Appears in the 4th International Conference on Autonomous Agents, June 2000

Émile: Marshalling Passions in Training and Education

Jonathan Gratch
University of Southern California Information Sciences Institute
4676 Admiralty Way, Marina del Rey, CA 90292
gratch@isi.edu

ABSTRACT

Emotional reasoning can be an important contribution to automated tutoring and training systems. This paper describes Émile, a model of emotional reasoning that builds upon existing approaches and significantly generalizes and extends their capabilities. The main contribution is to show how an explicit planning model allows a more general treatment of several stages of the reasoning process. The model supports educational applications by allowing agents to appraise the emotional significance of events as they relate to students' (or their own) plans and goals, model and predict the emotional state of others, and alter behavior accordingly.

1 INTRODUCTION

Emotional computers may seem an oxymoron but recent years have seen a flurry of computation accounts of emotion in a variety of applications. This paper describes Émile, a model of emotional reasoning that extends and significantly generalizes prior work. Émile illustrates how an explicit planning model supports a more general treatment of emotional reasoning (specifically with respect to 'prospective' emotions, the treatment of inter-agent interactions, the derivation of emotional intensity, and management of the dynamics of emotional expression). The model allows agents to appraise the emotional significance of events as they relate to plans and goals, model and predict the emotional state of other agents, and alter behavior accordingly. Émile is particularly suited to applications where agents must plan as well as react (as in Rickel and Johnson, 1998; Hill *et al.,* 1997; Pell *et al.,* 1998).

But why model emotion? Emotions have received the most attention for their strength in creating engaging and believable characters in entertainment and art. Though somewhat less obvious, emotional reasoning can play a key role in educational tools. Psychological and neuroscience research indicates that emotions have a powerful influence on human behavior, both through their use as a non-verbal communication channel (Oatley and Johnson-Laird, 1987) and through their sway over planning and action selection (de Souza, 1987; Damasio, 1994). Recent work on tutoring systems has begun to address the communicative function by, for example, making tutoring systems respond to a student's emotional 'language' and generating emotion cues during instruction (Lester *et al.,* 1999; Johnson *et al.,* 2000). Research into military training simulators has also begun to address the influence of emotions on behavior (Hudlicka and Billingsley, 1999). Indeed, a panel of the National Research Council conducted a survey of this topic and concluded that the lack of "behavioral moderators" such as emotion substantially limits the capabilities of existing military simulations (NRC, 1998), leading, in part, to the creation of a new research institute to investigate such issues.[1]

This article takes as given that modeling affect is important, and sets forth one proposal for how this might be accomplished in the context of dynamic, interactive, and non-scripted training environments. The goal is to create autonomous agents that populate so-called "constructive simulations" where the trajectory of a training exercise arises from the interactions between participants (real or synthetic). This imposes strong requirements on the emotional models. Agents must automatically appraise the environment, derive a plausible emotional state, and modulate their behavior accordingly. Training simulations place other requirements that differ from entertainment-oriented applications. On one hand, the focus is more constrained as one only needs to model typical human behavior, not Micky Mouse as well, and not the richness of character necessary for entertainment applications. On the other hand, as the goal is to create operational systems, there is more need for generality, scalability, and the ability to support the constant flux in equipment and tactics.

Émile builds on prior computational models of emotion, notably Neal Reilly's (1996) Em algorithm and Clark Elliott's (1992) Affective Reasoner. Like many models it traces its lineage back to the influential theory proposed by Ortony, Clore, and Collins (1988). Émile differs in emphasizing the role of plans in emotional reasoning, and thus draws significantly on the work of Lazarus (1991), Oatley and Johnson-Laird (1987), Sloman (1987) and Beaudoin (1995). As with Sloman, Beaudoin, and Neal Reilly, my focus is on understanding the interaction of emotion, decision-making and action selection (so-called "broad-agents"). I therefore describe emotional reasoning in the context of other forms of reasoning: the overall system addresses the issues of how to develop and execute plans to achieve goals, model the plans and emotional responses of other (human or synthetic) agents, guide the presentation of information through emotional expressions or gestures, and influence the process of plan generation and action selection.

2 PLANNING

Many psychological theories of emotion emphasize the relationship between plans and emotions. Nevertheless, the vast majority of computational approaches do not explicitly represent plans (even approaches derived from psychological theories that do). This is partly pragmatic – when they were developed, planning algorithms had a number of limitations that precluded their easy

[1] The Institute for Creative Technology: www.isi.edu/uarc.html

inclusion into real-time intelligent agents (Agre and Chapman, 1987). Instead, prior accounts used "reactive planning" approaches that maintain an implicit model of plans. Such systems execute actions in the service of goals, but don't represent how actions contribute to goal achievement, can't predict future states of the world, and don't recognize interactions between steps in different plans – three properties that are critical for supporting a more general model. Planning research has advanced significantly in recent years and planning algorithms now reside at the core of a number of complex real-time autonomous systems in domains such as intelligent tutoring (Rickel and Johnson, 1998), spacecraft navigation (Pell *et al.*), information retrieval (Knoblock, 1995), and command and control modeling (Hill, et al.). The time seems ripe to reexamine the commitment to purely reactive models.

Adopting a plan-based approach has some key advantages. By maintaining an explicit representation of an agent's plans one can easily reason about future possible outcomes – a key requirement for handling emotions like hope and fear that involve future expectations. Explicit representations allow one to detect interactions between plans, for example, as when the plans of one agent are incompatible with those of another – a key requirement for handling emotions like anger or reproach which typically involve multiple actors. Planning algorithms have general mechanisms for making these assessments and we can leverage this generality in creating a model of emotional reasoning.

