Modeling Social Emotions and Social Attributions

Jonathan Gratch

Wenji Mao

Stacy Marsella

1. Introduction

Emotions play a crucial role in mediating human social relationships (Davidson, Scherer, & Goldsmith, 2003). Whether articulated through body movements, voice, deed, or through the ways we justify our actions, human relationships are laden with emotion. Emotion can act as a signal, communicating information about the sender's mental state, indicating their future actions, and indirectly inducing emotions in the mind of observers. Emotion can also act as a mental process, altering how people see the world, how they form decisions, and how they respond to the environment. In our work we seek to develop testable computational models that emphasize the relationship between emotion and cognition (Gratch & Marsella, 2001; Marsella & Gratch, 2003). In this chapter, we focus on emotions that have a social component: the rage arising from a perceived offence, the guilt we feel after harming another. Such emotions arise from *social* explanations involving judgments not only of causality but intention and free will (Shaver, 1985). These explanations underlie how we act on and make sense of the social world. In short, they lie at the heart of social intelligence. With the advance of multi-agent systems, user interfaces, and human-like agents, it is increasingly important to reason about this uniquely human-centric form of social inference. Here we relate recent progress in modeling such socio-emotional judgments.

Modeling emotions is a relatively recent focus in artificial intelligence and cognitive modeling and deserves some motivation. Although such models can ideally inform our

understanding of human behavior, we see the development of computational models of emotion as a core research focus that will facilitate advances in the large array of computational systems that model, interpret or influence human behavior. On the one hand, modeling applications must account for how people behave when experiencing intense emotion including disaster preparedness (e.g., when modeling how crowds react in a disaster (Silverman, 2002)), training (e.g., when modeling how military units respond in a battle (Gratch & Marsella, 2003)), and even macro-economic models (e.g., when modeling the economic impact of traumatic events such as 9/11 or the SARS epidemic). On the other hand, many applications presume the ability to correctly interpret the beliefs, motives and intentions underlying human behavior (such as tutoring systems, dialogue systems, mixed-initiative planning systems, or systems that learn from observation) and could benefit from a model of how emotion motivates action, distorts perception and inference, and communicates information about mental state. Emotions play a powerful role in social influence, a better understanding of which would benefit applications that attempt to shape human behavior, such as psychotherapy applications (Marsella, Johnson, & LaBore, 2000; Rothbaum et al., 1999), tutoring systems (Lester, Stone, & Stelling, 1999; Ryokai, Vaucelle, & Cassell, in press; Shaw, Johnson, & Ganeshan, 1999), and marketing applications (André, Rist, Mulken, & Klesen, 2000; Cassell, Bickmore, Campbell, Vilhjálmsson, & Yan, 2000). Lastly, models of emotion may give insight into building models of intelligent behavior *in general*. Several authors have argued that emotional influences that seem irrational on the surface have important social and cognitive functions that would be required by any intelligent system (Damasio, 1994; Minsky, 1986; Oatley & Johnson-Laird, 1987; Simon, 1967; Sloman & Croucher, 1981). For example, social emotions such as anger and guilt may reflect a mechanism that improves group utility by minimizing social conflicts, and thereby explains peoples "irrational" choices in social games such as prison's dilemma (Frank, 1988). Similarly, "delusional" coping strategies such as wishful thinking may reflect a rational mechanism that is more accurately accounting for certain social costs (Mele, 2001).

Virtual Humans and "Broad" Cognitive Models

Though, much of cognitive science and cognitive modeling has focused on accurately modeling relatively narrow psychological phenomena, our work is part of a growing trend to demonstrate cognitive models within the context of "broad agents" that must simultaneously exhibit multiple aspects of human behavior (Anderson & Lebiere, 2003). Arguably, the most ambitious of such efforts focus on the problem of developing *virtual* humans, intelligent systems with a human-like graphical manifestation. Building a virtual human is a multidisciplinary effort, joining traditional artificial intelligence problems with a range of issues from computer graphics to social science. Virtual humans must act and react in their simulated environment, drawing on the disciplines of automated reasoning and planning. To hold a conversation, they must exploit the full gamut of natural language research, from speech recognition and natural language understanding to natural language generation and speech synthesis. Providing human bodies that can be controlled in real time delves into computer graphics and animation. And because a virtual human looks like a human, people readily detect and are disturbed by discrepancies from human norms. Thus, virtual human research must draw heavily on psychology and communication theory to appropriately convey nonverbal behavior, emotion, and personality. Through their breadth and integrated nature, virtual humans provide a unique tool for assessing cognitive models.

In developing computational models of emotional phenomena, we focus on models that can influence and exploit the wide range of capabilities that a virtual human provides. In particular, we have used emotion models to mediate the cognitive and communicative behavior of virtual humans in the context of the Mission Rehearsal Exercise (MRE) training system. In this system, students can engage in face-to-face spoken interaction with the virtual humans in high-stress social settings (Figure 1a) (Gratch, 2000; Gratch & Marsella, 2001; Marsella & Gratch, 2002, 2003; Rickel et al., 2002). Emotional models help create the non-verbal communicative behavior and cognitive biases one might expect if trainees were interacting with real people in similar high-stress settings. Our scenarios focus on dialogue and group decision-making, rather than physical action, so the focus of our emotional models is on cognitive source of emotions, emotion's influence on

cognition (decision-making, planning, and beliefs) and external verbal and non-verbal communicative behavior that reflect the virtual human's emotional state.

Figure 1: Two applications that use virtual humans to teach people to cope with emotionally-charged social situations. The image on the left illustrates the first author interacting through natural language with the MRE system, designed to teach leadership skills. The image on the left is from Carmen's Bright Ideas (Marsella, Johnson, & LaBore, 2003), developed by the third author, and designed to teach coping skills to parents of pediatric cancer patients.

Social Emotions

Allowing naïve users to freely interact with a broad cognitive model can quickly reveal its limitations, and the work described here is motivated by the following example of "novel" emotional reasoning on the part of our virtual humans. In the Mission Rehearsal Exercise, trainees have the opportunity to make bad decisions. In one instance, a human user issued a particular flawed order to his virtual subordinate. The subordinate suggested a better alternative, but when this was rejected, the subordinate, in turn, ordered lower level units to execute the flawed order. Rather than blaming the trainee, however, the virtual human paradoxically displayed anger at the subordinate characters that executed the plan. In contrast, human observers universally assign blame to the trainee, as the subordinate was clearly following orders and even attempted to negotiate for a different outcome. The virtual human's "novel" attribution of blame was traced to some simplifying assumptions in the model: the model assigns blame to whoever actually executes an act with undesirable consequence. In this case, however, the action was clearly coerced. Such results indicate an impoverished capacity to judge credit or blame in a social context. How we addressed this limitation is the subject of the second half of this chapter.

Overview

This chapter provides an overview of EMA, our current model of emotion, and then describes our efforts to extend the model with respect to its ability to reason about social (multi-agent) actions. Section 2 gives a review of cognitive appraisal theory, the theoretical underpinning of our model. Section 3 outlines our current computational approach. Section 4 contrasts our model with related work and describes some limitations. Section 5 discusses how we can extend the model to better account for attributions of social credit and blame. Section 6 ends with some concluding remarks.

2. Cognitive Appraisal Theory (a review)

Motivated by the need to model the influence of emotion on symbolic reasoning, we draw theoretical inspiration from cognitive appraisal theory, a theory that emphasizes the cognitive and symbolic influences of emotion and the underlying processes that lead to this influence (K. R. Scherer, Schorr, & Johnstone, 2001) in contrast to models that emphasize lower-level processes such as drives and physiological effects (Velásquez, 1998). In particular, our work is informed by Smith and Lazarus' cognitive-motivational-emotive theory (Smith & Lazarus, 1990).

Appraisal theories argue that emotion arises from two basic processes: appraisal and coping. Appraisal is the process by which a person assesses their overall relationship with the environment, including not only current conditions, but events that led to this state and future prospects. Appraisal theories argue that appraisal, although not a deliberative process in of itself, is informed by cognitive processes and, in particular, those process involved in understanding and interacting with the environment (e.g., planning, explanation, perception, memory, linguistic processes). Appraisal maps characteristics of these disparate processes into a common set of intermediate terms called *appraisal variables*. These variables serve as an intermediate description of the person-environment relationship and mediate between stimuli and response. Appraisal variables characterize the significance of events from an individual's perspective. Events do not have significance in of themselves, but only by virtue of their interpretation in the context of an individual's beliefs, desires and intention, and past events.

Coping determines how the organism responds to the appraised significance of events, preferring different responses depending on how events are appraised (Peacock & Wong, 1990). For example, events appraised as undesirable but controllable motivate people to develop and execute plans to reverse these circumstances. On the other hand, events appraised as uncontrollable lead people towards denial or resignation. Appraisal theories typically characterize the wide range of human coping responses into two classes. *Problem-focused coping* strategies attempt to change the environment. *Emotion-focused coping*) (Lazarus, 1991) are inner-directed strategies that alter one's mental stance toward the circumstances, for example, by discounting a potential threat or abandoning a cherished goal.

The ultimate effect of these strategies is a change in a person's interpretation of his or her relationship with the environment, which can lead to new (re-)appraisals. Thus, coping, cognition and appraisal are tightly coupled, interacting and unfolding over time (Lazarus, 1991; K. Scherer, 1984): an agent may "feel" distress for an event (appraisal), which motivates the shifting of blame (coping), which leads to anger (re-appraisal). A key challenge for a computational model is to capture this dynamics.

3. A Computational Model of Appraisal and Coping

EMA is a computational model of emotion processing that we have been developing and refining over the last few years (Gratch, 2000; Gratch & Marsella, 2001, 2004a; Marsella & Gratch, 2003). EMA is implemented within the Soar, a general architecture for developing cognitive models (Newell, 1990). Here, we sketch the basic outlines of the model and some of the details of its Soar implementation. Soar is intended to model the mixture of parallel and sequential reasoning that has been posited to underlie human cognition and can be seen as a blackboard model. It provides an unstructured working memory (in terms of objects with attributes and values that can be other objects). Persistent changes to working memory are made by operators that are proposed in parallel but selected sequentially and are intended to model the sequential bottleneck of deliberative reasoning. Elaboration rules fire rapidly and in parallel and make transitory elaborations to working

memory. Soar also provides a model of learning via a chunking mechanism and a model of universal subgoaling, though these last two features do not play a role in our current model.

