible utterances on the user’s beliefs, using the belief revision system and a model of the user’s beliefs. If the system believes that simply informing the user of some belief would be insufficient to get them to change their beliefs, then it will also ‘spontaneously’ construct an explanation or justification for that belief, and present the user with a justified proposal. The construction of the appropriate justification also uses the same predictive mechanism, so the system will eventually generate an utterance (or utterances) which it believes will cause the target belief change in the user. Of course, the system’s assumptions about the user’s beliefs may be incorrect, so this result is not guaranteed.

If the user rejects the system’s proposal then the system must first reason about whether it still believes it itself. If it does then the conflict between

\(^2\)Mann and Thomson have made a similar suggestion for using the ‘evidence’ relation in texts (Mann & Thomson, 1987).
system and user beliefs will motivate further negotiation, involving further justifications, constructed using the same predictive mechanism. Explanations are thus not just supplied on request from the user, but are generated whenever the system expects (or has received) a rejection of its proposals.

Negotiations in the system are motivated by the goal of achieving mutual belief (or mutual acceptance) of some proposal. This may be a proposal of the system, based on its partial and provisional conclusions, or a proposal of the user (possibly in response to a system question). Negotiation may involve any belief which falls within the overlap of the areas of expertise of the system and user (Cawsey et al., 1992b), and is not limited to the system's 'final' problem solution. Through negotiation, participants establish an agreed common context for their developing individual and joint problem solving.

Currently the model of the user's beliefs is only updated using simple techniques (e.g., believing user believes what she says). Richer user modelling techniques will be required to obtain full benefit from the approach outlined (as discussed in papers in Wahlster and Kobsa, 1989, for example). However, even with a very simple (and sparse) user model negotiation is possible. For example, given no prior beliefs about the user, the system may assume in the first instance that its proposals will be accepted by the user, and if rejected may provide an explanation (as illustrated in the example below). A richer user model will allow more efficient communication, with more 'spontaneous' and appropriate explanations, but is not required for effective negotiation.

4 Example System Dialogue

In this section we will go through a simple dialogue fragment generated by the system (based on the human dialogue fragment illustrated above), illustrating the belief revision and explanation construction processes involved. At the beginning of our example the system has been told that the user wants documents relating to material after 1974. The system infers that there is therefore a 'date restriction', and thus that the history database is a possible choice. The system records the justifications for its beliefs, and also the endorsement for unjustified (e.g., communicated) beliefs, for example, '2cs' refers to beliefs that are communicated strongly by another participant. In this simple example it is these endorsements (and the principle of minimal
change) that determine how the system will choose to revise its beliefs. The system’s beliefs and inference rules are represented as follows:

**System Beliefs:**

1. \( \text{doc-content(post-74)} \rightarrow \text{doc-restriction(date)} : [\text{endorsement, 1cs}] \)
2. \( \text{doc-restriction(date)} \rightarrow \text{database(history)} : [\text{endorsement, default}] \)
3. \( \text{doc-content(post-74)} : [\text{endorsement, 2cs}] \)
4. \( \text{doc-restriction(date)} : [\text{from 1,3}] \)
5. \( \text{database(history)} : [\text{from 1,2,3}] \)

**System’s beliefs about user’s beliefs:**

1. \( \text{believes(user, doc-content(post-74))} : [\text{endorsement, 2cs}] \)

Initially the system has the goal of convincing the user that the history database may be appropriate (in order to achieve mutual belief concerning the appropriate database). The system predicts, based on its model of the user’s beliefs, that simply informing the user will be sufficient. The belief is communicated weakly, since the system is not strongly committed to it\(^3\), using the following formal language:

\[ S: \text{tellref(system, user, database(history), weak)}. \]

But the user responds by strongly rejecting this suggestion:

\[ \text{User: tellref(user, system, not database(history), strong)}. \]

The system now updates its model of the user’s beliefs with the information that the user does not believe the history database is appropriate. The system then reasons about whether it should change its own beliefs, but (based on its belief revision mechanism) chooses to hold onto its existing beliefs. (This is partly because the system does not believe the user is knowledgeable about appropriate databases.) The system recognises that there is

---

\(^3\)A participant may be strongly or weakly committed to a belief. In our framework this refers to how easy it would be to disbelieve the given belief. This degree of commitment is expressed in communications.
now a conflict between its beliefs and those of the user, and is motivated to try to resolve that conflict. It does this by constructing a justification or explanation which it believes (based on its new model of the user’s beliefs) will cause the user to accept the original proposal:

S: tellwhy(system, user, database(history) because
doc-content(post-74) → doc-restriction(date) strong &
doc-restriction(date) → database(history) weak)

In other words, the system says that the reason why the history database is believed suitable is that wanting post-74 material is a kind of date restriction, and the history database is suitable if there is a date restriction. The user now decides to accept this suggestion, though only weakly:

User: tellref(user, system, database(history), weak)

This example is obviously very simple, but illustrates the basic belief preference and negotiation mechanisms that we are concerned with. The system has to convince the user that the history database is appropriate, by providing justifications which it believes will cause the change in beliefs. However, in the process of the negotiation the system may change its own beliefs – the user may in turn try and convince the system that they are wrong. In this case the explanation and negotiation process may influence the developing problem solving process itself.

5 Conclusion: Collaborative Problem Solving via Negotiation

In this paper we have outlined a radical view of the role of explanations in collaborative problem solving. Explanations have a dynamic role in the problem solving process itself, and aren’t simply supplied to a passive user on request. They are constructed to achieve belief changes in others, and hence agreement on a (partial) problem solution. Negotiations involving explanations are a vital part of the interaction with the user, and influence the ongoing problem solving, as the system’s own beliefs may change in the light of the interaction. Explanations may be provided unprompted, based on
predicted problems; in response to apparent problems in acceptance (ie belief conflicts); or in response to direct requests for explanation. The model applies to negotiations between a single system and user, or between a collection of system and user 'agents' all collaborating in some problem solving process. Each agent autonomously determines what to believe itself, but may revise its beliefs in the face of new evidence.

It is unclear as yet whether the approach will prove a practical (or appropriate) solution to the development of collaborative expert problem solvers. In order to construct appropriate justifications for the user the system must have an adequate model of their likely prior (and changing) beliefs. In order to reason about the user's suggestions and explanations, the system must have fairly wide general knowledge, so it can integrate the user's suggestions into its own belief and justification structures. And for both, the system needs a fairly complex belief revision system, able to reason about conflicts and preferences among sets of beliefs.

The framework is also based strongly on human-human negotiation and cooperation. It is unclear whether the same framework is necessarily appropriate for human-machine negotiation. However, we believe that models of human dialogue provide, at the least, a rich source of insights which may enhance such human-machine dialogues. Through the development of such models, new paradigms for human-machine negotiation may emerge, enabling practical systems of the future. We are currently exploring the limitations and possibilities of the approach through computational implementation and experimentation, with some of our early results described in (Cawsey et al, 1992b).

References


Karzenty, L. and Falzon, P. (this volume) Spontaneous Explanations in Cooperative Validation Dialogs.
