Explaining Emotions

Paul O'Rorke
ororke@ics.uci.edu

Andrew Ortony
ortony@ils.nwu.edu

Technical Report 92–22

Revised
March 20, 1992

Work on this project was supported in part by grant number IRI-8813048 from the National Science Foundation.

Supported in part by grant number IRI-8812699 from the National Science Foundation and in part by Andersen Consulting through Northwestern University's Institute for the Learning Sciences.

Explaining Emotions

Paul O'Rorke[1]
(ororke@ics.uci.edu)
Phone and Fax: (714) 854-2894
Department of Information and Computer Science
University of California, Irvine, CA 92717

Andrew Ortony[2]
(ortony@ils.nwu.edu)
Phone: (708) 491-3500
Institute for the Learning Sciences
Northwestern University
1890 Maple Avenue
Evanston, IL 60201

March 20, 1992

[1]Work on this project was supported in part by grant number IRI-8813048 from the National Science Foundation.

[2]Supported in part by grant number IRI-8812699 from the National Science Foundation and in part by Andersen Consulting through Northwestern University's Institute for the Learning Sciences.

Abstract

Emotions and cognition are inextricably intertwined. Feelings influence thoughts and actions, which in turn give rise to new emotional reactions. We claim that people infer emotional states in others using common-sense psychological theories of the interactions between emotions, cognition, and action. We present a situation calculus theory of emotion elicitation representating knowledge underlying common-sense causal reasoning involving emotions. We show how the theory can be used to construct explanations of emotional states. The method for constructing explanations is based on the notion of abduction. This method has been implemented in a computer program called AMAL. The results of computational experiments using AMAL to construct explanations of examples based on cases taken from a diary study of emotions indicate that the abductive approach to explanatory reasoning about emotions offers significant advantages. We found that the majority of the diary study examples cannot be explained using deduction alone, but they can be explained by making abductive inferences. These inferences provide useful information relevant to emotional states.

Contents

List of Figures

List of Tables

1 Introduction

Explaining people's actions often requires reasoning about emotions. This is because experiences frequently give rise to emotional states which in turn make some actions more likely than others. For example, if we see someone striking another person, we may explain the aggression as being a result of anger. As well as reasoning about actions induced by emotional states, we can reason about emotional states themselves. In the right context, we might reason that a person was angry because we knew that he or she had been insulted. Explaining emotional states often requires reasoning about the cognitive antecedents of emotions. This paper focuses on explanations of this kind.

Although people appear to generate explanations involving emotions effortlessly, the question of how one might compute such explanations remains a difficult open question, just as the more general question of how to automate commonsense reasoning remains open. We present a computational model of the construction of explanations of emotions. The model is comprised of two main components. The first component is a situation calculus theory of emotion elicitation. The second component is a method for constructing explanations. The representation of emotion eliciting conditions is inspired by a theory of the cognitive structure of emotions proposed by Ortony, Clore, and Collins (1988). In addition to codifying a set of general rules of emotion elicitation, we have also codified a large collection of cases based on diary study data provided by Turner (1985). We have implemented a computer program that constructs explanations of emotions arising in these scenarios. The program constructs explanations based on a first order logical abduction method. We describe these components in some detail in later sections. In the remainder of the introduction, we provide some background information on the theory of emotions and the diary study.

1.1 A theory of emotions

The theory of the cognitive structure of emotions employed in the research we describe views emotions as valenced reactions to events, agents and their actions, and objects. It specifies a total of 22 emotion types summarized in an abbreviated form in Table 1. We provide only a brief sketch here. A full description can be found in Ortony, et al. (1988).

The emotion types are essentially just classes of eliciting conditions, but each emotion type is labelled with a word or phrase – generally an English emotion word corresponding to a relatively neutral example of an emotion fitting the type. The simplest emotions are the "well-being" emotions *joy* and *distress*. These are an individual's positive and negative reactions to desirable or undesirable events.

The "fortunes of others" group covers four emotion types: *happy-for*, *gloating*, *resentment*, and *sorry-for*. Each type in this group is a combination of pleasure or displeasure over an event further specialized as being presumed to be desirable or undesirable for

Group	Specification	Types (name)
Well-Being	appraisal of event	pleased (joy) displeased (distress)
Fortunes-of-Others	presumed value of an event affecting another	pleased about an event desirable for another (happy-for) pleased about an event undesirable for another (gloating) displeased about an event desirable for another (resentment) displeased about an event undesirable for another (sorry-for)
Prospect-based	appraisal of a prospective event	pleased about a prospective desirable event (hope) pleased about a confirmed desirable event (satisfaction) pleased about a disconfirmed undesirable event (relief) displeased about a prospective undesirable event (fear) displeased about a confirmed undesirable event (fears-confirmed) displeased about a disconfirmed desirable event (disappointment)
Attribution	appraisal of an agent's action	approving of one's own action (pride) approving of another's action (admiration) disapproving of one's own action (shame) disapproving of another's action (reproach)
Attraction	appraisal of an object	liking an appealing object (love) disliking an unappealing object (hate)
Well-Being/ Attribution	compound emotions	admiration + joy → gratitude reproach + distress → anger pride + joy → gratification shame + distress → remorse

Table 1: Emotion Types

another person.

The "prospect-based" group includes six emotion types: *hope*, *satisfaction*, *relief*, *fear*, *fears-confirmed*, and *disappointment*. Each type is a reaction to a desirable or undesirable event that is still pending or that has been confirmed or disconfirmed.

The "attribution" group covers four types: *pride*, *admiration*, *shame*, and *reproach*. Each attribution emotion type is a (positive or negative) reaction to either one's own or another's action.

The "attraction" group is a structureless group of reactions to objects. The two emotions in this group are the momentary feelings (as opposed to stable dispositions) of liking or disliking.

The final group is comprised of four compounds of "well being" $\times$ "attribution" emotion types. These compound emotions do not correspond to the co-occurrence of their component emotions. Rather, each compound's eliciting conditions are the union of the component's eliciting conditions. For example, the eliciting conditions for anger combine the eliciting conditions for reproach with those for distress.

In general, eliciting conditions are specified in terms of variables that contribute toward increasing the intensity of emotions. The theory specifies global variables that affect all emotions, and local variables that affect subsets of emotions. The variables have values and weights associated with them, and the theory claims that an emotion is experienced only when certain levels, the emotion thresholds, are exceeded.

For anger, the variables affecting intensity are:

- the degree of judged blameworthiness,
- the degree of deviation from personal or role-based expectations,
- the degree to which the event is undesirable.

The first variable, blameworthiness, is the evaluation of an action against the standards of the judger. The second variable, deviations from expectations, gauges the extent to which the action is unexpected of the agent. The third variable reflects an evaluation of the event (perpetrated by the agent) in terms of its impact upon personal goals.

1.2 A diary study of emotions

We use data taken from a diary study of emotions (Turner, 1985) as a source of examples. Most of the subjects who participated in the study were sophomores at the University of Illinois at Champaign-Urbana. They were asked to describe emotional experiences that occurred within the previous 24 hours. They typed answers to a computerized questionaire containing questions about which emotion they felt, the event giving rise to the emotion, the people involved, the goals affected, and so on. Over 1000 descriptions of emotion episodes were collected, compiled, and recorded on magnetic media.

We chose to use this diary study as a source of examples because, while nearly every emotion type is represented, the situations and events described in the entries tend to cluster into a relatively small number of stereotypical scenarios. This is a natural consequence of the importance of examinations, dating, and so on in the emotional lives of undergraduate students. We were thus able to focus on aspects of the theory and computational model most relevant to emotions, rather than being distracted by problems having to do with representing a wide range of domain specific knowledge.

2 A situation calculus theory of emotion elicitation

In this section, we describe a representation language designed to support the construction of explanations involving emotions. The language is based on work on two major contributions to knowledge representation, the situation calculus (McCarthy, 1968) and conceptual dependency (Schank, 1972).

2.1 Situation calculus

The situation calculus provides us with a language for expressing causal laws relating actions and physical situations. This first order logical language was originally invented by John McCarthy. We employ a version of the situation calculus incorporating improvements by Green (1969), McCarthy and Hayes (1969), Fikes and Nilsson (1971), Kowalski (1979), Elkan (1990), and Reiter (1991).

Situations are represented by terms. Fluents are statements that may or may not be true in a given situation. Partial descriptions of the state of affairs in a given situation S state that fluents such as P hold in S: *holds(P,S)*.

Actions are functions that map situations representing the state of the world before the execution of the action into situations representing the state of the world afterward. The situation resulting from applying action A to state S is designated by the term *do(A,S)*.

Actions are defined by specifying their preconditions and effects. Preconditions are divided into *action* and *fluent* preconditions (Reiter, 1991). Action preconditions are fluents that must hold in order for the action to be possible in a given situation. Fluent preconditions specify what must hold in order for individual effects to follow upon execution of an action.