A plan-based approach also allows a richer model of how cognition influence one's emotional state. Most of us have experienced a flash of insight in our research that leaves us with intense feelings of hope, only to be crestfallen seconds later by the realization of some crucial flaw. We can model this dynamic by relating emotional appraisals to the current state of plans in an agent's memory. As plans grow and change through the planning process, so too the emotional state will change as a reflection of this process – in a sense providing a window into an agent's mental processes.

3 ÉMILE

Émile – after Rousseau's (1762) treatise on education – consists of five separate stages of processing, each of which is informed by plan representations. First, Émile must represent plans and manipulate this representation to determine which actions will further its goals. Second, it must qualitatively appraise how events (mental and physical) relate to its plans and goals. Third, it must assign a quantity to the appraisal. Next, it must integrate a variety of appraisals into an overall emotional state. Finally, it must use appraisals to guide action selection and planning.

Whereas Émile is less developed that some models, it serves as a comprehensive examination of how plan representations can inform each state of emotional processing. The specific detail of how this is realized is less important than the generality that a plan-based view brings to the problem.

3.1 Plan representation

For plan-based appraisal I adopt the "classical" model of plans as a set of STRIPS operators (Fikes and Nilsson, 1971) plus a variety of constraints. I adopt this representation, not because this is ideal, but because it has received the most study and numerous planning algorithms use it (or minor variants). STRIPS operators represent the actions that an agent may take in the world and consist of a set

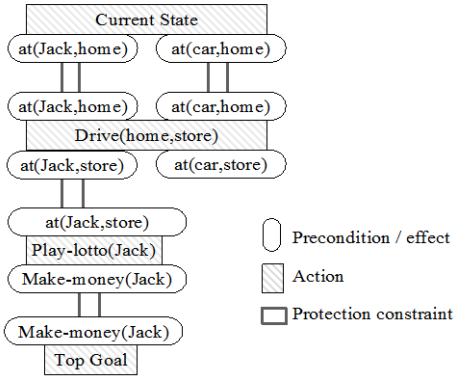

Figure 1: An example plan

of preconditions that must hold if the operation is to be performed, and a set of effects that describe how the world would change if the operator were executed. Plans consist of a set of actions and a variety of constraints: such as temporal constraints (this operator must be executed before that operator; this goal is needed by this deadline), binding constraints (this variable in this effect has the same value as that variable in that precondition), and protection constraints (this effect must stay true during some time interval). Protection constraints are typically used to keep track of the causal structure of a plan. For example, I may wish to drive to the market so that I might play the lottery (Figure 1). In planning terminology, "being at the market" is a precondition of the play-lotto action and it is *established* by an effect of the drive action. Planning algorithms keep track of this establishment relationship by explicitly noting that the effect of the drive action is used to achieve the precondition of the play-lotto action and must be protected (i.e., I shouldn't leave the market before buying the ticket). Protection constraints can also be used to represent maintenance goals (i.e. some fact must be maintained over some interval of time).

Planning algorithms differ in many details, but all can be viewed as incrementally refining some initial plan into one that achieves a set of goals. A variety of planners could be integrated with the model, however I assume that the planner can interleave planning and execution (Knoblock, 1995, Gratch, 1998) and explicitly represent plans of other agents in memory and reason about interactions between such plans (Gratch, 1998, Wilkins and Myers, 1998).

3.2 Appraisal

Models of emotion typically use the term *appraisal* to refer to the process of (qualitatively) evaluating the emotional significance of events. Several theories argue that appraisal involves relating events to goals and expectations (Lazarus, 1991; Ortony *et al*, 1988; Frijda and Swagerman, 1987). For example, when observing a football score, opposing fans observe the same event but have quite different emotional reactions because of their differing goals.

Many computational models of emotion are inspired by Ortony *et al*.'s, detailed theory of appraisal. Émile builds on Clark Elliott's account of this psychological theory, called construal theory, and illustrates how it can be generalized through an explicit accounting of plans. Construal theory assesses the relationship between events and an agent's disposition (described by its goals, social standards, and preferences) through a set of knowledge structures called *construal frames*. These frames do two jobs. They first determine

whether a relationship exists. If so, they characterize the relationship in terms of a set of features called *emotion-eliciting conditions*. Emotions are define in terms of these features, which include desirability (is the event desirable to the agent), expectation status (does the event confirm or disconfirm an expectation), evaluation (does the event uphold or violate a standard of behavior), etc.

3.2.1 Plan-based Appraisal

A plan-based perspective allows a major shift in the organization of the appraisal process. Construal theory, like many models of appraisal, focuses on events. An agent recognizes an event (e.g., action occurrence or state change), and matches it against construal frames to appraise its relation to the agent's goals, standards, and preferences. Figure 2 illustrates a construal frame that determines how an event that contains a touchdown relates to an agent's 'heroic-finish' goal (Elliott, 1992 p. 77). Émile adds a level of indirection that significantly generalizes this process. Rather than appraising events directly, Émile appraises the state of plans in memory. Events influence plans indirectly through the activities of the planner, allowing Émile to disassociate the two functions performed by construal frames. The relationship between events and an agent's disposition is derived more generally by a general-purpose planning algorithm. This perspective also allows other factors besides external events to drive the appraisal process. For example, allowing an agent to sit and think (elaborate its plans) will initiate new appraisals through changes in plan memory.

