Knowledge Acquisition Within the Affective Reasoner*

Clark Elliott

Institute for the Learning Sciences
Northwestern University
Evanston, Illinois, 60201
email: elliott@ils.nwu.edu
and
DePaul University
Department of Computer Science
243 South Wabash Ave.
Chicago, Illinois, 60604

Abstract

Comprehensive intellectual modeling is a very difficult, if not impossible task[Van88, Ric79, Ric83]. On the other hand, many researchers have pointed out the increasing need for architectures that support reasoning in and about societies of autonomous agents, both human and machine [Ric91, AAA91, GRWH90, DLC89], which in turn implies the need for models of other agents. As a solution to this difficult problem we propose using affective personality representation as a tractable way of establishing the perspectives of individual agents within a society. A brief description of the underlying cognitive theory is presented along with our implementation of these ideas in the Affective Reasoner. This program, designed to simulate and reason about the affective life of autonomous agents, is discussed with respect to affective knowledge acquisition.

1 Introduction

In this paper we discuss the need for reasoning about the motivations of other agents in an uncertain world. We take the stand that it may be more useful to ask the question, *How do I address the agent's concerns?* then it is to ask the question, *What is the agent thinking about?* Using feedback from affective states to monitor successful interaction (and otherwise) may well be a simpler process than modeling the intellectual processes of other agents. As an example, consider the modeling of students in a teaching environment. A teacher may spend significant energy looking for affective feedback such as "the light in the student's eyes" while consuming little or no resources in intellectual modeling of the student. Since such feedback is continuous, and relatively cheap, even "dumb" teaching strategies such as presenting alternate explanations by rote may become effective: boredom, frustration, anger, etc. can be used to abort poor teaching paths; joy, admiration, challenge, etc. can be used to continue fruitful ones.

How hard *is* affective modeling? Can it be done within a content domain? Even if we are able to do so, is there any payoff with respect to explaining the behavior of other agents? An interesting discussion can be generated from the claim that knowledge about the *motivations* of other agents may be had from clues in the affective lexicon established for a given domain. This supposes that (1) an affective lexicon *can* be established for the domain, that (2) we are able to abduce affective states from actions on the part of agents, and that (3) we can tie affect-eliciting conditions in the "world" to these states through suppositions about the way particular agents view that world.

In section 2 we give a brief description of the underlying cognitive theory taken from [OCC88], and the implementation of this theory in the Affective Reasoner. We then show how this theory of the eliciting conditions for emotions is extended to include both action generation and simple personal-

*Preparation of this article was supported in part by grant IRI-8812699 awarded to Andrew Ortony by the National Science Foundation, and and in part by Andersen Consulting through Northwestern University's Institute for the Learning Sciences. Citataion appears in bibliography.

ity reasoning. In section 3 we show how current trends in AI will be putting increasing stress on the need for successful interaction between autonomous agents, despite the knowledge that this is very difficult to achieve. We make the argument that affective reasoning can provide a context-independent framework for communicating between diverse entities and in diverse domains, and that all intelligent agents must have some capacity for acquiring this affective knowledge.

In section 4 we discuss the method the Affective Reasoner uses for constructing and maintaining multiple personality-of-other representations, and how it uses these for explaining the behavior of agents in the system. We discuss a global repository for known construals of events and how this may be used to hypothesize about states present in agents as they respond to events in their world. To illustrate these ideas let us imagine that we see a man jumping up and down after the completion of a horse race. We might suppose that he is jumping for joy after having won a bet. On the other hand we might also hypothesize that he is jumping up and down from acute anxiety after having lost a great deal of money. In either case, however, we would add *betting on horses* and *observing the outcomes of races* to the concerns we are supposing that he has. This would be part of our representation of his personality. Still later we might discover, empirically, that the man has no interest in betting at all, and that he was jumping up and down in rage because he has just discovered that his wife, a compulsive gambler, has broken a promise never to go to the race track again.