3.1 EMA Overview

A central tenant in cognitive appraisal theories in general, and Smith and Lazarus' work in particular, is that appraisal and coping center around a person's *interpretation* of their relationship with the environment. This interpretation is constructed by cognitive processes, maintained in a working memory, summarized by appraisal variables and altered by coping responses. To capture this interpretative process in computational terms, we have found it most natural to build on decision-theoretic planning representations (e.g., (Blythe, 1999)) and on methods that explicitly model commitments to beliefs and intentions (Bratman, 1990; Grosz & Kraus, 1996). Planning representations provide a concise description of the causal relationship between events and states, key for assessing the relevance of events to an agent's goals and for forming causal attributions. The appraisal variables of desirability and likelihood find natural analogues in the concepts of utility and probability as characterized by decision-theoretic methods. In addition to inferences about causality, attributions of blame or credit involve reasoning if the causal agent intended or foresaw the consequences of their actions, most naturally represented by explicit representations of beliefs and intentions. As we will see, commitments to beliefs and intentions also play a key role in assigning social blame and credit. Admittedly, these methods and representational commitments have issues from the standpoint of cognitive plausibility, but taken together they form a first-approximation of the type of reasoning that underlies cognitive appraisal.
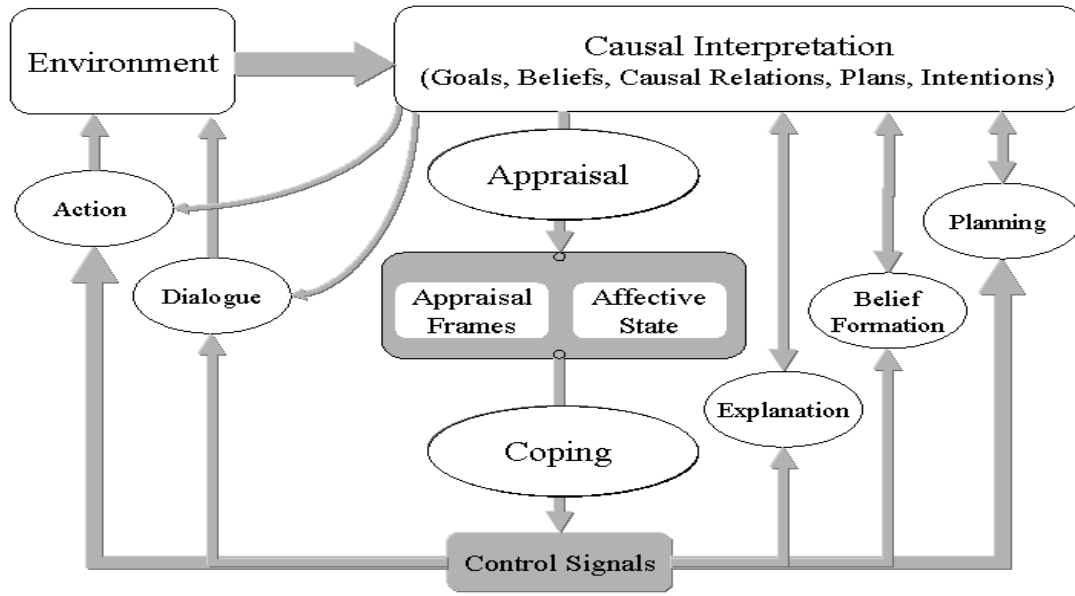
Figure 2: EMA's reinterpretation of Smith and Lazarus

In EMA, the agent's current interpretation of its "agent-environment relationship" is reified by an explicit representation of beliefs, desires, intentions, plans and probabilities that correspond to the agent's working memory. Following a blackboard-type model, this representation encodes as the input, intermediate results and output of reasoning process that mediate between the agent's goals and its physical and social environment (e.g., perception, planning, explanation, and natural language processing). These incremental processes are implemented as Soar operators, though we use the more general term *cognitive operators* to refer to these processes and adopt the term *causal interpretation* to refer to this collection of data structures to emphasize the importance of causal reasoning as well as the interpretative (subjective) character of the appraisal process. At any point in time, the causal interpretation encodes the agent's current view of the agent-environment relationship, an interpretation that may subsequently change with further observation or inference. EMA treats appraisal as a set of feature detectors that map features of the causal interpretation into appraisal variables. For example, an effect of an action that threatens a desired goal would be assessed as a potential undesirable event. Coping acts by creating control signals that prioritize or trigger the processing of cognitive operators, guiding them to overturn or maintain features of the causal interpretation that yield high-intensity appraisals. For example, coping may resign the agent to the threat by

abandoning the desired goal. Figure 2 illustrates a reinterpretation of Smith and Lazarus'
cognitive-motivational-emotive system consistent with this view.

Figure 3 illustrates the representation of a causal interpretation. In the figure, an agent has
a single goal (affiliation) that is threatened by the recent departure of a friend (the past
action "friend departs" has one effect that deletes the "affiliation" state). This goal might
be re-established if the agent "joins a club." Appraisal assesses every instance of an act
facilitating or inhibiting a fluent in the causal interpretation. In the figure, the interpreta-
tion encodes two "events," the threat to the currently satisfied goal of affiliation, and the
potential re-establishment of affiliation in the future.

Each event is appraised in terms of several appraisal variables by domain-independent
functions that examine the syntactic structure of the causal interpretation:

- Perspective: from whose perspective is the event judged
- Desirability: what is the utility of the event if it comes to pass, from the perspective
  taken (i.e., does it causally advance or inhibit a state of some utility)
- Likelihood: how probable is the outcome of the event
- Causal attribution: who deserves credit or blame (i.e., what entity performed the action
  leading to the desirable/undesirable outcome)
- Temporal status: is this past, present, or future
- Controllability: can the outcome be altered by actions under control of the agent whose
  perspective is taken
- Changeability: can the outcome be altered by some other causal agent

Each appraised event is mapped into an emotion instance of some type and intensity, fol-
lowing the scheme proposed by Ortony et al (Ortony, Clore, & Collins, 1988). A simple
activation-based focus of attention model computes a current emotional state based on
most-recently accessed emotion instances.

Coping determines how one responds to the appraised significance of events. Coping strategies are proposed to maintain desirable or overturn undesirable in-focus emotion instances. Coping strategies essentially work in the reverse direction of appraisal, identifying the precursors of emotion in the causal interpretation that should be maintained or altered (e.g., beliefs, desires, intentions, expectations). Strategies include:

- Action: select an action for execution
- Planning: form an intention to perform some act (the planner uses such intentions to drive its plan generation)
- Seek instrumental support: ask someone that is in control of an outcome for help
- Procrastination: wait for an external event to change the current circumstances
- Positive reinterpretation: increase utility of positive side-effect of an act with a negative outcome
- Resignation: drop a threatened intention
- Denial: lower the probability of a pending undesirable outcome
- Mental disengagement: lower utility of desired state
- Shift blame: shift responsibility for an action toward some other agent
- Seek/suppress information: form a positive or negative intention to monitor some pending or unknown state

Strategies give input to the cognitive processes that actually execute these directives. For example, planful coping will generate an intention to perform the join "join club" action, which in turn leads to the planning system to generate and execute a valid plan to accomplish this act. Alternatively, coping strategies might abandon the goal, lower the goal's importance, or re-assess who is to blame.
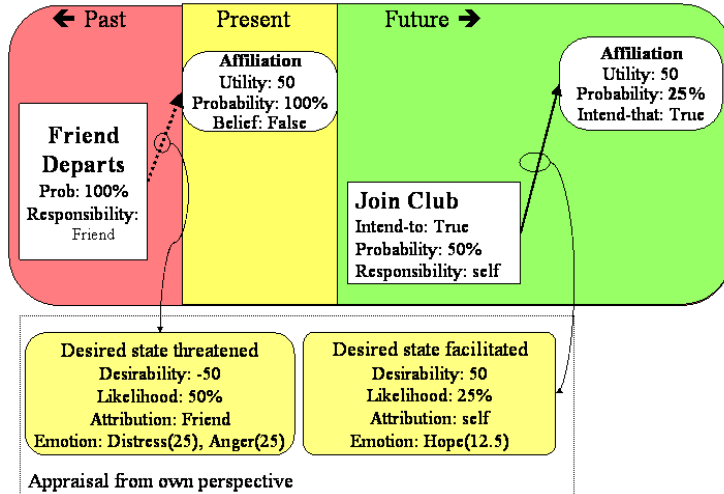
Figure 3: An example causal interpretation

Not every strategy applies to a given stressor (e.g., an agent cannot engage in problem directed coping if it is unaware of an action that impacts the situation), however multiple strategies can apply. EMA proposes these in parallel but adopts strategies sequentially. EMA adopts a small set of search control rules to resolve ties. In particular, the model prefers problem-directed strategies if control is appraised as high (take action, plan, seek information), procrastination if changeability is high, and emotion-focus strategies if control and changeability is low.

1. Construct and maintain a causal interpretation of ongoing world events in terms of beliefs, desires plans and intentions.

2. Generate multiple appraisal frames that characterize features of the causal interpretation in terms of appraisal variables

3. Map individual appraisal frames into individual instances of emotion

4. Aggregate instances and identify current emotional state.

5. Propose and adopt a coping strategy in response to the current emotional state

Figure 4: Stages in EMA'S emotional reasoning

In developing EMA's model of coping, we have moved away from the broad distinctions of problem-focused and emotion-focused strategies. Formally representing coping requires a certain crispness that is otherwise lacking in the problem-focused/emotion-focused distinction. In particular, much of what counts as problem-focused coping in the clinical literature is really inner-directed in an emotion-focused sense. For example, one might form an intention to achieve a desired state – and feel better as a consequence – without ever acting on the intention. Thus, by performing cognitive acts like planning, one can improve ones interpretation of circumstances without actually changing the physical environment.