The fact that an action A is possible in a situation S is represented as *poss(A, S)*. If P_1 to P_n are action preconditions for doing A, this is represented:

$$poss(A, S) \leftarrow holds(P_1, S) \wedge ... \wedge holds(P_n, S).$$

The effects of actions are represented using "add" and "delete" statements (Fikes & Nilsson, 1971) specifying fluents added or deleted upon execution of an action. Both positive and negative effects are encoded in conditional "cause" statements of the form:

$$causes(A, F, S) \leftarrow holds(F_1, S) \wedge ... \wedge holds(F_2, S).$$

where each F_i is a fluent precondition for action A causing fluent F in situation S. Positive and negative effects are inferred through the following law of direct effects:

$$holds(F, do(A, S)) \leftarrow causes(A, F, S) \wedge poss(A, S).$$

Frame axioms state that nothing changes unless it is explicitly stated that it changes as a result of some action. We use the following frame axiom schema:

$$holds(P, do(A, S)) \leftarrow \overline{causes(A, \overline{P}, S)} \wedge holds(P, S) \wedge poss(A, S).$$

This frame axiom states that the fluent P will hold after execution of an action if it held before and the action did not cancel it.

The rewrite rules listed below facilitate explanations involving chains of fluents by linking several equivalent representations:

$$\begin{aligned} holds(F(Args), S) &\Rightarrow F(Args, S) \\ holds(holds(F(Args), S_1), S_2) &\Rightarrow holds(F(Args), S_1) \\ holds(F(Args, S_1), S_2) &\Rightarrow F(Args, S_1) \end{aligned}$$

The first rule establishes the equivalence between a fluent with a situation as an argument and the corresponding situationless fluent holding in the same situation. The second rule "unwraps" embedded *holds* statements. The truth of a *holds* statement depends only on the situation it applies to. The third rule is a consequence of the first two equivalences.

We provide the predicates *agent, precedes,* and *precondition.* They express important constraints and help to control inference. The *agent* predicate is used to identify or extract the agent of a given action. The *precedes* predicate applies to two arguments:

$$precedes(S_0, do(A, S)).$$

This statement is read: "The situation S_0 precedes the situation resulting from the execution of action A in situation S." It is true if $S_0 = S$ or if $precedes(S_0, S)$ is true. The *precondition* predicate applies to two arguments, an action and a fluent:

$$precondition(A, C).$$

This statement is true if C is an action precondition of A.

2.2 Conceptual dependency

Primitive actions provided by conceptual dependency (*ptrans, move, atrans, propel, grasp, ingest, expel, mtrans,* and *attend*) are encoded in our situation calculus representation as functions mapping situations to new situations. The arguments of the functions correspond to roles. Variants of the functional representations are used when the value of some argument is unknown or unimportant.

For example, the function *atrans* is used to represent transfers of ownership. The most explicit form of *atrans* has arguments for the agent responsible for the transaction, the object in question, the new owner, and the previous owner. In many cases, the previous owner is the agent. We use a three argument version of *atrans* in such cases. A two argument version is used when the agent is the recipient and the previous owner is irrelevant.

We show how knowledge about actions is encoded using an example of *ptrans*. It is only possible for an agent P to move an object T from one location F to a destination D if the thing T is *at* the initial location F:

$$poss(ptrans(P, D, F, T), S) \leftarrow holds(at(T, F), S).$$

This is an example of an action precondition. The effects of *ptrans* are as follows. A physical transfer of a thing T to a destination D causes the thing to be *at* the destination:

$$causes(ptrans(P, D, F, T), at(T, D), S).$$

This is a positive effect of the transfer. A negative effect is that the thing is no longer at its original location:

$$causes(ptrans(P, D, F, T), \overline{at(T, F)}, S).$$

Note that the positive and negative effects are both unconditional; there are no fluent preconditions in this formulation of *ptrans*.

Additional actions required by the examples that we have encoded include *abuse, attack, breakup, call, close, die, excel, fight, gossip, hurt, insult, kill, open,* and *score.* Preconditions and effects of actions are encoded using fluents such as *alive, dead, did, have, rested, single,* and *unfaithful.* These actions and their preconditions and effects are represented using general laws similar to those shown above in the example of *ptrans*.

2.3 Emotion eliciting conditions

Eliciting conditions for emotions are encoded in a collection of rules for all emotion types except likes and dislikes. The rules are an initial attempt to represent the emotion eliciting conditions proposed by Ortony, et al., and sketched in Section 1.1. Simplifying assumptions and limitations of this initial representation are discussed in Section 4.3. We present the rules in pairs corresponding to complementary emotions.

People may experience joy over a fluent in a situation if they want it and it holds; but they may experience distress if they want a fluent that holds not to hold:

$$joy(P, F, S) \leftarrow wants(P, F, S) \wedge holds(F, S).$$

$$distress(P, F, S) \leftarrow wants(P, \overline{F}, S) \wedge holds(F, S).$$

A person may experience neither joy nor distress in the event that a fluent holds, if he or she desires neither the fluent nor its negation. Even if we grant the law of the excluded middle for a fluent, it is still possible that a person is indifferent to it.

A person may be happy for another if he or she experiences joy over a fluent presumed to be desirable for the other. We express this in terms of joy over the other's joy:

$$happy\_for(P_1, P_2, F, S) \leftarrow joy(P_1, joy(P_2, F, S_0), S).$$

Note that the desire for the fluent is implicit in the embedded joy. Although the rule does not encode the fact that a person is usually happy for another before or while the other is happy, the rule does reflect the fact that they may be happy in different situations (at different times).

A person may be sorry for another if he or she experiences distress over a fluent presumably undesirable for the other. We express this in terms of distress over the other's distress:

$$sorry\_for(P_1, P_2, F, S) \leftarrow distress(P_1, distress(P_2, F, S_0), S).$$

The undesirability of the fluent is implicit in the embedded distress. The two people may be distressed in different situations and no temporal constraints are placed on these situations in the present formalization.

A person may gloat over a fluent that gives them joy that (they believe) is not wanted by another. We express this in terms of joy over another's distress:

$$gloats(P_1, P_2, F, S) \leftarrow joy(P_1, distress(P_2, F, S_0), S).$$

People may resent another person if they are distressed about an event that pleases the other person. We express this in terms of distress over another's joy:

$$resents(P_1, P_2, F, S) \leftarrow distress(P_1, joy(P_2, F, S_0), S).$$

Again, the desirability of the event for the other is implicit in the embedded distress and joy and we currently do not require any particular temporal order for the relevant situations.

The *hopes* rule captures the idea that people may experience hope if they want a fluent and anticipate it:

$$hopes(P, F, S) \leftarrow wants(P, F, S) \wedge anticipates(P, F, S).$$

People may experience fear if they want an anticipated fluent not to obtain:

$$fears(P, F, S) \leftarrow wants(P, \overline{F}, S) \wedge anticipates(P, F, S).$$

These rules allow for hopes and fears even if, unknown to the person, the hoped-for or feared fluent in fact already holds.

Although many examples of hopes and fears involve expectations, we use the predicate *anticipates* in order to suggest the notion of "entertaining the prospect of" a state of affairs. The purpose of this is to avoid suggesting that hoped-for and feared events necessarily have a high subjective probability. We also want to avoid suggesting that they always occur in the future.

Satisfaction occurs when a hoped-for fluent holds:

$$satisfied(P, F, S) \leftarrow precedes(S_0, S) \wedge hopes(P, F, S_0) \wedge holds(F, S).$$

Fears are confirmed when feared fluents hold:

$$fears\_confirmed(P, F, S) \leftarrow precedes(S_0, S) \wedge fears(P, F, S_0) \wedge holds(F, S).$$

We require the fear to precede its confirmation and we expect the hope to occur before it is satisfied.

Relief may be experienced when the negation of a feared fluent holds:

$$relieved(P, \overline{F}, S) \leftarrow precedes(S_0, S) \wedge fears(P, F, S_0) \wedge holds(\overline{F}, S).$$

$$relieved(P, \overline{F}, S) \leftarrow fears(P, F, S_0) \wedge holds(\overline{F}, S).$$

We give two rules for relief because fear usually occurs before the fluent holds, but sometimes relief occurs in the absence of prior fear (as when a person discovers that a missed flight crashed).

Disappointment occurs when the negation of a hoped-for fluent holds:

$$disappointed(P, \overline{F}, S) \leftarrow precedes(S_0, S) \wedge hopes(P, F, S_0) \wedge holds(\overline{F}, S).$$

$$disappointed(P, \overline{F}, S) \leftarrow hopes(P, F, S_0) \wedge holds(\overline{F}, S).$$

The fluent is usually hoped-for in a situation that occurs in advance of the present situation but disappointment (e.g., at a missed opportunity) may occur in the absence of prior hope.

Pride and shame can occur for individuals or groups. An agent experiences pride over praiseworthy actions executed either by the agent or by another member of a "cognitive unit" (Heider, 1958) containing the agent. Agents may experience shame if they do a blameworthy act or if another member of their cognitive unit does a blameworthy act:

$$\begin{aligned}
proud(P, A, S) &\leftarrow agent(A, P) \wedge holds(did(A), S) \wedge praiseworthy(A). \\
proud(P_1, P_2, A, S) &\leftarrow agent(A, P_2) \wedge holds(did(A), S) \wedge \\
&\quad\; praiseworthy(A) \wedge cognitive_unit(P_1, P_2). \\
shame(P, A, S) &\leftarrow agent(A, P) \wedge holds(did(A), S) \wedge blameworthy(A). \\
shame(P_1, P_2, A, S) &\leftarrow agent(A, P_2) \wedge holds(did(A), S) \wedge \\
&\quad\; blameworthy(A) \wedge cognitive_unit(P_1, P_2).
\end{aligned}$$

The predicates *praiseworthy* and *blameworthy* are intended to reflect personal standards rather than normative or social standards, except insofar as the judging person subscribes to such standards.