Forward reasoning, such as the above, from eliciting conditions to affective states is combined with (backwards) heuristic classification of action sets and event features into those same states. An overview of the implementation of this two-pronged approach in the Affective Reasoner is given in the context of knowledge acquisition about personalities.

2 Aspects of the Cognitive Theory

In the Affective Reasoner we divide the affective personality component into two parts: The *Interpretive Personality* mechanism, i.e. the interpretation engine and emotion rules that produces affective states in response to events in the world, discussed in section 2.1, and the *Action Personality* mechanism, i.e. the action generation engine which produces reactions to these affective states, discussed in section 2.2.

2.1 Interpretive Personality: The Construal Phase

One of the essential tenets of our research is that events by themselves are of little interest to a society of complex agents. What *is* of interest is how these events are interpreted. Consequently, a great deal of emphasis is placed on the *construal* mechanism, i.e. the matching and binding process that filters out events not relevant to an agent, and gives meaning to those that are. A brief outline of the underlying theory of the construal process and our implementation of it in the Affective Reasoner follows.

Our platform makes use of a base of the twenty-two emotion categories specified in [OCC88]. Each of these states has a set of eliciting conditions in a domain. When the eliciting conditions are met, and various thresholds have been crossed, these states will follow. A key element of the theory is that these eliciting conditions are strictly relative to the interpretation of the individual agent. That is, eliciting conditions are only causal when they are of *concern* to the responding agent. The type of the response is also dependent upon the nature of these concerns. For example, suppose that my team wins a basketball game with a basket at the buzzer, and your team loses. I may experience *joy* at the event, whereas you may experience *distress.* In both cases the event is the same; it is only the *construal* of the event which is different.

The emotion classes under consideration are *joy* and *distress* over the blocking and achievement of goals; *happiness-for, pity, gloating* and *resentment* over the fortunes of other agents with respect to their goals; *hope* and *fear* over the prospects of some expected future blocking or achievement of a goal; *satisfaction, fears-confirmed, relief* and *disappointment* over the (dis)confirmation of some formerly future goal having been blocked or achieved; *admira-*

tion and *reproach* toward another agent who has upheld or violated a standard; *pride* and *shame* toward one's self when having upheld or violated a standard; *liking/loving* and *disliking/hating* with respect to the attractiveness of some object or agent-as-object; and finally *remorse*(self) and *anger* (other) when goals have been blocked through the violation of some standard and *gratification* (self) and *gratitude* (other) when they have been achieved through the upholding of a standard.

Figure 1 illustrates the different sources for these affective states in the Affective Reasoner. Briefly, a hierarchy of frames containing pattern-matching variables and procedural attachment of working memory to their slots represents the concerns of the individual agents. When an event occurs, agents appraise its relevance by matching it against the corresponding frames. Should an event be of importance to one or more components of the affective machinery of an agent the event is bound to a (set of) interpretation(s). Affect-eliciting relations are created and contain such fields as desirability-for-self, desirability-for-other, confirmation status of the event, goal-blockage, and so forth. Next, heuristic processing rooted in the cognitive theory is performed to decide on the nature of compound and complex states. Such questions as, *When do distress and reproach become anger?* and *Is it possible to love and hate someone at the same time?* are answered here.

The resulting relations are used to determine the affective states generated, and the full set of bindings is used to determine both the exact nature of those states (e.g. the intensity levels of the states) and, later, the nature of the response actions. When states such as *hope* and *fear* are generated, they also generate expectations with respect to future events[1]. States may be generated on behalf of some other agent. This requires having an internal representation of the personality of the other agent and is discussed at length in section 4. The domain-independent rules contain the backbone of the theory. These are the essential eliciting-condition-to-affective-state rules such as *When one of your goals is thwarted you may feel distress,* and *When another agent upholds one of your standards and helps you to achieve a goal you may feel gratitude towards that agent.* Lastly, the domain-dependent rules are those which are affected by the current state of working memory. For example, if someone else succeeds, your affective response will depend upon the current state of your relationship (e.g., are they an enemy or a friend?).