A plan-based perspective also simplifies the second function of construal frames - extracting emotion-eliciting conditions. Though these conditions provide a useful structure to the appraisal process, construal theory does not provide much guidance in how to compute them (instead, deriving them from domain-specific rules). By basing appraisals on the state of plans in memory, I show how to define these conditions in terms of the structure of plan memory. Thus, Émile replaces a large number of domain-specific construal frames needed by construal theory with a small number of domain-independent rules. Domain-specific information, for the most part, can be restricted to the operator descriptions (the domain theory) from which plans are built, and which we typically need anyway to inform planning and action selection.

Construal theory forms appraisals with respect to an agent's goals, social standards (norms of behavior), and preferences (the appealingness of domain objects). In plan-based appraisal, I define these terms as relations over plans. Goals correspond to an agent's top-level goals as well as any subgoals that arise in the plans developed to achieve top-level goals. Standards are viewed as constraints on behavior (i.e., constraints on 'socially acceptable' plans) and may encode some domain-specific information. For example, the standard "thou shalt not kill" could be represented as the constraint that plans executed by an agent must not contain an action who's effect leaves another agent dead. I depart from construal theory with regard to preferences, adopting the view of Lazarus (1991) that preferences are encoded in terms of goals: it is not that Émile prefers to be in a room with a desirable woman, but that he has a goal of mutual affection or self-affirmation. Such preferences are encoded as utility values associated with goals.[2]

[2] In general, construal theory's encoding of preferences in terms of objects is problematic as it suggests such objects are good/bad regardless of context. In contrast, it is easy to imagine that if the desirable woman likes Émile's best friend, that Émile would prefer to be with his friend or the woman in isolation, but not together. This "non-additivity," I argue, is better explained in terms of plans and goals: Émile's "acquisition" plan is threatened by his friend's presence.

FRAME: heroic-finish-goal, isa football-goal

event type:	touchdown
time-left:	?time-left
team-1:	?t1
team-2:	?t2
team-1-score	?t1s
team-2-score	?t2s
predicate:	?time-left < 0.01
predicate	?t1s ^= ?t2s
predicate:	?t1 = Northwestern or ?t2 = Northwestern
blocked:	(?t1 = Northwestern and ?t1s< ?t2s) or (?t2 = Northwestern and ?t2s < ?t1s)

Figure 2: An example of a construal frame

In plan-based appraisal, emotion-eliciting conditions are derived through a set of rules that reference syntactic features in plan memory. Rather than providing a single appraisal for the entire plan memory, Émile forms numerous local appraisals that are later integrated into the agent's overall emotional state (Section 3.4). For example, Émile appraises the local characteristics of each goal and subgoal in plan memory. The condition extracting rules are connected to a simple truth-maintenance system, allowing appraisals to change as the planner manipulates plans in memory. I discuss the derivation of four of these eliciting conditions to give a flavor of how plan representations support the appraisal process.

Self: Construal theory allows an agent to reason not just about its own emotions, but the emotions of other agents. The 'self' condition specifies whose perspective is being used to form the appraisal. For example, if Jack knows (or represents a guess about) Steve's plans and goals, Jack can appraise both how Steve's actions impact his own plans and goals, as well as how Steve might interpret Jack's actions.

Desire-self: This condition summarizes if some local characteristics of plan memory are desirable to the agent named in the 'self' condition. These characteristics are assessed to be either desirable or undesirable. For example, when appraising some subgoal, local characteristics include whether there is a plan to achieve it, and if so, whether there are any threats to this plan. Obviously, the agent who formed the subgoal would desire it to be achieved. Since based on plans, rather than events, these characteristics are expectations (e.g., I believe that I have a successful plan to achieve this subgoal). These expectations can be subsequently confirmed or disconfirmed as evaluated by the next emotion eliciting condition. Specifically, a (sub)goal's characteristics are assessed to be desirable if some effect in plan memory establishes the (sub)goal and no intervening effect (possibly) unestablishes it. Otherwise, its condition is undesirable.

Status: The status condition characterizes the state of the expectation underlying a "desire-self" assessment. Plans allow more flexibility than a simple binary distinction between confirmation and dis-confirmation, however to remain consistent with construal theory I draw a line that seems consistent with Elliott's intent. A (sub)goal's characteristics are appraised to be desirable but unconfirmed if the (sub)goal has an unthreatened establisher (it is estab-

lished and no possibly intervening action unestablishes it). This desire is confirmed once the establishing effect occurs. A (sub)goal's characteristics are appraised as unconfirmed and undesirable if the (sub)goal doesn't have an establisher or the establisher is threatened. This state is confirmed if the probability of generating a plan drops to zero, or the threat occurs, respectively.

Evaluation: This specifies if the local plan characteristics contain a praiseworthy or blameworthy act. This involves reasoning about standards of behavior and Émile currently only models a single standard: "thou shalt not introduce threats into someone else's plans." This is too simplistic as it avoids the issue of intent, but it is enough to get things started. In terms of the planning model described above, this standard is violated if a protection constraint associated with an agent's (sug)goal is threatened by an action associated with another agent. It is relatively easy to model other standards. For example, it could be considered praiseworthy if one agent proposes an action that achieves another agent's goals.

3.2.2 Emotion Classification

Émile uses emotion-eliciting conditions to classify which emotion arises from an appraisal. Following Ortony *et al.*, emotions are defined as relations over emotion-eliciting conditions. For example, the model associates an instance of joy with a character Steve if there is an event where Self is Steve, Desire-self is desirable, and Status is confirmed. Many, possibly conflicting, appraisals are generated simultaneously from plan memory and must be later integrated to infer the agent's overall state. By integrating subsets of these appraisals, and Émile agent could 'contextualize' its feelings: Steve feels happy but is troubled by his plan to go surfing.