A person may admire another person if the other person does something praiseworthy, but a person may feel reproach toward another if the other does something blameworthy:

$$admire(P_1, P_2, A, S) \leftarrow agent(A, P_2) \wedge holds(did(A), S) \wedge praiseworthy(A).$$

$$reproach(P_1, P_2, A, S) \leftarrow agent(A, P_2) \wedge holds(did(A), S) \wedge blameworthy(A).$$

Compound emotion types are comprised of the eliciting conditions of components taken from the "well being" and "attribution" groups. We do not include the component emotions in the eliciting conditions in order to avoid suggesting that the component emotions are necessarily felt as part of feeling the compound. Instead we collect the eliciting conditions of the components and simplify them, eliminating redundancies.

Gratitude is a compound of the eliciting conditions of joy and admiration. A person may be grateful toward another person if the other person does a praiseworthy action that causes a desirable fluent to hold.

$$\begin{aligned}
grateful(P_1, P_2, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad\; precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge \\
&\quad\; praiseworthy(A) \wedge wants(P_1, F, S_1) \wedge holds(F, S_1).
\end{aligned}$$

The *angry_at* emotion type focuses on anger at other agents. It is a compound comprised of the eliciting conditions of reproach and distress. A person may be angry at another if an undesirable fluent is caused by a blameworthy action taken by the other person:

$$\begin{aligned}
angry_at(P_1, P_2, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad\; precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge \\
&\quad\; blameworthy(A) \wedge wants(P_1, \overline{F}, S_1) \wedge holds(F, S_1).
\end{aligned}$$

We distinguish this from angry-about, which because it focuses on the undesirability of the situation is better thought of as a special case of *distress* — frustration (typically at goal blockage).

Gratification is a compound emotion comprised of the eliciting conditions of pride and joy. A person may be gratified if he or she does a praiseworthy action that results in a desirable fluent.

$$\begin{aligned}
gratified(P, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge \\
&\quad wants(P, F, S_1) \wedge holds(F, S_1) \wedge praiseworthy(A). \\
gratified(P_1, P_2, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge cognitive_unit(P_1, P_2) \\
&\quad wants(P_1, F, S_1) \wedge holds(F, S_1) \wedge praiseworthy(A).
\end{aligned}$$

Since there is a cognitive unit variant of pride, there is also a variant of gratification. This variant of *gratified* is closely related to *grateful.*

Remorse is a compound emotion comprised of the eliciting conditions of shame and distress. People may be remorseful if they do a blameworthy action that results in an undesirable fluent:

$$\begin{aligned}
remorseful(P, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge \\
&\quad wants(P, \overline{F}, S_1) \wedge holds(F, S_1) \wedge blameworthy(A). \\
remorseful(P_1, P_2, A, S_1) &\leftarrow agent(A, P_2) \wedge holds(did(A), S_1) \wedge \\
&\quad precedes(S_0, S_1) \wedge causes(A, F, S_0) \wedge cognitive_unit(P_1, P_2) \\
&\quad wants(P_1, \overline{F}, S_1) \wedge holds(F, S_1) \wedge blameworthy(A).
\end{aligned}$$

The second rule above provides the eliciting conditions of a cognitive-unit variant of *remorseful.* These conditions are derived from the corresponding variant of *shame.* This variant is closely related to *angry-at.*

The eliciting conditions given in this section represent a first attempt at formalizing necessary, but not necessarily sufficient, conditions for the elicitation of the corresponding emotions. (This is why we say "a person *may* experience emotion X under conditions Y.") In some cases, additional conditions may be required to more fully capture common-sense knowledge about emotion elicitation. It may be that a person's disposition towards another person should play a role in explaining the elicitation of "fortunes of others" emotions. For example, perhaps the eliciting conditions of *happy_for* should include a requirement that a person may be happy for another if that person does not dislike the other.

3 Explaining emotions

In this section, we show how our situation calculus representation of emotion elicitation can be used to explain emotional states. First we sketch our general approach to constructing explanations, then we give several examples to be discussed later.

3.1 Constructing explanations

Our approach to constructing explanations is based on work in Artificial Intelligence and Cognitive Science on computational methods employing Charles Sanders Peirce's notion of abduction (Peirce, 1931–1958). Peirce used the term abduction as a name for a particular form of explanatory hypothesis generation. His description was basically:

> *The surprising fact C is observed;*
> *But if A were true, C would be a matter of course,*
> *hence there is reason to suspect that A is true.*

In other words, if there is a causal or logical reason A for C, and C is observed, then one might conjecture that A is true in an effort to explain C.

Since Peirce's original formulation, many variants of this form of reasoning have also come to be referred to as abduction. We focus on a logical view of abduction advocated by Poole (e.g., Poole, Goebel, & Aleliunas, 1987). In this approach, observations O are explained given some background knowledge expressed as a logical theory T by finding some hypotheses H such that

$$H \wedge T \vdash O.$$

In other words, if the hypotheses are assumed, the observation follows by way of general laws and other facts given in the background knowledge.

We use an abduction engine — a computer program that automatically constructs explanations. The explanations of observations are based on general background knowledge and knowledge of particular cases provided to the program in a machine readable form. An input/output characterization of the program is given in Figure 1.

The particular abduction machinery that we use is based on an early approach to mechanizing abduction (Pople, 1973). The abduction engine is part of an explanation-based learning system called AMAL (Abductive MAcro Learner). It is implemented in PROLOG.[1] A simplified sketch of the procedure followed by the abduction engine is shown in Table 2.

[1]PROLOG was chosen because the basic operation involved in constructing explanations, abductive inference, is related to backward chaining. PROLOG provides basic operations such as unification that are an essential part of backward chaining.

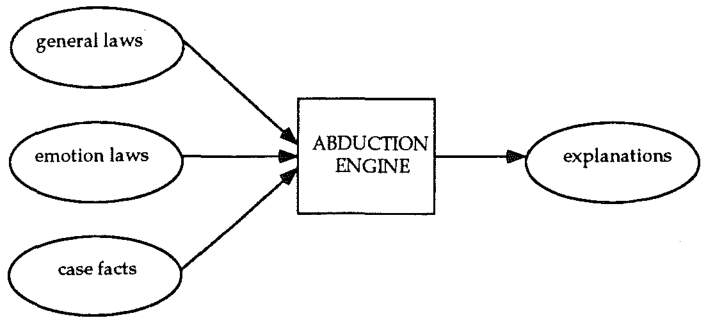

Figure 1: Inputs and Outputs of the Abduction Engine

The abduction engine attempts to construct explanations of given observations using general laws and specific facts. In the implementation, explanations are proofs represented as AND trees. Observations to be explained correspond to conclusions of the proofs (roots of the trees). General laws are encoded as rules and these are used to generate the proofs through a process based upon backward chaining.

The mechanization of abduction embodied in AMAL is comprised of three steps (Table 2). The first step corresponds to backward chaining as it is implemented in PROLOG interpreters. The observation is treated as a query. There may be a number of open questions kept in the query list Q. Initially, there is only one query — to explain the given observation. The search process attempts to ground the explanation (tree) in known facts. Where possible, leaves of the partial proofs are identified with facts in the data-base. If this is not possible, rules are used to extend the partial explanation, replacing existing queries with new queries.

The second step begins when backward chaining fails. In this step, the remaining unexplained facts and hypotheses are examined and some of them are assumed to be "the same." Unlike the previous step, this inference is not deductively sound, but errors are recoverable through backtracking. In terms of Table 2, at the beginning of this stage Q is the empty list, U is a non-nil list of unexplained statements, and the explanation is incomplete. The algorithm continues by first selecting an arbitrary unexplained statement u from U. If u can be identified (unified) with any other statement in U, then the pair is replaced in U with their identification. The identification step ends when no more queries in U are pairwise identifiable.

One advantage of "identification" is that it simplifies explanations. It is based on the *synthesis* operator advocated by Pople (1973) and justified in terms of Occam's razor. Another advantage of identification assumptions is that they often introduce new information. The identification of two previously unrelated statements in different parts of an explanation often causes sharing of information between the separate branches of the explanation. In the implementation, statements are identified by unifying two well-formed-formulae.

Table 2: A Procedural Sketch of the Abduction Engine

Given: a query-list Q containing observations to be explained;
Find: explanations for the queries, using abductive inferences.

1. BACKWARD CHAINING:

   While the query list Q is not empty, do:

   (a) Select the first query q and remove it from Q.
   (b) If q is directly explainable then unify q with the fact that explains it.
   (c) If q is indirectly explainable using a rule, then use the rule to generate new queries and add them to Q, else add q to the list U of unexplained queries.

2. IDENTIFICATION:

   While there are more unifiable pairs of queries in U, unify and replace pairs.

3. ASSUMPTION:

   While there are unexplainable, unidentifiable queries in U,

   (a) Select the first query u and remove it from U.
   (b) If u is assumable, then assume u, else fail and backtrack.

This can cause variables in both formulae to take on new bindings. The new bindings then propagate to other statements that share the same variables.

The third abduction step tests whether remaining questions can be assumed. We provide a domain-specific "assumability" test. This test is applied to each of the queries u in list U. If u is not assumable, then the current attempt to find an explanation is aborted and backtracking is invoked in order to continue the search for acceptable explanations.

