

Empirically Building and Evaluating a Probabilistic Model of User Affect

Cristina Conati¹, Heather Maclaren²

Final Draft

Department of Computer Science
University of British Columbia
Vancouver, BC, Canada V6T 1Z4
¹conati@cs.ubc.ca, ²maclarhr@gmail.com

Abstract

We present a probabilistic model of user affect designed to allow an intelligent agent to recognise multiple user emotions during the interaction with an educational computer game. Our model is based on a probabilistic framework that deals with the high level of uncertainty involved in recognizing a variety of user emotions by combining in a Dynamic Bayesian Network information on both the causes and effects of emotional reactions. The part of the framework that reasons from causes to emotions (*diagnostic model*) implements a theoretical model of affect, the OCC model, which accounts for how emotions are caused by one's appraisal of the current context in terms of one's goals and preferences. The advantage of using the OCC model is that it provides an affective agent with explicit information not only on *which emotions* a user feels but also *why*, thus increasing the agent's capability to effectively respond to the users' emotions. The challenge is that building the model requires having mechanisms to assess user goals and how the environment fits them, a form of *plan recognition*. In this paper, we illustrate how we built the predictive part of the affective model by combining general theories with empirical studies to adapt the theories to our target application domain. We then present results on the model's accuracy, showing that the model achieves good accuracy on several of the target emotions. We also discuss the model's limitations, to open the ground for the next stage of the work, i.e., complementing the model with diagnostic information.

Keywords: affective computing, dynamic Bayesian networks, evaluation, user modeling.

Introduction

Recent years have seen a flourishing of research directed towards adding an affective component to human–computer dialogue. One key element of this endeavour is the computer's capability to recognize the user's emotional states during the interaction, which requires a model of the user's affect. Humans use different sources of information to assess a person's emotions, including causal information on both context and the person's relevant traits, and symptomatic information on the person's visible bodily reactions. However, this information is often incomplete and even contradictory, making assessment of emotion a task riddled with uncertainty.

To handle this uncertainty, Conati (2002) proposed a probabilistic framework for affective user modeling that integrates in a Dynamic Decision Network (DDN) (Dean & Kanazawa, 1989), information on both the possible causes of the user's affective reaction and its observable effects. Leveraging any information available on the user's emotional state is crucial, because the different sources of evidence are often ambiguous, and their reliability varies significantly according to both the user and each particular interaction.

In this paper, we illustrate how we used the framework proposed in (Conati, 2002) to build a model of user affect during interaction with an educational computer game. The long–term goal is to employ this user model to guide adaptive system interventions aimed at improving the overall success of the student's educational experience with the game. We focus on the part of the model that assesses user emotions from causes by relying on the OCC theory, a theoretical model that accounts for how emotions derive from one's appraisal of the current context (Ortony, Clore, & Collins, 1988). Our goal is to understand how accurate an affective assessment can be achieved with causal information that can be gathered non-intrusively from naturally occurring interaction events, vs. when it becomes worthwhile using the potentially more intrusive/expensive technology necessary to gather diagnostic information from other user's behaviors.

Although there is still no hard evidence that taking user affect into account can substantially improve human–computer interaction in general, there are several studies indicating that maintaining positive student affect is beneficial in educational settings. Craig et al. (2004) reported that flow (the state of being fully immersed in and focused on the interaction) and confusion were positively correlated with learning, whereas boredom was negatively correlated. Linnenbrink and Pintrich (2002) found that while most students experience some confusion when confronted with information that does not fit their current knowledge, those in a generally positive affective state will adapt their known concepts to assimilate it, whereas students in a generally negative affective state will reject the new knowledge. Cordova and Lepper (1996) found that learners exposed to motivationally embellished educational software (Lepper et al., 1993) had higher levels of intrinsic motivation. As a result, they become more deeply engaged by the interaction, and learn more in a fixed period of time. Zakharov et al. (2008) compared two versions of a pedagogical agent for a database tutor, one which tailors its responses to the valence of the student affective state (positive vs. negative) and one which does not, and report that students found the interventions of the first agent more useful and appropriate.

We believe that the benefits of taking user affect into account are even stronger for educational activities that rely heavily on the student's direct involvement in the learning process, such as those supplied by educational computer games. An educational game tries to increase the learner's motivation by embedding pedagogical activities in highly engaging, game-like interactions. While there has been growing interest in computer games as educational tools, results on their effectiveness are mixed. There is evidence that these games can increase student engagement and motivation (e.g., Alessi et al. 2001, Lee et al. 2004), but the results on their pedagogical potential are limited (e.g., Klawe 1998, Vogel 2004, Van Eck 2007) and there is also evidence that, for some students, educational games can induce worse affect than more traditional

e-learning tools (Rodrigo et al 2008). One possible cause of these results is that most existing educational games are designed based on a one-size-fits-all approach rather than being able to respond to the specific needs of individual students. To overcome this limitation, we are designing emotionally intelligent pedagogical agents that, as part of game playing, generate tailored interventions aimed at both stimulating student learning, as well as maintaining a high level of student affective engagement (Conati & Klawe, 2002). The affective model we describe in this paper is meant to be used by our pedagogical agents, together with a model of student learning, to generate these interventions.

In order to improve the agent's ability to tailor its responses to a student's specific needs, our affective model assesses not only which emotions a user is feeling, but also *why*. To do so, it relies on a computational representation of the OCC theory of emotions, which sees emotions as the reactions of an individual towards current states of the world and towards their causes, shaped by the individual's goals and preferences. While having an affective model based on the OCC theory yields great potential for rich adaptive interaction, building the model is challenging because it requires having mechanisms to assess user goals and how the environment fits them, i.e., a form of *plan recognition*. In this paper, we illustrate how we built the OCC-based model with an iterative process of design and evaluation, relying on several empirical studies to adapt the model's theoretical underpinnings to our target domain. We then present results on the model accuracy, showing that the model achieves good accuracy on several of the target emotions. We also discuss its limitations, to open the ground for the next stage of the work, i.e., complementing the model with diagnostic information from the user's affective reactions¹.

While there has been substantial work on using the OCC theory for implementing *generative* emotion models to direct the affective behaviours of artificial agents (e.g. Gratch 2000, Prendinger and Ishizuka 2002, Gebhard 2005), there have been only few preliminary, not validated attempts to use the OCC theory to recognize user emotions in real-time during interaction (Streit et al. 2004, Chalfoun et al. 2006). Another approach that has used causal information to detect user affect relies on machine learning and sees causal information as one of the many dimensions used to learn patterns from data to emotions. While this approach has shown promising results in terms of model accuracy (e.g. Kapoor and Picard 2005), it loses the ability to use the casual information as an explanation of the user's reactions.

Another distinguishing feature of our work is that we consider multiple, rapidly changing emotions that possibly overlap and conflict, as often experienced by students playing an educational games. In contrast, most work on affect recognition has focused on detecting one specific emotion (e.g., Kapoor and Picard 2005, Healey and Picard 2005), lower-level affective measures such as affective valence and arousal (e.g., Prendinger et al. 2005, Litman and Forbes-Riley 2004) or overall emotional predisposition over a complete interaction (Mandryk et al. 2006, Yannakakis et al. 2008) Other work that, like ours, has dealt with multiple emotions has only attempted to capture one specific emotion at a time, rather than potentially overlapping emotions (D'Mello et al. 2008).

The structure of this paper is as follows. In Section 1, we describe the general framework we used to build our probabilistic model of user affect. In Section 2, we introduce Prime Climb, the educational computer game we used as a test-bed application for model development. Section 3 illustrates the predictive part of the affective model. In Section 4, we introduce our technique for model evaluation and apply it to test the predictive model. In Section 6, we discuss related work,

¹ This paper is an extension of the work described in (Zhou and Conati 2002 and Conati and MacLaren 2007). Here we give a more detailed account of the empirical studies underlying our approach. We also provide new results on accuracy, based on more extensive cross-validation and including a specific comparison of model performance when the user goals are given and when they have to be assessed by the model. Previously published results relate only to the model with the goals provided as evidence (Conati and MacLaren 2004).

and in Section 6 we conclude with a discussion of the research presented as well as ideas for future work.

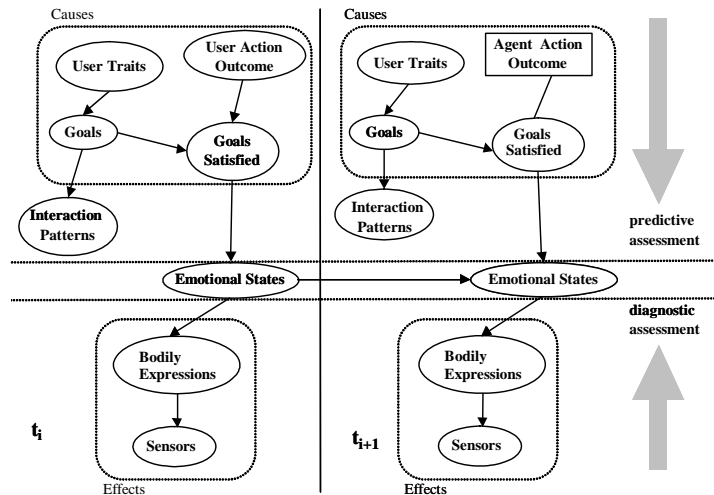


Figure 1. Two time-slices of the DDN for affective modeling

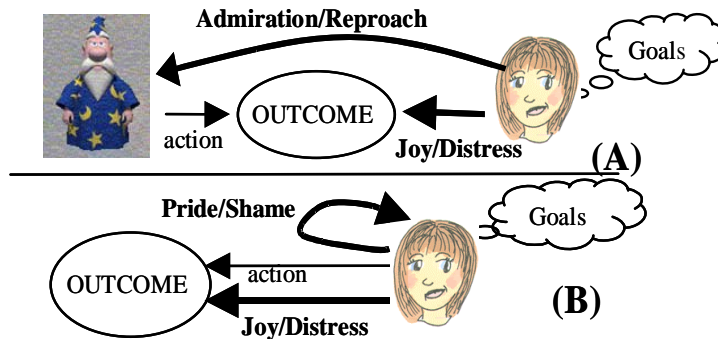
1 A Dynamic Decision Network for Emotion Recognition

A DDN is a graph where nodes represent either stochastic variables of interest or points where an agent needs to make deliberate decisions. Arcs in the graph capture the direct probabilistic relationships between the nodes, including temporal dependencies between the evolving values of dynamic variables. Each node has an associated probability distribution representing the conditional probability of each of its possible values, given the values of its parent nodes. As evidence on one or more network variables becomes available, *ad hoc* algorithms update the posterior probabilities of all the other variables, given the observed values.

Figure 1 shows a high-level representation of two time-slices in the theoretical DDN-based framework for affective modeling proposed in (Conati, 2002). Each time slice represents the model’s variables at a particular point in time. For illustration purposes, the nodes in Figure 1 represent classes of variables instead of individual variables in the DDN. As the figure shows, the network is designed to combine evidence on both the causes and effects of emotional reactions, to compensate for the fact that often evidence on causes or effects alone is insufficient to accurately assess the user’s emotional state. While in the following sections we focus on the implementation and evaluation of the causal part of the model, here we give a general description of the complete framework to illustrate our overall approach.

The subnetwork above the nodes *Emotional States* is the predictive component of the framework. It represents the relations between possible causes and emotional states as described in the OCC cognitive theory of emotions. According to this theory, emotions derive from cognitive appraisal of the current situation, which consists of events, agents, and objects. The outcome of the appraisal depends on how the situation fits with one’s goals and preferences. For instance, depending on whether the current event (e.g., the outcome of an action in Figure 2) does or does not fit with one’s goals, that person will feel either *joy* or *distress* toward the event (see Figure 2, A and B). Correspondingly, if the current event is caused by a third-party agent, that person will feel *admiration* or *reproach* toward the agent (see Figure 2A); if that agent is oneself,

the person will feel either *pride* or *shame* (see Figure 2B). Based on this structure, the OCC theory defines 22 different emotions, described in terms of their valence and the entity they relate



to.

Figure 2. **Example emotions in the OCC theory**

We adopted this particular theory of emotion for our affective modeling framework because its clear and intuitive representation of the causal nature of emotional reactions lends itself well to devise computational models that can assess *why* a user feels given emotions in addition to *which* these emotions are. This more fine-grained information can enhance the capability of an interactive agent to adequately respond to a user affect. For instance, if the agent can recognize that the user feels a negative emotion because of something wrong she has done (*shame* by OCC definition) it can decide to provide hints aimed at making the user feel better about her performance. If the agent recognizes that the negative feelings are caused by its own behavior (*reproach* by OCC definition) it may decide to take actions that allow it to make amends with the user. The agent can also *explain* the rationale underlying its interventions, a feature that can contribute to increase the user’s trust and confidence in the system (Jameson 2005). These specific interventions are not possible with approaches that detect emotions with no explicit knowledge of their reasons (e.g. D’Mello et al. 2008).