3.2 Soar Implementation

The overall model consists of the repeated application of the five stages listed in Figure 4. Note that similar stages have been suggested by other cognitive modeling architectures. In particular, they are analogous to the standard problem solving cycle used in the Soar architecture (1990), of which we take advantage in our Soar implementation. Here we describe these stages in some detail.

3.2.1 Construct Causal Interpretation

The causal interpretation is a structured representation built atop Soar's working memory. This representation can be viewed as an explicit representation of a partial order plan in

the sense of (Ambros-Ingerson & Steel, 1988). Certain working memory elements correspond to actions that are liked to precondition and effect objects. Other objects represent relationships between actions such as establishment relations (this action establishes a precondition of that action), threat relations (this action has an effect that disables a precondition of that action), and ordering relations (this action should be executed before that action). There is also an explicit representation of beliefs, desires and intentions (e.g, actions have attributes indicating if they are intended, states have attributes representing their worth to the agent and if they are believed to be true in the current world).

The causal interpretation is constructed sequentially through the application of operators (a process analogous to deliberation). These operators adjust the causal interpretation at a micro level. For example, an update-belief operator will change the belief associated with a single state object. An add-step operator will add a signal step to the current plan, and so forth.

3.2.2 Appraise the Causal Interpretation

Appraisal is performed by elaboration rules that trigger automatically and in parallel based on changes to working memory. For example, if an add-step operator adds a new operator to the plan, elaboration rules automatically fire to assess the significance of this new action from the perspective of the agent's goals: Does the action have an effect that facilitates or inhibits certain desired states? How does this action impact the likelihood of goal achievement, etc. These conclusions are represented by explicit appraisal frames stored in working memory. A separate frame exists for each state object represented in working memory and these are automatically created or modified as a side effect of operators manipulating the causal interpretation.

3.2.3 Construct Emotion Instances

Emotion instances are generated automatically and in parallel from appraisal rules operating on the appraisal variables listed in each appraisal frame. One or more objects representing an emotion type and intensity are associated with the appraisal frame that generates them. The emotion type of the instance is determined by a fixed mapping based on

the configuration of appraisal variables. For example, a frame with low desirability and high likelihood would yield to intense Fear.

3.2.4 Determine Emotional State

EMA uses an activation-based sub-symbolic process, modeled outside of the Soar architecture and loosely motivated by ACT-R, to identify a particular emotional instance to exhibit and cope with. This activation is based on two factors: 1) how recently cognitive structures associated by the instance were "touched" by a Soar operator, and 2) how congruent the instance is to the other emotion instances in memory (this latter factor is intended to account for mood-congruent effects of emotion). For the activation factor, each time a Soar operator accesses an element of the causal interpretation that has an associated appraisal frame, this frame is assigned an activation level equal to its intensity (this currently decays to zero upon the next application of a Soar operator). For example, an "add-step" operator would tend to activate an instance of hope that the step will address the threat and fear that the goal is threatened. For the congruence factor, EMA communicates the type and intensity of all current instances to a module that decays their intensity according to a fixed rate and sums the intensities of instances of a given type into an overall score that can be viewed as the agent's mood (e.g., there is an overall Fear score that consists of the sum of the intensities of each instance of Fear). A small fraction of this mood vector is added to the activation-level of activated instances. The instance with the most activation becomes the emotion to be displayed and coped with.

3.2.5 Propose and Adopt a Coping Strategy

Soar elaboration rules propose individual coping strategies that could potentially address the emotion instance identified in the previous stage. The strategy itself is implemented by a Soar operator and each of these operators is proposed in parallel but only one is ultimately selected by Soar to sequentially apply.

3.3 Limitations and Related Work

EMA relates to a number of past appraisal models of emotion. Although we are perhaps the first to provide an integrated account of coping, computational accounts of appraisal have advanced considerably over the years. In terms of these models, EMA contributes

primarily to the problem of developing general and domain-independent algorithms to support appraisal, and by extending the range of appraisal variables amenable to a computational treatment. Early appraisal models focused on the mapping between appraisal variables and behavior and largely ignored how these variables might be derived, instead requiring domain-specific schemes to derive their value variables. For example, Elliott's (1992) Affective Reasoner, based on the OCC model (1988), required a number of domain specific rules to appraise events. A typical rule would be that a goal at a football match is desirable if the agent favors the team that scored. More recent approaches have moved toward more abstract reasoning frameworks, largely building on traditional artificial intelligence techniques. For example El Nasr and colleagues (2000) use markov-decision processes (MDP) to provide a very general framework for characterizing the desirability of actions and events. An advantage of this method is that it can represent indirect consequences of actions by examining their impact on future reward (as encoded in the MDP), but it retains the key limitations of such models: they can only represent a relatively small number of state transitions and assume fixed goals. The closest approach to what we propose here is WILL (Moffat & Frijda, 1995) that ties appraisal variables to an explicit model of plans (which capture the causal relationships between actions and effects), although they, also, did not address the issue of blame/credit attributions, or how coping might alter this interpretation. We build on these prior models, extending them to provide better characterizations of causality and the subjective nature of appraisal that facilitates coping.

There are several obvious limitations in the current model. The model could be viewed as over-emphasizing the importance of task-oriented goals. Many psychological theories refer to more abstract concepts such as ego-involvement (Lazarus, 1991). Other theories, for example, the theory of Ortony, Clore and Collins (1988), emphasize the importance of social norms or standards in addition to goal processing. For example, fornication may satisfy a personal goal but violate a social standard. Our approach is to represent social standards by (dis-utility) utility over states or actions that (violate) uphold the standard, which we have found this sufficient in practice. Perhaps the largest deficiency of the model concerns the impoverished reasoning underlying causal attributions (and social

reasoning in general), which we will address in the second half of this chapter. Currently the model assumes the executor of an act deserves responsibility for its outcomes, but this can lead to nonsensical conclusions in the case of social actions. We address this limitation in the next section.

4. Modeling Social Attributions

EMA must be extended with respect to its ability to form social attributions of blame and credit. Currently, an entity is assumed credit/blameworthy for an outcome if it actually performed the act. While this works well in single-entity scenarios, in multi-agent settings it can often fall short. For example, when someone is coerced by another to perform an undesirable act, people tend to blame the coercer rather than the actor. People also excuse social blame in circumstances where the act was unintentional or the outcome unanticipated. Failing to account for these mitigating circumstances can lead EMA to produce nonsensical appraisals. The following example from one of our training exercises is illustrative. In the exercise, a trainee (acting as the commander of a platoon) ordered his sergeant (played by a virtual human) to adopt a course of action that the sergeant agent considered highly undesirable. The command was such that it could not be executed directly by the sergeant, but rather the sergeant had to, in turn, order his subordinates to perform the act. The current model assigned blame to the subordinates as they actually performed the undesirable action with the result that the sergeant became angry at his subordinates, *even though he commanded them to perform the offensive act.* Clearly, such results indicate an impoverished ability to assign social credit and blame.

To address this limitation we turn to social psychology. This is in contrast to most computational work on blame assignment that, inspired by philosophy or law, emphasizes proscriptive approaches that try to identify "ideal" principles of responsibility (e.g., the legal code or philosophical principles) and ideal mechanisms to reason about these, typically

contradictory principles (e.g., non-monotonic or case-based reasoning) (McCarty, 1997). As our primary goal is to inform the design of realistic virtual humans that mimic human

communicative and social behavior, our work differs from these models in emphasizing descriptive rather than proscriptive models.

Our extension of EMA is motivated by psychological *attribution theory*, specifically the work of Weiner (Weiner, 1995) and Shaver (Shaver, 1985), as their symbolic approaches mesh well with our existing approach. Indeed, Lazarus pointed to Shaver as a natural complement to his own theory. In these theories, the assignment of credit or blame is a multi-step process initiated by events with positive or negative consequences and mediated by several intermediate variables. First one assesses *causality*, distinguishing between personal versus impersonal causality (i.e., is causal agent a person or a force of nature). If personal, the judgment proceeds by assessing key factors: did the actor *foresee* its occurrence; was it the actor's *intention* to produce the outcome; was the actor forced under *coercion* (e.g., was the actor acting under orders)? As the last step of the process, proper degree of credit or blame is assigned to the responsible agent.[1]

---

[1] Note that we did not strictly follow the process model of Shaver in our approach. As it is explained in later sections, we model the same basic inferences but relax the strict sequential nature of his model. This generalization follows more naturally from the model and, indeed, has been argued for by subsequent theorists (e.g., Weiner).
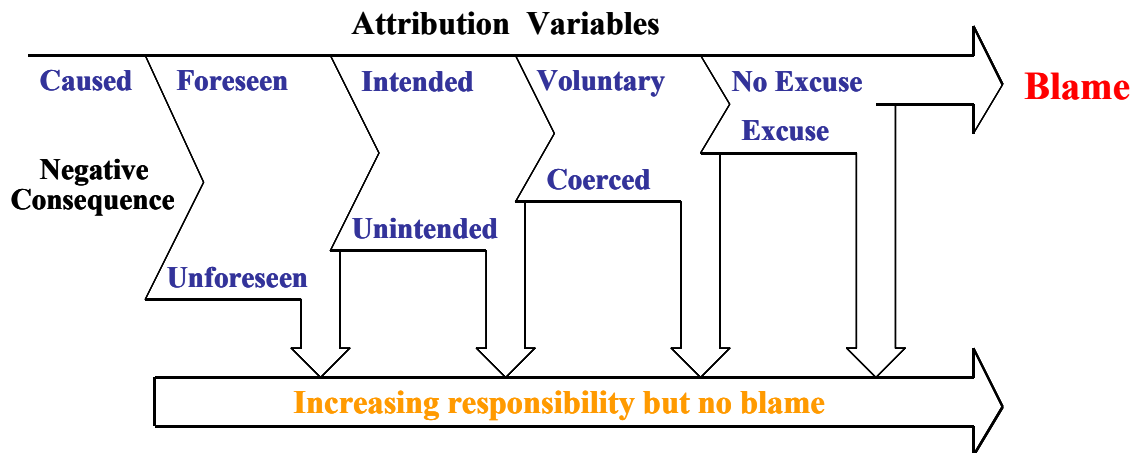
Figure 5: Process model of blame/credit attribution (adapted and simplified from Shaver)

We extend EMA by incorporating these mediating factors (foreseeability, coercion, etc.) into our assignment of causal attribution. The variables mediating blame in these models are readily derived by representations underlying appraisal and we show how planning and dialogue processing can inform and alter these assessments. Causality and intention map to our representations of action, beliefs, desires and intentions. Coercion requires a representation of social relationships and understanding of the extent to which it limits one's range of options. For example, one may be ordered to carry out a task but to satisfy the order, there may be alternatives that vary in blame or creditworthiness. In the remainder of this section, we describe this extension in detail.