Figure 3 extends the example in Figure 1. Steve has now told Jack that he's taking the car to the beach and Jack has incorporated this information into plan memory. The planner's threat detection processes recognize a potential goal violation: If Steve takes the car it violates Jack's constraint that the car stays home till he can take it to the store. The appraisal mechanisms make several inferences from the characteristics of the 'car at home' goal. Whereas Jack was previously hopeful that this subgoal would be achieved, he is now fearful that it will be violated. Furthermore, Jack now resents Steve as he has violated a social standard by threatening his plans.

3.3 Intensity

Given qualitative appraisals, Émile must assess their intensity. There is fear and there is FEAR, but how to automatically recognize the difference? Here too, Émile draws heavily on the explicit plan representation to extend the generality of this computation. Ortony *et al.* propose up to eight intensity variables for each emotion type (e.g. goal importance, unexpectedness, arousal) and Elliott and Siegal (1993) have continued to expand this list. Following Neal Reilly (1996) I adopt a much simpler model using two intensity variables: probability of goal attainment and goal importance. This is sufficient to illustrate the use of plan representations and can be readily extended to account for other factors.

3.3.1 Probability of Goal attainment

Émile incorporates the view of Oatley and Johnson-Laird and Neal Reilly that emotions are related to changes in the perceived probability of goal attainment. Basing intensity on probability has the advantage that (almost) everyone agrees on what it means, and

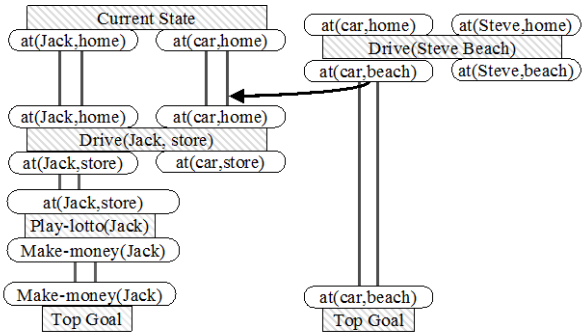

Figure 3: An example of inter-plan threats

they can (in theory at least) be learned from experience. Furthermore, probability assessment can be (and I argue should be) based on one's current representation of the plans that could bring that goal about. Thus we can model the change in probability assessment through cognition (i.e. as the planner works on its goals).

The main advantage over Neal Reilly's approach is that leveraging off of an explicit plan representation allows more generality in how one computes this dynamic probability. Many planning researchers have considered how to compute the probability of goal attainment from plan representations (Russell and Wefald, 1991; Blythe and Veloso, 1996; Onder and Pollack, 1999). In general, the computation must consider not only the probability that a given plan will succeed, but also the more problematic question of how likely one can derive a plan for a given goal or subgoal.

The current model uses a simpler approach to deriving these probabilities based on a strong independence assumption (all joint probabilities are modeled as the product of their constituent probabilities). This suffices as a first approximation and greatly simplifies a number of equations. It is also clear that people used flawed probability models (Tversky and Kahneman, 1983). The probability that a goal will be achieved depends on how one intends to achieve it (current plans), but ultimately the computation grounds out in a number of 'base-probabilities' that must either be supplied by a domain expert or learned in the course of problem solving.

Specifically, one must specify two classes of probabilities. First one must supply the probability that an effect of an action will be achieved if the action is executed. The *execution probability* of a given effect is denoted as $P_{EX}(\text{eff})$. Second, one must derive the probability that an unplanned for goal can be successfully achieved (meaning we must both generate a plan and successfully execute it). The unplanned-for probability of a goal is denoted as $P_{UNP}(\text{goal})$. Simple models can treat these base-probabilities as constants. More sophisticated models could account for other factors. For example, the probability of me getting a date might be conditional on if I took a bath that morning, or the time remaining before the date. Some of the above mentioned methods can account for such conditional probabilities.

The initial probability assessment of a top-level goal is its base probability. This becomes refined as the planner refines its plans for the goal. The probability model propagates base probabilities through the plans in memory using simple rules that key off of certain syntactic properties of plans. A (sub)goal is considered *established* if the planner has identified some effect in plan memory that could achieve it. This *establisher* is considered *threatened* if some other effect – called the *threat* – possibly undoes it before

the (sub)goal is needed. The *action* associated with an effect has an execution *state* (an action may be pending or initiated). Émile's planning model assumes actions have duration so an action may be initiated and its effects observed somewhat later. An effect is *satisfied* as long as it is observed to be true in the environment.

Probability of an effect: P(eff)

IF state(action(eff)) = -Initiated THEN
  $P(eff) = P_{EX}(eff) \prod P(precondition(action(eff))$

IF state(action(eff)) = Initiated AND -satisfied(eff) THEN
  $P(eff) = P_{EX}(eff)$

IF state(action(eff)) = Initiated AND satisfied(eff) THEN
  $P(eff) = 1$

IF state(action(eff)) = Initiated AND -satisfied(eff) THEN
  $P(eff)=0$

Probability of a goal/precondition: Pr(goal)

IF -established(goal) THEN
  $P(goal) = P_{UNP}(goal)$

IF established(goal) AND -threatened(goal) THEN
  $P(goal) = P(establisher(goal))$

IF established(goal) AND threatened(goal) THEN
  $P(goal) = P(establisher(goal))[1 - P(threat(goal))] + P_{UNP}(goal)P(threat(goal))$

As with appraisal rules, probability rules are associated with a truth maintenance system, updating automatically as plans in memory change.