Another distinguishing feature of the OCC model is that it captures emotions that are instantaneous reactions to the world, as opposed to affective states such as *frustration*, *boredom*, *confusion* and *flow* that some emotion researchers may classify as *moods*, i.e., states that are less specific than simple emotions, less likely to be triggered by a particular stimulus or event, and longer-lasting (Thayer 1989). Several education technology researchers have focused their efforts on these affective states because there is evidence that they arise during learning (e.g. Craig et al 2004). We see these longer-term affective states as being complementary to the ones captured by the OCC model in that the instantaneous emotions can contribute to create the longer term affective states (Merabian 1996). Ideally a model of user affect should be able to capture all these different affective dimensions and their relationships, as (Gebhard 2005) proposed for a generative model of affect, but for the time being we decided to focus on capturing and reacting to instantaneous emotions as an indirect way to positively affect the user’s longer term affective states.

To apply the OCC theory to emotion recognition during human-computer interaction, our DDN includes variables for goals that a user may have during the interaction with a system that includes an intelligent agent, (nodes *Goals*² in Figure 1). The events subject to the user’s appraisal are any visible interface outcomes generated by the user’s or the agent’s action (nodes

² We currently represent players preferences in terms of goals, as suggested in (Gratch, 2000).

User Action Outcome and *Agent Action Outcome* in Figure 1)³. Agent action outcomes are represented as decision variables in the framework, indicating points where the agent decides how to intervene in the interaction. The desirability of an event in relation to the user's goals is represented by the node class *Goals Satisfied*, which in turn influences the user's *Emotional States*.

The user's goals are a key element of the OCC model, but assessing these goals is not trivial, especially when eliciting them with queries to the user during the interaction would be too intrusive, as is the case during game playing. Thus, our DDN also includes nodes to infer user goals from indirect evidence. User goals can depend on *User Traits* such as personality (Costa & McCrae, 1992). Also, user goals can influence user *Interaction Patterns*, which in turn can be inferred by observing the outcomes of individual user actions. Thus, observations of both the relevant user traits and action outcomes can provide the DDN with indirect evidence for assessing user goals.

The sub-network below the nodes *Emotional States* is the diagnostic part of the affective modeling framework, representing the interaction between emotional states and their observable effects. *Emotional States* directly influence user *Bodily Expressions*, which in turn affect the output of *Sensors* that can detect them. Because in many situations a single sensor cannot reliably identify a specific emotional state, our framework is designed to modularly combine any available sensor information, and gracefully degrade in the presence of partial or noisy information.

In Figure 1, the links between emotion nodes in different time-slices indicate how the corresponding variables evolve over time. These links model, for example, the fact that a user is more likely to feel a given emotion at time t_{i+1} if the user felt it at time t_i . A new time-slice is added to the network whenever either the user or the agent performs an action (this happens, for instance, between every 3 and 10 seconds in the framework application we describe in the following sections); the new slice represents the state of the world just after the corresponding action occurred. In a DDN, only the time-slices that directly influence the current state need to be maintained. We currently assume that maintaining two time-slices is sufficient to capture the relevant temporal dependencies in our framework. Since at the moment the only temporal variables in the framework are the emotion variables, this assumption implies that the user's emotions at any given time depend only on the last game action and his or her emotional state in the previous slice, while the effects of earlier actions on the current emotional state are channelled through the emotional state in the previous time-slice. This assumption would be invalid in situations where a sequence of actions directly causes a particular emotional reaction, rather than influencing it via a chain of subsequent emotional states. Our framework also assumes that a user's high-level goals do not change over time, as indicated by the lack of a link between the *Goals* node at time t_i and the *Goals* node at t_{i+1} in Figure 1. Both assumptions derive from our philosophy for tackling the complexity of modeling affect: start with reasonably simplified models and increment them as limitations are uncovered by empirical evaluations.

Having described the general framework underlying our model of user affect, we will now present the educational game we used to apply and test the framework.

³ We explicitly model action outcomes rather than action themselves because one individual action may generate several effects at once, each of which may be appraised in relation to a different goal (we provide examples of this scenario later in the paper). We don't need to include action nodes in addition to action outcome because we assume that the visible effects of an action are deterministic. Thus, the occurrence of an action is implicitly represented by the description of its outcomes.

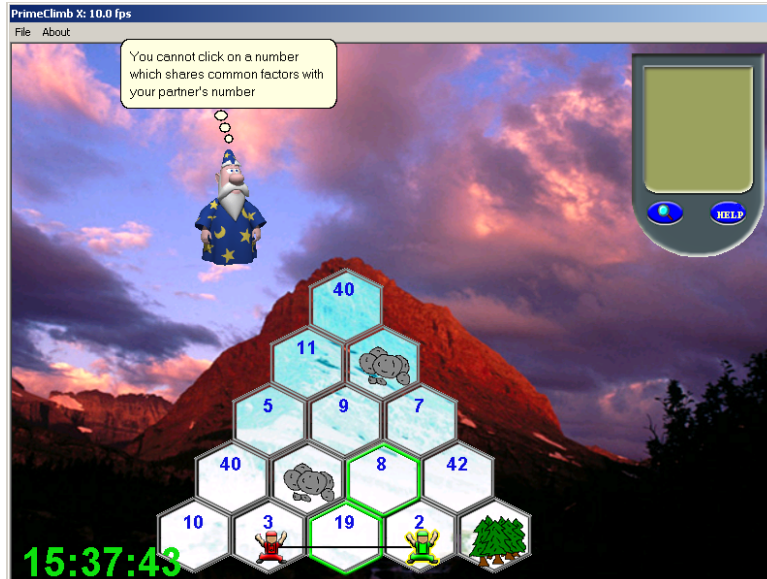


Figure 3. The Prime Climb Interface

2 The Prime Climb Educational Game

As a test-bed for the general affective modeling framework described in the previous section, we used Prime Climb, an educational game designed by the EGEMS group at the University of British Columbia to help 6th and 7th grade students practise number factorization. Figure 3 shows a screenshot of Prime Climb. Two players must cooperate to climb a series of mountains that are divided in numbered sectors. Each player should move to a number that does not share any common factors with her partner’s number, otherwise she falls. Prime Climb provides two tools to help students: (i) a *magnifying glass* that the student can use to see a number’s factorization (accessible by clicking on the magnifying glass icon at the bottom-left corner of the hand-held device in Figure 3); (ii) a *help box* (accessible by clicking on the *help* icon at the bottom-right corner of the hand-held device) that allows the student to ask for advice, which is provided by the pedagogical agent we are building for the game.

The pedagogical agent is an autonomous agent that provides individualized support, both on demand and unsolicited, when the student does not seem to be learning from the game (Conati & Zhao, 2004). To decide when to intervene and what hints to provide, the agent relies on a probabilistic model of the player’s factorization knowledge which is continuously updated during the player’s interaction with the game. When the probabilities in the model of student learning indicate that the player is missing key pieces of knowledge to learn from her current move, the pedagogical agent provides hints designed to stimulate the student to reason about the relevant domain knowledge. This can happen even after a student’s correct move, if the underlying student model predicts that the successful move was based on luck rather than knowledge. When the player falls, the agent provides hints in three incremental levels of detail, examples of which are shown in Figure 4:

1. At the most general level, the agent’s hints include reminders to think about number factorization (as shown in Figure 4a) or to think about common factors when climbing.
2. At a second level, the agent suggests that the player uses the magnifying glass to see a number’s factorization, e.g. “Do you need help? Use the Magnifying glass” (Figure 4b).
3. The hints in the last level include examples of common factors between two numbers (as shown in Figure 4c) and of number factorizations.

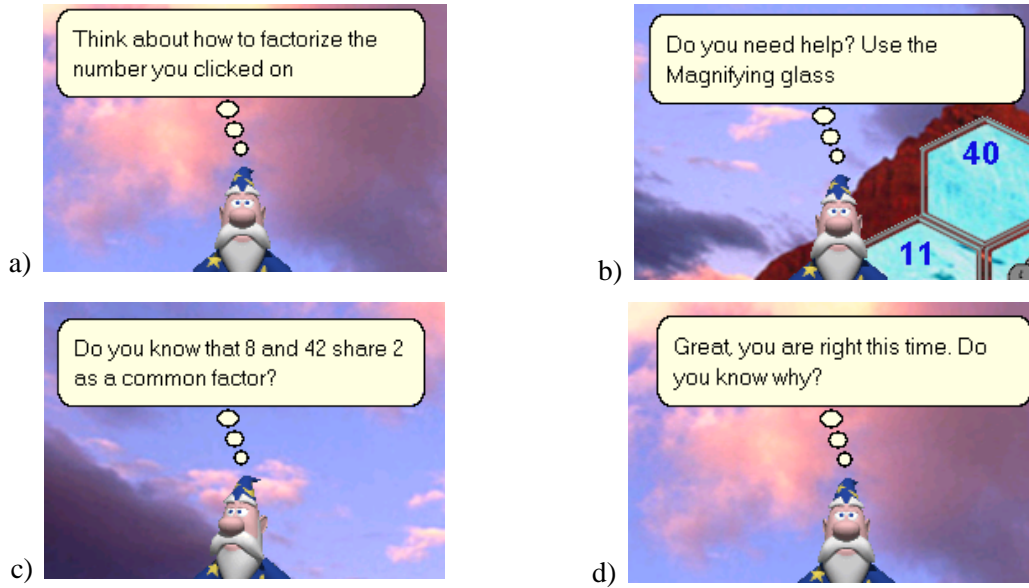


Figure 4. Merlin providing different levels of hint. (a) a general level hint, (b) a second level hint, (c) a final level hint, and (d) a hint after a successful move.

When the player makes a successful move that it is believed to be due to a lucky guess rather than knowledge, the agent attempts to stimulate reasoning about the target domain by asking the player if she knows why the move she has just made was correct (Figure 4d). The agent also occasionally attempts to encourage the student by congratulating her when she is successful.

To avoid interfering with the student’s level of engagement while playing the game, we used the framework described in the previous section to build an affective user model for Prime Climb that the agent can use to decide when and how to intervene. This user model produces a real-time assessment of the player’s emotions during interaction with Prime Climb, and will eventually be integrated with the model of student learning to inform the agent’s pedagogical decisions.

3 Building the Predictive Component of the Prime Climb Affective Model

In this section, we describe how we instantiated the predictive component of the affective modeling framework in Figure 1 to model the affective states of a Prime Climb player. We present two sub-network structures within the predictive component; the first assesses the student’s goals (*goal assessment sub-network*); the second models the student’s appraisal of game events in relation to those goals, to produce an assessment of the student’s current affective state (*appraisal sub-network*).

3.1 Instantiation of the Goal Assessment Sub-network

Figure 5 shows the structure of the sub-network that assesses student goals. Because all of the variables in this sub-network are observable either during or after the interaction with Prime Climb, we identified relevant individual variables and built the corresponding conditional probability tables (CPTs) using data collected through a series of Wizard of Oz studies where pairs of students interacted with the game while an experimenter controlled the pedagogical agent. Here we give a high-level description of this process. For more details see (Zhou & Conati, 2003).

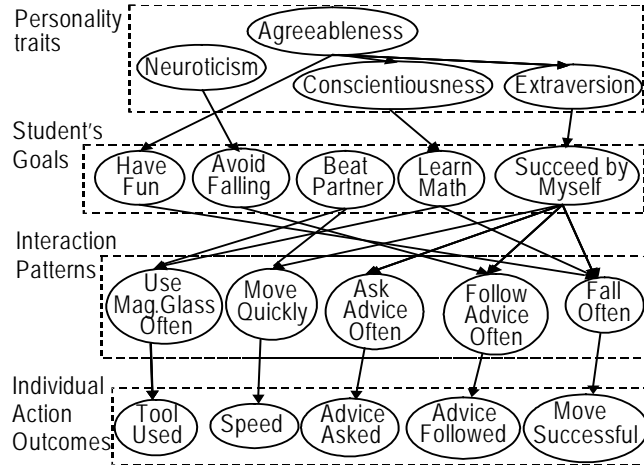


Figure 5. Sub-network for Goal Assessment

Information to instantiate variables representing student goals was collected via a post-game questionnaire in which students could express the goals they had while playing the game. We identified five high-level goals in our user studies, represented in the model by the following binary variables (with Boolean values indicating having/not having a given goal): *Have Fun*, *Avoid Falling*, *Beat Partner*, *Learn Math*, and *Succeed By Myself*⁴. We also found that students can have more than one of these goals at the same time. For this reason, the affective model represents each goal through a dedicated node rather than as one of the five mutually exclusive values on a single variable.

Because personality is known to influence one's goals and behaviours (Costa & McCrae, 1992), our model contains nodes and links representing student personality types and their relation to student goals in playing Prime Climb. We used the personality types suggested by the Five-Factor Model (Costa & McCrae, 1992), in which personality traits are structured as five domains – *neuroticism*, *extraversion*, *openness*, *agreeableness* and *conscientiousness*. Data to instantiate the prior and conditional probabilities involving these variables was collected through a personality test specifically designed for children (Graziano et al., 1997). As Figure 5 shows, our affective model currently includes variables for only four of the five domains, because our study data showed that *openness* was not directly relevant to our task. All of the personality variables are binary, with Boolean values representing whether or not a student belongs to a given personality domain.