As can be seen, the affective states fall into three major categories: those concerned with the effect of events on the *goals* of an agent, those concerned with *standards and principles* invoked by actions of some agent, and those concerned with *tastes and preferences* with respect to objects, including other agents treated as objects. Another way to view these categories is as an agent's assessment of the *desirability* of some event, of the *praiseworthiness* of some action, or of the *attractiveness* of some object.

An important point to note is that *no event has meaning to an agent until after it has been filtered through the content theory representing the concerns and interpretations of that agent.* Without going into the philosophical foundations of this argument it should be evident that people work this way: extraneous cognitive and perceptual information is filtered out of the input stream, and the events that do pass this filter are passed along with the interpretation of why they are relevant. Obviously not all interpretations are relevant to the human affective machinery (e.g. the information that the light needs to be turned on is not likely to cause an emotional response), but the point is simply that a significant amount of binding to affective inference structures occurs at the time eliciting events are assessed. For example, when someone is identified as being important with respect to an agent's goals, they are also identified as *friend, enemy,* or *neither.*

From the point of view of the agents then, no event actually "happens" until it has been given an interpretation. It follows then that if one is to do reasoning about motivations of other agents (and postulating that motivation is intimately tied to affect) then one must have a mechanism for *seeing events through the eyes of another agent.* Establishing a method for acquiring and indexing enough information to do this effectively is at the heart of this paper.

[1] Conversely, the event that generates *any* affective state may also represent some (dis)confirmation of a prior set of expectations. This is a nasty matching problems and raise theoretical questions such as, *At what point does some future event become (dis)confirmed, by default, through lack of any related action?*