3.3.2 Emotional Importance

Goal importance is the second factor influencing emotion intensity. A goal's importance to an agent is divided into intrinsic and extrinsic components (following Sloman). Intrinsic importance is the reward (utility) an agent receives from achieving the goal. An agent might view "having money" as an end in itself, and assign intrinsic importance to obtaining it. Goals may also further other ends (subgoals). A goal's extrinsic worth relates to how it furthers other intrinsic goals. Sloman and Beaudoin define extrinsic worth in terms of syntactic characteristics of the plan (e.g., the depth the goal falls in the plan hierarchy, the number of operators that could achieve the goal, etc.). In my view, this syntactic characterization amounts to a heuristic for assessing how much a subgoal's achievement contributes to the probability of attaining intrinsic goals. Rather, I explicitly define this contribution in terms of the change in the probability of intrinsic goal achievement: the importance of a subgoal is the sum of the intrinsic importance of goals it helps establish, weighted by how much its establishment adds to the probability each of these intrinsic goals will be achieved. Again, plan representations are key in this computation.

Extrinsic importance depends on two factors. First, we must identify all of the intrinsic goals impacted by a subgoal. Second, we must identify how much the subgoal impacts each of these intrinsic goals. The set of impacted goals is simply the set of goals with intrinsic importance that are directly or indirectly connected to the subgoal via plans (in the transitive closure of the establishment relation). Computing the probabilistic contribution of the subgoal to each impacted goal has been studied in the planning community (e.g., Onder and Pollack, 1999). I adopt a much simpler (but not necessarily accurate) computation that exploits the assumed independence of precondition probabilities assumed above. The extrinsic importance of a subgoal s is defined as the sum of the importance of each impacted goal weighted by the change in probability that attaining the subgoal would have on it:

$$Import(s)=Im_{intrinsic}(s) + \sum_{g \in Impact(s)} Im_{intrinsic}(g)\frac{P(g)[1-P_{UNP}(s)]}{P(s)}$$

This captures the notion that difficult to achieve goals are more important (i.e. unplanned-for probability is low).

3.3.3 Emotional Intensity

The intensity of an appraisal is based on its importance and the how the current plans in memory impact the probability of goal achievement. One could augment these variables with other factors (Elliott and Siegal, 1993). To keep the model simple, however, I propose the following intensity rules, which could be subsequently refined. Émile currently implements only five emotion types:

$$Intensity_{HOPE}(goal) = Import(goal)\ P(goal)$$

$$Intensity_{JOY}(goal) = Import(goal)$$

$$Intensity_{FEAR}(goal) = Import(goal)[1-P(goal)]$$

$$Intensity_{DISTRESS}(goal) = Import(goal)$$

$$Intensity_{ANGER}(goal) = Import(goal)P(threat)$$

Hope arises from a belief that something good might happen. In Émile, this translates into the fact that a goal has been established but not yet achieved. The intensity is based on the goal's importance and its probability of achievement. Joy arises when something good has happened. This translates into the fact that a goal has been successfully achieved. Intensity is tied to the goal's importance. Fear arises from a belief that something bad may happen. In Émile, this means some goal is unestablished, or its establisher is threatened. Intensity is based on the goal's importance and it's probability of failure. Distress arises when some fear has been confirmed. This translates into the fact that a goal has been prevented from occurring. (It's establishing plan was threatened and the threat occurred). The intensity of distress is directly proportional to the goal's importance. Anger arises when some agent has violated a social principle. The current model has a single standard: don't introduce threats into other agents' plans. Anger becomes more intense if the goal is important and the threat is likely to be realized. A more general treatment of standards would associate importance with standards as well as goals.

This approach can be extended to other emotion types, or made subtler through the introduction of more intensity variables (though see the discussion in the final section). For example, emotions such as surprise or relief can be defined in terms of the change in probability of goal attainment (its derivative with respect to time) as in Neal Reilly's Em. Because Émile explicitly models the process of planning, its easy to incorporate 'planning effort' into the equation, as suggested by Ortony *et al.*

3.4 Integration and Dynamics

At any moment in time, the appraisal mechanism will have produced a number of appraisals from the current plan structure. Two questions that remain are how appraisals change over time and how to integrate several (possibly conflicting) assessments.

Again, the plan-based view provides a general (albeit partial) model of the dynamics of emotion. Whenever plan memory changes,

Émile automatically revises its appraisal and probability assessments (through the actions of the truth maintenance system that manages these rules). For example, an agent may have an important and unestablished goal that leads to an appraisal of fear. After some effort the planner discovers a workable plan, causing the fear appraisal to retract and a hope appraisal to be asserted. The model is partial because it may lead to somewhat jarring transitions and doesn't address the issue that emotional responses tend to decay over time. It also doesn't specify how the various and possibly conflicting appraisals are integrated into an overall emotional state. Finally, it doesn't specify how to integrate cognitive appraisals (those that derive from plans and goals) with other influences on emotion, such as bodily states.

I draw on Velásquez's (1997) Cathaxis model to mitigate some of the issues of dynamics and integration. Cathaxis is inspired by ethological and neuroscience theories and has been used to model emotional responses in robotic systems. Cathaxis is based on an energy metaphor. Different appraisals act as energy elicitors that excite or inhibit different emotional states, and decay over time. I use an instantiation of a general class of models that Velásquez describes. This stage, unlike the previous ones, requires a number of free parameters that don't have obvious 'objective' assignments.