During the studies, we also collected log files of the interactions, to mine the possible relationships between student goals (assessed via the goal post-questionnaire) and interaction behaviours. Our data indicated several dependencies between student goals and playing behaviour. The interaction patterns we identified to be relevant for inferring student goals included: (1) a tendency to make moves quickly or slowly (represented by the node *Move Quickly*); (2) a tendency to use the magnifying glass often or not (node *Use Mag. Glass Often*); (3) a tendency to ask the agent for advice often or not (node *Ask Advice Often*); (4) a tendency to follow the agent's advice often or not (node *Follow Advice Often*); (5) a tendency to fall often (node *Fall Often*). All the Interaction Pattern nodes are binary, with Boolean values indicating the presence/absence of a given pattern.

⁴ The goal *Beat Partner* is inconsistent with the collaborative nature of the game, but it is not surprising given findings indicating that certain personality types tend to be competitive even during collaborative interactions.

The probabilistic dependencies among goals, personalities, interaction patterns and individual student actions were established through correlation analysis between the personality test results, the goal questionnaire results and student actions logged during the interactions (Zhou & Conati, 2003). Figure 5 shows the resulting sub-network, incorporating both positive and negative correlations. The bottom level specifies how interaction patterns are recognized from the relative frequency of individual action outcomes (Zhou & Conati, 2003).

We originally intended to represent different degrees of personality type and goal priority by using multiple values in the corresponding nodes. However, we did not have enough data to populate the larger CPTs that this would generate, thus all the nodes in the goal assessment sub-network are binary. This simplification has not proven to be particularly detrimental to the performance of the goal assessment sub-network, as we will see in the model evaluation section to come. However, the resulting assumption that all goals have the same priority when present, together with the assumption that goals do not change over time, do affect the accuracy of the model’s assessment of student emotions, as we will also discuss in the evaluation section.

3.2 Instantiating the Appraisal Subnetwork

Figure 6 and Figure 7 show the details of the two types of time-slices used in the part of the network representing the *appraisal mechanism* (i.e., how the mapping between student goals and game states influences student emotions). Figure 6 shows the appraisal time-slice that is added to the affective model whenever the student performs an action. Figure 7 shows the time-slice added to the affective model whenever the pedagogical agent intervenes. Note that, for clarity purposes, Figure 6 and Figure 7 do not include the personality and interaction nodes used for goal assessment. The reader can refer to Figure 1 for an integrated picture of the goal assessment and appraisal sub-networks. For both types of appraisal time-slices, we specified an initial network structure based on the general OCC appraisal mechanism and our intuition, and then refined the structure by using empirical data collected from user studies designed for this task (described in Section 3.2.2.2). In this section, we first describe the structure of the initial network (corresponding in both figures to the solid-line nodes and links). We then describe the parts of the sub-network that were refined using empirical data (dashed-line components in both figures).

3.2.1 Initial Structure

The appraisal sub-network currently represents only 6 of the 22 emotions defined in the OCC model. They are *joy/distress* for the current state of the game, *pride/shame* of the student toward herself, and *admiration/reproach* toward the agent. These six particular emotions were chosen because we observed them often during pilot studies with Prime Climb, thus they seemed highly relevant for directing the actions of the Prime Climb pedagogical agent. While other emotions in,

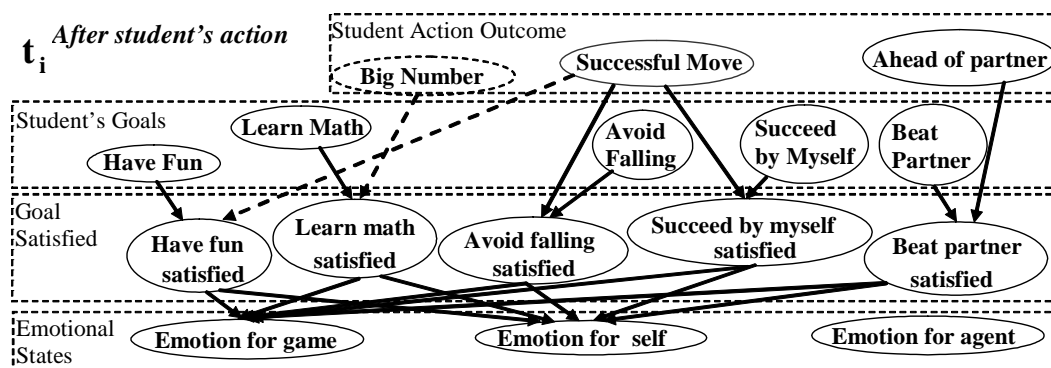


Figure 6. Sub-network time-slice for appraisal of student actions

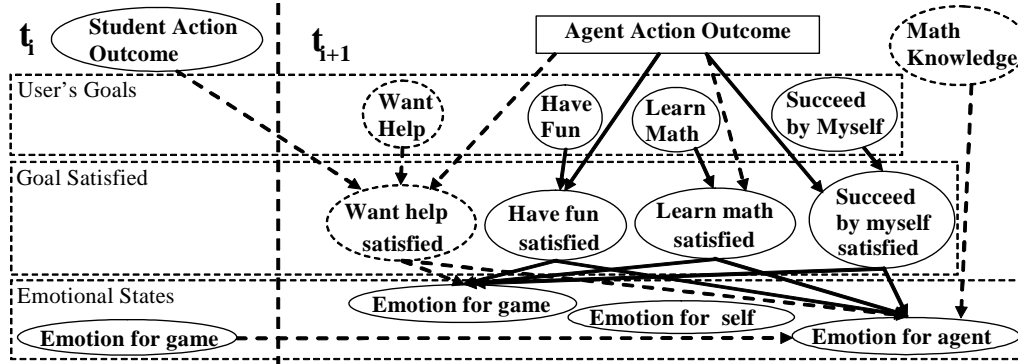


Figure 7. Sub-network time-slice for appraisal of agent interventions

the OCC model may be relevant, for instance emotions toward one’s partner during game play, we decided to start with a relatively simple model and progress to more complex ones only after having ascertained the viability of our approach.

Each of the three emotion pairs included in the model is represented by a binary node — *emotion-for-game*, *emotion-for-self* and *emotion-for-agent*, respectively (see nodes in the *Emotional States* level in Figure 6 and Figure 7) — where binary values represent the probability that the student is feeling one of the two emotions in the corresponding pair. This structure was chosen because, while the two emotions in each pair are mutually exclusive and are thus best represented by a binary node, students may simultaneously feel emotions in the different pairs, requiring a separate node for each pair.

Following the OCC appraisal model, a student’s emotional state depends on whether her goals are satisfied or not during game playing. In the appraisal network, goal satisfaction is explicitly represented by a *Goal Satisfied* node for each goal in the goal assessment network (see nodes in the *Goal Satisfied* level in Figure 6 and Figure 7). The links between *Goal Satisfied* nodes and the emotion nodes are defined as follows. We assume that the outcome of every relevant agent or student action is subject to student appraisal. Thus, each *Goal Satisfied* node influences *emotion-for-game* in every slice. Whether a *Goal Satisfied* node influences *emotion-for-self* or *emotion-for-agent* in a given slice depends upon whether the slice was generated, respectively, by a student action (Figure 6) or an agent’s action (Figure 7). Each *Goal Satisfied* node has three possible values: true, false, and neutral. The CPTs for emotion nodes are defined so that the probability of each positive emotion is proportional to the number of true *Goal Satisfied* nodes.

The probability of each *Goal Satisfied* node depends on whether the outcome of the current student or agent action matches the corresponding student’s goal. In the initial appraisal sub-network, the links and CPTs between *Goal* nodes, the outcome of student or agent actions, and *Goal Satisfied* nodes were based on our intuition, and defined connections that are quite obvious.

Let’s start by looking at these links in the time slice for the appraisal of student action outcomes (Figure 6). Initially, this time slice included only student moves on the Prime Climb mountains as actions that trigger the appraisal mechanism, because the other two possible student game actions (using the magnifying glass and asking for help), were not seen very often during our studies and thus we did not have a clear sense of how they might influence student affect. The solid binary nodes in the *Student Action Outcome* level at the top of Figure 6 represent two different aspects of the outcome of a student move that we had observed to trigger student emotional reactions during game playing. The node *Successful Move* indicates whether the student’s move was successful or not. The node *Ahead of Partner* indicates whether or not the move brought the student to be ahead of her partner on the mountain. We encoded some intuitive dependencies between these action outcomes and student goals in the initial appraisal network. For instance, if the student has the goal *Avoid Falling*, a successful move likely satisfies it, while

a fall likely does not. If the student has the goal *Beat Partner*, only a move that brings the player ahead of the partner on the mountain is likely to satisfy this goal.

In the initial version of the time slice that models the appraisal of agent actions, the decision node *Agent Action Outcome*⁵ was a four-valued node that represented the types of intervention that the agent could produce (see Section 2). These interventions were represented by the following decision values: (1) generate a hint at the first or second level of detail (e.g., a reminder to think about common factors when climbing or a suggestion to use the magnifying glass) after a student's fall; (2) generate a hint with example following a fall; (3) generate a hint to stimulate reasoning after a successful climb; (4) generate an encouragement. We used the same node value for the first two levels of hints following a fall because we observed in previous studies that the students tended to express similar reactions to these hints, thus we hypothesised that the students were appraising the hints the same manner. Reflecting this similarity in the decision node enabled us to reduce the complexity of that part of the network structure.

We encoded the following intuitive dependencies between agent actions and the satisfaction of student goals in this time slice. If the student has the goal *Have Fun*, providing encouragement will likely satisfy this goal, whereas providing any of the other more pedagogically oriented hints likely will not. If the student has the goal *Succeed By Myself*, providing any of the pedagogically oriented hints likely will not satisfy this goal (when the agent provides encouragement, goal satisfaction is neutral).

For the intuitive links described above, the conditional probabilities of the *Goal Satisfied* nodes were set by assigning: (1) a high probability that goal satisfaction is true when the student has the goal and an event that satisfies it occurs; (2) a high probability that goal satisfaction is false when the student has the goal and the opposite of a satisfying event occurs; and (3) a high probability of goal satisfaction being neutral when the student does not have the goal.

However, we had no good intuition of how various student game actions would be appraised in relation to the goals *Have Fun* and *Learn Math*. We also had no good intuition of how agent actions would be appraised for the goal *Learn Math*, since the appraisal should reflect the student's perception of whether he/she learned math rather than whether this was what actually happened. Thus, we decided to base these appraisals of student and agent game actions on empirical data.

3.2.2 User Study to Refine the Appraisal Sub-network

The primary goal of the study described in this section was to collect data from students to refine the model's event appraisal mechanism described in the previous section. We also wanted to reuse data from this study to evaluate the resulting predictive model, as we will describe in Section 4. Thus, we set up the study to obtain, among other things, a reliable measure of the user's affective states during the interaction, for comparison with the model's assessment. In this section we first describe the method for labelling student affective states that we used in the study. Next, we describe the general study design.

3.2.2.1 Collecting Affective Self-Reports While Playing Prime Climb

Collecting reliable data on actual users' emotions during real-time interactions is difficult, especially when the emotions are ephemeral and can change many times during the interaction, as we observed to often be the case with Prime Climb.

When emotions are varied and rapidly changing, it is hard for the users to describe them by using post-treatment self-reports, as was done for instance in (Lisetti & Nasoz, 2004; Peter &

⁵ We keep the label "Agent Action Outcome" for consistency with the slice for Student Action Outcome. In practice however, agent actions and their outcomes coincide because, unlike student actions, every agent action has a single outcome.

Herbon, 2006) to label valence/arousal or the occurrence of one specific emotion. When D’Mello et al (2008) tried this method to discriminate among various user emotions during interaction with an educational system, post-treatment user self-reports had one of the lowest intercoder reliability as compared to other collection methods.

Another commonly used method to label user emotions is to record participants using a video-camera and then ask observers to review the video to produce annotations of emotions visibly expressed during the interaction. This method has been shown to work well to measure different levels of a single emotion such as interest (Kapoor & Picard, 2005), or to recognize clearly separated emotions (D’Mello et al. 2008). However, when we tried to use it in our research, we found that observers often had a hard time discriminating equally-valenced feelings in our emotion set (e.g., discriminating between reproach toward the agent and distress toward the game). Thus, we decided to collect emotions self-reports directly from students *during* the interaction, similarly to (deVincente & Pain, 1999). In this work, a slider-based interface was used to get university students to volunteer information on their motivational state while interacting with an intelligent tutoring system. One of the study’s outcomes was that students did not volunteer information frequently (an average of 3.5 times for an interaction of about 15 minutes). Since we are dealing with much younger subjects, we were concerned that this phenomenon would be even more pronounced if we used the same approach based on volunteered student reports. Thus, we modified the method so that it can elicit self-reports from the students more frequently and provide sufficient data for model construction and evaluation.