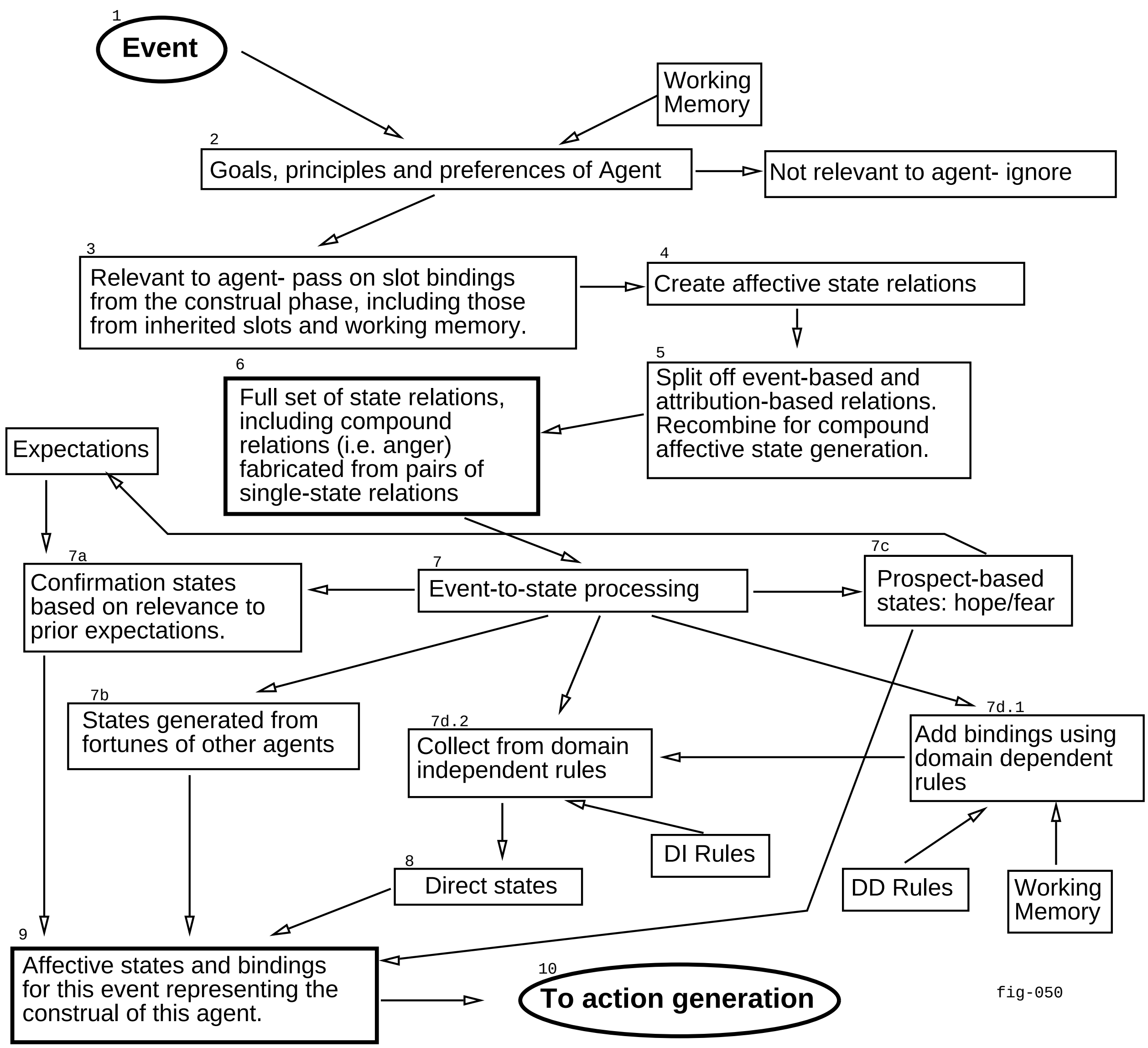

Figure 1: Structure of the Mapping from Event to Affective States

2.2 Response Personality: The Action Generation Phase

The response action personality component performs the task of generating appropriate actions for affective states within the context of some event. We break this response-space into three dimensions. The first is the twenty-two emotion classes, the second is a theoretical breakdown into twenty action response categories such as *somatic* responses, *behavioral-toward-animate-, other-agent-directed attribution-, suppression-, problem-oriented-* responses and so forth. The third is the actions themselves. For example, in response to one of the *fear* emotions one might have a spontaneous somatic *trembling* response, or he might have a planned problem-oriented *attack* response. To get an idea of the scope of the problem, consider this: There are 440 emotion-class / action-category duples. Each of these duples may average 3 actions each, giving us 1320 actions tracked by the system. Each action may be simple (such as a token for *laugh*), it may contain variables (such as in the template *obsessively focus on the opinion of ?other-agent*), or it may be a full mini-plan (such as *attack other agent*). Other complicating issues are the selection of more than one action at once (such as when one smiles and tells another good news), multiple and conflicting emotions, sets of individually correct but mutually incompatible actions, the intensity of the response and the duration of the response.

The Affective Reasoner tracks the 440 different emotion-class / action-category duples, and supports an average of about three actions for each class, of varying complexity. In addition it reasons about the selection of an appropriate responses within a given category, allows selection from multiple response categories, and resolves conflicts between incompatible actions. Because of the many-to-many nature of the mapping from emotions to actions, a case-based heuristic classification system is used for acquiring knowledge about the meaning of an agent's actions. This is discussed further in section 4.1.

3 Human-Computer Interaction

There is a growing interest in treating computers and humans as part of a larger community of cognitive agents. In the working notes for the 1991 AAAI fall symposium on Knowledge and Action at Social and Organizational Levels a number of authors address this issue [AAA91]. Hovy, for example, talks about a multi-part functional unit of car and driver scheduling tune-ups with a multi-part functional unit of garage and mechanic, where instigation may actually be performed by the car computer or the garage computer. Hanson discusses the merits of using multiple autonomous agents for problem solving, but then stresses that issues such as distrust inevitably arise. This argument is extended to include both human and automated systems. Hughes discusses the creation of artificial primate societies and the analysis of the politics that naturally arise in such systems, with the implication that simple societies such as this may be simpler models of those in which humans operate.