The previous stages result in a set of appraisals associated with different agents and with differing emotional labels (joy, hope, etc.) and intensities. As a first step, as long as these appraisals persist, Émile decays their intensity by a constant rate. The integration model adds these decaying intensities into different buckets based on their emotional label. Thus, if one has several separate appraisals of "fear", these are added together into an overall fear intensity. If one has a model of other sources of excitation (e.g., the influence of hunger) these could be added to the buckets as well. Collectively, these buckets correspond to the current activation vector. The activation vector at time t is denoted $\mathbf{A_t}$. According to the Cathaxis model, the activation of a given emotion is excited by some emotional states (e.g., joy excites hope) and inhibited by others. These influences can be expressed as an $n \times n$ matrix, $\mathbf{M}$, where n is the number of emotional labels and the diagonal elements have a value of zero. For example, $\mathbf{M}_{joy,fear}$=-0.5 corresponds to the notion that joy has a modest negative influence on fear. The overall intensity of each emotion at time t, represented in vector form $\mathbf{E_t}$ is defined as:

$$\mathbf{E}_{t+1} = \alpha \cdot \mathbf{E}_t + (1-\alpha)\left[\mathbf{A}_{t+1} + \mathbf{M} \times \mathbf{E}_t\right]$$

The parameter α controls the smoothness of transitions. As a last step, the intensity vector $\mathbf{E}$ is passed through a sigmoid function to map the emotional state to a value from zero to one. This normalized value is then used to influence behavior.

3.5 Influencing Behavior

An agent's current emotional state can influence behavior by altering its communication, planning, and action selection. Communication may be altered in several ways. In the application I next describe, the emotional state is used to select gestures, facial expressions and to alter speech generation (based on a model by Cahn, 1990). Modeling the emotional state of other agents can guide discourse between agents. For example, an agent might ask another agent what's wrong if they unexpectedly frown.

The emotional state can influence the plan generation process in several ways as well. Appraisals can be used to provide search control to the planning algorithm (as suggested by Damasio). For example, one could focus planning effort on the portions of the plan generating the most intense appraisals. I have also experimented with using the emotional state to alter how the planner treats interactions between plans (Gratch, 1998). For example, the planner may develop a plan for one goal that defeats a plan for a less important goal. If the agent is experiencing strong negative emotions, it may not care about these interactions. However if it is experiencing positive emotions, it may try to ensure that both goals can be successfully achieved. This same idea can be applied to social situations. For example, if I'm angry with another agent, I might wish to bias my plan generation to defeating their goals.

In its current incarnation, Émile only learns about the activities of other agents through perception of their actions, or through communication events, as when Steve tells Jack his plan to go to the beach. In general, I would like the system to guess the goals and plans of other entities in the world through some form of plan recognition. Even when agents communicate, there may be ambiguity in their utterances. Clearly, the emotional state could impact the plan recognition or disambiguation process. For example, a fearful agent might assume the worst when choosing among multiple interpretations. The emotional state could also influence probability assessment, for example, by allowing a joyful agent to be hopelessly optimistic when assessing the probability of goal attainment.

Finally, emotional state can bias action selection. For example, Rousseau and Hayes-Roth (1997) describe an approach that views emotion and personality as a point in some multi-dimensional space. Actions are assigned to points in this space and an agent chooses whatever action is closest to the current emotional state.

4 ILLUSTRATION

Émile is being applied to the problem of modeling agents that populate a military simulator and exhibit realistic emotional responses to battlefield situations (building on Hill *et al.,* 1997). This work is quite preliminary so I describe an artificial application that illustrates Émile's basic capabilities. I've incorporated Émile into a version of Steve (a pedagogical agent developed by Rickel and Johnson, 1998) by, in essence, removing Steve's brain and replacing it with the plan based model. The Steve software provides a representation of a body, including a model of perception, motor commands, and a text-to-speech interface.

The following annotated exchange comes from a sample interaction between two Émile agents who have been given differing goals, domain knowledge, and personality parameters. Jack's goal is to make money and has a conscientious personality. Steve's goal is to surf, and tends to be rude. Both develop different plans but have to contend with a shared resource. The model of how personality influences planning and how discourse is organized is outside the scope of this paper but is discussed in Gratch (1999). Figure 4 illustrates a snapshot of the visual presentation, including each agent's current emotional state and plan memory.

Jack: I want to make-some-big-money. *[Looks concerned, scratches his head, then, after devising a plan looks hopeful.]*

Steve: I want to catch-some-waves. *[Looks concerned, scratches head, and continues to look concerned. Surfing is important to Steve and he cannot devise a plan.]*