This process of abduction (deductive explanation followed by two non-deductive abduction steps) continues until either (1) the list U becomes empty, or (2) no explanation is acceptable. In the case of (1), all queries have been either deductively explained or abductively assumed. In the case of (2), all potential explanation trees contain queries that cannot be deductively explained and at least one of those queries in every such explanation is not assumable.

In general, many explanations are possible and it is important to constrain the search to avoid large numbers of implausible hypotheses and explanations. In early experiments, we found that the abduction engine conjectured large numbers of implausible causal relationships. This problem was solved by disallowing assumptions of the following forms in step three of the algorithm:

$$preconditions(A, F).$$

$$causes(A, F, S).$$

In other words, the abduction engine was not allowed to assume that an arbitrary fluent might be a precondition for an action, nor was it allowed to assume unprovable cause-effect relationships between actions and fluents.

3.2 Examples of explanations of emotions

In this section, we use examples to illustrate the construction of explanations of emotional states. We describe the process of codifying examples. We show how explanations are produced by the abduction engine sketched earlier. We give examples of explanations produced by the computer program and describe how to interpret them. In a later section, we will use these examples to illustrate features of the situation calculus of emotion elicitation and the abductive approach to explaining emotions.

Example 3.2.1 The first example is based on the following "case," a simplified version of a scenario taken from Turner's diary study.

> *Mary went home to see her family.*
> *She ate a home-cooked meal.*
> *She wanted to stay home.*
> *Mary was happy.*

We hand code the case into the inputs shown in Table 3. The first input states that Mary wants to be at home in the situation that follows after she went home and ingested a home-cooked meal. Abbreviations at the bottom of the table are used for the relevant situations. Note that this codification of the example is crude in the sense that we have not attempted to capture much of the information associated with Mary visiting her family. The underscores and the use of a constant for "home_cooked_meal" also signify simplification aimed at avoiding having to deal with issues related to quality of food and different methods of food preparation. Such subtleties are lost. We strive only to capture basic facts of the case. The case specifies that Mary is happy in a situation following certain actions. It specifies that she wants to be at home in this situation but it does not specify the fluent that she is happy about. In the query about why Mary is experiencing joy, the situation is specified but the fluent is left blank (using a "don't care" variable designated in PROLOG as an underscore "_").

Table 3 also shows an explanation produced by the abduction engine. The explanation is in the form of a tree. The first line is the root of the tree; indented lines are branches. The first level of indentation shows the propositions immediately supporting the root. The second level of indentation shows their supporters, and so on. The deepest levels of indentation correspond to leaves of branches of the explanation tree.

This explanation of Mary's happiness is interpreted as follows. In the first line, we see that the fluent she is happy about has been identified as the fact that she is at home in

Table 3: An Explanation for Joy

```
Case Facts
      wants(mary, at(mary, home), s2).
Query
      why(joy(mary, _, s2))
Explanation
      joy(mary, at(mary, home), s2)
            wants(mary, at(mary, home), s2)
            holds(at(mary, home), s2)
                  not causes(ingest(mary, home_cooked_meal), not at(mary, home), s1)
                  holds(at(mary, home), s1)
                        causes(ptrans(mary, home), at(mary, home), s0)
                        poss(ptrans(mary, home), s0)
                  poss(ingest(mary, home_cooked_meal), s1)
                        [holds(have(mary, home_cooked_meal), s1)]
Abbreviations
      s1=do(ptrans(mary, home), s0)
      s2=do(ingest(mary, home_cooked_meal), s1)
```

the given situation. Mary's location prior to going home was not specified in the given case fact and neither is it determined during the construction of the explanation. Her joy is supported by the second line, which was part of the input. A case fact stating that Mary wanted to be at home was given. The fact that she is at home is explained by the remainder of the tree. Mary is at home because she went there and the fact that she ingested a meal did nothing to cancel this result. The explanation that Mary is at home after she ate the meal rests on an assumption that it was possible to eat it because she had it after she went home. Abductive inferences (assumptions) are distinguished from leaves of explanation trees that are known to be true by enclosing them in boxes as in Table 3.

The explanation was constructed by the abduction engine using the situation calculus theory of emotion elicitation described earlier. The system is not allowed to explain the initial query (why was Mary happy?) directly, even though it is a known fact. Instead, it finds reasons by backward chaining on rules of situation calculus and emotion elicitation rules. In this case, the rule specifying the eliciting conditions for joy applied. Backward chaining on this rule generated two new queries: does Mary want something — something that holds in the given situation? A given case fact stated that Mary wants to be at home. This fact was used to "ground out" one of the queries. The query about how the desired state of affairs came to pass (how Mary came to be at home) was answered by backward chaining on a law of situation calculus. The frame axiom stating that effects

Table 4: An Example for Happy_For

```
Case Facts
    NONE
Query
    why(happy_for(mary, roommate(mary), at(roommate(mary), europe), s1))
Explanation
    happy_for(mary, roommate(mary), at(roommate(mary), europe), s1)
        joy(mary, joy(roommate(mary), at(roommate(mary), europe), s2), s1)
            wants(mary, joy(roommate(mary), at(roommate(mary), europe), s2), s1)
                [d_likes(mary, roommate(mary))]
            joy(roommate(mary), at(roommate(mary), europe), s2)
                [wants(roommate(mary), at(roommate(mary), europe), s2)]
                holds(at(roommate(mary), europe), s2)
                    causes(ptrans(roommate(mary), europe), at(roommate(mary), europe), _29530)
                    poss(ptrans(roommate(mary), europe), _29530)
Abbreviations
    s1=do(ptrans(roommate(mary), europe), s0)
    s2=do(ptrans(roommate(mary), europe), _29530)
```

Table 5: An Example for Gloating

```
Case Facts
    NONE
Query
    why(gloats(john, guy, did(expel(guy,vomit)), s2))
Explanation
    gloats(john, guy, did(expel(guy, vomit)), s2)
        joy(john, distress(guy, did(expel(guy, vomit)), do(expel(guy, vomit), _21375)), s2)
            [wants(john, distress(guy, did(expel(guy, vomit)), do(expel(guy, vomit), _21375)), s2)]
            distress(guy, did(expel(guy, vomit)), do(expel(guy, vomit), _21375))
                wants(guy, did(not expel(guy, vomit)), do(expel(guy, vomit), _21375))
                holds(did(expel(guy, vomit)), do(expel(guy, vomit), _21375))
                    causes(expel(guy, vomit), did(expel(guy, vomit), _21375))
                    [poss(expel(guy, vomit), _21375)]
Abbreviations
    s1=do(expel(guy,vomit), s0)
    s2=do(hear(john, expel(guy,vomit)), s1)
```

Table 6: An Example for Hope and Satisfaction

```
Case Facts
    wants(mary,did(score(mary, high,art_history_exam)),_)
    holds(did(score(mary, high, art_history_exam)), s1)
Query
    why(satisfied(mary,did(score(mary,high,art_history_exam)),_))
Explanation
    satisfied(mary, did(score(mary, high, art_history_exam)), s1)
        precedes(_16540, s1)
        hopes(mary, did(score(mary, high, art_history_exam)), _16540)
            wants(mary, did(score(mary, high, art_history_exam)), _16540)
            [anticipates(mary, did(score(mary, high, art_history_exam)), _16540)]
        holds(did(score(mary, high, art_history_exam)), s1)
            causes(study_for(mary, art_history_exam),
                        did(score(mary, high, art_history_exam)), _16540)
            poss(study_for(mary, art_history_exam), _16540)
Abbreviations
    s1=do(study_for(mary, art_history_exam), _281)
```

Example 3.2.4 Many diary study cases involve emotions in response to examinations. The following example, based on one such case, illustrates the emotion types *satisfied* and *hopes*.

Mary wanted to get a high mark on her art history exam.
Mary did get a high mark on her art history exam.
Mary was satisfied.

This example is encoded in Table 6. The given facts of the case state that Mary wants to score highly and that this desired fluent holds in a situation resulting from her studying. The explanation of Mary's satisfaction states that she hoped for a high score and she did get the hoped-for score. That she hoped for a high score is explained in terms of her known desire for a high score together with an assumption that she anticipated this achievement.

Example 3.2.5 The following case provides examples of *relief* and *fear*.

Mary wanted to go to sleep.
Karen returned.
T.C. finally left her place.
Mary was relieved.

Table 7: Explanations of Relief and Fear

```
Case Facts
    wants(mary, sleep(mary), _)
Query
    why(relieved(mary, not at(tc, home(mary)), s2))
Explanation
    relieved(mary, not at(tc, home(mary)), s2)
        precedes(s1, s2)
        fears(mary, at(tc, home(mary)), s1)
            [wants(mary, not at(tc, home(mary)), s1)]
            [anticipates(mary, at(tc, home(mary)), s1)]
        holds(not at(tc, home(mary)), s2)
            causes(ptrans(tc, _29887, home(mary), tc), not at(tc, home(mary)), s1)
            poss(ptrans(tc, _29887, home(mary), tc), s1)
                holds(at(tc, home(mary)), s1)
                    not causes(ptrans(karen, home(mary)), not at(tc, home(mary)), s0)
                    [holds(at(tc, home(mary)), s0)]
                    poss(ptrans(karen, home(mary)), s0)
Abbreviations
    s1=do(ptrans(karen, home(mary)), s0)
    s2=do(ptrans(tc, _29887, home(mary), tc), s1)
```

The case is encoded as shown in Table 7. The case facts say Mary wants sleep. The query asks why Mary is relieved that T.C. is not at her home in the situation that results after T.C.'s departure. T.C.'s departure occurred in the situation resulting from Karen's return.