In addition, there is current interest in distributed problem solvers using loosely coupled agents [GRWH90, DLC89]. Such systems are bound to proliferate in the future as a result of the connection of special purpose computers and software packages via network links. A natural outgrowth of this is the interaction of human and computer expert partners. Along these lines Rich points out that many goals are too complex to plan for outright, and that under such circumstances goals and subgoals may have to be continually revised as action proceeds [Ric91]. In large projects, especially in those making creative demands on both human and automated contributing agents we see the need for both increasing autonomy and effective communication across domains. We have the view that in such complex and changing "worlds" agents must be able to respond to each other and to events in that world in a dynamic and timely fashion, and that emotions reasoning will lay a foundation for this.

In like manner, as computers start to produce results when given less well-defined, more creative tasks, increased stress and importance will be placed on the human-computer interface. Both the computer *and* the human user will have to be able to reassess goals as the problem state is altered. There are two specific features in such an environment that call for the use of affective control components: (1) in an unpredictable world there is need for continual assessment of the state of that world with respect to the concerns of the observing agent, and (2) local control of action initiation is necessary to respond in a timely manner to both serendipitous opportunities and immediate preservation requirements.

4 Acquiring Knowledge about Personalities

In the Affective Reasoner we use an independent reasoning mechanism for generating explanations about affective states and the actions that result from them. To do this we must store, and modify, internal representations of all other agents about whom we wish to reason. We must also have some theory of the meaning of expressive actions with respect to affective states. We may reason in both directions, that is, the one that says, *given the eliciting conditions we can surmise the affective states* and the one that says, *given the actions we can explain them in terms of the affective state,* but it should be clear that in our system, and in the human world as well, we can make no direct observation that will fully explain what an agent is *feeling.* Emotions are notoriously hard to track, even for the person who may be having them. In the degenerate case identical actions may even be caused by conflicting emotions which are caused by identical eliciting conditions. To wit: Sam is weeping tears of real pain because he is sad that his favorite uncle has died. He is also weeping with joy because he is so happy that the death of his uncle means he will inherit a million dollars and he can now pay for leukemia treatments for his daughter.

In our system, the more interaction the reasoner has with an agent the better his representation of that agent's concerns becomes. As time progresses predictions about the agent's reactions improve and explanations are better. Reasoning time also decreases. We use the above-mentioned two-pronged approach to developing accurate personality representations.

4.1 Heuristic Classification

The first of these approaches uses PROTOS, an exemplar-based knowledge acquisition system developed by Ray Bareiss [Bar89, PBH90]. PROTOS observes events and response actions and is taught to classify these as suggesting particular affective states. As more cases become available, certain features become more important in determining the likelihood of a certain state. Composite sets of features become prototypical cases. For example, if a shaking fist is exclusively associated with anger in some domain then this action feature will come to be known as highly indicative of anger. As another example, if the event of getting a speeding ticket is always associated with remorse or anger, then that event feature comes to be correlated with the blocking of goals and the violation of standards.

But, while this is useful information for reasoning backwards from artifacts to affective states, it does little to help us with constructing personality representations. To do this we must ask the question, *What interpretations of eliciting conditions would cause this state to occur?* For example, assuming that we have established that anger is probably present for *Vince* because he is shaking his fist, we must then look for a construal of the preceding event which might cause *Vince* to have this emotion. Once we have found a plausible candidate we can add this to our Interpretive Personality representation for Vince. Now, if in the future this event occurs again, in a similar instantiation, we might correctly explain that Vince is angry, even though he might not be shaking his fist as an expression of this.

Similarly, if we have established a high correlation between getting a ticket and the blocking of goals and violation of standards, then we might be able to correctly classify a getting-ticket event as one of anger or remorse, even if the action-response features are unknown to us. Once we have established that getting tickets usually leads to anger or remorse, we will adopt construals that lead to these states in the personality representations of agents in our system. Having done this we are then able to directly derive affective states by using these representations for *filtering this type of event through the eyes of other agents.* This is explained further in the next section.