Following (deVincente & Pain, 1999), we provide an emotion-report dialog box permanently present on the side of the Prime Climb game window, for students to volunteer self-reports on their emotional states (see Figure 8 and Figure 9). However, the dialog box also pops up whenever either one of the following conditions is satisfied: (1) the student has not entered any emotion in the permanent dialog box for a period of time longer than a set threshold or (2) the underlying affective model detects a relevant change (also based on a set threshold) in what it believes to be the student’s emotional state. The pop-up dialog box is necessary because a preliminary study confirmed our fears that students do not volunteer enough emotion self-reports via the permanent dialog box (Conati, 2004). The thresholds that influence the appearance of the pop-up box were adjusted through pilot studies to balance the amount of data that it allows us to collect and the level of interference that it generates during game playing (Conati, 2004).

As Figure 9 shows, the emotion dialog box only elicits information on two of the three pairs of emotions targeted by our model (emotions towards the game and emotions towards the agent). We chose this design because we felt that dealing with three pairs of emotions would be too confusing for our young subjects, and because sixth and seventh grade teachers suggested that students would have more problems in reporting emotions toward themselves than toward the game or the agent.

Data from a post-questionnaire on interface acceptance, which 20 students filled in as part of a study to test the final version of the self-report mechanism, showed good user acceptance (Conati, 2004). For instance, the students’ average ratings (on a Likert scale where 1 = strongly disagree, and 5 = strongly agree) for the statement “The popup dialog box interfered with my game playing” was 2.8 (st. dev. 1.4), while the average ratings for “It bothered me having to tell the system how I feel” was 2.1 (st. dev. 1.1). We also found that the negative emotions self-reports were only a small fraction of the self-reports generated by the students who reported annoyance with the dialog box. These results suggest that, even when subjects expressed annoyance with the dialog box, this annoyance did not necessarily translate into annoyance with the game or the agent. These findings are quite encouraging for researchers interested in evaluating affective models, because they indicate that subjects can tolerate to some extent the interference caused by the artefacts designed to elicit their emotions.

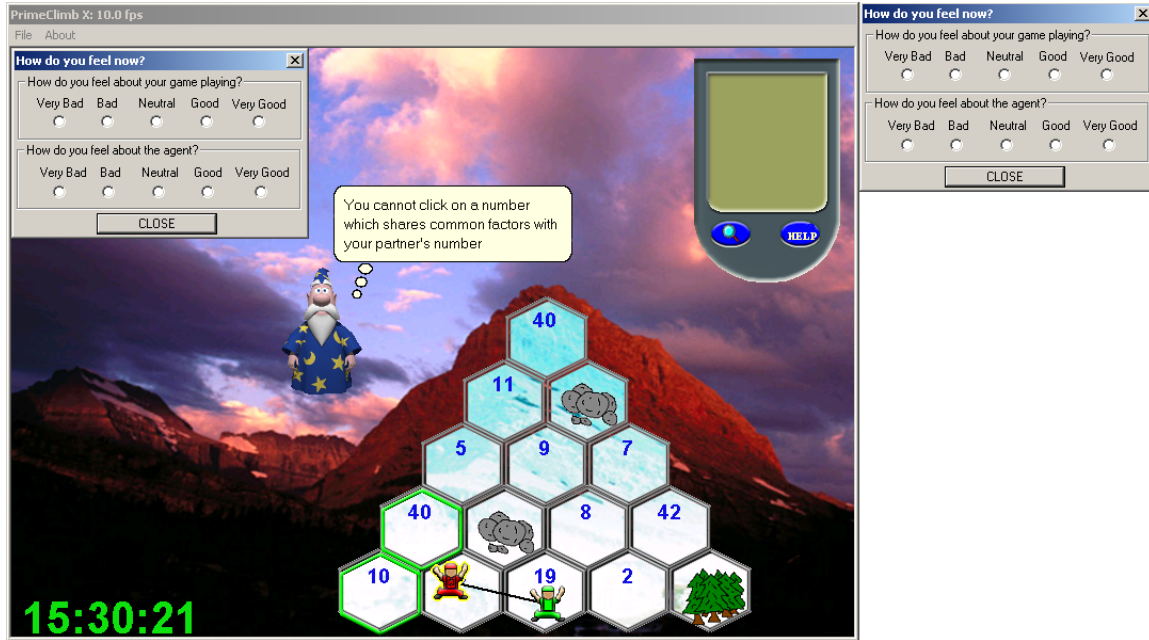


Figure 8. Interface with both the permanent and pop-up emotion-reporting dialog box

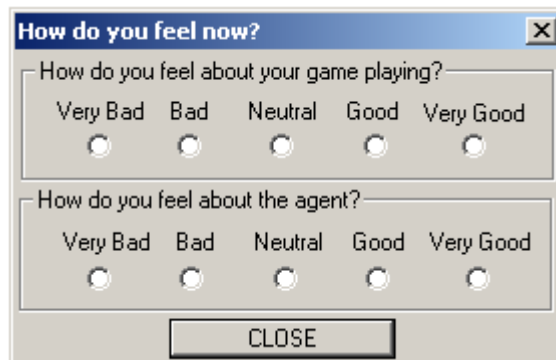


Figure 9. The dialog box presented to the students

3.2.2.2 Study Design and Materials

Sixty-six 6th and 7th grade students from three local schools participated in our study. The study took place in the schools, with the constraint that each study session had to be held during a class period (40 minutes) to avoid disrupting regular class schedules. Because of limited computer availability, we could only run two students at time. The two students were excused from the class for that period and joined the experimenters in a room provided by the school for the experiment. Each session was designed to last at most 30 minutes so that there would be sufficient time for students to get to the study room and return to their class for the next period. Students first took a pre-test on factorization knowledge. Next, they were told that they would be playing a computer game, and received a demo of Prime Climb with the emotion self-report mechanism. They were told that the game contained a computer-based agent that was trying to understand their needs and help them play the game better. The students were encouraged to provide their feelings whenever their emotions changed so that the agent could take them into account when providing help. We did not deem it necessary to provide any further way to engage the students in the task because, from the several studies we had already run on Prime Climb, it



Figure 10. The study setup

was apparent that the mere fact of playing a computer game during school time was sufficient to greatly engage the students, at least for the short period of playing time necessary for the study. This first phase of the experiment lasted at most 10 minutes.

Next, participants played Prime Climb for about 10 minutes. Each student played with an experimenter as a climbing companion. Due to time and space constraints, we had to run two experimenter/student pairs in parallel in the same room, sitting side by side, as shown in Figure 10.

We did not make students play together because we wanted to avoid the extreme emotions toward the playing partner that we often observed with that set-up, given that our affective model currently does not model these emotions. To further limit the impact of students' feelings toward their climbing companion, experimenters were instructed to play as neutrally as possible, trying to avoid making mistakes (although mistakes did happen on some of the mountains with larger numbers) and to avoid leading the climb too much. Furthermore, students did not know which of the two experimenters they were playing with. They were told that the game randomly assigned their partner, so that the partner was not necessarily the experimenter sitting across from them. This measure reduced the student's tendency to make eye contact or attempt to verbally communicate with the experimenter.

To reduce as much as possible the distraction generated by having students sitting side by side, students were reminded before the beginning of each game that they would not be playing with each other. Experimenters noted that some students did glance across to check the progress of the other student's game, but on most occasions this occurred at the end of a game level while they were waiting for the next level to load. However, if one student was observed to be particularly disruptive and disturbed the other student in the room, then the data from both students was discarded. In practice, this happened once or twice in each school. The discarded students were not included in the count of student participants mentioned above.

During game playing, the Prime Climb agent generated pedagogical interventions to help the student learn from the game, by relying on the model of student learning mentioned in Section 2 (Conati & Zhao, 2004). All of the agent's and student's actions were captured by the version of the affective model with the initial appraisal network described in Section 3.2.1. The model was updated in real time to direct the appearance of the pop-up dialog box, as described in Section 3.2.2.1, but the pedagogical agent did not use it to direct its interventions. It should be noted that this is the reason why we can re-use log files from this study to evaluate various versions of the affective model, as we describe in Section 4. It should also be noted that the study is not a Wizard of Oz, since the agent acts autonomously, even if it does not use the affective model, and the experimenter plays the role of a human player.

Log files of the interaction recorded all of the events that occurred within the game, the student's reported emotions, and the corresponding model assessments. After game playing, students completed a post-test on number factorization (collected for purposes not related to this research) and four post-questionnaires: one on interface acceptance, one to indicate the goals they

had during game playing (including 13 questions of the type “I wanted to GOAL while playing Prime Climb”) and the two post-questionnaires (appraisal questionnaires) on the events that affect *Have Fun* and *Learn Math*, including 15 questions of the type described below. Having students answer these goal-related questions during game playing would probably generate more accurate information, but it would have been too disruptive in our set-up, given that the players were already been interrupted to provide the emotion self-reports. We felt that students could still reason about their goals right after the study’s rather short game playing sessions, because the goals in question are quite high-level and not as transient and ephemeral as emotions are (although they can still change during interaction, as we will see in a later section).

Each post- test and post-questionnaire was designed and pilot-tested so that this final phase of the experiment lasted at most 10 minutes. As with the emotion self-report collection mechanism, we did our best to strike a balance between the amount of data we needed for model construction/evaluation and avoiding student fatigue that could make the data unreliable. We now provide a description of the two appraisal questionnaires because it is necessary to understand how we used the data to refine the appraisal component of the predictive user model.

Each appraisal post-questionnaires contained a list of statements of the type ‘*I learnt math/had fun when <game event>*’. Students rated each statement using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). The game events considered in the questionnaires included the following:

For *Have Fun*

- *Student-generated events*: a successful move; a fall (unsuccessful move); using the magnifying glass; using the help box; reaching the top of the mountain.

For *Learn Math*

- *Student-generated events*: same as above, plus following the agent’s advice, and encountering big numbers.
- *Agent-generated events*: reminder to think about common factors when climbing (level 1 hint); suggestion to use the magnifying glass (level 2 hint);

When considering which events to include in the questionnaires, we aimed to investigate as many game events as possible that we thought could influence the satisfaction of student goals, while keeping the questionnaire at a reasonable length. For instance, in *Agent-generated events* we included separate events for hints from the first and second levels of hint detail (as described in Section 2) rather than a single event (as specified in the *Agent Action Outcome* node described in Section 3.2.1), since data on these two levels of hint would enable us to test our initial hypothesis that these agent actions were appraised in a similar manner by the student. However, we did not include hints that provided examples of number factorization or computation of common factors, or hints generated after a successful move, because the agent version used in the study rarely generated these hints and thus most students would not be able to rate them. In addition, we restricted the number of hints included in the questionnaire to one hint per level of detail, and therefore did not include the reminder to think about number factorization.

While some of the *Student-generated events* were included in the questionnaires because they pertained to the three basic Prime Climb actions (a successful/unsuccessful move, using the magnifying glass, using the help box), other events such as *encountering big numbers* and *reaching the top of the mountain* were included based on anecdotal evidence from experimenters who had run previous user studies; we had not included them in the initial networks because the anecdotal evidence was insufficient to insert them in a meaningful way. To limit the length of the questionnaire, we did not ask students about events that already satisfied other goals within the model (e.g., the student being ahead of her partner, encouragement by the agent).

3.2.3 Data-based Refinement of the Appraisal Sub-network

In both Figure 6 and Figure 7, dashed lines indicate the parts of the network defined using the empirical data from the study described in the previous section. We will now describe how we generated all of the refinements to the appraisal time-slices beginning with the time-slice used to appraise the outcome of the student's action.

3.2.3.1 Appraisal of Student Action Outcomes

The students' answers to the appraisal questionnaires indicated that all of the tested student-generated events were relevant to some degree. In order to determine which events made a difference to the model's assessment, we scored all possible network structures derived from including these events, using their log marginal likelihood (Heckerman, 1999). The mutually exclusive events *successful move/fall* were represented via the binary values of the node *Successful Move* in Figure 6, while the events *using the magnifying glass*, *using the help box*, *reaching the top of the mountain* and *moving to a big number* were each represented by a new binary node in the time-slice for appraising student actions.

We found that the structure with the best score was the one encoding the following appraisal relations, in addition to those represented in the original network: (1) whether the student's move was successful or not influenced satisfaction of the goal *Have Fun*; and (2) whether the student encountered a big number influenced satisfaction of the goal *Learn Math*. The dashed components in Figure 6 show how these relations are encoded in the appraisal time-slice. The new binary node *Big Number* is linked to satisfaction of the goal *Learn Math* while the existing node *Successful Move* is linked to satisfaction of the goal *Have Fun*. We used frequencies from the questionnaire answers to set the CPTs for these new links. We based our definition of a big number on the large numbers frequently incorrectly factorized in students' pre-tests in our studies.