The automatically constructed explanation assumes that Mary fears T.C.'s presence in her home because she wants T.C. not to be in her home but she anticipates that he will be there. A deeper explanation connecting this desire and anticipation to Mary's desire for restful sleep should be possible. For example, the presence of T.C. might interfere with Mary's sleep.

The explanation in Table 7 states that Mary is relieved because T.C. is no longer at her home. The explanation of his absence does not include the possibility that he may have been driven away by Karen's return. But it is interesting for another reason. It illustrates the use of causal laws to infer negative fluents relevant to emotional reactions. In this case, since T.C. moved from Mary's home to another location, he is no longer *at* Mary's home.

Example 3.2.6 The following example, based on a diary study case involving a dating scenario, illustrates the *angry_at* emotion type. The example involves the breakup of a

Table 8: An Explanation of Anger

```
Case Facts
    wants(kim,single(kim),_)
    wants(john,not(single(john)),_)
Query
    why(angry_at(john,kim,breakup(kim,kim,john),s1))
Explanation
    angry_at(john, kim, breakup(kim, kim, john), s1)
        agent(breakup(kim, kim, john), kim)
        holds(did(breakup(kim, kim, john)), s1)
            causes(breakup(kim, kim, john), did(breakup(kim, kim, john)), s)
            poss(breakup(kim, kim, john), s)
                [holds(couple(kim, john), s)]
        precedes(s, s1)
        causes(breakup(kim, kim, john), single(john), s)
        [blameworthy(breakup(kim, kim, john))]
        wants(john, not single(john), s1)
        holds(single(john), s1)
            causes(breakup(kim, kim, john), single(john), s)
            poss(breakup(kim, kim, john), s)
                [holds(couple(kim, john), s)]
Abbreviations
    s1 = do(breakup(kim, kim, john), s)
```

couple of college students named Kim and John.

> *Kim wanted to breakup with John.*
> *John didn't want to breakup with Kim.*
> *They broke up.*
> *John is angry at Kim.*

The example is encoded as shown in Table 8. The query encodes the question "Why is John angry with Kim over the breakup?" The explanation states that John is angry with Kim because Kim initiated the breakup and it caused John to be single. The fact that John doesn't want to be single was given but the remainder of the explanation involves two assumptions. The first is that Kim and John were a couple prior to the breakup. The second assumption is that John views Kim's breaking up with him to be *blameworthy*.

Example 3.2.7 Our last example involves the emotion type for gratification. It is based on a diary study case revolving around a particularly important examination, the graduate record examination (GRE).

John got a high score on his GRE
John wants to go to grad school.
John is elated.

John's "elation" is encoded as an example of the emotion type *gratified* in Table 9. John's desire to "go to" graduate school is encoded as a desire to enroll. Besides the facts of the case, general knowledge about admissions is provided. In particular, the fact that matriculation is an action that causes enrollment is given, but a precondition of matriculation is that a student must be admitted. A high score on the GRE is a precondition of admission to graduate school.

The explanation of John's gratification is that scoring highly on examinations is praiseworthy and furthermore, John desires to be admitted to graduate school and he will be admitted as a result of his high score. The explanation states that John wants to be admitted since this is a precondition of an action (matriculation) that will result in achievement of a known goal (enrollment).

4 Discussion

The previous sections described a representation for knowledge about emotion elicitation and a computer program that constructs explanations based on cases taken from a diary study about emotion episodes. In this section, we discuss some of the strengths and weaknesses of our explanatory reasoner and our representation of knowledge about emotion elicitation.

4.1 Advantages of abductive reasoning about emotions

We claim that abduction is superior to deduction as a basis for explanatory reasoning about emotions because it subsumes deduction which, on its own, will fail when a proof cannot be derived from a given set of facts. The primary advantage of abduction is that it allows for the possibility that assumptions may be required to complete explanations, so that an explanation of a given observation can be proposed even when it does not follow logically from given facts.

It is unreasonable to expect all the information needed to construct explanations involving emotions to be provided in advance. It is particularly unlikely that all the relevant information about someone's mental states will be provided. Indeed, we would like to acquire this sort of information by inference, and abductive inference allows us to do so

Table 9: An Explanation for Gratification

```
Case Facts
    wants(john,did(enroll(john,grad_school)),_)
Query
    why(gratified(john,score(john,high,gre), s1))
Explanation
    gratified(john, score(john, high, gre), s1)
        agent(score(john, high, gre), john)
        holds(did(score(john, high, gre)), s1)
            causes(score(john, high, gre), did(score(john, high, gre)), s0)
            poss(score(john, high, gre), s0)
        precedes(s0, s1)
        causes(score(john, high, gre), did(score(john, high, gre)), s0)
        wants(john, did(score(john, high, gre)), s1)
            precondition(admit(grad_school, john), did(score(john, high, gre)))
            causes(admit(grad_school, john), did(admit(grad_school, john)), s1)
            wants(john, did(admit(grad_school, john)), s1)
                precondition(matriculate(john, grad_school), did(admit(grad_school, john)))
                causes(matriculate(john, grad_school), did(enroll(john, grad_school)), s1)
                wants(john, did(enroll(john, grad_school)), s1)
        holds(did(score(john, high, gre)), s1)
            causes(score(john, high, gre), did(score(john, high, gre)), s0)
            poss(score(john, high, gre), s0)
        praiseworthy(score(john, high, gre))
Abbreviations
    s1=do(score(john, high, gre), s0)
```

during the construction of explanations. Abduction may be viewed as a search for plausible hypotheses that help explain observations.

The majority of the cases in the diary study data require assumptions. In this regard, the examples that we have discussed are representative. The kinds of assumptions needed included missing preconditions, goals, prospects, and judgements.

Examples of preconditions inferred by abductive inference included the following. In the example of *joy* (Example 3.2.1) the explanation was completed with an assumption that Mary had possession of a home-cooked meal. This explained how it was possible for her to ingest it. In the example of *gloating* (Example 3.2.3) it was assumed that it was possible for John's neighbor to vomit. In the example of *relief* and *fear* (Example 3.2.5) the assumption that T.C. was at Mary's home in the initial situation helped explain why he was there after Karen came home. In Example 3.2.6 for *anger* it was assumed that Kim and John were a couple before she broke up with John and he became angry.

The reader may wonder how it is possible to make these assumptions given that the abduction engine is not allowed to assume that an arbitrary fluent might be a precondition for an action (Section 3.1). The reason is that assumptions of the form *preconditions(A,F)* are prohibited, while assumptions of the form *holds(F,S)* are permitted. In other words, we allow a conjecture that a condition is true in a given situation but we disallow a conjecture that the condition is a precondition of an action.

Some inferences of missing preconditions can be made deductively, rather than abductively. We saw several examples of necessary preconditions in the emotion cases. It is necessary for two people to form a couple in order to break up, it is necessary to have something prior to eating it, and so on. If we provide the reasoning system with facts stating that such preconditions are necessary, it should be possible to replace these abductive inferences with deductive inferences.

Recent research on the relationship between abduction and other forms of reasoning (Konolige, in press) shows that there is a close relationship between abduction and an alternative consistency and minimization-based approach. It is often possible to translate abduction into the alternative approach by rewriting a logical theory adding "closure statements," for example, statements to the effect that the known preconditions or causes are the only ones. While it may be worthwhile to add such closure information to deal with special cases such as necessary preconditions, in general the abductive approach is superior because it does not require complete knowledge of causation, and causal closures need not be computed and asserted.

Examples of abductive assumptions about goals occur frequently. Example 3.2.5 required an assumption that Mary wanted T.C. to go somewhere else in order to explain Mary's fear that T.C. would be at her home. Assumptions about other's goals occur in explaining "fortunes of others" emotions. In Example 3.2.2 for *happy_for*, an assumption was made about Mary's roommate's desire to be in Europe. The example of *gloating* (Example 3.2.3) required an assumption that John wanted his neighbor to be distressed.

Abductive assumptions about other mental states include assumptions about whether agents anticipate events. In Example 3.2.4 for *satisfied* and *hopes*, it was necessary to assume that Mary anticipated a high score on her art history exam, in order to explain her satisfaction upon achieving a high score. In the example of *relief* (Example 3.2.5), it was necessary to assume that Mary anticipated T.C.'s continued (unwelcome) presence in her home.

Assumptions about judgements of blameworthiness and praiseworthiness are also important in explaining attribution emotions and compound emotions. For example, in the *anger* case, the assumption that Kim's breaking up with John was blameworthy was made in order to explain why John was angry at Kim.

None of the explanations constructed in these examples could have been constructed by the abduction engine without its abductive inference capability, given the background knowledge and codifications of the cases provided with the observations to be explained. Given the same information, a purely deductive PROLOG-style interpreter would have failed to find an explanation. Admittedly, the knowledge base could in principle be extended so that some assumptions could be eliminated and replaced by deductive inferences. We have already pointed out that abduction can be translated into a deductive approach based on closure plus minimization by translating the theory and adding closure statements for causes. Similarly, some assumptions could be eliminated by adding implications for (necessary) preconditions. Knowledge of ethics and standards of behavior could be provided, reducing the number of assumptions in explanations requiring judgements of blameworthiness and praiseworthiness. But it is not likely that all relevant preconditions, causes, desires, prospects, and judgements can be provided in advance. The abductive approach is well suited to the domain of emotion-relevant reasoning since it does not require complete knowledge and closure axioms.