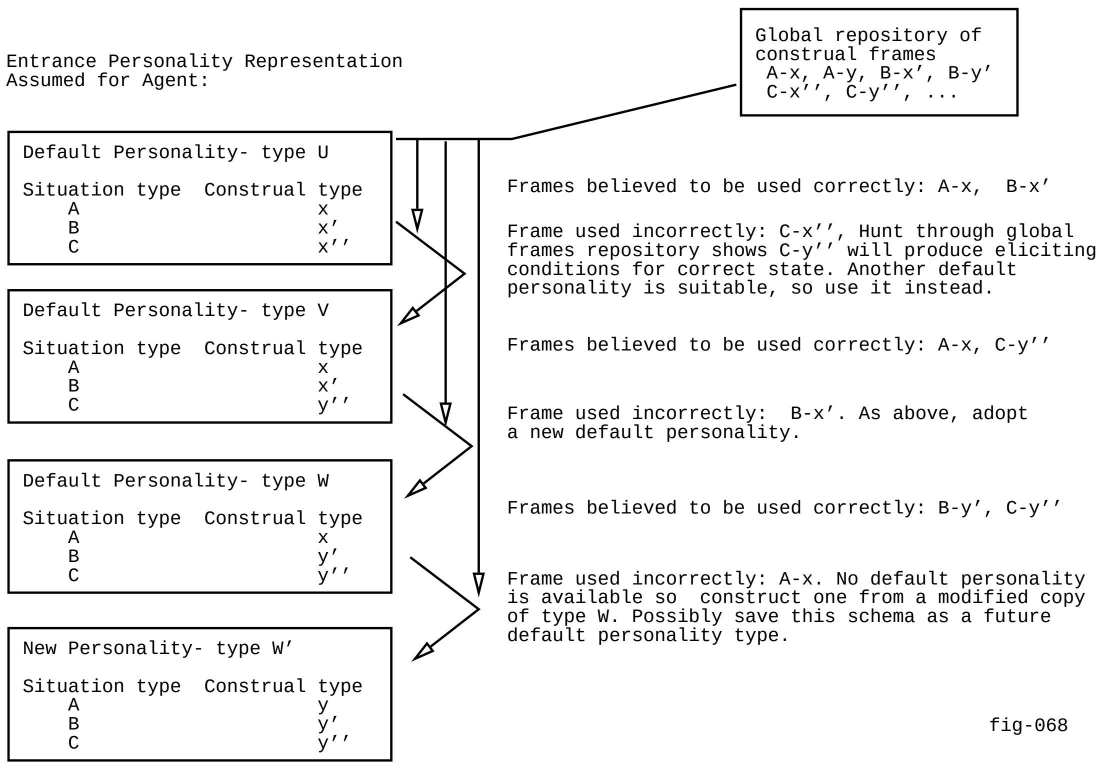

Figure 2: Selecting and Adapting an Interpretive Personality Structure for Another Agent

4.2 Building Interpretive Personality Representations of Other Agents

Building the frame-hierarchies that represent the concerns of other agents is an iterative process. Figure 2 diagrams this process. We start with a default Interpretive Personality representation provided by the system. As long as interpretations of events filtered through this construal database are consistent with the observed features, and as long as the "teacher" of the system has not complained, then no changes are made. It is not until some construal is found to be in error that we make changes to the database.

In the example given here, the original default personality is given as the abstract type $\mathcal{U}$. This produces the suitable construal of x for an event of type A (e.g. event type A might be getting a speeding ticket, construal x for A might be having a retain-money goal blocked), and the suitable construal, x', for an event of type B. A construal of type x'' is *not* suitable for event type C however, presumably because it does not lead to the affective state believed to have resulted from the event. This initiates a search in the global repository of construal forms for an interpretation that will lead to the known affective state. We find a construal form, y'', for event type C which does produce the correct affective state, and we replace (or possibly complement) the existing construal form x''.

