ABSTRACT

MCQUIGGAN, SCOTT WILLIAM. An Inductive Framework for Affect Recognition and Expression in Interactive Learning Environments. (Under the direction of James C. Lester.)

Recent years have seen a growing recognition of the importance of affective reasoning in human-computer interaction. Because affect plays an important role in cognitive functions, such as perception and decision-making, the prospect of modeling user affect and enabling interactive systems to respond appropriately holds much appeal for a broad range of applications. Affective reasoning is particularly promising for educational applications because of the strong connections between affect and learning. If it were possible to accurately detect frustration, monitor changes in efficacy, and predict students’ affective states, interactive learning environments could more effectively tailor problem-solving episodes. However, constructing computational models of affect recognition and affect expression is challenging because of the need to devise solutions that are accurate, efficient, and capable of making early predictions.

To this end we propose CARE, an inductive framework for affect recognition and expression. CARE learns models of affect from observation of human-computer and human-human interaction. First, in training sessions, users perform a series of tasks in interactive environments while CARE monitors reports of users’ affective experiences. In addition, CARE monitors user actions, world state, and physiological responses. Second, CARE induces models of affect from observed data with machine learning techniques that include decision trees, naive Bayes classifiers, support vector machines, Bayesian networks, and n-grams. Third, at runtime, CARE-induced models monitor user actions, world state, and physiological responses to predict user affective states. In a series of studies involving more than four hundred subjects, the CARE framework has successfully been used to perform a number of affect prediction tasks, including emotional state prediction, self-efficacy,
and metacognitive monitoring prediction. It has also been used to induce models of empathy for virtual agents in interactive learning environments. Results suggest that CARE-induced affect models satisfy the real-time requirements of interactive systems and provide a solid foundation for empirically informed affective reasoning.
An Inductive Framework for Affect Recognition and Expression in Interactive Learning Environments

by
Scott W. McQuiggan

A dissertation submitted to the Graduate Faculty of
North Carolina State University
In partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Computer Science
Raleigh, North Carolina

2009

APPROVED BY:

______________________________  ________________________________
James C. Lester                John Nietfeld
Chair of Advisory Committee

______________________________  ________________________________
Munindar P. Singh              R. Michael Young
Dedication

This work is dedicate to my wife, Jamie, whose loving encouragement made it possible,

and to Mother and Pops who have always been just down the right field line.
Biography

Scott William McQuiggan was born in Elizabethtown, Pennsylvania, to his parents, William and Carol. He attended Elizabethtown Area High School, graduating in 1999. He then headed to Selinsgrove, Pennsylvania to attend Susquehanna University finally settling on a major in Computer Science. Prior to graduation Scott was fortunate to spend significant time gaining valuable experience in industry. He has worked as a research assistant at the Pennsylvania State Data Center from May 2001 - May 2002 and held a co-op position with the Defense Information Systems Agency from May 2002 - July 2003. Scott obtained his Bachelor of Science degree from Susquehanna University in 2003. That same year, Scott headed south, relocating to the research triangle in North Carolina. In the fall of 2003, Scott began his graduate studies at North Carolina State University. As a graduate student, Scott joined the IntelliMedia Center for Intelligent Systems, led by his advisor, Dr. James C. Lester. In 2005, Scott completed his Masters of Science degree, successfully defending his thesis, An Inductive Approach to Modeling Affective Reasoning in Interactive Synthetic Agents, before entering the Ph.D. program at North Carolina State University. He married his wife, Jamie, in October, 2006. While completing his graduate studies and research Scott has been working as a technical student at SAS Institute Inc. in the JMP Division, since 2004.

Scott’s research continues to pursue an interdisciplinary agenda combining artificial intelligence, education, educational psychology, human factors and human-computer interaction, and psychology. His primary interests center on investigating student-interaction with new learning technologies such as intelligent tutoring systems, game-based learning environments, and narrative-centered learning environments. His future endeavors most likely will focus on such applications, using artificial intelligence and machine learning techniques to tailor experiences to individual learners.
Acknowledgements

I would like to first thank the members of my committee, John Nietfeld, Munindar Singh and Michael Young for their academic enthusiasm and support.

I am immensely thankful for the opportunity to be a student of my advisor, James C. Lester, who has given continuous guidance, support, and confidence throughout my graduate studies, research, and beyond. Dr. Lester has bestowed many lessons in academia and insightful discussions for which I am genuinely grateful.