The ability to generate hypotheses and make assumptions enables our abductive approach to

- assume that implicit preconditions held in an effort to explain how an action led to an effect,
- assume that agents must have had certain desires in order to explain their emotional reactions or actions,
- assume that agents must have anticipated certain events in order to explain emotional reactions, and
- assume that actions are considered praiseworthy or blameworthy on the basis of emotional reactions to those actions.

Abduction thus has advantages important for explanatory reasoning about emotions.

4.2 Advantages of the situation calculus for emotions

In this section, we summarize the advantages of our situation calculus theory of emotion elicitation. Our goal in this section is to spell out some requirements imposed by the theory of the cognitive structure of emotions and to show how they are met by our representation of knowledge about emotion elicitation.

Since emotions are positively and negatively valenced reactions, negations and opposites are important in the emotion eliciting conditions. Indeed, each emotion is paired with an emotion with a complementary valence. In addition, pairs of opposing extrinsic predicates play an important role in reasoning about emotions. For example, the opposing predicates *praiseworthy* and *blameworthy* play a crucial role in the emotion theory. Our representation language allows for both positive and negative fluents and actions. The negation of a fluent F is (also) a fluent. The negation of an action A is the action of not executing A. The term $do(\overline{A})$ is considered to be an alias for $\overline{do(A)}$.

The situation calculus account of actions captures important causal information clearly needed in constructing explanations involving emotions. Situation calculus provides for a causal theory of actions that includes both positive and negative effects and preconditions. Both the positive and negative effects are mediated by a single law of direct effects:

$$holds(F, do(A, S)) \leftarrow causes(A, F, S) \wedge poss(A, S).$$

This law states that a fluent holds in the situation resulting from the execution of an action if the action was possible to begin with and if the action causes the fluent.

Emotion types (represented as fluents) are not caused directly by actions, in the sense that they do not appear as direct effects encoded in *causes* statements. Instead, they are caused indirectly; the theory specifies eliciting conditions that contain actions and other fluents. We saw several instances of actions causing effects that resulted in emotional reactions. The conceptual dependency primitive for physical transfers serves to illustrate. The first law below specifies the precondition that some thing, T, must start out *at* an origin, *From*, before an agent, P, can change its position to a destination, *To*, in a situation, S:

$$poss(ptrans(P, To, From, T), S) \leftarrow holds(at(T, From), S).$$

$$causes(ptrans(P, To, From, T), at(T, To), S).$$

$$causes(ptrans(P, To, From, T), \overline{at(T, From)}, S).$$

In the first emotion example, Mary was happy to be at home. The fact that she was at home was caused by the fact that she went there. This is an instance of the second statement above (about the positive effects of *ptrans*). We saw an example of a negative consequence of an action engendering an emotional reaction in the case of Mary's relief when T.C. vacated her home. Given that T.C. moved to another location, the cancellation law above was used to infer (and explain) the fact that T.C. was no longer at Mary's home.

Frame axioms capture the notion that fluents persist unless explicitly altered by actions. The emotion eliciting conditions use this to advantage: they do not require fluents to be caused by the action most recently executed. Frame axioms are employed to propagate fluents caused by one action through successive actions into later states (provided they are not cancelled by intervening actions). In the initial example, the reason Mary is at home is that she went there earlier and the fact that she ingested a meal did nothing to cancel this result.

The advantage of the frame axiom we have adapted is that there is no need to have a separate frame axiom for every relation. Instead one only needs a single frame axiom. Kowalski (1979) points out that earlier versions of frame axioms can be had by forming macros from the very general frame axiom and specific statements about what is deleted.[3]

A number of examples that we have studied suggest that people do not make a strong distinction between actions and fluents in the sense that they often want an action to be done without focusing on any explicit effect caused by the action. Actions that are done for their own sake because they are intrinsically enjoyable rather than to achieve other goals are good examples (e.g., chatting on the phone, skiing, watching one's favorite sports team). In response to this observation, we introduced a function *did* that maps actions to fluents and we added a causal law relating corresponding actions and fluents:

$$causes(A, did(A), S).$$

The intuition behind this fluent for actions is that, if nothing else, doing an action at least causes it to be done. (If action A is executed in situation S, then it causes the fluent $did(A)$ to be true in the resulting situation.) This law allows us to refer to actions through fluents and not just as mappings between situations. This is useful for actions that are done for their own sake but more generally, the *did* fluent is useful whenever we do not wish to focus on a specific effect of an action but rather on the action itself.

This feature is important for reasoning about emotions because many examples of emotional reactions defined in the emotion elicitation rules in terms of fluents are naturally expressed as responses to actions or other events. In the example of *gloat*ing, John's unfortunate neighbor vomited. The ejection of contents of the neighbor's stomach through his mouth resulted in a relocation of said contents. But John's neighbor's distress was in reaction to his vomiting rather than its effect (the new location of the contents of his stomach).

One advantage of the *holds* relation introduced by Kowalski is that it avoids the need for an extra state parameter for all relations. The disadvantage of the use of *holds* (as opposed to including extra arguments for situations) is that it requires an extra predicate in the sense that it requires us to embed fluents in *holds* statements. Besides being aesthetically

[3]Interestingly, this can be done by explanation-based learning (DeJong & Mooney, 1986; Mitchell, Keller, & Kedar-Cabelli, 1986) .

undesirable in some situations, the use of *holds* can increase the branching factor of some explanatory searches, since facts and clauses tend to be indexed and fetched by the top level predicate. Our representation language allows the use of both representations. Sometimes it is more convenient to use situations as arguments of fluents. In other cases, it is preferable to associate fluents with situations using the *holds* predicate. It is important to link these alternative representations.

The emotion types occurring in the heads of the rules specifying emotion eliciting conditions are expressed as fluents with situational arguments. The bodies of the rules contain *holds* statements whose arguments are fluents eliciting emotional reactions. It is important to ensure that the alternate representations are viewed as equivalent so that inferences are not lost. We enforce the equivalence by mapping complex representations to relatively simple canonical forms using rewrite rules. This facilitates emotional chains. For example, when explaining Mary's joy, a new query is generated in an effort to explain her roommate's joyful reaction to being in Europe. The query is initially in the form:

$$holds(joy(roommate(mary), at(roommate(mary), europe), s2), s1).$$

but this is immediately simplified, using rewrite rules mapping fluents into a canonical form to strip off a superfluous *holds* predicate and an unnecessary situation $s1$:

$$joy(roommate(mary), at(roommate(mary), europe), s2).$$

This simplified version of the query invokes the emotion eliciting rule for *joy* producing the explanation shown in Table 4

Chains of emotional reactions occur frequently. Our representation provides for such chains because emotions are represented as fluents that take fluents as arguments and because fluents appear in the eliciting conditions of emotions. Like other "fortunes of others" emotions, *happy_for* is an emotional chain reaction. Fortunes of others emotions are reactions to events, but instead of focusing exclusively on the event, they also focus on another's emotional reaction to that event. In the example of *happy_for*, the fact that Mary's roomate was going to Europe was not so important to Mary as her roommate's happiness. Mary's roommate's joy engenders Mary's joy.

Goals play a large role in emotion elicitation, so it is important to be capable of reasoning about goals in constructing explanations involving emotions. The following general laws facilitate reasoning about goals:

$$wants(P, did(A), S) \leftarrow causes(A, F, S) \wedge diff(F, did(A)) \wedge wants(P, F, S).$$

$$wants(P, F, S) \leftarrow precondition(A, F) \wedge causes(A, G, S) \wedge wants(P, G, S).$$

The first rule states that a person may want an action to be done if some effect caused by the action is desired. Note the use of the predicate *diff*, which ensures that the effect of the action A is different from $did(A)$. The second statement allows for the fact that an action may be directed at satisfying goals that contribute to the eventual achievement of longer term goals. In particular, a person may want a fluent to hold if it is a precondition of an action that causes another desired fluent. A good example of reasoning about such chains of desires occurs in the GRE case. In that case, John's joy is explained in terms of his desire to be admitted to graduate school. John wants to be admitted since this is a precondition of matriculation and matriculation results in achievement of enrollment (a known goal).

Situation calculus provides support for temporal reasoning. A situation

$$do(a_n, do(a_{n-1}, ...do(a_1, s_0))$$

defines a temporal sequence of situations $s_0, s_1, ..., s_n$ where for $i = 1$ to n, $s_i = do(a_i, s_{i-1})$ and s_{i-1} *precedes* s_i. Temporal precedence is used in eliciting conditions for the prospect-based emotions *satisfied, fears_confirmed, relieved,* and *disappointed.* These emotions are reactions to the confirmation or disconfirmation of a hoped-for or feared fluent. The precedence constraints apply when the relevant fluents are hoped-for or feared in a situation prior to confirmation or disconfirmation. The compound emotions *grateful, angry_at, gratified,* and *remorseful* also employ temporal constraints. Each of these emotions is a reaction to an action and a fluent caused by the action. Two situations are relevant in these emotion types, the situation when the action causes the fluent, and the ensuing situation associated with the emotional reaction to the fluent. The eliciting conditions use the predicate *precedes* to ensure that the temporal precedence constraint between these situations is satisfied. We saw examples in the cases involving anger (Example 3.2.6) and gratification (Example 3.2.7). In the case of *satisfaction*, Mary hoped that she would get a high score prior to taking her examination. Her desire was satisfied in the situation afterward. Another example occurred in the case of *relief* (Example 3.2.5). In the situation prior to his leaving, Mary feared that T.C. would be at her home. She experienced relief after he left.