Rather than do this outright, however, we first search for an existing personality type that contains both the forms already used correctly and the new one as well. In the diagram, we are able to adopt a personality of type $\mathcal{V}$ that has these properties. Following a similar episode involving the event type B we have to adopt a personality of type $\mathcal{W}$. Finally, we have to create a new personality type, $\mathcal{W}'$ to model this agent. In the Affective Reasoner default personality types may be *paranoid, stingy, altruistic,* etc.

Note that we do not make any attempt to directly model the emotion-to-action component of an agent's personality. This has to do with the one-to-many mapping characteristics of action generation, and has not been found to be generally useful. On the other hand we *do* do some modeling of the relationship between an agent's affective states and his actions through the heuristic classification system. For example, we might know that *Vince* usually shakes his fist when he is angry, but *Tom* simply grows quiet.

5 Conclusion

Affective personality representation is practical and useful. We can use it to reason effectively in diverse domains and about diverse agents that have contrasting views of the world. Given the state of AI we are likely to see an increased use of systems which place extensive demands on the survival and productivity of autonomous agents in unpredictable and ill-defined worlds. One way to reason about other agents in such worlds is by generating explanations for their activities based on suppositions about their emotional states. The Affective Reasoner program we have developed exhibits some of these features.

References

[AAA91] Working notes for aaai fall symposium on knowledge and action at the social and organizational levels. Distriubted to Workshop Attendees, November 1991. References to the work of Clark Elliott, Eduard Hovy, Robin Hanson and Lucian Hughes.

[Bar89] Ray Bareiss. *Exemplar-Based Knowledge Acquisition, A Unified Approach to Concept Representation, Classification, and Learning.* Academic Press, Inc., 1989.

[DLC89] Edmund H. Durfee, Victor R. Lesser, and Daniel D. Corkill. Cooperative distributed problem solving. In *The Handbook of Artificial Intelligence*, volume 4, pages 83–146. Addison-Wesley, 1989.

[EHL92] Clark Elliott, Lawrence Henschen, and James Lu. Adapting a cognitive modeling platform for use in the control of automated reasoning. unpublished, 1992.

[Ell91] Clark Elliott. Taxiworld: Computing emotions. In *Notes of the Center for Neural Engineering Student Workshop on Emotions*, pages 117–137. Uinversity of Southern California, 1991.

[Ell92] Clark Elliott. Knowledge acquisition within the Affective Reasoner. In *Notes for the AAAI Spring Symposium: Cognitive Aspects of Knowledge Acquisition*, pages 64–71, Stanford University, Palo Alto, CA, 1992. American Association of Artificial Intelligence. THIS PAPER.

[EO92] Clark Elliott and Andrew Ortony. The affective reasoner: Modeling emotions in a multi-agent system. Unpublished Manuscript, April 1992.

[FS87] Nico Frijda and Jaap Swagerman. Can computers feel? theory and design of an emotional system. *Cognition & Emotion*, 1(3):235–257, 1987.

[GRWH90] Les Gasser and Jr. Randall W. Hill. Coordinated problem solvers. *Annual Review of Computer Science*, 4:203–253, 1990.

[OCC88] Andrew Ortony, Gerald L. Clore, and Allan Collins. *The Cognitive Structure of Emotions.* Cambridge University Press, 1988.

[PBH90] Bruce W. Porter, Ray Bareiss, and Robert C. Holte. Concept learning and heuristic classification in weak-theory domains. *Artificial Intelligence*, 45:229–263, 1990.

[Ric79] Elaine Rich. User modeling via stereotypes. *Cognitive Science*, 3:329–354, 1979.

[Ric83] Elaine Rich. Users are individuals: individualizing user models. *International Journal of Man-Machine Studies*, 18:199–214, 1983.

[Ric91] Elaine Rich. Planning, reacting and communicating. In press, 1991.

[Van88] Kurt Van Lehn. Student modeling. In Martha C. Polson and J. Jeffery Richardson, editors, *Foundations of Intelligent Tutoring Systems*. Lawrence Erlbaum Associates, 1988.