4.1. Computational Representation

4.1.1 Actions and Consequences

EMA represents causal information through a hierarchical plan representation. Actions consist of a set of propositional preconditions and effects. Each action step is either a primitive action (i.e., an action that can be directly executed by some agent) or an abstract action. An abstract action may be decomposed hierarchically in multiple ways and each alternative consists of a sequence of primitive or abstract sub-actions. The desirability of action effects (i.e., effects having positive/negative significance to an agent) is represented by utility values (Blythe, 1999) and the likelihood of preconditions and effects is represented by probability values.

A *non-decision node* (or *And-node*) is an abstract action that can only be decomposed in one way. A *decision node* (or *Or-node*), on the other hand, can be decomposed in more than one way. In a decision node, an agent needs to make a decision and select among different options. If a decision node A can be decomposed in different ways $a_1, a_2, ..., a_n$, we will refer to $a_1, a_2, ..., a_n$ as *alternatives* of each other. Clearly, a primitive action is a non-decision node, while an abstract action can be either a non-decision node or a decision node.

Consequences or outcomes (we use the terms as exchangeable in this chapter) of actions are represented as a set of primitive action effects. The *consequence set* of an action A is defined recursively from leaf nodes (i.e., primitive actions) in plan structure to an action A as follows. Consequences of a primitive action are those effects with non-zero utility, and all the consequences of a primitive action are certain. For an abstract action, if the abstract action is a non-decision node, then the consequence set of the abstract action is the union of the consequences of its sub-actions. If the abstract action is a decision node, we need to differentiate two kinds of consequences. If a consequence p of a decision node occurs among all the alternatives, we call p a *certain consequence* of the decision node; otherwise p is an *uncertain consequence* of the node.

In addition, each action step is associated with a *performer* (i.e., the agent that performs the action) and an agent who has *authority* over its execution. The performer cannot execute the action until authorization is given by the authority. This represents the hierarchical organizational structure of social agents.

4.1.2 Attribution Variables

Weiner and Shaver define the attribution process in terms of a set of key variables:[2]

Causality refers to the connection between actions and the effects they produce. In our approach, causal knowledge is encoded via *hierarchical task representation*. Interde-

---

[2] Note that these models differ in terminology. Here we adopt the terminology of Shaver.

pendencies between actions are represented as a set of causal links and threat relations. Each causal link specifies that an effect of an action achieves a particular goal that is a precondition of another action. Threat relations specify that an effect of an action threatens a causal link by making the goal unachievable before it is needed.

Foreseeability refers to an agent's foreknowledge about actions and consequences. We use *know* and *bring-about* to represent foreseeability. If an agent knows an action brings about certain consequence before its execution, then the agent foresees the action brings about the consequence.

Intention is generally conceived as a commitment to work toward certain act or outcome. Intending an act (i.e., *act intention*) is distinguished from intending an outcome of an act (i.e., *outcome intention*) in that the former concerns actions while the latter concerns consequences of actions. Most theories argue that outcome intention rather than act intention is the key factor in determining accountability and intended outcome usually deserves more elevated accountability judgments (Weiner, 1986, 2001). We use *intend* with *do* to represent act intention and *intend* with *achieve* for outcome intention. Since our work is applied to rich social context, comparing with (Bratman, 1987; Grosz & Kraus, 1996), we include indirect intentions in our work. For example, an agent intends an action or a consequence, but may not be the actor himself/herself (i.e., by intending another agent to act or achieve the consequence), or an agent intends to act but is coerced to do so.

Similar difference exists in *coercion*. An agent may be coerced to act (i.e., *act coercion*) yet not be coerced to achieve any outcome of the action (i.e., *outcome coercion*), depending on whether the agent has choices in achieving different outcomes among alternatives. It is important to differentiate act coercion and outcome coercion, because it is the latter that actually influences our judgment of behavior, and is used to determine the *responsible agent*. We use *coerce* with *do* to represent act coercion and *coerce* with *achieve* for outcome coercion. In the case of outcome coercion, the responsible agent for a specific outcome is the performer or the authority of an action, but the action may not be the primitive one that directly leads to the outcome.

4.1.3 Representational Primitives

In modeling Shaver and Weiner's attribution theory, we need to map attribution variables into representational features of an agent's causal interpretation. Here we define a number of specific primitive features that support this mapping.

x and y are different agents. A and B are actions and p is a proposition. The following primitives are adopted in system.

(1) *and-node*(A): A is a non-decision node in plan structure.

(2) *or-node*(A): A is a decision node in plan structure.

(3) *alternative*(A, B): A and B are alternatives of performing the same higher-level action.

(4) *effect*(A): Effect set of a primitive action A.

(5) *certain-consq*(A): Certain consequence set of A.

(6) *uncertain-consq*(A): Uncertain consequence set of an abstract action A.

(7) *performer*(A): performing agent of A.

(8) *authority*(A): authorizing agent of A.

(9) *know*(x, p): x knows p.

(10) *intend*(x, p): x intends p.

(11) *coerce*(y, x, p): y coerce x the proposition p.

(12) *want*(x, p): x wants p.

(13) *by*(A, p): By acting A to achieve p.

(14) *bring-about*(A, p): A brings about p.

(15) *do*(x, A): x does A.

(16) *achieve*(x, p): x achieves p.

(17) *responsible*(p): Responsible agent for p.

(18) *superior*(y, x): y is a superior of x.

4.1.4 Axioms

We identify the interrelations of attribution variables, expressed as *axioms*. The axioms are used either explicitly as *commonsense* inference rules for deriving key attribution values, or implicitly to keep the consistency between different inference rules.

x and y are different agents. A is an action and p is a proposition. The following *axioms* hold from a rational agent's perspective (To simplify the logical expressions, we omit the universal quantifiers in this chapter, and substitute A for $do(*, A)$ and p for $achieve(*, p)$ here).

(1) $\exists y(coerce(y, x, A)) \Rightarrow intend(x, A)$

(2) $intend(x, A) \wedge \neg(\exists y(coerce(y, x, A))) \Rightarrow \exists p(p \in certain\text{-}consq(A) \wedge intend(x, p))$

(3) $intend(x, p) \Rightarrow \exists A(p \in certain\text{-}consq(A) \wedge intend(x, A))$

(4) $intend(x, by(A, p)) \Rightarrow know(x, bring\text{-}about(A, p))$

The *first* axiom shows that act coercion entails act intention. It means if an agent is coerced an action A by another agent, then the coerced agent intends A[3]. The second and the third axioms show the relations between act intention and outcome intention. The *second* one means if an agent intends an action A and the agent is not coerced to do so (i.e. A is a voluntary act), then the same agent must intend at least one consequence of A. The *third* means if an agent intends a consequence p, the same agent must intend at least one action that has p as a consequence[4]. Note that in both axioms, intending an action or a consequence includes the case that an agent intends another agent to act or achieve the consequence. The *last* one shows the relation between intention and foreseeability. It means if an agent intends acting A to achieve a consequence p, the same agent must know that A brings about p.

---

[3] The notion of intention in this axiom is not identical to the typical implication of intention in literatures, as here it is applied to coercive situations.

[4] This axiom is not true in general cases, as the agent may not know that an action brings about p. Here we apply it within the restrictive context of after action evaluation, where actions have been executed and the consequence has occurred.

4.1.5 Attribution Rules

Social credit assignment focuses on consequences with personal significance to an agent. This evaluation is always from the perspective of a perceiving agent and based on the attribution values acquired by the individual perceiver. As different perceivers have different preferences, different observations, and different knowledge and beliefs, it may well be the case that for the same situation, different perceivers form different judgments.

Nevertheless, the attribution process and rules are general, and applied uniformly to different perceivers. Following Weiner (Weiner, 2001), we use *coercion* to determine the responsible agent for credit or blameworthiness, and *intention* and *foreseeability* in assigning the intensity of credit/blame.

If an action performed by an agent brings about *positive/negative* consequence, and the agent is *not coerced* to achieve the consequence, then *credit/blame* is assigned to the *performer* of the action. Otherwise, assign credit/blame to the *authority*. If the authority is also coerced, the process needs to be traced further to find the *responsible agent* for the consequence. The *back-tracing algorithm* for finding the responsible agent will be given later.

Rule 1: If <*consequence*> of <*action*> is *positive*/*negative* and
       <*performer*> is *not coerced* the <*consequence*>
    Then Assign *credit*/*blame* to the <*performer*>

Rule 2: If <*consequence*> of <*action*> is *positive*/*negative* and
       <*performer*> is *coerced* the <*consequence*>
    Then Assign *credit*/*blame* to the <*responsible* agent>

We adopt a simple categorical model of intensity assignment, though one could readily extend the model to a numeric value by incorporating probabilistic rules of inference. If the responsible agent intends the consequence while acting, the intensity assigned is *high*. If the responsible agent does not foresee the consequence, the intensity is *low*.

4.2 Commonsense Inference

Judgments of causality, foreseeability, intentionality and coercion are informed by dialogue and causal evidence. Some theories have formally addressed subsets of this judgment task. For example, (Sadek, 1990) addresses the relationship between dialogue and inferences of belief and intention. These theories have not tended to consider coercion. Rather than trying to synthesize and extend such theories, we introduce small number of commonsense rules that, via a justification-based truth maintenance system (*JTMS*), allow agents to make inferences based on this evidence.