While at North Carolina State University I have had the great fortune to learn and interact with a number of faculty and staff including: Barbara Adams, Ginny Adams, Carol Allen, Jon Doyle, Marty Dulberg, Christopher Healey, Thomas Honeycutt, Purush Iyer, David Kaber, Kay Leager, Margery Page, Robert Rodman, Carla Savage, Heather Smolensky, Robert St. Amant, Alan Tharp, and David Thuente.

I appreciate the generosity and help offered by the members of the IntelliMedia Center for Intelligent Systems. I would like to especially thank Kristy Boyer, Cohan Carlos, Rachael Dwight, Julius Goth, Eunyoung Ha, Seung Lee, Sunyoung Lee, Bradford Mott, Jenny Navoraphan, Rob Phillips, Jennifer Robison, Jonathan Rowe, Lucy Shores, and Michael Wallis. Discussions with fellow students have often led to a better understanding of my own work and its application. For this I owe thanks to Leo Bae, Yuna Cheong, Kevin Damm, Neha Jain, Arnav Jhala, Joseph Grafsgaard, James Niehaus, Karl Nilsson, Ben Rose, Jim Thomas, and Tommy Vernieri. As a research assistant I have had privilege to work with Kristin Hoffmann, Gwynn Morris, John Nietfeld, Hiller Spires, and Kimberly Turner and interact with Sammie Carter, Mustapha Jawara, Vernell Jones, and Gayle Merrill. I also appreciate the support of the eighth-grade science teachers who participated in Crystal Island, Ada Lopez, Nelda Phillips, and Betty Welsh, the administration and staff of Centennial Campus Middle School, especially Elwood Peters, and the administration and staff of Cane River Middle School, especially Beverly Brown.
I am grateful to Boris Roussev, who instilled in me much confidence and motivation at Susquehanna University, leading me to pursue the graduate studies I have come to treasure so deeply. Also, at Susquehanna University I am also grateful for the guidance and teaching of Kenneth Brakke, James Handlan, and William Miller.

I would like to additionally thank SAS Institute, Inc., particularly the JMP™ division, Richard Potter, and the development team for allowing me to gain valuable experience in the development of first-class statistical software while prioritizing academics throughout the course of this work.

I would like to express my sincere appreciation to my sister, Kelly, my brother, Steven, and the rest of my family and friends who have, at some time or another, readily listened to me explain some aspect of my research. Furthermore, it is with the loving support, ready advice and unending encouragement gifted to me by my parents, William and Carol McQuiggan that has enabled me to pursue my interests and flourish in whatever makes me happy.

Finally, I would like to express my gratitude for the unconditional understanding, patience, and emotional support of my best friend and wife, Jamie.
# Table of Contents

List of Figures ................................................................................................................................. x
List of Tables ........................................................................................................................................ xii

1. Introduction ........................................................................................................................................ 1
   1.1 Motivation ........................................................................................................................................ 6
   1.2 Affect Recognition and Expression Challenges ............................................................................. 8
   1.3 The CARE Inductive Approach ........................................................................................................ 11
   1.4 Contributions .................................................................................................................................. 12
   1.5 Summary of Results ....................................................................................................................... 15
   1.6 Dissertation Organization ............................................................................................................... 22

2. Background and Related Work .......................................................................................................... 23
   2.1 Appraisal Theory .............................................................................................................................. 23
       2.1.1 Overview .................................................................................................................................. 24
       2.1.2 Smith and Lazarus’ Model ........................................................................................................ 25
       2.1.3 Other Structured Appraisal Theory Models .............................................................................. 28
   2.2 Affective Computing ....................................................................................................................... 32
       2.2.1 Affect Recognition .................................................................................................................... 32
       2.2.2 Affect Expression ...................................................................................................................... 36

3. An Inductive Framework for Affect Recognition and Expression .................................................... 43
   3.1 Defining Affect Recognition and Expression .................................................................................. 43
   3.2 Appraisal Theory as the Theoretical Basis for CARE ..................................................................... 44
   3.3 The Inductive Approach .................................................................................................................. 48
   3.4 Inductive Modeling Techniques .................................................................................................... 51
       3.3.1 Naïve Bayes ............................................................................................................................... 52
       3.3.2 Decision Trees ........................................................................................................................... 52
       3.3.3 Support Vector Machines ........................................................................................................ 53
       3.3.4 N-grams .................................................................................................................................... 53