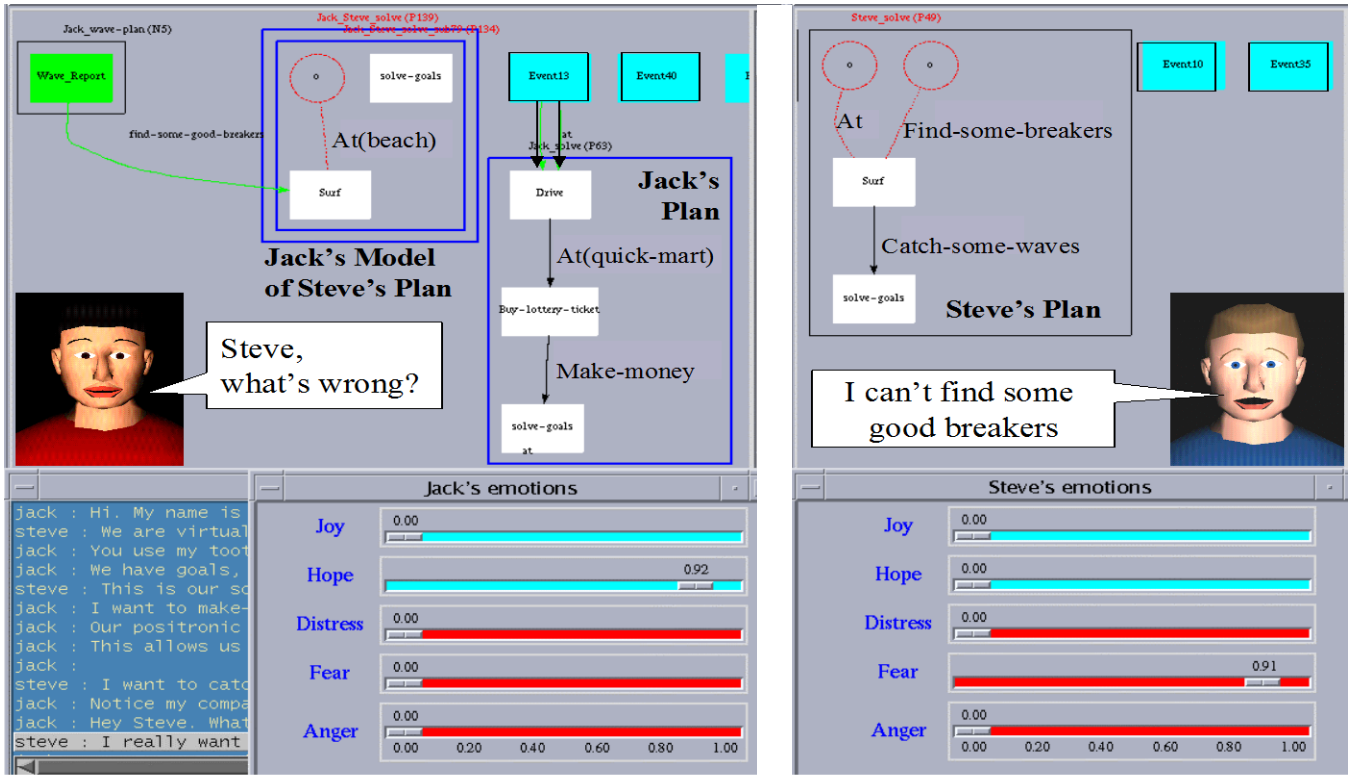

Figure 4. An exchange between two Émile agents

Jack: [*Perceives Steve's emotional state and generates an information request.]* Hey Steve, what's wrong?

Steve: *[Locates the appraisal generating the most intense negative emotional excitation. Communicates the associated plan fragment in a distressed tone of voice.]* I want to catch some waves but can't find any good breakers.

Jack: *[Incorporates Steve's plan fragment into plan memory and locates relevant information. Jack has knowledge of a wave report that establishes Steve's blocked subgoal]* Steve, does it help that someone did say there's some great waves near the pier?

Steve: *[Incorporates the communicated plan fragment. Completes a plan to go surfing and looks hopeful.]*

Jack: *[Perceives Steve's change in expression and seeks to confirm his expectation that the information he provided helped Steve.]* So that information helped?

Steve: *[Handles Jack's information request.]* Yes Jack. I plan to drive the car to the beach, then I plan to surf-my-brains-out.

Jack: *[Incorporates Steve's plan fragment and finds a conflict with his own plans. Based on personality, Jack attempts to negotiate a fair solution.]* Wait a second. Our plans conflict. I plan to drive the car to the quicky-mart then I plan to buy a-lottery-ticket.

Steve: *[Incorporates Jack's plan fragment and recognizes the same interaction. Based on personality model, Steve responds to interaction differently. Steve exits stage right.]* Later dude, I'm driving the car to the beach.

Jack: *[Perceives that car has departed without him. Looks angry. Says in angry voice:]* I want to kill-my-roommate.

5 SUMMARY

Obviously, several important issues are finessed in this demonstration. The agents do not understand English. Rather they communicate symbolic representations of plans and a simple generation routine converts this to English-like speech for the benefit of the observer. The agents do not have general emotion recognition capabilities (as in Picard, 1997). Rather, they directly perceive high-level predicates like *expression(Steve, Anger)*.

More fundamentally, the treatment of appraisal is too simplistic for several potential applications. The model does a reasonable job of accounting for emotions that arise from a concrete task. It provides a less than adequate account of interpersonal emotions (where one needs to factor in issues like intent or locus-of control), "life goals" (where the connection between events and goals is rather tenuous), and rich coping strategies (as in dehumanizing someone you have wronged to mitigate one's guilt). I am currently considering proposals by Lazarus (1991), who uses a two-stage appraisal process to account for such phenomena.

Nevertheless, the system is still quite flexible. Leveraging off of a planner's general reasoning eases some of the knowledge engineering complexities. The discourse and personality knowledge is also fairly general. Thus, Jack tells Steve about the wave report because of a general rule that states: If you are friends with another agent, and you have information that is relevant to them, and you do not believe that they are aware of that information, than communicate the information. The assessment of relevance is based on a general model (desJardin and Wolverton, 1998).