3.2.3.2 Appraisal of Agent Actions Outcomes

The data-based refinement of the time-slice added after an agent action consisted of two stages:

Stage 1. First, we used the students' questionnaire items related to the influence of the agent's actions on the goal *Learn Math* to test our hypothesis that hints from the first two levels of hint detail following a fall should be represented by a single value within the node *Agent Action Outcome*. To do this we created two candidate network structures to represent the student's appraisal of the agent's action with regards to the goal *Learn Math*. These structures were identical except for the number of possible values in the node *Agent Action Outcome*. The first structure contained an *Agent Action Outcome* node with the four values we had initially specified using subjective heuristics. The second structure contained an *Agent Action Outcome* node where the unique value representing two levels of hint had been replaced by two separate values: *generate a hint at the first level* and *generate a hint at the second level*. We used the students' questionnaire answers to produce a log marginal likelihood score for each network structure. We found that the structure containing the *Agent Action Outcome* node with our original set of four values received the highest score and thus we retained it in the refined model.

We then refined the model by creating a link between *Agent Action Outcome* and satisfaction of the goal *Learn Math*. For the pedagogical agent actions that had been included in the questionnaire, we used the frequencies from the questionnaire answers to generate the corresponding CPT values. For the pedagogical agent actions that had not been included in the questionnaire due to their rarity (e.g., hints with examples following a fall, a hint following a successful climb) we set the CPT values for goal satisfaction to equal probability for true and false. For encouragement by the agent, a non-pedagogical action, we set satisfaction to neutral.

However, a preliminary evaluation of these changes showed that the model was underestimating students’ admiration toward the agent, suggesting that the model still contained sources of inaccuracy related to appraisal of agent actions. We therefore moved to a second stage of data analysis to determine these problems.

Stage 2. This stage makes use of the emotion self-reports that students generated during the study using the emotion dialogue box (Section 3.2.2.1). To better understand the relations between agent behaviours and student emotions, we analyzed the study log files to identify particular situations in the game in which students tended to report experiencing *Admiration* or *Reproach* toward the agent. Our data confirmed that encouragement by the agent generates students’ admiration (45% *Admiration* reports against 7% reports of *Reproach* and 48% neutral reports), although we cannot tell whether this happens through the satisfaction of the goal *Have Fun* as we have encoded in initial time-slice for appraisal of agent actions. It also showed that students who are generally successful usually report either *Admiration* or *Neutral* feelings towards the agent, regardless of their goals (53% *Admiration* reports against 8% reports of *Reproach* and 39% neutral reports). This finding suggests that the students’ positive feelings toward the game will positively influence their attitude towards the agent. We translated this finding into the model by adding a link from the student’s emotion towards the game in the previous time-slice to the student’s emotion towards the agent (as shown by the dashed line at the bottom of Figure 7).

Finally, we looked at situations in which students fell repeatedly and either received help or did not. Analysis of these situations revealed that approximately one-half of the students who reported *Admiration* when the agent intervened after they fell had declared the goal *Succeed By Myself*. Also, about one-half of the students who reported *Reproach* when the agent did not intervene had declared that goal. This result seems to indicate that, although some of the students may want to succeed by themselves in general, they may also want help in especially critical situations (e.g., when they fall repeatedly). That is, in these situations some students may reduce the priority of wanting to succeed by themselves in favour of wanting help. The data also revealed students who had not declared the goal *Succeed By Myself*, but when they began to fall they demonstrated annoyance when the agent intervened. That is, in these situations, the students demonstrated that they preferred to succeed by themselves rather than wanting help. These observations invalidate two of the choices previously made in the model implementation: (1) to ignore goal priority; and (2) to assume that goals are static during the interaction. Because we currently don’t have enough data to model goal evolution in a principled way, we only addressed the implementation of multiple priority levels to model the relation between *Succeed By Myself* and wanting help. We changed the model as follows.

First, we added an additional goal, *Want Help*. Note that we did not represent this goal as one of the two values of the node *Succeed by Myself* because, as we discussed above, these goals are not necessarily mutually exclusive. For some students, they seem to represent a general vs. local attitude toward receiving help during game playing, and thus they may co-exist, although with different, possibly shifting priorities. The satisfaction of *Want Help* is dependent on two factors: the outcome of the student’s move (i.e., a successful climb or a fall) and the agent’s action. When the student falls, *Want Help* can only be satisfied if the agent provides help. If the student does not fall, then satisfaction is neutral.

Second, we tried to determine which traits influenced the students’ attitudes towards receiving

Table 1. Confusion matrix comparing students’ math knowledge with whether they wanted help.

		Math Knowledge	
		High	Low
Want Help	Yes	13	4
	No	4	9

help during repeated falls. The only factor that seemed to play a role was students' math knowledge, a factor that we measured using pre-tests on factorization as part of our standard study design. Table 1 shows a confusion matrix comparing the students' math knowledge and whether they demonstrated that they wanted help when falling repeatedly. We classified the students' math knowledge as 'high' if they correctly answered 50% or more of the questions on the factorization pre-test, otherwise the math knowledge was classified as 'low'. As the matrix shows, high math knowledge is associated with wanting help, whereas low math knowledge is associated with not wanting help. A Fisher test (Fisher, 1935) between the students' pre-test scores and whether they demonstrated that they wanted help after repeated falls showed a significant relationship (Fisher score = 0.025). Although this relationship seems backward, the results agree with the findings of Baker et al. (2004) that students with lower pre-test scores are more likely to want to succeed via trial and error than think about domain knowledge, whereas students with higher pre-test scores are more likely to want to learn from the resources available in the system, including provision of help. Given the above findings, a new node, representing prior math knowledge, was used to influence the priorities a student gives to the goals *Succeed By Myself* and *Want Help*.

We added a link from the new node, *Math Knowledge*, to *emotion-for-agent* (as shown in Figure 7). As we mentioned earlier in Section 3.2, the CPT for *emotion-for-agent* was defined so that the probability of the student feeling *Admiration* was proportional to the number of true *Goal Satisfied* nodes. We refined the CPT in *emotion-for-agent* so that, if the student had high math knowledge, then the influence of the node *Succeed by Myself Satisfied* on the probability of *Admiration* was lower than the influence of the other *Goal Satisfied* nodes. If the student had low math knowledge, then the influence of the node *Want Help Satisfied* on the probability of *Admiration* was lower instead.

Our third and final change to the model was to refine the decision node representing the available agent's actions so that it included the agent choosing not to intervene. All *Goal Satisfied* nodes other than *Succeed By Myself* and *Want Help* were given a neutral satisfaction for this new action. *Want Help* was discussed earlier; *Succeed By Myself* was given a small probability of satisfaction to reflect possible mild positive feelings towards the agent for not interrupting in general, rather than at specific events.

4 Evaluation of the Affective Model

We evaluated the predictive part of the affective model in two stages. First, we evaluated the model's event-appraisal mechanism, independently from the performance of the sub-network for goal assessment. To do so, we assumed that the students' answers to the goals post-questionnaire were an accurate representation of the goals that they had during game playing. We used these answers to set the values of goal nodes in the appraisal network, rather than relying on the model's own assessment of student goals. Second, we tested the complete predictive network by repeating the evaluation with the model's own assessments of student goals during the interaction.

Before describing the results of each of the two evaluations, we illustrate the common evaluation method we used. We measured the performance of each model to be tested via a simulator that replays the event logs from the study described in Section 3.2.2.2 with that model. The simulator includes the execution of an additional '*no agent action*' event after each student action that was not followed by an agent intervention. This "no agent action" event had not been recorded in the original log files because its relevance was discovered through the data analysis for model refinement described in Section 3.2.3.2.

We performed cross-validation on model accuracy by using the following well-known random resampling method (Mitchell, 1997). We divided the set of students into a training set and a test

set of equal size using random selection. We then used the data from the students in the training set to train the necessary CPTs in the model, and ran the event logs of the students in the test set through the simulator to produce a measure of model accuracy (computed as we describe below). We then randomly divided the original set of students into a new training set and test set, and performed the same evaluation steps again. In total, we performed this procedure 100 times⁶.

For each test set, we measured model accuracy by computing how often the model’s assessment agreed with the student’s reported emotions at corresponding times. To enable the comparison, we translated both the students’ reports for each emotion pair (e.g., *Joy/Distress*) and the model’s probability over the emotion node corresponding to that pair (e.g., the node *emotion-for-game* in Figure 6) into 2 values: positive (indicating the element with positive valence in the pair, e.g., *Joy*) and negative (indicating the element with negative valence, e.g., *Distress*). A student report was classified as positive if it was higher than ‘neutral’ in the dialog box, and as negative if it was lower. The model’s assessment was classified as positive if the probability of the corresponding emotion node was higher than a set threshold, and negative otherwise. The threshold value of 0.65 was determined using the data from an earlier empirical evaluation (Conati & Maclaren, 2004).

For each emotion pair, we report individual model accuracies in detecting the positive and the negative emotions. These correspond to standard measures of true positive rate (or *sensitivity* of the classifier) and true negative rate (or *specificity* of the classifier), deemed as appropriate to discuss the accuracy of user models when the model’s performances in detecting positive vs. negative data points have different implications in terms of the adaptive behaviours the model is designed to support (e.g., VanLehn & Niu, 2001). This is generally the case in affective user modeling, where failing to detect and act upon a user’s negative affective state can have greater repercussions on the interaction than failing to detect a positive affective state. Since there is a trade-off between sensitivity and specificity, we also need a measure of combined accuracy that gives a better sense of the model’s overall performance. Two common choices include:

- the percentage of cases correctly classified over all the test instances (regardless of the class they belong to), also known as *micro-average*.
- the average of the accuracies for each class, also known as *macro-average*. In our case, the macro-average corresponds to the average of the specificity and sensitivity of the model for each emotion pair.

Micro averages are a commonly used measure for reporting the accuracy of a classifier over multiple classes, but they produce a somewhat biased picture in the presence of classes with an unbalanced distribution of instances, as the accuracy over classes with few data-points is overshadowed by the accuracy over larger classes. Macro averages are considered an adequate way to overcome this short-coming (e.g., Esuli et al., 2006; Sebastiani, 2002). The rationale is that a macro average allows one to give fair weight to classes with few instances, when it is important that these few instances are correctly detected. This is exactly the case for our model: in our studies, for each emotion pair we had far fewer negative than positive data points (see Table 2 in Section 4.1), however, it is crucial for the model to detect these negative emotions

⁶ We used this method for cross-validation because we did not have enough data (especially negative data-points, as we will see in a later section) to perform a traditional N-fold cross-validation, where the N test/training pairs are non-overlapping partitions of the data and N is large enough to allow for measures of statistical significance. The drawback of random re-sampling is that the test/training sets it generates are not independent and thus violate one of the assumptions required by standard tests for statistical significance. Thus, although random resampling is commonly used in machine learning to deal with limited data (Mitchell, 1997), any statistically significant results that it generates should be interpreted as significant trends.

since they may compromise the player's overall attitude towards the game. When using micro averages, the accuracy over these cases is washed out by the accuracy over the much larger number of data points with positive affect. Given the nature of our dataset, we argue that macro averages are a more appropriate measure of model's overall accuracy than micro averages, however in the subsequent sections we report both measures for the sake of completeness.

One suggested reason for the small number of negative reports is that the type of Prime Climb interaction in our study does not generate negative emotions very often, because it was shaped to focus on the player's emotions towards the game and towards the agents when the second player has a very stable and neutral behaviour. Given that both the game and the agent were designed to provide students with an enjoyable and profitable experience, few negative datapoints are a reasonable and desirable outcome (no negative emotions would be ideal but hard to achieve without perfect knowledge of what students want during game playing). While we could induce more negative data points on purpose by having the agent generate deliberately disturbing behaviours, we argue that the students' negative feelings in this case would not be indicative of the real emotions that they would experience during interactions with our target agent, i.e. an agent that aims to make students learn. We will discuss this problem and make some suggestions as to other ways to collect more reports of negative emotions later in our conclusions in Section 6.

Finally, it should be noted that making a binary prediction from the model's assessment is guaranteed to disagree with any 'neutral' reports given by the students. The only way to fix this problem in the predictive network would be to add a third value to each emotion node that represents neutrality with respect to that emotion type. However, altering the emotion nodes' CPTs to include this additional value would not be trivial because the OCC model does not provide an obvious definition of neutrality. An alternative is to catch at least some instances of neutrality in the diagnostic part of the model. We found that 65 student reports were neutral for both *emotion-for-game* and *emotion-for-agent* (63% and 58% of the neutral *emotion-for-game* and *emotion-for-agent*, respectively). Because neutrality on both emotions corresponds to a low level of emotional arousal, this state should be easily picked up by adequate physiological sensors in the diagnostic part of the model (see Figure 1). This is a clear example of a situation where the observed evidence of a student's emotional state can be combined with predictive assessment to generate more accurate predictions on student affect.

4.1 Evaluation of the Event Appraisal Sub-network

Table 2 shows the results of using the mechanism discussed in the previous section to evaluate the refined appraisal network from Section 3.2.3. As we mentioned earlier, in order to eliminate possible confounding factors deriving from inaccuracies in the goal assessment network, the values of goal nodes were directly derived from the students' answers in the goal post-questionnaire. Although the goal *Want Help* was added to the model after the study and thus did not have a pre-dedicated item in the post-questionnaire, we were able to derive its value from the questionnaire item '*I wanted help when I became stuck*', originally used together with another item to assess the goal *Succeed By Myself*.