4. Learning Models of Affect for Recognition and Expression ............................................................... 54
   4.1 Defining Affect Model Learning ....................................................................................................... 54
   4.2 Affect Modeling Corpora ................................................................................................................ 55
   4.3 Designing Corpus Acquisition Tasks ............................................................................................... 59
   4.4 Model Induction ............................................................................................................................... 61
       4.2.1 Data Construction ....................................................................................................................... 62
       4.2.2 Data Cleansing ........................................................................................................................... 62
       4.2.3 Model Learning .......................................................................................................................... 62
5. Implementationed Interactive Environments for Education and Entertainment ...... 67
  5.1 Online Tutorial System ................................................................. 67
  5.2 Treasure Hunt ........................................................................ 69
  5.3 Crystal Island Learning Environment ........................................ 73
    5.3.1 Narrative-Centered Learning ............................................... 73
    5.3.2 Student Motivation in Narrative Learning .............................. 76
  5.3.3 Crystal Island - Version 1.0 ................................................. 78
  5.3.4 Crystal Island – Version 2.0 ................................................ 80

6. Modeling Self-Efficacy .................................................................. 87
  6.1 Self-Efficacy Background, Related Work, and Motivations .......... 87
  6.2 Self-Efficacy Modeling in an Online Tutorial System ............... 91
    6.2.1 Method ............................................................................. 91
    6.2.2 Procedure ....................................................................... 92
    6.2.3 Results ........................................................................... 94
    6.2.4 Discussion ...................................................................... 99
    6.2.5 Study Limitations ............................................................ 100
  6.3 Self-Efficacy Modeling in Crystal Island, Version 1.0 .............. 101
    6.3.1 Method .......................................................................... 102
    6.3.2 Procedure ....................................................................... 103
    6.3.3 Results ........................................................................... 105
    6.3.4 Discussion ...................................................................... 108
    6.3.5 Study Limitations ............................................................ 111
  6.4 Summary of Self-Efficacy Studies .............................................. 111

7. Modeling Emotion ....................................................................... 115
  7.1 Emotion Background, Related Work, and Motivations ............. 115
  7.2 Affective State Modeling in Crystal Island, Version 1.0 .......... 118
    7.2.1 Method .......................................................................... 118
    7.2.2 Procedure ....................................................................... 118
    7.2.3 Results ........................................................................... 119
    7.2.4 Discussion ...................................................................... 120
    7.2.5 Study Limitations ............................................................ 121
  7.3 Frustration Modeling in Crystal Island, Version 1.0 ............... 121
    7.3.1 Early Prediction Methods ................................................... 122
    7.3.2 Study Method .................................................................. 124
    7.3.3 Procedure ....................................................................... 124
    7.3.4 Results ........................................................................... 125
    7.3.5 Discussion ...................................................................... 128
    7.3.6 Study Limitations ............................................................ 128
7.4 Summary of Emotion Modeling Studies

8. Modeling Goal Monitoring

8.1 Goal Monitoring Background, Related Work, and Motivations

8.2 Modeling Goal Monitoring in Crystal Island, Version 1.0
   8.2.1 Method
   8.2.2 Procedure
   8.2.3 Results
   8.2.4 Discussion
   8.2.5 Study Limitations

8.3 Modeling Goal Monitoring in Crystal Island, Version 2.0
   8.3.1 Method
   8.3.2 Procedure
   8.3.3 Results
   8.3.4 Discussion
   8.3.5 Study Limitations

8.4 Summary of Self-Efficacy Studies

9. Modeling Empathy

9.1 Empathy Background, Related Work, and Motivations

9.2 Modeling Empathy in Treasure Hunt
   9.2.1 Method
   9.2.2 Procedure
   9.2.3 Results
   9.2.4 Discussion
   9.2.5 Study Limitations

9.3 Modeling Parallel and Reactive Empathy in Crystal Island, Version 1.0
   9.3.1 Method
   9.3.2 Procedure
   9.3.3 Results
   9.3.4 Discussion
   9.3.5 Study Limitations

9.4 Summary of Empathy Studies

10. Affective Experience and Outcomes with Interactive Environments

10.1 Background, Related Work, and Motivations
   10.1.1 Presence
   10.1.2 Empathy
   10.1.3 Psychological Assessment Instrumentation

10.2 Evaluating Perceived Empathetic Accuracy of Induced Empathy Models
   10.2.1 Method
   10.2.2 Procedure
   10.2.3 Results