4.2.1 Dialogue Inference

Conversational dialogue between agents is a rich source of information for deriving values of attribution variables. In a conversational dialogue, a *speaker* and a *hearer* take turns alternatively. When a *speech act* (Austin, 1962; Searle, 1969, 1979) is performed, a perceiving agent (who can be one of the participating agents or another agent) makes inferences based on observed conversation and current beliefs. As the conversation proceeds, beliefs are formed and updated accordingly.

Assume conversations between agents are *grounded* (Traum & Allen, 1994) and they conform to Grice's maxims of *Quality*[5] and *Relevance*[6] (Grice, 1975). Background information (agents' social roles, relationship, etc) is also important, for example, an order can be successfully issued only to a subordinate, but a request can be made of any agent.

x and y are different agents. p and q are propositions and t is time. For our purpose, we analyze following speech acts that help infer agents' desires, intentions, foreknowledge and choices in acting.

(1)   *inform*(x, y, p, t): x informs y that p at t.

(2)   *request*(x, y, p, t): x requests y that p at t.

---

[5] The Quality maxim states that one ought to provide true information in conversation.

[6] The Relevance maxim states that one's contribution to conversation ought to be pertinent in context.

(3) *order*(*x*, *y*, *p*, *t*): *x* orders *y* that *p* at *t*.

(4) *accept*(*x*, *p*, *t*): *x* accepts *p* at *t*.

(5) *reject*(*x*, *p*, *t*): *x* rejects *p* at *t*.

(6) *counter-propose*(*x*, *p*, *q*, *t*): *x* counters *p* and proposes *q* at *t*.

We have designed commonsense rules that allow perceiving agents to infer from dialogue patterns. These rules are general. Hence, they can be combined flexibly and applied to variable-length dialogue sequences with multiple participants.

Let *z* be a perceiving agent. If at time *t1*, a speaker (*s*) *informs* a hearer (*h*) that *p*, then after *t1*, a perceiving agent can infer that both the speaker and the hearer know that *p* as long as there is no intervening contradictory belief.

Rule 3: inform(*s*, *h*, *p*, *t1*) ∧ *t1*<*t3* ∧ ¬(∃*t2*)(*t1*<*t2*<*t3* ∧ believe(*z*, ¬know(*s*, *p*)∨¬know(*h*, *p*), *t2*)) ⇒ believe(*z*, know(*s*, *p*)∧know(*h*, *p*), *t3*)

A *request* gives evidence of the speaker's *desire* (or *want*). An order gives evidence of the speaker's *intend*.

Rule 4: request(*s*, *p*, *t1*) ∧ *t1*<*t3* ∧ ¬(∃*t2*)(*t1*<*t2*<*t3* ∧ believe(*z*, ¬want(*s*, *p*), *t2*) ⇒ believe(*z*, want(*s*, *p*), *t3*)

Rule 5: order(*s*, *p*, *t1*) ∧ *t1*<*t3* ∧ ¬(∃*t2*)(*t1*<*t2*<*t3* ∧ believe(*z*, ¬intend(*s*, *p*), *t2*)) ⇒ believe(*z*, intend(*s*, *p*), *t3*)

The hearer may *accept, reject* or *counter-propose*. If the speaker wants (or intends) and the hearer *accepts*, it can be inferred that the hearer intends. An agent can accept via speech or action execution. If the hearer accepts what the superior wants (or intends), there is evidence of coercion.

Rule 6: believe(z, want/intend(s, p), t1) ∧ accept(h, p, t2) ∧ ¬superior(s, h) ∧ t1<t2<t4 ∧

¬(∃t3)(t2<t3<t4 ∧ believe(z, ¬intend(h, p), t3)) ⇒ believe(z, intend(h, p), t4)

Rule 7: believe(z, want/intend(s, p), t1) ∧ accept(h, p, t2) ∧ superior(s, h) ∧ t1<t2<t4 ∧

¬(∃t3)(t2<t3<t4 ∧ believe(z, ¬coerce(s, h, p), t3)) ⇒ believe(z, coerce(s, h, p), t4)

In the rules above, if act coercion is true, act intention can be deduced from *Axiom 1*.

If the speaker wants (or intends) and the hearer *reject*s, infer that the hearer does not intend.

Rule 8: believe(z, want/intend(s, p), t1) ∧ reject(h, p, t2) ∧ t1<t2<t4 ∧ ¬(∃t3)(t2<t3<t4 ∧

believe(z, intend(h, p), t3)) ⇒ believe(z, ¬intend(h, p), t4)

If the hearer *counters* acting *A* and *proposes* acting *B* instead, both the speaker and the
hearer are believed to know that *A* and *B* are alternatives. It is also believed that the hearer
does not want *A* and wants *B* instead.

Rule 9: counter-propose(h, do(h, A), do(h, B), t1) ∧ t1<t3 ∧ ¬(∃t2)(t1<t2<t3 ∧ believe(z,

¬know(h, alternative(A, B))∨¬know(s, alternative(A, B)), t2)) ⇒ believe(z,

know(h, alternative(A, B))∧know(s, alternative(A, B)), t3)

Rule 10: counter-propose(h, p, q, t1) ∧ t1<t3 ∧ ¬(∃t2)(t1<t2<t3 ∧ (believe(z, want(h,

p)∨¬want(h, q), t2))) ⇒ believe(z, ¬want(h, p)∧want(h, q), t3)

If the speaker has *known* that two actions are *alternatives* and still *requests* (or *orders*)
one of them, infer that the speaker wants (or intends) the chosen action instead of the al-
ternative. The beliefs that the speaker wants (or intends) the chosen action can be de-
duced from *Rules 4&5*.

Rule 11: believe(z, know(s, alternative(A, B)), $t1$) $\wedge$ request/order(s, do(h, A), $t2$) $\wedge$ $t1<t2<t4$

$\wedge$ $\neg(\exists t3)(t2<t3<t4$ $\wedge$ believe(z, want(s, do(h, B)), $t3$)) $\Rightarrow$ believe(z, $\neg$want/intend(s, do(h, B)), $t4$)

4.2.2 Causal Inference

Causal knowledge encoded in plan representation also helps derive values of attribution variables. Different agent may have access to different plans in memory. While plans are specific to certain domain, the structure and features of plans can be described using domain-independent terms such as action types, alternatives and action effects. We adopt the hierarchical task formalism that differentiates action types, explicitly expresses consequences of alternatives, and separates certain consequences of an action from its uncertain ones.

An agent's *foreknowledge* can be derived simply by checking primitive action effects. If a consequence p is an effect of a primitive action A, then the agents involved (i.e., the performer and the authority) should know that A brings about p.

Rule 12: p$\in$effect(A) $\Rightarrow$ believe(z, know(performer(A), bring-about(A, p)))

p$\in$effect(A) $\Rightarrow$ believe(z, know(authority(A), bring-about(A, p)))

Outcome intent can be partially inferred from evidence of act intent and comparative features of consequence sets of action alternatives. According to *Axiom 2*, if an agent intends a voluntary action A, the agent must intend at least one consequence of A. If A has only one consequence p, then the agent is believed to intend p. In more general cases, when an action has multiple consequences, in order to identify whether a specific outcome is intended or not, a perceiver may examine *alternatives* the agent intends and does not intend, and compare the consequences of intended and unintended alternatives.

If an agent intends an action A voluntarily and does intend its alternative B, we can infer that the agent either intends (at least) one consequence that only occurs in A or does not

intend (at least) one consequence that only occurs in *B*, or both. If the consequence set of *A* is a subset of that of *B*, the rule can be simplified. As there is no consequence of *A* not occurring in the consequence set of *B*, we can infer that the agent does not intend (at least) one consequence that only occurs in *B*. In particular, if there is only one consequence *p* of *B* that does not occur in the consequence set of *A*, infer that the agent does not intend *p*.

Rule 13: believe(z, intend(x, A)∧¬intend(x, B)∧¬(∃y(superior(y, x)∧coerce(y, x, A)))) ∧ alternative(A, B) ∧ certain-consq(A)⊂certain-consq(B) ⇒ ∃p(p∉certain-consq(A) ∧ p∈certain-consq(B) ∧ believe(z, ¬intend(x, p)))

On the other hand, given the same context that an agent intends an action *A* and does not intend its alternative *B*, if the consequence set of *B* is a subset of that of *A*, infer that the agent intends (at least) one consequence that only occurs in *A*. In particular, if there is only one consequence *p* of *A* that does not occur in the consequence set of *B*, the agent must intend *p*.

Rule 14: believe(z, intend(x, A)∧¬intend(x, B)∧¬(∃y(superior(y, x)∧coerce(y, x, A)))) ∧ alternative(A, B) ∧ certain-consq(B)⊂certain-consq(A) ⇒ ∃p(p∈certain-consq(A) ∧ p∉certain-consq(B) ∧ believe(z, intend(x, p)))

Outcome coercion can be properly inferred from evidence of act coercion and consequence sets of different action types. In a non-decision node (i.e., *and-node*), if an agent is coerced to act, the agent is also coerced to achieve the consequences of subsequent actions, for the agent has no other choice.

Rule 15: ∃y(superior(y, x) ∧ believe(z, coerce(y, x, A)) ∧ and-node(A) ∧ p∈certain-consq(A) ⇒ believe(z, coerce(y, x, p)))

In a decision node (i.e., *or-node*), however, an agent must make a decision amongst multiple choices. Even if an agent is coerced to act, it does not follow that the agent is co-

erced to achieve a specific consequence of subsequent actions. In order to infer outcome coercion, we examine the choices at a decision node. If an outcome is a certain consequence of every alternative, then it is unavoidable and thus outcome coercion is true. Otherwise, if an outcome is an uncertain consequence of the alternatives, then the agent has the option to choose an alternative to avoid this outcome and thus outcome coercion is false. Our definition of consequence set ensures the consistency when the rules are applied to actions at different levels of plan structure.