In order to assess how well our model performed compared to a simpler approach, we calculated the baseline accuracy of predicting the emotion with the highest probability based on the frequency of emotions occurring in the students' reports. Because our data set has a much higher number of positive data-points for each emotion pair (see Table 2) the baseline model always predicts *Joy* and *Admiration* and would thus have an accuracy of 100% in predicting these emotions, but 0% in predicting *Distress* and *Reproach*. The baseline accuracies, including micro averages and macro averages are also shown in Table 2.

Table 2: Model accuracy when goals are given as evidence

Emotion	Accuracy (%)				Data-points
	Model Predictions		Baseline		
	Mean	Std.Dev.	Mean	Std.Dev.	
Joy	69.59*	5.69	100	0.00	170
Distress	62.30*	16.26	0.00	0.00	14
Macro Average	65.95*	8.08	50.00	0.00	
Micro Average	68.72*	5.13	91.27	2.88	
Admiration	67.42*	6.58	100	0.00	127
Reproach	38.66*	9.31	0.00	0.00	28
Macro Average	53.04*	5.27	50.00	0.00	
Micro Average	60.95*	5.56	77.11	6.41	

* Significant increase/decrease from baseline accuracy

The model's macro average accuracy in predicting students' emotions towards the game is significantly⁷ higher than the baseline, with a large effect size ($t(99)=19.66, p<.001, d=2.79$). However, the model's macro average accuracy in predicting students' emotions towards the agent was reduced by the poor performance in predicting *Reproach*, and thus, although there is a significant improvement over baseline accuracy ($t(99)=6.03, p<.001, d=.82$), the effect size is smaller than for emotions towards the game. For both emotion pairs, the micro average of our model is significantly lower than the micro average of the baseline ($t(99)=-43.28, p<.001, d=-5.42$ for emotions towards the game and $t(99)=-23.84, p<.001, d=-2.69$ for emotions towards the agent), because the baseline's excellent sensitivity off-balances its non-existent performance on the negative data points. This result appears in all of the comparisons we perform in the rest of the paper; still it would be hard to argue for a model that cannot catch negative affect. Thus, while we will continue to report model micro averages throughout the paper, we will base all further discussions on accuracy solely on sensitivity, specificity, and corresponding macro average.

To understand the reasons for the rather poor model's performance on detecting *Reproach*, we engaged in a detailed analysis of the model's assessment in relation to the interactions simulated from the log files. This analysis showed that approximately 50% of the misclassified *Reproach* data-points, and approximately 28% of the misclassified *Admiration* data-points, were due to the fact that the students' declarations for the goal *Want Help* at the end of a game session did not seem to consistently match with whether they were trying to achieve this goal during the game. Five students did not declare the goal *Want Help*, but they reported *Reproach* towards the agent when they began to fall and the agent did not intervene, suggesting that they did want help in these situations. Three students declared the goal *Want Help* but then reported *Admiration* instead of *Reproach* when they fell repeatedly and the agent did not intervene, suggesting that they actually did not mind the lack of help. These findings confirm what we had already seen

⁷ To compare the model's performance to the baseline we used t-tests to determine the statistical significance of the differences (throughout the paper, we use .05 for significance and .1 for marginal significance). To compare the macro average accuracies we used a one-sample t-test, in all other cases we used a paired-samples t-test. We also measured effect sizes (the magnitude of the differences), using Cohen's *d* (Cohen, 1988), to determine the practical significance of the differences. We consider $d > .8$ to be a large effect, $.8 > d > .5$ to be a medium effect, and $d < .5$ to be a small effect, as per Cohen's standard.

from the data analysis in Section 3.2.3.2, i.e., that goal priority can change during the interaction. As we discussed in that section, we currently don't have enough data to model goal evolution in a principled way, and thus our model still includes the incorrect assumption of static goals and cannot model correctly those students who have shifting goals. The influence of this assumption on the *Reproach* inaccuracy reported here is amplified by the fact that goal nodes were set to deterministic values based on evidence from student questionnaires. Deterministic values have a higher negative influence than probability distributions on model assessment when they do not actually reflect the current student's goals.

A second factor that explains an additional 25% of the misclassified *Reproach* data-points is that using only previous math knowledge to help assess the relative priority some students gave to succeeding by themselves vs. receiving help incorrectly modeled four students. In each case the students reacted as expected for the goals they had declared, for example, one student had declared the goal *Succeed By Myself*, and had subsequently reported *Reproach* when the agent intervened after a repeated fall. However, in each case, the math knowledge of the student indicated that the model should give a low priority to the goal that was not satisfied by the agent's action. Thus the negative impact of giving help (or not giving help, in some cases) was underestimated. This result indicates that there are other traits that should be taken into account to correctly model priority shifts for some individuals.

4.2 Evaluation of the predictive model using Model's assessment of student goals

Our assessments of the accuracy of the predictive affective model have thus far been limited to the appraisal sub-network. That is, we have used student answers to the goal post-questionnaire as evidence for setting goal nodes in the appraisal network, so as to separate it from the model's goal assessment mechanism and isolate inaccuracies in the event-appraisal mechanism. However, information on students' goals will not be available when using the model in real-time during game playing. Instead, the model's own probabilistic assessments of the students' goals will be used. In order to determine how well the complete predictive model will perform during real-time interactions, we evaluated it by using the simulator described in Section 4 and allowing the model to use its own probabilistic assessments of the students' goals instead of the evidence from the students' post-questionnaires. We used the frequencies of goals declared by students from a previous study (Zhou & Conati, 2003) to help inform the model's goal assessments by creating population priors for each goal being assessed⁸.

Table 3 compares the accuracy of the model using goal evidence and using goal assessment with population priors. As the table shows, the model's performance using goal assessments increased significantly for *Distress* ($t(99)=4.82$, $p<.001$, $d=.55$) and for *Reproach* ($t(99)=5.29$, $p<.001$, $d=.75$)⁹, although in both cases the increase had only a medium effect size due to the high standard deviation, and the accuracy for *Reproach* is still below 50%. The model's performance for *Joy* and for *Admiration* decreased slightly, but not significantly ($p>.40$ in both cases). For the combined performance measures, Table 3 shows significant increases in both the model's macro average for emotions towards the game ($t(99)=5.06$, $p<.001$, $d=.50$) and the model's macro average for emotions towards the agent ($t(99)=4.48$, $p<.001$, $d=.61$), with a medium effect size in each case. Thus, from a practical standpoint, the trends for *Distress* and

⁸ Since goal nodes are not root nodes, population priors are included by (i) adding a fictitious root node as an additional parent for each goal node in the goal-assessment network; (ii) setting the CPT of the root node to the population prior for the goal.

⁹ All measures of statistical significance comparing different versions of the affective model are based on a two-tailed paired samples t-test with $p < 0.05$, $df=99$

Table 3. Comparing accuracy of the affective model when using goal evidence vs. goal population priors.

Emotion	Accuracy (%)						Data-points
	Goal evidence		Population priors		Baseline		
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.	
Joy	69.59*	5.69	68.73*	7.37	100	0.00	170
Distress	62.30*	16.26	70.34 [†] *	12.90	0.00	0.00	14
Macro Average	65.95*	8.08	69.53 [†] *	6.00	50.00	0.00	
Micro Average	68.72*	5.13	68.63 [†] *	6.38	91.27	2.88	
Admiration	67.42*	6.58	66.07*	9.70	100	0.00	127
Reproach	38.66*	9.31	46.69 [†] *	12.05	0.00	0.00	28
Macro Average	53.04*	5.27	56.38 [†] *	5.61	50.00	0.00	
Micro Average	60.95*	5.56	61.72*	7.03	77.11	6.41	

* Significant increase/decrease from baseline accuracy ; [†] Significant increase/decrease from model with goal evidence

Reproach are in favour of the probabilistic goal assessment, and the decreases for *Joy* and *Admiration* are small.

The good performance of the goal assessment mechanism should not come as a surprise given our previous discussion on the dynamic nature of student goals. When goal nodes are set as evidence, their values are fixed throughout the interaction, no matter what students do during game playing. When they are associated with a probability distribution, the distribution changes as evidence on student actions comes into the goal–assessment network. Although these changes cannot fully reflect changes in the students’ actual goals because the network models goal nodes as static, the changes in the probability distributions still approximate the students’ changing goals better than the immutable evidence values. Thus, for instance, the assessment of *Reproach* improves because using probabilities for the goals *Succeed By Myself* and *Want Help* allows the model to correctly assess some of the students who declared that they did not have the goal *Want Help*, but during the interaction displayed behaviours showing the contrary. These were cases where the model’s assessment was already very close to the threshold value used to classify *Admiration* vs. *Reproach*, thus the fluctuations in the goal probabilities caused by the student’s interface actions were sufficient to steer the model toward the right assessment, despite the lack of a mechanism to model changing goals. The rest of the cases were still misclassified, indicating that the model still requires an ability to assess changes in student goal priorities to achieve higher accuracy.

To summarize, our predictive model of user affect based on the OCC appraisal theory of emotions can achieve reasonable accuracy on three of the four emotions tested in our study, by relying only on evidence coming from student interface actions. However, two simplifying assumptions made in the model, i.e., that student goals remain the same throughout the game session and that they all have same priority, limit model accuracy, especially in detecting negative feelings towards the Prime Climb agent. Thus, a possible direction for future work is to investigate effective ways to remove these two assumptions, with special attention to constructing a clearer picture of how the user’s goal priorities fluctuate during game sessions. We expect, however, that this task will be very difficult, given that we are essentially trying to do *plan recognition* (one of AI’s notoriously challenging problems) in a highly dynamic environment. All in all, we should not be surprised that the predictive part of the model by itself has limitations. Even humans often need to integrate information on both potential causes and visible effects of the interlocutor’s emotional reactions to compensate for the limited reliability of each source in recognizing emotion. Therefore, before attempting to further refine the predictive part of the model, we are currently investigating complementing the model with diagnostic information on student observable reactions, as we further discuss in the last section of the paper.

5 Related Work

5.1 Models of user affect based on the OCC theory

There have been various preliminary attempts to use the OCC theory in predictive models of user affect. One example is the work by Streit et al., (2004). They propose using the OCC theory for an affective user model embedded within the multi-modal dialog system SmartKom, which recommends products and services based on the user's goals, likes, dislikes, and standards. The system is implemented using logical rules, and uses abduction to infer user goals from the user's reactions to the system's generated dialog. Since the proposed affective model was still quite preliminary, the authors do not report any evidence of the effectiveness of their approach. Chalfoun et al., (2006) propose using the OCC theory to model student affective reactions upon receiving the results of completing a web-based quiz. They assume that students have either one of two goals: (1) to achieve an expected mark in a post-treatment quiz; or (2) to achieve a mark above the passing mark in that quiz. However, they bypass the problem of goal assessment, and instead learn a decision tree to predict affective reactions from data on student sex, personality and test score. The authors report a prediction accuracy of 84% for their approach, although they do not provide details on how this accuracy was computed.

Theories of affect like the OCC model have more often been used in generative models that direct the affective behaviour of virtual agents such as in the ALMA project (Gebhard, 2005), in Émile (Gratch, 2000), and in the model produced by Dias & Paiva (2005). Some of these computational models of affect have the potential to transfer to modeling the affect of users. The authors of FLAME (Seif El-Nasr et al., 2000) say that their affective framework could support a user model once it incorporates additional factors such as individual differences. In addition, (Gratch & Marsella, 2004) consider extending their computational framework of appraisal & coping, EMA, by modeling the link from appraisal to bodily expression as future work. Finally, (Elliot et al., 1999) discuss how the Affective Reasoner, a rule-based framework to build agents that respond emotionally, could also be used to model users' affect.

5.2 Models of user affect that integrate causes and effects

Probabilistic approaches based on Bayesian Networks have become quite popular to model user affect from both causes and effects. Since the initial proposal of a probabilistic model that combines predictive and diagnostic inference to form a single affective assessment (Conati, 2002), several models have followed this approach. Like our model, the Bayesian network produced by Bosma & André (2004) is intended for use by a pedagogical agent within an educational game. Following our framework, the network combines contextual information on the current state of the game with diagnostic information including user's eyebrows position, heart rate and evidence collected from skin conductance. The goal is to produce an assessment of general arousal and valence, and then use it to disambiguate the utterances students generate when playing the game. The authors report significant correlations between the physiological signals used and valence/arousal, but do not evaluate the accuracy of their complete model. Li & Ji (2003) produced a Dynamic Bayesian Network (DBN) for use in intelligent user assistance systems (e.g. monitoring car drivers to detect potentially dangerous conditions such as fatigue). The model combines contextual information with evidence in the form of head gestures, hand gestures, and eye-movements to produce an assessment of affective states that include 'fatigue', 'confused', and 'frustration'. The authors do not test the accuracy of their model with real users. Instead, they evaluate its efficiency by measuring the number of time-slices required by the DBN to identify an affective state that was continuously expressed using simulated sensor data. Kapoor & Picard (2005) propose a unified Bayesian approach based on a mixture of Gaussian classifiers to detect various levels of interest in children interacting with an educational game by integrating

facial recognition, posture recognition and information on the state of the game. When evaluated on classifying states of *interested* vs *uninterested*, this approach reached an excellent accuracy of 86.55%. McGuiggan et al. (2007) achieved 89% accuracy in detecting student frustration with a narrative-based educational system from user goal, progress on goals, heart rate and skin conductance. While this approach is similar to ours in that it uses goal-related factors as a source of causal information for affect assessment, goals are explicitly given to students during the interaction. Thus, this approach does not need to deal with the problem of goal assessment as our work does.