As discussed in Section 2.3, in some cases of relief and disappointment, the attendant fears and hopes violate the constraint requiring them to occur prior to the relief and disappointment. Additional eliciting conditions allow for this, but since it is the exception rather than the rule, priority is given to the interpretation that includes the temporal constraint.

Our situation calculus does not force a temporal ordering on all events. In the eliciting conditions of *happy_for*, there is no time constraint between the situations when the two agents are happy. In the example of *happy_for* (Examples 3.2.2), Mary's roommate's emotional reaction and Mary's reaction are allowed to occur in different situations. Using separate situations is useful, for example, if some intervening action results in someone being informed of another's earlier good fortune. Avoiding temporal constraints on the situations is useful because in some cases the usual temporal order is reversed. For example, one might be happy for another in anticipation of the other's happiness (e.g., upon learning that the other won a lottery even before the lucky winner knew it).

Cognitive units play an important role in some emotions and are provided for in the situation calculus theory of emotions. A special predicate is used for cognitive units in the eliciting conditions of the *attribution* emotions *pride* and *shame*. This predicate is also used in the given "background knowledge" to encode groups that may form cognitive units. This is an attempt to capture the idea that people can form (relatively stable) cognitive units with their family members and close friends.

Studies of a number of examples suggest that many goals are shared by members of the same cognitive unit. People want good things to happen not just to themselves but also to the others in their cognitive unit. They want to avoid bad things and they do not want bad things to happen to people in their cognitive unit. For example, everyone wants to excel, and they want people in their cognitive unit to excel, too. People generally do not want to experience harm, and they do not want other members of their group to be harmed either. This sort of general law is represented using conditional *wants*:

$$wants(P, \overline{harmed(Q)}) \leftarrow cognitive\_unit(P, Q).$$

4.3 Limitations of the situation calculus of emotion elicitation

In this section, we discuss the main limitations of the situation calculus of emotion elicitation. These include limitations inherent in the situation calculus itself and limitations in the theory of emotion elicitation.

A major limitation of the present study is that we did not attempt to represent or reason about intensities of emotions. It is important to extend the representation presented here to include intensities. Many emotion types are represented in natural language by a number of emotion tokens. Many tokens indicate specific relative intensities of a particular emotion type, for example, "annoyance," "exasperation," and "rage" denote increasingly intense subtypes of anger. Ortony, et al., (1988) suggest that emotions only occur when their intensities are driven above thresholds. The approach taken in the present paper is to use predicates that are true or false in place of these continuous, real valued variables. This approach may be viewed as a crude first approximation. We speculate that methods developed in AI research on qualitative reasoning about physical systems (e.g., Forbus 1984; Kuipers 1986; Weld & de Kleer 1990) could be applied to the problem of representing and reasoning about intensities of emotions.

Another limitation of the emotion eliciting conditions is that they are phrased in terms of facts rather than beliefs. This is because we wanted to avoid having to reason about beliefs and knowledge. But such reasoning is clearly relevant to emotions. Consider the eliciting condition for satisfaction. It states that a person is satisfied if the person hoped for a fluent earlier and now it holds. It seems clear that the eliciting condition is too simple.

It should be phrased in terms of the person's epistemic state. For example, the rule for satisfaction might be rewritten: a person may be satisfied if that person *believes* that some hoped for fluent holds. Logics of belief and knowledge should help address such issues, but they are beset with their own complexities (e.g., referential opacity) that might introduce more problems than solutions.

The most important difficulty for the situation calculus underlying our representation is the famous qualification problem (McCarthy, 1977). The problem is that it is difficult, if not impossible, to specify all the preconditions relevant to the successful execution of an action. We do allow assumptions about the possibility of actions which means that we can explain how an action occurred without knowing all the preconditions that might have made it possible. At present, we do not attempt explanations of inaction, so we do not have to deal with the difficult problem of inferring preconditions that failed thus preventing an action from occurring. We do not claim to have solved the qualification problem but we believe our representation and reasoning methods are no more limited by it than are other approaches.

Some examples in the diary study are beyond the scope of our current methods because they require reasoning about actions not taken and the resulting negative effects. In one case, a woman's roommate fails to pay their phone bill. This triggers anger and fear. She is afraid that she will get a bad credit rating and that her phone will be disconnected. In another example, a student expresses anger because his mother failed to send his records to a dentist and he can't get his teeth cleaned without them. Several cases involve students worrying about poor grades caused by not studying for examinations. The situation calculus says little about negative actions. Perhaps extensions are warranted along lines such as the following causal law for non-actions:

$$causes(\overline{A}, \overline{F}, S) \leftarrow causes(A, F, S).$$

The idea is to express the intuition that not doing an action can cause its effects not to hold. This particular law may be logically inadequate but it still seems important to find a way to express the idea that the failure to do an action can often be said to be the reason its effects do not hold. The following is a consequence of the proposed causal law for non-action and existing laws for indirect wants:

$$wants(P, did(\overline{A}), S) \leftarrow causes(A, F, S) \land diff(F, did(\overline{A})) \land wants(P, \overline{F}, S).$$

The idea here is that one may not want an action to be done if it causes an undesirable effect.

Time is important in reasoning about many emotions, especially the prospect-based emotions *hope, satisfaction, relief, fear, fears-confirmed,* and *disappointment.* Our situation calculus deals with temporal precedence but ignores all other temporal relationships such as simultaneity. In the situation calculus, information is stored in an initial state and then propogated to later states via frame axioms. This is a problem in reasoning about emotional reactions to ongoing events. For example, a woman can be grateful to her husband for giving her a massage while he is giving it to her, but limitations of the situation calculus prevent a formulation of the eliciting conditions for gratitude during an ongoing action. The use of representation and reasoning methods suggested by Allen (1981) might help address this limitation.

Actions are viewed as discrete, opaque transitions. Situation calculus provides no tools for describing what happens while an action is in progress; it provides no tools for describing continuous processes like the gradual dissipation of anger. Again, representation and reasoning techniques developed in research on qualitative physics (e.g., Bobrow, 1985; Hobbs, 1985) might help overcome this limitation.

Strict logical implication often fails to capture the reality of relationships between events, actions, and possible effects. Many contributions of actions toward the achievement of goals involved in examples drawn from the diary study are uncertain in the sense that the action is not guaranteed to achieve the goal. Often, actions facilitate or increase the probability that the goal will be achieved. Several examples in the case study data describe student's emotional reactions to the granting of extensions on due dates of assignments. Besides temporal reasoning, these examples require probabilistic reasoning, in the sense that the granting of extensions increases the likelihood of successful completion of projects. Instead of encoding these weak causal relationships as implications, qualitative representations of conditional probabilities could be associated with cause-effect relationships. Plausibility information such as probabilities could enter more directly into the eliciting conditions of several emotions. In particular, the *hopes* and *fears* emotion types depend on entertaining the possibility that the hoped-for or feared fluent will occur. The intensity of hope or fear (and its plausibility) depends in part on the subjective likelihood of the prospective event (Ortony, et al., 1988).

4.4 The problem of evaluating explanations

The most important limitation of the method of constructing explanations implemented here is that it does not address important questions of plausibility. Such questions include the following:

How can one avoid a combinatorial explosion of explanations, many of which are completely implausible? Evaluation of plausibility cannot wait until all possible explanations have been produced. Sometimes, infinitely many explanations are possible, so some evaluation must be done during explanation construction.

The machine-generated explanations we have presented here were generated using depth-first search. Most were the first acceptable explanations generated but in some cases the initial explanations were rejected by the user. Alternative explanations for a given example are often compatible, but in general alternate explanations will include mutually inconsistent competitors. This raises the question of how one should decide what to believe and what to disbelieve when conflicts arise between alternate explanations? Methods for weighing the evidence would help decide which explanation is more plausible in many cases. But in other cases, it might be prudent to delay making a decision (Josephson, 1990). Or one could take some action aimed at acquiring new information that might help resolve the conflict.

Finally, how should one decide when to assume a hypothesis that would explain given observations? Currently, we rely on simple heuristics that specify inadmissable assumptions (see Section 3.1). When these fail, the abduction engine falls back on the user. The user is asked to validate each assumption and to accept each explanation. If the user rejects an assumption or explanation, backtracking occurs and the abduction engine seeks the next alternative.

5 Related and future work

A different approach to formalizing commonsense reasoning about emotions is presented in Sanders (1989). Her work takes a deductive approach, using a deontic logic of emotions. The logic focuses on a cluster of emotions involving evaluations of actions — including what we have called admiration, reproach, remorse, and anger. The evaluation of actions is ethical, and involves reasoning about obligation, prohibition, and permission. The logic was used to solve problems involving actions associated with ownership and possession of property (e.g., giving, lending, buying, and stealing) by proving theorems. For example, the fact that Jack will be angry was proved given the following:

Jack went to the supermarket.
He parked his car in a legal parking place.
When he came out, it was gone.