Rule 16: $\exists y(\text{superior}(y, x) \wedge \text{believe}(z, \text{coerce}(y, x, A)) \wedge \text{or-node}(A) \wedge p \in \text{certain-consq}(A)$

   $\Rightarrow \text{believe}(z, \text{coerce}(y, x, p)))$

   $\exists y(\text{superior}(y, x) \wedge \text{believe}(z, \text{coerce}(y, x, A)) \wedge \text{or-node}(A) \wedge p \in \text{uncertain-consq}(A) \Rightarrow \text{believe}(z, \neg\text{coerce}(y, x, p)))$

4.3. Back-Tracing Algorithm

We have developed a *back-tracing algorithm* for evaluating the responsible agent for a specific consequence. The evaluation process starts from the primitive action that directly causes a consequence with positive or negative utility. Since coercion may occur in more than one level in hierarchical plan structure, the process must trace from the primitive action to the higher-level actions. We use a back-tracing algorithm to find the responsible agent. The algorithm takes as input some desirable or undesirable consequence of a primitive action (*step 1*) and works up the task hierarchy[7]. During each pass through the main loop (*step 2*), the algorithm initially assigns default values to the variables (*step 2.2*). Then apply dialog rules to infer variable values at the current level (*step 2.3*). If there is evidence that the performer was coerced to act (*step 2.4*), the algorithm proceeds by applying plan inference rules (*step 2.5*). If there is outcome coercion (*step 2.6*), the authority is deemed responsible (*step 2.7*). If current action is not the root node in plan structure and outcome coercion is true, the algorithm enters next loop and evaluates the next level up in the task hierarchy.

---

[7] Given that the evaluating agent is aware of the task hierarchy.

After the execution of the algorithm, the responsible agent for the outcome is determined. Meanwhile, through applying inference rules, the algorithm also acquires values of intention and foreknowledge about the agents. The variable values are then used by the attribution rules (*Rules 1&2*) to assign credit or blame to the responsible agent with proper intensity.

Events may lead to more than one desirable/undesirable consequence. For evaluating multiple consequences, we can apply the algorithm the same way, focusing on one consequence each time during its execution. Then, to form an overall judgment, the results can be aggregated and grouped by the responsible agents.

Backtrace (*consq*, *plan structure*):

1. *parent* = *A*, where *consq* is an effect of action *A*

2. DO

    2.1   *node* = *parent*

    2.2   coerce(authority(*node*), performer(*node*), *node*) = *unknown*

            coerce(authority(*node*), performer(*node*), *consq*) = *unknown*

            responsible(*consq*) = performer(*node*)

    2.3   Search dialog history on *node* and apply *dialog inference rules*

    2.4   IF coerce(authority(*node*), performer(*node*), *node*) THEN

    2.5   apply *plan inference rules* on *node*

    2.6   IF coerce(authority(*node*), performer(*node*), *consq*) THEN

    2.7   responsible(*consq*) = authority(*node*)

4.4.   Illustrative Example

The need to extend EMA was motivated by a number of odd social attributions generated by agents in the Mission Rehearsal Exercise (*MRE*) leadership training system (Rickel et al., 2002), to which EMA was applied. By extending EMA with a more realistic social attribution process, we eliminated the obvious departures of the model from normal human behavior. Here we illustrate how the model operates on one of these previous defects. The example arises from the following extract of dialogue taken from an actual run
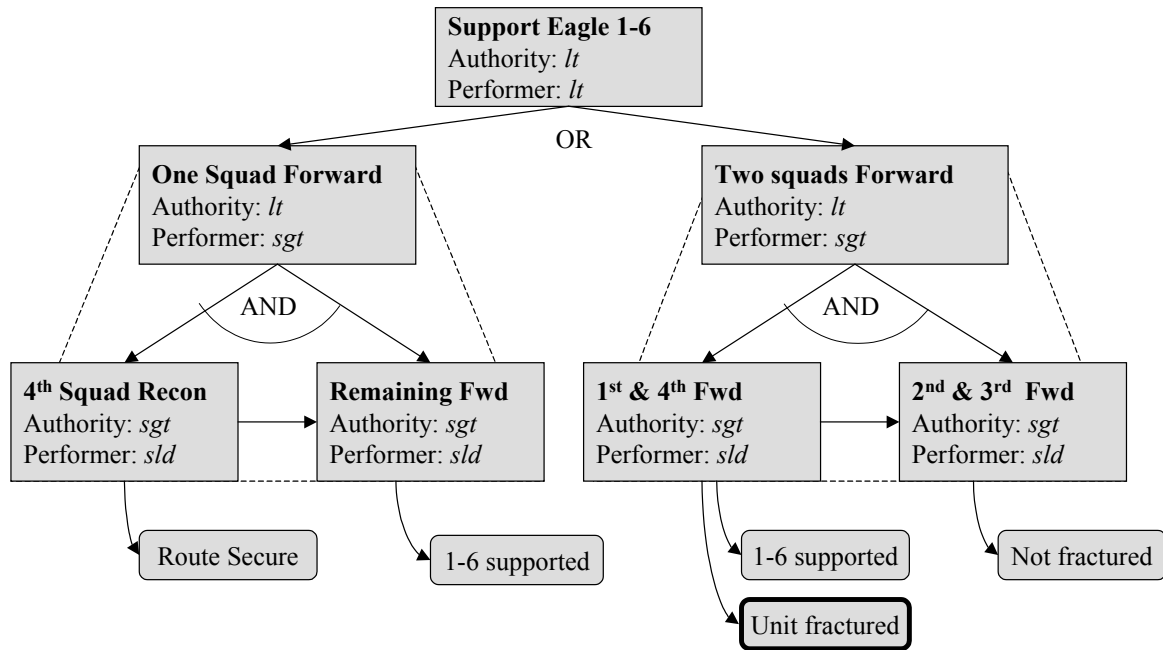
Figure 6: Team plan from the sergeant's perspective

of the system. Details on how this negotiation is automatically generated and how natural language is mapped into speech acts can be found in (Traum, Rickel, Gratch, & Marsella, 2003):

Student:	Sergeant. Send two squads forward.
Sergeant:	That is a bad idea, sir. We shouldn't split our forces. Instead we should send one squad to recon forward.
Student:	Send two squads forward.
Sergeant:	Against my recommendation, sir. Lopez! Send first and fourth squads to Eagle 1-6's location.
Lopez:	Yes, sir. Squads! Mount up!

We focus on three social actors, the *student*, the *sergeant* and the *squad leader* (Lopez), who act as a team in this example. The student is a human trainee and acts as an authority over the sergeant. The squad leader acts as a subordinate of the sergeant. Conversations between agents are represented within the system as speech acts and a dialogue history as in the *MRE*. Figure 6 illustrates the causal knowledge underlying the example.

Take the sergeant's perspective as an example. The sergeant perceives the conversation between the actors and task execution. Dialogue history includes the following acts, ordered by the time the speakers addressed them (*std, sgt* and *sld* stand for the student, the sergeant and the squad leader, respectively. *t1<t2<…<t6*).

(1) order(*std*, do(*sgt, two-sqds-fwd*), *t1*)

(2) inform(*sgt, std*, bring-about(*two-sqds-fwd, unit-fractured*), *t2*)

(3) counter-propose(*sgt*, do(*sgt, two-sqds-fwd*), do(*sgt, one-sqd-fwd*), *t3*)

(4) order(*std*, do(*sgt, two-sqds-fwd*), *t4*)

(5) accept(*sgt*, do(*sgt, two-sqds-fwd*), *t5*)

(6) order(*sgt*, do(*sld, 1st-and-4th-to-celic*), *t6*)

To simplify the example, we illustrate part of the task structure from *MRE* scenario and evaluate one of the *negative* consequences, though we can generally apply the approach in the chapter to more complex judgments. The sergeant has access to a partial plan, where *one squad forward* and *two squads forward* are two choices of action *support eagle-1-6*. *One squad forward* is composed of two primitive actions, *4th squad (recon) forward* and *remaining (squads) forward*. *Two squads forward* consists of *1st and 4th (squads) to celic* and *2nd and 3rd (squads) to celic*. Two action effects are salient to the sergeant, (*eagle*) *1-6 supported* and *unit fractured*. *1-6 supported* is a desirable team goal. Assume the sergeant assigns negative utility to *unit fractured* and this consequence serves as input to the back-tracing algorithm. We illustrate how to find the blameworthy agent given the sergeant's task knowledge and observations.

Loop 1: The algorithm starts from primitive action *1st-and-4th-to-celic*, of which *unit-fractured* is an effect. The sergeant perceived that the *squad leader* performed the action.

Step 2.2: Initially, coerce(*sgt, sld, 1st-and-4th-to-celic*) and coerce(*sgt, sld, unit-fractured*) are unknown. Assign the *squad leader* to the responsible agent.

Step 2.3: Relevant dialogue history is *act 6*. Since the sergeant *ordered* the squad leader the act, apply *Rule 5*. The algorithm infers that the sergeant believes he *intended* the squad leader to act. Since the squad leader *accepted* by executing the action and the ser-

geant is the *superior*, apply *Rule 7*. The sergeant believes that he *coerced* the squad leader to act.

Step 2.4–2.5: Since coerce(*sgt, sld, 1ˢᵗ-and-4ᵗʰ-to-celic*) is true and the primitive action is an *and-node* in the plan structure, apply *Rule 15* . The sergeant believes he coerced the squad leader to fracture the unit. Since *unit-fractured* is an effect of the primitive action, apply *Rule 12*. The sergeant believes that both he and the squad leader *knew* the action bringing about *unit-fractured*.

Step 2.6–2.7: Since coerce(*sgt, sld, unit-fractured*) is true, assign the sergeant to the responsible agent. The *sergeant* believes that he is responsible for *unit-fractured* and he has the *foreknowledge* while acting.

Since parent node is *not* the *root* of plan structure and outcome coercion is *true*, the algorithm enters next loop.