Alternative rule-based approaches to integrate diagnostic and causal information have been proposed by Hudlicka and McNeese (2002) to assess level of anxiety in combat pilots during a mission, and by Zakharov et al. (2008) to assess valence of student affect during interaction with an intelligent tutoring system. The rules proposed by Hudlicka & McNeese (2002) specify how to combine predictive factors such as general properties of the mission at hand, events that happen during the mission, and pilot's traits (such as personality, experience and expertise) with information on the pilot's heart rate. A very preliminary evaluation of the framework was conducted on a sample set of simulated users, i.e. made-up users with scripted behaviours desirable for testing. The rules in Zakharov et al. (2008) generate a preliminary assessment of the student affective valence based on the history of the student's answers (correct vs. incorrect) with the system. Information from expression-recognition software is then used to validate the assessment. While this model has not been fully evaluated, the authors provide evidence that students found the interventions of the ITS that included affective model more appropriate than the intervention of the ITS without the model.

5.3 Models based on diagnostic assessment

Investigating potential sources of affective data for diagnostic assessment has been the focus of several research groups. Many of the results on the links between physiological signals and emotions were achieved in controlled laboratory conditions, where subjects were either asked to express the desired affect (e.g, Picard et al 2001) or this affect was induced via ad-hoc emotion-eliciting procedures (e.g, Kim and Andre 2006, Scheirer et al 2002, Lang et al. 1993). Experiments in less controlled conditions have mainly focused on detecting a single emotion, affect valence and arousal or overall emotional predisposition over the course of a complete interaction, as opposed to the multiple instantaneous emotions that are the focus of our work. For instance, by integrating measurements from five physiological sensors, three video-cameras and a microphone Healey & Picard (2005) were able to predict with 89% accuracy four levels of anxiety for subjects who experienced a sequence of different driving conditions. D'Mello et al (2006) used sitting posture monitored via pressure-detecting sensors on the learner's chairs to discriminate between low and high engagement during the interaction with a tutoring system. Litman and Forbes-Riley (2004) evaluated acoustic-prosodic and lexical features of student speech as data sources to assess the affective valence (positive, negative and neutral) of students engaged in dialogues with a tutoring system. Prendinger et al., (2005) detect user valence and arousal from measures of heart beat and skin conductance, and present preliminary evidence that an agent producing empathic responses based on this assessment can reduce the level of user stress during job interviews with an animated agent. Mandryk et al., (2006) used skin conductance, heart rate variability, EMG (to measure jaw-clenching), and respiration sensors to measure the overall emotional dispositions of players interacting with a video-game, as labeled by players' post-episode subjective ratings of 'fun', 'challenge', 'boredom', and 'frustration'. Yannakakis et al (2008) achieved 65% accuracy in predicting user preference between two gaming experiences from measures related to player heart rate during the interaction

One notable exception of research that like ours focuses on recognizing multiple emotions that vary during the course of an interaction is the work by D'Mello et al., (2008) This work looked at

students' interaction with a dialogue-based intelligent tutoring system, and investigated dialogue features as predictors of boredom, confusion, flow and frustration, obtaining a peak accuracy of 54%. The main difference between the work of D'Mello et al. (2008) and ours is that they treat their target emotions as mutually exclusive, while we try to capture potentially overlapping emotions, thus adding an additional level of complexity to the modeling task. Furthermore, because the approach in (D'Mello et al. 2008) does not include an explicit representation of the causes of student affect, it provides less information than our approach for an agent to decide how to best react to the student emotions.

6 Discussion and Conclusions

In this paper, we presented and evaluated an affective user model designed to detect multiple emotions of players interacting with Prime Climb, an educational game for number factorization. The model is to be used by an intelligent pedagogical agent that attempts to improve how students learn from the game while still maintaining the high level of positive emotional engagement that is one of the key assets of game-based education. The model relies on a general framework for affective modeling that tackles the high level of uncertainty in emotion recognition by probabilistically combining information on both causes and effects of users' emotional reactions (Conati, 2002). In this paper, we have focused on a detailed evaluation of the part of the model that predicts affect from possible causes, identified with states of the user interaction with Prime Climb. This evaluation is important to assess if and why it may be worthwhile using the potentially more intrusive technology necessary to infer emotions via diagnostic reasoning from the user affective reactions, in addition to causal information derived primarily from user interface actions. While approaches to combining diagnostic and predictive inference have received substantial attention from researchers interested in affective user modeling, to our knowledge ours is the first attempt to provide a detailed evaluation of this technique, even if limited to the causal component of the approach. Furthermore, ours is one of the few affective models targeting the recognition of multiple, possibly overlapping emotions. Most existing models focus on assessing measures of affective valence and arousal, one individual emotion, or multiple non-overlapping emotions. Because environments like educational games tend to trigger multiple, possibly overlapping and rapidly changing emotions, we argue that recognizing these emotions can improve the effectiveness of a pedagogical agent for game-based learning by improving the precision of the agent's interventions. For the same reason, we argue that it is important to provide the agent with a model that can explicitly assess the reasons for the user's emotional reactions. If this knowledge is available, the agent can not only react to the user states by addressing its causes, but it can potentially explain the rationale underlying its interventions, possibly improving the user's trust and confidence in the system (Jameson 2005).

In the paper, we illustrated how we provided the predictive component of the affective model with this level of granularity by relying on a well known emotion theory, the OCC model of cognitive appraisal (Ortony et al., 1988). The details of the implementation have been based as much as possible on data from real users. Because of the model's complexity, collecting reliable data for model construction was an extremely laborious process requiring several user studies. In the paper, we have summarized some of these studies, to give the reader a sense of the scope of the work and the challenges it entailed. We have not always been able to overcome these challenges at best, and thus our resulting model has shortcomings that affect its accuracy. Still, we believe that our results are both very promising and informative for the future development of this and other research in affective user modeling.

We evaluated our model on four of the six emotions that it can assess: joy or distress toward the game; admiration or reproach towards the agent. We showed that the predictive part of the affective model alone can already achieve good accuracy on the mutually exclusive emotions

towards the game. Accuracy for *Joy* was 69%, for *Distress* was 70%, and the macro average of 69% was statistically significantly better than the baseline accuracy of 50% achieved by always predicting the most likely emotion (*Joy*). The macro average for emotions toward the agent is still significantly better than the baseline, but practically very close to it (56%). This is mostly due to the model's problems in detecting *Reproach* (47% accuracy), while *Admiration* reaches an accuracy of 66%.

An important aspect of our results is that they have been achieved via a model that includes an assessment of user goals, crucial for performing predictive inference based on the OCC theory. We have shown that model with goal assessment outperformed the model with data on user goals given as evidence. Goal recognition is one of the hard problems in AI and thus one of the bottlenecks for a more widespread use of the OCC theory for affective user modeling. Ours is the only work that has shown with hard data the feasibility of this approach. Still, and not surprisingly, goal recognition is one of the limiting factors of our model's accuracy. In the paper, we discussed how the poor performance on *Reproach* is largely due to two goal-related model shortcomings: its inability to assess goals that dynamically change during the interaction, and the fact that we don't properly model goal priority in the presence of multiple goals.

In order to refine the model so as to remove these assumptions, we would first need to collect empirical data to understand why and how student goal priorities may change during game playing. Data on goal priorities could be recorded either via a self-report mechanism similar to the one we use to collect emotion self-reports, or by post-session annotations by experts. However both of these options have inherent difficulties. Asking students to identify their own goal priorities, even if asking about a reduced set of at most two goals, is likely to cause confusion as to what is being asked. For experts, it is likely that attempting to annotate student goals during the interaction would be rather difficult and laborious, given the novelty of the interaction and the fact that goals are often related to a non-trivial combination of factors including student personality, goals and interaction patterns. Thus, although this direction of investigation is possible, it also contains some very difficult challenges.

Instead, we are investigating the addition to our model of diagnostic evidence from physiological sensors, to overcome the limitations of its predictive component. Our framework is already set up to support flexible combinations of sensors given the modularity of its diagnostic component (Conati, 2002). We are planning to look next at the addition of sensors that help assess affective valence (such as Electromyography, or EMG, to monitor movements of the user corrugators muscles and Blood Volume Pulse, or BVP, to monitor changes in heart rate) and affective arousal (such as sensors to monitor Skin Conductance). A preliminary attempt to include a BVP signal in our model showed that the sensors commonly used in this research are too sensitive to noise when donned on highly active children (Conati et al. 2003). Thus, we plan to investigate the usage of Beats per Minute (BMP) sensors that are advertised to be especially suitable for usage with young subjects¹⁰.

Other future work relates to better addressing two other major challenges we encountered during model construction. The first challenge relates to reliably recording students' affective states during game playing. As we mentioned in a previous section, in our research using judges to produce affective labels from video-recordings of the interaction is very difficult because of the requirement to distinguish separate feelings towards the game and towards the agent. We could not try to ask students to recall their feelings by viewing a replay of the interaction after game playing because of time constraints. Thus, we introduced the mechanism for obtaining emotions self-reports during game playing. While we have evidence that this mechanism is not overly intrusive on average, it does introduce an extraneous element in the interaction that may have unwarranted side effects for some students. Furthermore, it does not allow us to obtain data simultaneously on all of the emotion pairs we aim to assess and also on affective arousal. In the

¹⁰ See for instance, <http://www.dataharvest.com/Products/easysense/sensors/heart.htm>

future, we plan to pilot test asking students to generate the self-reports after game playing and, if they prove able to deal with the task, we will explore ways to extend our study sessions to include this alternative method for collecting affective labels.

The second challenge relates to collecting data-points for negative emotions. Throughout our studies, students have generated far fewer reports of negative emotions (*Distress* or *Reproach*) than positive emotions. One suggested reason for the small number of negative reports is the nature of our test-bed application, because Prime Climb is indeed designed to provide students with a fun and engaging activity. We could induce negative emotions on purpose during game-playing, but we argue that the students' negative feelings in this case would not be indicative of the real emotions that they would experience during real interactions. Possible solutions that we are planning to explore to overcome this challenge include: (i) find ways to have students interact longer with the game; (ii) have students play with each other. This second solution has the double advantage of giving us twice as much data for each playing session, and being likely to generate more and stronger emotional episodes, given what we have seen when students play together. However, it requires that we add to the model the capability of assessing emotions toward a partner, which is one of the next steps of this research

But the most important challenge that we need to address in the development of this research is to prove that having a sophisticated model for the assessment for players' individual emotions is worth the effort. That is, we need to show that the Prime Climb pedagogical agent can indeed benefit from having detailed information on user affect. Although our model currently underestimates the student's feelings of *Reproach*, its accuracy in assessing *Joy*, *Distress* and *Admiration* is high enough for us to consider an indirect evaluation to determine whether the model as it is would contribute to the pedagogical effectiveness of Prime Climb. We have begun to investigate ways to combine the assessment of the affective model and the model of student knowledge (Manske & Conati, 2005) into a decision theoretic framework that will allow the Prime Climb pedagogical agent to decide how to intervene so as to maximize the trade-off between student learning and engagement. Once we have completed this task, we can compare the overall effectiveness of the pedagogical agent with and without affective assessments. Once we have improved the accuracy for identifying feelings of *Reproach*, we can also run ablation studies to test our assertion that the more detailed information the agent has on the student affect, the better it can help the student interacting effectively with Prime Climb.

References

- Alessi, S.M., Trollip, S.R. (2001). *Multimedia for Learning: Methods and Development*, 3rd ed. Allyn & Bacon, Needham Heights.
- Baker, R. S., Corbett, A. T., Koedinger, K. R., & Wagner, A. Z. (2004). Off-Task Behavior in the Cognitive Tutor Classroom: When Students "Game the System". *CHI'04: Conference on Human Factors in Computing systems*, Vienna, Austria, 383-390
- Baker, R., Rodrigo, M.T., Xolocotzin, U.E. (2007): The Dynamics of Affective Transitions in Simulation Problem-Solving Environments. *ACII 2007*, Lisbon, Portugal, 666-677
- Bosma, W., & André, E. (2004). Exploiting Emotions to Disambiguate Dialogue Acts. *IUI'04, International Conference on Intelligent User Interfaces*, Madeira, Portugal, 85-92.
- Chalfoun, P., Haffar, S., & Frasson, C. (2006). Predicting the Emotional Reaction of the Learner with a Machine Learning Technique. *Workshop on Motivational and Affective Issues in ITS, ITS'06, International Conference on Intelligent Tutoring Systems*, Jhongli, Taiwan.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Hillsdale: Lawrence Earlbaum associates.
- Conati, C. (2002). Probabilistic Assessment of Users' Emotions in Educational Games. *Journal of Applied Artificial Intelligence, special issue on "Merging Cognition and Affect in HCI"*, 16(7-8), 555-575.