It is not clear whether the theorems were proved automatically or by hand, so questions of complexity of inference and control of search in the deontic logic remain unanswered. We have argued that abduction offers advantages over deduction alone when applied to the task of constructing explanations involving emotions. Furthermore, our situation calculus of emotion elicitation is more comprehensive than the deontic logic for emotions in that it covers more emotion types. But our approach could benefit from the treatment of ethical evaluations. We hope to undertake a detailed comparison and integration of the best parts of the two approaches in future work.

A number of theories of the cognitive determinants of emotions exist (e.g., Roseman, 1984). In principle, situation calculus could be used to codify these alternative theories. The abductive method we propose for explanatory reasoning about emotions does not depend on the particular emotion theory used.

Recent research on abduction addresses the issues of search control and plausibility mentioned earlier. Stickel (Hobbs, Stickel, Appelt, & Martin, in press; Hobbs, Stickel, Martin, & Edwards, 1988) has suggested a hueristic approach to evaluating explanations in the context of natural language processing. Subsequent work by Charniak and Shimony (1990) gave Stickel's weighted abduction a probabilistic semantics. Still more recent work (Poole, 1991) incorporates probability directly into a logic-based approach to abduction. These methods promise to provide significantly improved abduction engines that could be used to construct explanations involving emotions, taking plausibility into account.

Work on explanations involving emotions is important for Artificial Intelligence (AI). Because emotions play a substantial role in our lives, emotion words occur frequently in natural language discourse. In Ortony, Clore, and Foss (1987) about 270 English words are identified as referring to genuine emotions from an initial pool of 600 words that frequently appear in the emotion research literature. In another study, 130 of these emotion words were distributed among the 22 emotion types discussed in the present paper. Some emotion words map to several different types, for example, "upset" is compatible with at least *angry_at*, *distress*, and *shame*. In addition, many words map to the same type. Encoding the relationship between the affective lexicon and the emotion types is an important topic for future research aimed at automatically processing natural language text involving emotions.

Natural language processing systems encountering text involving emotions will need to identify and reason about emotions felt by characters in the text. As pointed out by Charniak and McDermott (1986) in their introductory textbook on AI, motivation analysis and plan recognition frequently require reasoning about emotions. The present work focuses on explaining emotions in terms of eliciting situations. But while situations give rise to emotional reactions, emotions in turn give rise to goals and actions that change the state of the world. Applications such as plan recognition will require the representation of knowledge of causal connections between emotions and subsequent actions. For a brief description of a system for recognizing plans involving emotions, see Cain, O'Rorke, and Ortony (1989). This paper also describes how explanation-based learning techniques can be used to learn to recognize such plans. For a fuller discussion of reasoning about emotion-induced actions, see Elliott and Ortony (1992).

6 Conclusion

We have presented a first order logical theory of the cognitive antecedents of emotions and we have described an abductive method for explaining emotional states. We sketched a computer program, an abduction engine implemented in a program called AMAL, that uses the theory of emotion elicitation to construct explanations of emotions. We presented explanations of examples based on cases taken from a diary study of emotions. We examined the strengths and weaknesses of both the representation of knowledge of eliciting conditions and the method for constructing explanations.

The most important advantage of our approach to explanatory reasoning about emotions is that abduction allows us to construct explanations by generating hypotheses that fill gaps in the knowledge associated with cases where deduction fails. We found that the majority of the diary study examples could not be explained using deduction alone because they do not follow logically from the given facts. The abduction engine explained the emotions involved in these cases by making assumptions including valuable inferences about mental states such as desires, expectations, and the emotions of others.

Acknowledgments

Terry Turner kindly provided diary study data. Steven Morris, Clark Elliott, Tim Cain, David Aha, Tony Wieser, Stephanie Sage, Patrick Murphy, and Milton Epstein participated in various stages of this work.

Bibliography

Allen, J. F. (1981). Maintaining knowledge about temporal intervals. *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 221–226). Vancouver: Morgan Kaufman.

Bobrow, D. G. (Ed.). (1985). *Qualitative reasoning about physical systems.* Cambridge, MA: MIT Press.

Cain, T., O'Rorke, P., & Ortony, A. (1989). Learning to recognize plans involving affect. In A. M. Segre (Ed.), *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 209–211). Ithaca; NY: Morgan Kaufmann.

Charniak, E., & McDermott, D. V. (1986). *Introduction to Artificial Intelligence.* Reading, MA: Addison-Wesley.

Charniak, E., & Shimony, S. E. (1990). Probabilistic semantics for cost based abduction. *The Eighth National Conference on Artificial Intelligence* (pp. 106–111). Boston, MA: AAAI Press/ The MIT Press.

DeJong, G. F., & Mooney, R. (1986). Explanation-based learning: An alternative view. *Machine Learning,1*(2), 145-176.

Elkan, C. (1990). Incremental, approximate planning. *The Eighth National Conference on Artificial Intelligence* (pp. 145–150). Boston, MA: AAAI Press/The MIT Press.

Elliott, C., & Ortony, A. (1992). *Point of view: Modeling the emotions of others.* Manuscript submitted for publication.

Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence,2*, 189–208.

Forbus, K. D. (1984). Qualitative process theory. *Artificial Intelligence,24*, 85–168.

Green, C. C. (1969). Applications of theorem-proving to problem solving. In D. E. Walker, & L. M. Norton (Ed.), *International Joint Conference on Artificial Intelligence* (pp. 219–240). Washington, D. C.: Morgan Kaufmann.

Heider, F. (1958). *The psychology of interpersonal relations.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Hobbs, J. (Ed.). (1985). *Formal theories of the commonsense world.* Norwood, NJ: Ablex Publishing Corporation.

Hobbs, J. R., Stickel, M., Appelt, D., & Martin, P. (in press). Interpretation as abduction. *Artificial Intelligence.*

Hobbs, J. R., Stickel, M., Martin, P., & Edwards, D. (1988). Interpretation as abduction. *Proceedings of the Twenty Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 95–103). Buffalo, NY: The Association for Computational Linguistics.

Josephson, J. R. (1990). On the"logical form" of abduction. *Working notes of the AAAI spring symposium on automated abduction.* Stanford, CA: Department of Information and Computer Science, University of California, Irvine.

Konolige, K. (in press). Abduction vs. closure in causal theories. *Artificial Intelligence.*

Kowalski, R. (1979). *Logic for problem solving.* New York: Elsevier Science Publishing, Inc.

Kuipers, B. J. (1986). Qualitative simulation. *Artificial Intelligence,29*, 289–338.

McCarthy, J. (1968). Programs with common sense. In M. Minsky (Eds.), *Semantic Information Processing* (pp. 403–418). Cambridge, MA: MIT Press.

McCarthy, J. (1977). Epistemological problems of artificial intelligence. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence* (pp. 1038–1044). Cambridge, MA: Morgan Kaufmann.

McCarthy, J., & Hayes, P. (1969). Some philosophical problems from the standpoint of artificial intelligence. In *Machine Intelligence* (pp. 463-502). Edinburgh, UK: Edinburgh University Press.

Mitchell, T. M., Keller, R. M., & Kedar-Cabelli, S. T. (1986). Explanation-based generalization: A unifying view. *Machine Learning,1*(1), 47–80.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions.* New York: Cambridge University Press.

Ortony, A., Clore, G. L., & Foss, M. A. (1987). The referential structure of the affective lexicon. *Cognitive Science,11*(3), 361–384.

Peirce, C. S. S. (1931–1958). *Collected papers of Charles Sanders Peirce (1839–1914).* Cambridge, MA: Harvard University Press.

Poole, D. (1991). Representing diagnostic knowledge for probabilistic horn abduction. *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence* (pp. 1129–1135). Sydney, Australia: Morgan Kaufmann.

Poole, D. L., Goebel, R., & Aleliunas, R. (1987). Theorist: A logical reasoning system for defaults and diagnosis. In N. Cercone, & G. McCalla (Eds.), *The Knowledge Frontier: Essays in the Representation of Knowledge.* New York: Springer-Verlag.

Pople, H. E. (1973). On the mechanization of abductive logic. *Proceedings of the Third International Joint Conference on Artificial Intelligence* (pp. 147–152). Stanford, CA: Morgan Kaufmann.

Reiter, R. (1991). The frame problem in the situation calculus: A simple solution (sometimes) and a completeness result for goal regression. In V. Lifschitz (Eds.), *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy* (pp. 359–380). Boston: Academic Press, Inc.

Roseman, I. J. (1984). Emotions, relationships, and health. In P. Shaver (Eds.), *Review of personality and social psychology* (pp. 11–36). Beverly Hills, CA: Sage.

Sanders, K. E. (1989). A logic for emotions: a basis for reasoning about commonsense psychological knowledge. In E. Smith (Ed.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 357–363). Ann Arbor, MI: Lawrence Erlbaum and Associates.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language-understanding. *Cognitive Psychology,3*(4), 552–631.

Turner, T. J. (1985). *Diary study of emotions: Qualitative data.* Unpublished raw data. University of Illinois, Department of Psychology.

Weld, D. S., & de Kleer, J. (Ed.). (1990). *Readings in qualitative reasoning about physical systems.* San Mateo, CA: Morgan Kaufmann.