More importantly, by taking a plan-based perspective, Émile not only gains the generality discussed in this paper, but it also opens the way to integrating the system with a large body of techniques that have been developed for plan representations. Thus, it is relatively straightforward to augment Émile with the ability to provide concise descriptions of its plans (Young, 1999), recognize the plans of other agents (Lesh and Etzioni, 1995), as well as augment its planning capabilities with more efficient or more expressive planning algorithms as they become available. Clearly, a planning approach is not a cure-all. Planning algorithms must still wrestle with issues of efficiency and they don't provide any insight on non-cognitive influences on emotional state. Nonetheless, plan-based models like Émile have the potential to augment educational software such as tutoring systems and training simulations. Hopefully as this project continues, we will have a better sense of the extent that this potential may be realized.

Acknowledgements

I greatly benefited from discussions with Stacy Marsella and Paul Rosenboom. Thanks to Jeff Rickel for letting me steal his bodies. Clark Elliott, Jeff Rickel, Lewis Johnson, Celia Pearce, Kate LaBore, Wendy Hui Kyong, and Ricardo Dominguez gave greate comments, insights, and suggestions. This research was funded by the Army Research Institute under contract TAPC-ARI-BR

References

Agre, P. & Chapman, D. 1987. Pengi: an implementation of a theory of activity. AAAI97.

Beaudoin, L. 1995. *Goal Processing in Autonomous Agents*. Ph.D Thesis (CSRP-95-2), Univ. of Birmingham.

Blythe, J. & Veloso, M. 1996. Learning to improve uncertainty handling in a hybrid planning system. *AAAI Fall Symposium on Learning Complex Behavior*

Cahn, J. 1990. *Generating Expression in Synthesized Speech.* Ph. D Thesis, M.I.T.

Damasio, A. 1994. *Descartes' Error*. Avon Books, NY, NY.

de Souza, R. 1987. *The rationality of emotion.* MIT Press.

desJardins, M. & Wolverton, M. 1998. Coordinating planning activity and information flow in a distributed planning system. *AAAI Fall Symposium on Distributed Continual Planning.*

Elliott C. D. 1992. *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System*. Ph.D Thesis (TR#32), Northwestern University.

Elliott C. & Siegle, G., 1993. *AAAI Spring Symposium on Reasoning about Mental States.*

Fikes, R.E. & Nilsson, N. J. 1971. STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving. *Artificial Intelligence*, 2(3-4)

Frijda, N.H. & Swagerman, 1987 J. Can computers Feel? Theory and Design of an Emotional System. Cognition and Emotion, 1 (3) 235-257.

Gratch, J., 1998. Reasoning about multiple plans in dynamic multi-agent environments. AAAI Fall Symposium on Distributed Continual Planning, Orlando, FL.

Hill, R., Chen, J., Gratch, G., Rosenbloom, P., & Tambe, M. 1997. "Intelligent Agents for the Synthetic Battlefield," in AAAI-97/IAAI-97, pp. 1006-1012.

Hudlica, E., & Billingsley, J., 1999. Representing behavior moderators in military human performance models. *8th Con./ on Computer Generated Forces and Behavioral Representation.*

Johnson, L., Rickel, Jo. & Lester, J., 2000. Animated pedagogical agents: face-to-face interaction in interactive learning environments. *International* Journal of *Artificial Intelligence in Education. Forthcomming*

Knoblock, C. 1995. Planning, executing, sensing, and replanning for information gathering, IJCAI95.

Lazarus, R.S. 1991. *Emotion and Adaptation.* Oxford Press.

Lesh, N., and Etzioni, O. 1995. A sound and fast goal recognizer. *International Joint Conference on AI.*

Lester, J.C., Towns, S.G.,FitzGerald, P.J. 1999 Achieving Affective Impact: Visual Emotive Communication in Lifelike Pedagogical Agents. *International Journal of AI in Education, 10 (3-4)* pp. 278-291.

National Research Council (NRC), 1998. *Modeling Human and Organizational Behavior.* National Academy Press, Washington D.C.

Neal Reilly, W.S., 1996. *Believable Social and Emotional Agents*. Ph.D Thesis CMU-CS-96-138. Carnegie Melon Univ.

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

Onder, N. & Pollack, M. 1999. Conditional, probabilistic planning: a unifying algorithm and effective search control mechanisms. *AAAI99*

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

Pell, B. Bernard, D., Chien, S., Gat, E., Muscettola, N., Nayak, P. P., Wagner, M., Williams, B. 1998. An autonomous spacecraft agent prototype *Autonomous Robotics, 5.*

Picard, R. 1997. *Affective Computing.* MIT Press.

Rickel, J. & Johnson, L. 1998. Animated agents for procedural training in virtual reality: perception, cognition, and motor control. *Applied Artificial Intelligence, v 13.*

Integrating Pedagogical Capabilities in a Virtual Environment Agent. *First International Conference on Autonomous Agents.*

Rousseau, J., 1762. *Émile*. Reissued, Everyman Press, 1993.

Rousseau, D. & Hayes-Roth, B. 1997. Improvisational synthetic actors with flexible personalities. *Knowledge Systems Laboratory TR KSL97-10*

Russell, S., & Wefald, E. 1991. *Do the right thing: Studies in limited rationality*. MIT Press.

Sloman, A. 1987. Motives, mechanisms and emotions. *Cognition and Emotion*, 1, pp 217-234.

Tversky, A. & Kahneman, D. 1983. Extensional versus intuitive reasoning: the conjunction fallacy in probability judgement. *Psychological Review, v. 90 n.4.*

Velásquez, J. 1998. When robots weep: emotional memories and decision-making. *AAAI98.*

Wilkins, D, & Myers, K. 1998. A Multiagent Planning Architecture. *Conference on AI Planning Systems (AIPS98)*

Young, M. 1999. Cooperative plan identification: constructing concise and effective plan descriptions. *AAAI99.*