- Conati, C. (2004). How to Evaluate Models of User Affect? *ADS'04, Tutorial and Research Workshop on Affective Dialog Systems*, Kloster Irsee, Germany, 288-300.
- Conati, C., Chabbal, R., & Maclaren, H. (2003). A Study on Using Biometric Sensors for Detecting User Emotions in Educational Games. *Workshop on "Assessing and Adapting to User Attitude and Affects: Why, When, and How?"* in conjunction with UM'03, 9th International Conference on User Modeling, Pittsburgh, USA.
- Conati, C., & Klawe, M. (2002). Socially Intelligent Agents in Educational Games . K. Dautenhahn, A. Bond, D. Canamero & B. Edmonds (Eds.), *Socially Intelligent Agents - Creating Relationships with Computers and Robots*: Kluwer Academic Publishers, 213-220.
- Conati, C., & Maclaren, H. (2004). Evaluating a Probabilistic Model of Student Affect. *ITS'04, International Conference on Intelligent Tutoring Systems*, Maceio, Brazil, 55-66.
- Conati, C., & Zhao, X. (2004). Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Education Game. *IUI'04, International Conference on Intelligent User Interfaces*, Funchal, Madeira, Portugal, 6-13.
- Cordova, D., & Lepper, M. (1996). Intrinsic Motivation and the Process of Learning: Beneficial Effects of Contextualization, Personalization, and Choice. *Journal of Educational Psychology*, 88, 715-730.
- Costa, P. T., & McCrae, R. R. (1992). Four Ways Five Factors are Basic. *Personality and Individual Differences*, 13, 653-665.
- Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and Learning: An Exploratory Look into the Role of Affect in Learning with AutoTutor. *Journal of Educational Media*, 29, 241-250.
- Dean, T., & Kanazawa, K. (1989). A Model for Reasoning about Persistence and Causation. *Computational Intelligence*, 5(3), 142-150.
- deVincente, A., & Pain, H. (1999). Motivation Self-Report in ITS. *AIED'99, 9th International Conference on Artificial Intelligence in Education*, Le Mans, France, 651-659.
- Dias, J., & Paiva, A. (2005). Feeling and Reasoning: A Computational Model for Emotional Characters. *Progress in Artificial Intelligence, 12th Portuguese Conference on Artificial Intelligence*, EPIA 2005, Covilhã, Portugal, 127-140.
- D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting Affective States expressed through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue. *International Journal of Artificial Intelligence in Education*, 16, 3-28.
- D'Mello, S.K., P. Chipman, and A.C. Graesser (2006). Posture as a predictor of learner's affective engagement . *Proceedings of the 29th Annual Cognitive Science Society*, Stresa, Italy, 571-576.
- D'Mello, S.K., S. K., Craig, S.D., Witherspoon, A. W., McDaniel, B. T., and Graesser, A. C. (2008). Automatic Detection of Learner's Affect from Conversational Cues. *User Modeling and User-Adapted Interaction*, 2008. 18(1-2), 45-80.
- Elliot, C., Rickel, J., & Lester, J. C. (1999). Lifelike Pedagogical Agents and Affective Computing: An Exploratory Synthesis . M. Wooldridge & M. Veloso, M. (Eds.), *Artificial Intelligence Today*: Springer, 195-211.
- Esuli, A., Fagni, T., Sebastiani, F., & Aizerman, M. (2006). MP-Boost: A Multiple-Pivot Boosting Algorithm and its Application to Text Categorization. *13th International Symposium on String Processing and Information Retrieval (SPIRE'06)*, Glasgow, UK, 1-12.
- Fisher, R. A. (1935). *The Design of Experiment*. New York: Hafner.
- Gebhard, P. (2005). ALMA - A Layered Model of Affect. *4th International Joint Conference of Autonomous Agents & Multi-Agent Systems (AAMAS'05)*, Utrecht, The Netherlands, 29-36.
- Gratch, J. (2000). Emile: Marshalling Passions in Training and Education. *4th International Conference on Autonomous Agents*, Barcelona, Spain, 325--332.

- Gratch, J., & Marsella, S. (2004). A Domain Independent Framework for Modeling Emotion. *Journal of Cognitive Systems Research*, 5(4), 269-306.
- Graziano, W. G., Jensen-Campbell, L. A., & Finch, J. F. (1997). The Self as a Mediator Between Personality and Adjustment. *Journal of Personality and Social Psychology*, 73, 392-404.
- Healey, J. A., & Picard, R. W. (2005). Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2), 156-166.
- Heckerman, D. (1999). A Tutorial on Learning with Bayesian Networks . M. Jordan (Ed.), *Learning in Graphical Models*. Cambridge, MA: MIT Press. 301-354.
- Hudlicka, E., & McNeese, M. (2002). Assessment of User Affective and Belief States for Interface Adaptation: Application to an Air Force Pilot. *User Modeling and User Adapted Interaction*, 12(1), 1-47.
- Jain, A., & Zongker, D. (1997). Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 153-158.
- Jameson A. (2003). "Adaptive Interfaces and Agents" in *Human-Computer Interface Handbook*, eds J.A. Jacko and A. Sears, pp 305-330.
- Kapoor, A., & Picard, R. W. (2005). Multimodal Affect Recognition in Learning Environments. *13th Annual ACM International Conference on Multimedia*, Singapore, 677-682.
- Kim, J., & André, E. (2006). Emotion Recognition Using Physiological and Speech Signal in Short-Term Observation. *PIT'06, Perception and Interactive Technologies*, Kloster Irsee, Germany, 53-64.
- Klawe, M. (1998). When Does The Use Of Computer Games And Other Interactive Multimedia Software Help Students Learn Mathematics? *Technology and NCTM Standards 2000 Conference*, Arlington, VA.
- Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Look at Pictures: Affective, Facial, Visceral, and Behavioral Reactions. *Psychophysiology*, 30, 261-273.
- Lee, J., Luchini, K., Michael, B., Norris, C., Solloway, E.(2004). More than just fun and games: Assessing the value of educational video games in the classroom. *Proceedings of ACM SIGCHI 2004*, Vienna, Austria, pp. 1375-1378.
- Lepper, M., Woolverton, M., Mumme, D., & Gurtner, J.-L. (1993). Motivational Techniques of Expert Human Tutors: Lessons for the Design of Computer-based Tutors . S. P. Lajoie & S. J. Derry (Eds.), *Computers as Cognitive Tools*: Lawrence Erlbaum Associates, NJ, 75-105.
- Li, X., & Ji, Q. (2003). Active Affective State Detection and User Assistance. Workshop on "Assessing and Adapting to User Attitude and Affects: Why, When, and How?" in conjunction with *UM'03, 9th International Conference on User Modeling*, Pittsburgh, PA.
- Linnenbrink, E. A., & Pintrich, P. R. (2002). The Role of Motivational Beliefs in Conceptual Change . M. Limon & L. Mason (Eds.), *Reconsidering Conceptual Change: Issues in Theory and Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 115-135.
- Lisetti, C. L., & Nasoz, F. (2004). Using Non-invasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on Applied Signal Processing*, 11, 1672-1687.
- Litman, D.J. and K. Forbes-Riley (2004). Predicting Student Emotions in Computer-Human Tutoring Dialogues in *42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, 352-359.
- Mandryk, R. L., Inkpen, K. M., & Calvert, T. W. (2006). Using Psychophysiological Techniques to Measure User Experience with Entertainment Technologies. *Journal of Behavior and Information Technology (Special Issue on User Experience)*, 25, 141-158.
- Manske, M., & Conati, C. (2005). Modeling Learning in Educational Games. *AIED'05, 12th International Conference on AI in Education*, Amsterdam, The Netherlands, 411-418.

- Mitchell, T. (1997). *Machine Learning*: McGraw Hill.
- McQuiggan, S., S. Lee, and J. Lester (2007) Early Prediction of Student Frustration. *Second Int. Conf. on Affective Computing and Intelligent Interactions*, Lisbona, Portugal, 698-709.
- Mehrabian A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament . *Current Psychology*, 14, pages 261 – 292.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions*: Cambridge University Press.
- Peter, C., & Herbon, A. (2006). Emotion Representation and Physiology Assignments in Digital Systems. *Interacting with Computers*, 18(2), 139-170.
- Picard, R. W., Vyzas, E., & Healey, J. A. (2001). Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10), 1175-1191.
- Prendinger, H., Mori, J., & Ishizuka, M. (2005). Recognizing, Modeling, and Responding to Users' Affective States. *UM'05 10th International Conference on User Modeling*, Edinburgh, U.K., 60-69.
- Prendinger, H., Ishizuka, M. (2002). SCREAM: scripting emotion-based agent minds. *AAMAS 2002*, Bologna, Italy, 350-351.
- Qu, L., & Johnson, L. (2005). Detecting the Learner's Motivational States in an Interactive Learning Environment. *AIED'05 12th International Conference on Artificial Intelligence in Education.*, Amsterdam, The Netherlands, 547-554.
- Rodrigo, M.T., Baker, R., D'Mello, S. T. Gonzales et al. (2008). Comparing Learners' Affect While Using an Intelligent Tutoring System and a Simulation Problem Solving Game. *Intelligent Tutoring Systems 2008*, Montreal, Canada, 40-49.
- Scheirer, J., Fernandez, R., Klein, J., & Picard, R. W. (2002). Frustrating the User on Purpose: A Step Toward Building an Affective Computer. *Interacting with Computers*, 14(2), 93-118.
- Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), 1-47.
- Seif El-Nasr, M., Yen, J., & Ioerger, T. R. (2000). FLAME - Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems*, Barcelona, Spain, 219-257.
- Streit, M., Batliner, A., & Portele, T. (2004). Cognitive-Model-Based Interpretation of Emotions in a Multi-Model Dialog System. *ADS'04 Tutorial and Research Workshop on Affective Dialogue Systems*, Kloster Irsee, Germany. 65-76.
- Thayer, Robert E. (1989). *The biopsychology of mood and arousal*. New York, NY: Oxford University Press.
- Van Eck, R. (2007). "Building Artificially Intelligent Learning Games," in *Games and Simulations in Online Learning: Research and Development Frameworks*, D. Gibson, C. Aldrich, and M. Prensky, Eds.: *Information Science Pub.*, 2007, pp. 271-307.
- VanLehn, K., & Niu, Z. (2001). Bayesian Student Modeling, User Interfaces and Feedback: A Sensitivity Analysis. *International Journal of Artificial Intelligence in Education*, 12(2), 154-184.
- Vogel, J.J., Greenwood-Ericksen, A., Cannon-Bowers, J., Bowers, C.A.(2006).Using virtual reality with and without gaming attributes for academic achievement. *Journal of Research on Technology in Education* 39(1), 105–118.
- Vyzas, E., & Picard, R. W. (1998). Affective Pattern Classification. *AAAI 1998 Fall Symposium, Emotional and Intelligent: The Tangled Knot of Cognition*, 176-182.
- Yannakakis, G.N., J. Hallam, and H.H. Lund (2008). Entertainment Capture through Heart Rate Activity in Physical Interactive Playgrounds. *User Modeling and User-Adapted Interaction*, 18(1-2). 207-243.

- Zakharov, K., A. Mitrovic, and L. Johnston (2008). Towards Emotionally-Intelligent Pedagogical Agents. *Intelligent Tutoring Systems, 9th Int. Conf., ITS 2008*, Montreal, Canada, 19-28.
- Zhou, X., & Conati, C. (2003). Inferring User Goals from Personality and Behavior in a Causal Model of User Affect. *IUI'03, International Conference on Intelligent User Interfaces*, Miami, FL, 211-218

Authors' Vitae

Dr. Cristina Conati:

University of British Columbia, Department of Computer Science, 2366, Main Mall, Vancouver, BC, V6T 1Z4, cadana

Dr. Conati is an Associate Professor of Computer Science at the University of British Columbia. She received a M.Sc. in Computer Science at the University of Milan, Italy (1988), as well as a M.Sc. (1996) and Ph.D. (1999) in Artificial Intelligence at the University of Pittsburgh. Dr. Conati's areas of interest include Adaptive Interfaces, User Modeling, Affective Computing and Intelligent Tutoring Systems. She published over 50 strictly referred articles, and her research has received awards from the International Conference on User Modeling, the International Conference of AI in Education, the International Conference on Intelligent User Interfaces and the Journal of User Modeling and User Adapted Interaction.

Dr. Heather Maclaren:

Humanature Studios, Nexon Publishing North America, Vancouver, Canada

Heather Maclaren received both her bachelors degree in Computer Science and her Ph.D. in Computer Science from the University of York, England. Her Ph.D. was in the areas of inductive logic programming and user modeling. She continued to pursue her interests in machine learning and user modeling as a postdoctoral fellow under the supervision of Dr Cristina Conati at the University of British Columbia. Her contribution to the paper stems her work during this time. Since 2007 Dr. Maclaren has been a software engineer at Humanature Studios, applying her experience in user modeling and usability by working on rapid iterative prototyping of user interfaces for computer games.