Loop 2: The action is *two-sqds-fwd*, performed by the *sergeant*. Relevant dialogue history is *sequence 1–5*. A variety of beliefs can be inferred from commonsense rules by analyzing the task structure and conversation history. The results are given below.

(1) believe(*sgt*, intend(*std*, do(*sgt, two-sqds-fwd*)))   (*act 1* or *4, rule 5*)

(2) believe(*sgt*, know(*sgt*, bring-about(*two-sqds-fwd, unit-fractured*)))   (*act 2, rule 3*)

(3) believe(*sgt*, know(*std*, bring-about(*two-sqds-fwd, unit-fractured*)))   (*act 2, rule 3*)

(4) believe(*sgt*, know(*sgt*, alternative(*one-sqd-fwd, two-sqds-fwd*)))   (*act 3, rule 9*)

(5) believe(*sgt*, know(*std*, alternative(*one-sqd-fwd, two-sqds-fwd*)))   (*act 3, rule 9*)

(6) believe(*sgt*, ¬want(*sgt*, do(*sgt, two-sqds-fwd*)))   (*act 3, rule 10*)

(7) believe(*sgt*, want(*sgt*, do(*sgt, one-sqd-fwd*)))   (*act 3, rule 10*)

(8) believe(*sgt*, ¬intend(*std*, do(*sgt, one-sqd-fwd*)))   (*act 4, result 5, rule 11*)

(9) believe(*sgt*, coerce(*std, sgt, two-sqds-fwd*))   (*act 5, result 1, rule 7*)

(10) believe(*sgt*, coerce(*std, sgt, unit-fractured*))   (*act 5, result 9, rule 15*)

After *loop 2*, the sergeant believes the student coerced him to fracture the unit (*Result 10*). So the *student* is responsible for the outcome.

Loop 3: The action is *support-eagle-1-6,* performed by the student. There is no relevant dialogue in history. The initial values and the responsible agent are as default. There is no clear evidence of coercion, so the sergeant believes that the *student* is the responsible agent. Parent node is the *root* of plan. The algorithm terminates.

Now the sergeant also believes that the student intended to send two squads forward and did not intend to send one squad forward (*Results 1&8*). Since the consequence set of *one-sqd-fwd* (i.e., *1-6-supported*) is subset of that of *two-sqds-fwd* (i.e., *1-6-supported* and *unit-fractured*), apply *rule 14*. The sergeant believes that the student intended *unit-fractured* and foresaw the outcome (*Result 3*), so the *student* is to blame for *unit-fractured* with *high* intensity.

4.5. Discussion

By incorporating this richer model of causal attribution into EMA, the system now gives reasonable inferences on situations that arise in our current MRE application. As the work moves forward, several issues need further attention. We must incorporate probabilistic reasoning to deal with uncertainty in observations and judgment process. For modeling more complex multi-agent teamwork, we need to consider joint responsibility and sharing responsibility among teammates (the current model assumes one agent has sole responsibility) and less hierarchical relationships between social actors. Some inference rules are too restrictive and need to make better use of plan knowledge, particularly considering how preconditions and effects indirectly limit one's choices in acting. As our task representation has already encoded information about action preconditions and effects, this should be a natural extension of our existing methods.

A critical issue is formal evaluation. Although the work is based on psychological theory and seems to provide reasonable responses in practice, we would like to more systematically assess the veracity of the approach. This is a challenge given that social attributions are more variable than many phenomena studied by cognitive science, differing widely

both within and across individuals depending on non-observable factors like goals, beliefs, cultural norms, etc. And unlike work in decision making, there is no accepted normative model of such attributions or their dynamics that we can use as a gold standard for evaluating techniques. We would like to build on the "situational psychology" methodology we have used in evaluating the basic model (Gratch & Marsella, 2004a). Under this methodology, people are presented with a description of an evolving situation and queried as to their feelings and interpretations during several intermediate stages of the episode. For example, a subject is asked to imagine themselves in a stereotypical situation, such as an argument with their boss. They are asked how they would respond emotionally, how they appraise aspects of the situation and how they would cope. They are then given subsequent updates on the situation and asked how their emotions/coping would dynamically unfold in light of systematic variations in both expectations and perceived sense of control. Based on their evolving pattern of responses, subjects are scored as to how closely their reactions correspond to those of typical healthy adults. In our evaluation, we encode these evolving situations in EMA's domain language, run the scenarios, and compare EMA's appraisals and coping strategies to the responses indicated by the scale. In using this methodology to assess the extensions related to social attribution, we must identify or create a corpus of situations involving social attributions and compare the results of the mode against human data.

5. Evaluation

Given the broad influence emotions have over behavior, evaluating the effectiveness of such a general architecture presents some unique challenges. Emotional influences are manifested across a variety of levels and modalities. For instance, there are telltale physical signals: facial expressions, body language, and certain acoustic features of speech. There are also influences on cognitive processes, including coping behaviors such as wishful thinking, resignation, or blame-shifting. Unlike many phenomena studied by cognitive science, emotional responses are also highly variable, differing widely both within and across individuals depending on non-observable factors like goals, beliefs, cultural norms, etc. And unlike work in rational decision making, there is no accepted,

idealized model of emotional responses or their dynamics that we can use as a gold standard for evaluating techniques.

In evaluating our model, we adopt a multi-pronged approach, identifying certain specific functions that emotions play in humans and assessing the extent that the model reproduces those functions. Here we briefly summarize two recent evaluation studies, each illustrating this multi-pronged approach. In the first study, we address the question of process dynamics: does the model generate cognitive influences that are consistent with human data on the influences of emotion, specifically with regard to how emotion shapes perceptions and coping strategies, and how emotion and coping unfold over time. In the second, we address the question of behavioral influence: do external behaviors have the

Controllability
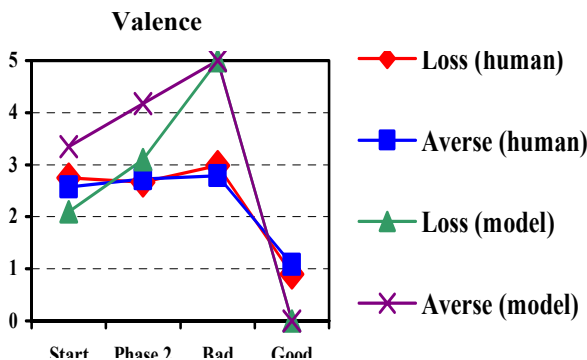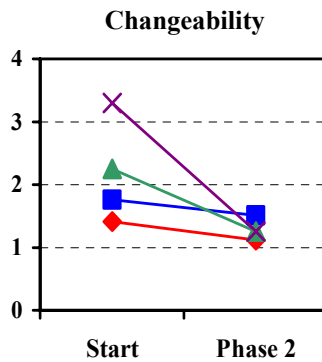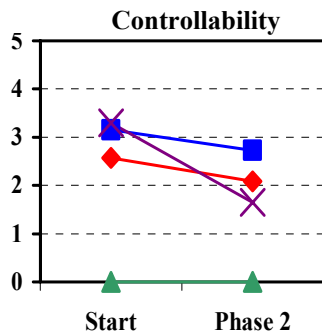
Coping Strategies

	Aversive	Loss
Phase 1	Seek information Take action	Suppress information Procrastinate Seek instrumental support
Phase 2	Mental disengagement Suppress information	Mental disengagement Suppress information Resignation Wishful thinking
Good	Accept responsibility	
Bad	Mental disengagement Suppress information	Mental disengagement Suppress information

Changeability

Valence

Loss (human)
Averse (human)
Loss (model)
Averse (model)

Figure 7. Some results from the emotion process evaluation. The experiment compares human and model responses to two emotion evoking scenarios ("aversive" and "loss"). Each scenario evolves over three phases, ending in either a good or bad outcome and subjects are queried as to their emotional state, appraisals and coping strategies after each phase. The model fits the basic trends of human subjects, though differs in specific ratings.

same social influence on a human subject that one person's emotion has on another person, specifically with regard to how emotional displays influence third-party judgments.

In the first study, we fit our model to a standard instrument used in the clinical psychological evaluation of a person's emotional and coping response to stressful situations, and in particular, how these responses evolve over time. In the Stress and Coping Process Questionnaire (Perrez & Reicherts, 1992), a subject is presented a stereotypical situation, such as an argument with their boss. They are asked how they would respond emotionally and how they would cope. They are then given subsequent updates on the situation and asked how their emotions/coping would dynamically unfold in light of systematic variations in both expectations and perceived sense of control. Based on their evolving pattern of responses, subjects are scored as to how closely their reactions correspond to those of normal healthy adults. In our evaluation, we encode these evolving situations in EMA's domain language, run the scenarios, and compare EMA's appraisals and coping strategies to the responses indicated by the scale. Figure 7 illustrates the basic results. The model matches the basic trends of normal human subjects, though differs in some particulars. See (Gratch & Marsella, 2004b) for details.
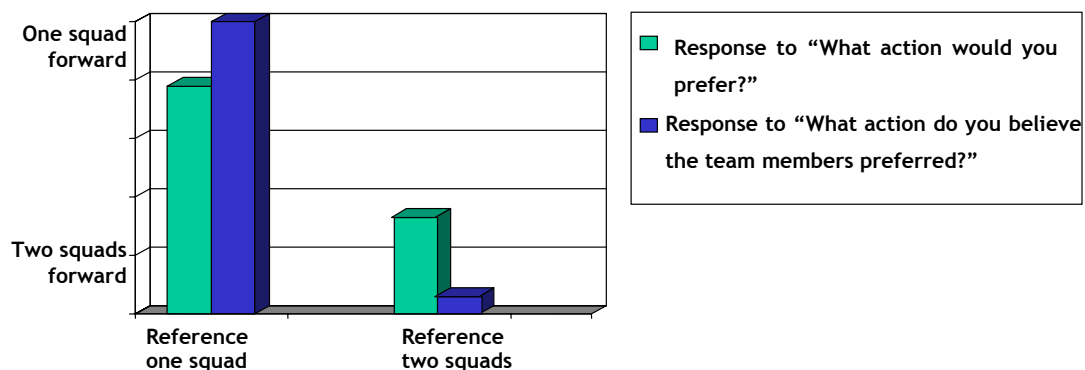

Figure 8: Study illustrates that the emotional displays of virtual characters can influence the decision making of human subjects. Consistent with the phenomenon of social referencing, when presented with an ambiguous decision, subjects inferred how bystanders appraised the situation through their emotional displays and factored this information into their decision.

For evaluating the social impact of our model, we are initially focusing on the phenomena of social referencing, whereby people, when presented with an ambiguous decision,

are influenced by appraisals of others (Campos, 1983). In our evaluation, we assess the ability of synthetic emotion displays to induce social referencing in human subjects in the context of the Mission Rehearsal Exercise. Subjects observe the disagreement described in Section 4.4 and are asked to indicate which course of action is better (sending two squads forward or sending one squad). As subjects have no military background, the correct action is ambiguous. Across two experimental conditions, we vary the emotional displays of the virtual team members that will ultimately have to carry out the order: in the "reference two squads" condition, the team members uniformly exhibit positive emotional displays when "two squads forward" is proposed and negative displays when "one squad forward" is proposed; vice versa for the "reference one squad" condition. The hypothesis is that human subjects both recognize that these displays indicate a preference and will be influenced to adopt a decision that is consistent with this preference. The results, Figure 8, support this hypothesis. See (Gratch & Marsella, 2004a) for more details.

Together, the results lend support to both the fidelity and social impact of the basic model. The extensions described in Section 4 have yet to be formally evaluated. The basic structure of this study will follow the basic structure of first study, though with material drawn from empirical studies of attribution theory.

6. Summary

EMA provides a general and comprehensive model of the processes underlying cognitive appraisal. In particular, we feel it is the first process model that explains how the appraisal of an event can change over time (by tying appraisal to an interpretation that can change with further inference) and is the first comprehensive attempt to model the range of human coping strategies. It is also one of the most comprehensive integrations of an appraisal model with other reasoning capabilities including planning, natural language processing, and non-verbal behavior. This chapter significantly extends the model's ability to reason about multi-agent situations by providing a cognitively plausible model of social blame and credit assignment based on social attribution theory.

Acknowledgements

References

Ambros-Ingerson, J., & Steel, S. (1988). *Integrating Planning, Execution and Monitoring.* Paper presented at the Seventh National Conference on Artificial Intelligence, St. Paul, MN.

Anderson, J. R., & Lebiere, C. (2003). The Newell Test for a theory of cognition. *Behavioral and Brain Sciences, 26*, 587-640.

André, E., Rist, T., Mulken, S. v., & Klesen, M. (2000). The Automated Design of Believable Dialogues for Animated Presentation Teams. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 220-255,). Cambridge, MA: MIT Press.

Austin, J. (1962). *How to Do Things with Words*: Harvard University Press.

Blythe, J. (1999, Summer). Decision Theoretic Planning. *AI Magazine, 20(2),* 37-54.

Bratman, M. (1987). *Intention, Plans and Practical Reason*: Harvard University Press.

Bratman, M. (1990). What is intention? In P. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in Communication*. Cambridge, MA: MIT Press.

Campos, J. J. (1983). The importance of affective communication in social referencing: a commentary on Feinman. *Merrill-Palmer Quarterly, 29*, 83-87.

Cassell, J., Bickmore, T., Campbell, L., Vilhjálmsson, H., & Yan, H. (2000). Human conversation as a system framework: Designing embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational Agents* (pp. 29-63). Boston: MIT Press.

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Avon Books.

Davidson, R. J., Scherer, K., & Goldsmith, H. H. (Eds.). (2003). *Handbook of Affective Sciences*. New York: Oxford University Press.

El Nasr, M. S., Yen, J., & Ioerger, T. (2000). FLAME: Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems, 3*(3), 219-257.

Elliott, C. (1992). *The affective reasoner: A process model of emotions in a multi-agent system* (Ph.D Dissertation No. 32). Northwestern, IL: Northwestern University Institute for the Learning Sciences.

Frank, R. (1988). *Passions with reason: the strategic role of the emotions*. New York, NY: W. W. Norton.

Gratch, J. (2000). *Émile: marshalling passions in training and education.* Paper presented at the Fourth International Conference on Intelligent Agents, Barcelona, Spain.

Gratch, J., & Marsella, S. (2001). *Tears and Fears: Modeling Emotions and Emotional Behaviors in Synthetic Agents.* Paper presented at the Fifth International Conference on Autonomous Agents, Montreal, Canada.

Gratch, J., & Marsella, S. (2003). Fight the way you train: the role and limits of emotions in training for combat. *Brown Journal of World Affairs, X(1)*(Summer/Fall).

Gratch, J., & Marsella, S. (2004a). *Evaluating a General Model of Emotional Appraisal and Coping.* Paper presented at the AAAI Symposium on Architectures for modeling emotion: cross-disciplinary foundations, Palo Alto, CA.

Gratch, J., & Marsella, S. (2004b). *Evaluating the modeling and use of emotion in virtual humans.* Paper presented at the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems, New York.

Grice, H. P. (1975). Logic and Conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3): Academic Press.

Grosz, B., & Kraus, S. (1996). Collaborative Plans for Complex Group Action. *Artificial Intelligence, 86*(2).

Lazarus, R. (1991). *Emotion and Adaptation*. NY: Oxford University Press.

Lester, J. C., Stone, B. A., & Stelling, G. D. (1999). Lifelike Pedagogical Agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments. *User Modeling and User-Adapted Instruction, 9*(1-2), 1-44.

Marsella, S., & Gratch, J. (2002). *A Step Toward Irrationality: Using Emotion to Change Belief.* Paper presented at the First International Joint Conference on Autonomous Agents and Multiagent Systems, Bologna, Italy.

Marsella, S., & Gratch, J. (2003). *Modeling coping behaviors in virtual humans: Don't worry, be happy.* Paper presented at the Second International Joint Conference on Autonomous Agents and Multi-agent Systems, Melbourne, Australia.

Marsella, S., Johnson, W. L., & LaBore, C. (2000). *Interactive Pedagogical Drama.* Paper presented at the Fourth International Conference on Autonomous Agents, Montreal, Canada.

Marsella, S., Johnson, W. L., & LaBore, C. (2003). *Interactive pedagogical drama for health interventions.* Paper presented at the Conference on Artificial Intelligence in Education, Sydney, Australia.

McCarty, L. (1997). *Some Arguments about Legal Arguments.* Paper presented at the 6th International Conference on Artificial Intelligence and Law, Melbourne, Australia.

Mele, A. R. (2001). *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.

Minsky, M. (1986). *The Society of Mind*. New York: Simon and Schuster.

Moffat, D., & Frijda, N. (1995). *Where there's a Will there's an agent.* Paper presented at the Workshop on Agent Theories, Architectures and Languages.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: Harvard University Press.

Oatley, K., & Johnson-Laird, P. N. (1987). Cognitive Theory of Emotions. *Cognition and Emotion, 1*(1).

Ortony, A., Clore, G., & Collins, A. (1988). *The Cognitive Structure of Emotions*: Cambridge University Press.

Peacock, E., & Wong, P. (1990). The stress appraisal measure (SAM): A multidimensional approach to cognitive appraisal. *Stress Medicine, 6*, 227-236.

Perrez, M., & Reicherts, M. (1992). *Stress, Coping, and Health*. Seattle, WA: Hogrefe and Huber Publishers.

Rickel, J., Marsella, S., Gratch, J., Hill, R., Traum, D., & Swartout, W. (2002). Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems, July/August,* 32-38.

Rothbaum, B. O., Hodges, L. F., Alarcon, R., Ready, D., Shahar, F., Graap, K., et al. (1999). Virtual Environment Exposure Therapy for PTSD Vietnam Veterans: A Case Study. *Journal of Traumatic Stress*, 263-272.

Ryokai, K., Vaucelle, C., & Cassell, J. (in press). Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of Computer Assisted Learning*.

Sadek, M. D. (1990). *Logical Task Modeling for Man-machine Dialogue.* Paper presented at the National Conference on Artificial Intelligence.

Scherer, K. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer & P. Ekman (Eds.), *Approaches to emotion* (pp. 293-317).

Scherer, K. R., Schorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal Processes in Emotion*: Oxford University Press.

Searle, J. R. (1969). *Speech Acts*: Cambridge University Press.

Searle, J. R. (1979). *Expression and Meaning*: Cambridge University Press.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. NY: Springer-Verlag.

Shaw, E., Johnson, W. L., & Ganeshan, R. (1999). *Pedagogical Agents on the Web.* Paper presented at the Proceedings of the Third International Conference on Autonomous Agents, Seattle, WA.

Silverman, B. G. (2002). Human Behavior Models for Game-Theoretic Agents: Case of Crowd Tipping. *CogSci Quarterly, Fall*.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74*, 29-39.

Sloman, A., & Croucher, M. (1981). *Why robots will have emotions.* Paper presented at the International Joint Conference on Artificial Intelligence, Vancouver, Canada.

Smith, C. A., & Lazarus, R. (1990). Emotion and Adaptation. In Pervin (Ed.), *Handbook of Personality: theory & research* (pp. 609-637). NY: Guilford Press.

Traum, D., & Allen, J. F. (1994). *Discourse Obligations in Dialogue Processing.* Paper presented at the 32nd Annual Meeting of the Association for Computational Linguistics.

Traum, D., Rickel, J., Gratch, J., & Marsella, S. (2003). *Negotiation over tasks in hybrid human-agent teams for simulation-based training.* Paper presented at the International Conference on Autonomous Agents and Multiagent Systems, Melbourne, Australia.

Velásquez, J. (1998). *When robots weep: emotional memories and decision-making.* Paper presented at the Fifteenth National Conference on Artificial Intelligence, Madison, WI.

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer.

Weiner, B. (1995). *The Judgment of Responsibility*: Guilford Press.

Weiner, B. (2001). Responsibility for Social Transgressions: An Attributional Analysis. In B. F. Malle, L. J. Moses & D. A. Baldwin (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*: The MIT Press.