viii
List of Figures

Figure 1.1. Affective Loop .......................................................... 2
Figure 1.2. Affective Loop Contributions .................................. 13
Figure 2.1. Appraisal Model of Emotions .............................. 26
Figure 2.2. OCC Model of Emotions ...................................... 29
Figure 2.3. Roseman’s Model of Emotions ............................ 31
Figure 3.1. Affect Recognition .............................................. 44
Figure 3.2. Affect Expression ................................................. 44
Figure 3.3. CARE Inductive Framework .............................. 49
Figure 4.1. Corpus Acquisition Methodology ....................... 60
Figure 5.1. Genetics Online Tutorial System ....................... 68
Figure 5.2. Genetics Online Tutorial System Problem Solving Environment .................. 68
Figure 5.3. Treasure Hunt .................................................. 69
Figure 5.4. An Excited Companion Agent in Treasure Hunt ...... 71
Figure 5.5. A Relaxed Companion Agent in Treasure Hunt ....... 71
Figure 5.6. A Frustrated Companion Agent in Treasure Hunt ...... 72
Figure 5.7. Crystal Island .................................................. 79
Figure 5.8. A Crystal Island Character ................................. 81
Figure 5.9. Crystal Island 2.0 Map ...................................... 82
Figure 5.10. Crystal Island 2.0 Cast of Characters .................. 83
Figure 5.11. Factsheet in Crystal Island 2.0 ............................ 84
Figure 5.12. Hammer Level Editor ........................................ 85
Figure 5.13. Virtual Books in Crystal Island ......................... 86
Figure 5.14. Web-based Posters in Crystal Island 2.0 ............. 86
Figure 6.1. Online Tutorial System Foundational Evaluation Data Flow .................. 94
Figure 6.2. ROC Curves for Naive Bayes and Decision Tree Models .................. 95
Figure 6.3. Heart Rate for High and Low Self-Efficacy Reports .......... 99
Figure 6.4. Wired-User Interacting with an Interactive Learning Environment ................. 104
Figure 6.5. Interactive Learning Environment Evaluation Data Flow ....... 105
Figure 6.6. ROC Curves for Naive Bayes and Decision Tree Models .................. 106
Figure 7.1. Model of Flow State ....................................... 117
Figure 7.2. Bigram Convergence Graph ................................ 126
Figure 8.1. Self-Regulated Learning Model ......................... 131
Figure 8.2. Monitoring .................................................... 132
Figure 8.3. Distribution of Monitoring Values for Study 1 and 2 .... 141
Figure 9.1. Empathy Construct ......................................... 145
Figure 9.2. Evaluation Data Flow ....................................... 150
Figure 9.3. Individual Empathizer IRI Results ....................... 152
Figure 9.4. Individual Empathizer IRI Subscale Results ............. 154
Figure 9.5. Average Empathizer IRI Subscale Results by Gender .......... 154
Figure 9.6. Affective State Frequencies from 25 Training Sessions ........... 155
Figure 9.7. Affective State Frequencies by Gender ................................................................. 155
Figure 9.8. Average Affective State Frequencies by Gender .................................................. 156
Figure 9.9. ROC Curves for Empathetic Assessment ................................................................. 156
Figure 9.10. ROC Curves for Empathetic Interpretation .......................................................... 157
Figure 9.11. Empathetic character schema .............................................................................. 165
Figure 10.1. Box plot of presence results (total PQ) from the first study .................................. 188
Figure 10.2. Box plot of presence results (total PQ) from the replicated study ....................... 189
Figure 10.3. Transitions from flow ......................................................................................... 200
Figure 10.4. Transitions from frustration ................................................................................ 200
Figure 10.5. Transitions from confusion and boredom by Dominant Goal Orientation ............ 203
Figure 10.6. Transitions from confusion and boredom by Presence ....................................... 205
Figure 10.7. Transitions from confusion and boredom by Agreeableness ............................. 205
Figure 10.8. Learning gains (pre- to post-test) by condition ...................................................... 213
Figure A.1. ROC Curves for physiological response modeling .................................................. 266
List of Tables

Table 1.1. Chronological Summary of Completed Studies .................................................. 16
Table 1.2. Studied Constructs, Instruments, Environments, and Subjects ......................... 20
Table 4.1. Representative observational attributes ................................................................ 57
Table 6.1. Model Accuracy Results ....................................................................................... 97
Table 6.2. Dynamic Decision Tree Models Compared to Baseline Models ................................ 97
Table 6.3. Significant Effects on Dynamic Self-Efficacy Models ........................................... 98
Table 6.4. Interactive Learning Environment Model Results .................................................. 107
Table 6.5. Model Results from Online Tutorial System and Interactive Learning Environment .... 107
Table 6.6. Baseline Model Comparisons .............................................................................. 108
Table 7.1. Classification results for decision tree and naive Bayes models ............................ 119
Table 7.2. Precision and recall analysis for the decision tree affect recognition model .......... 120
Table 7.3. n-gram Results .................................................................................................... 125
Table 7.4. Induced model results .......................................................................................... 126
Table 8.1. Modeling Metacognitive Monitoring Results ......................................................... 137
Table 8.2. Modeling Metacognitive Monitoring Results - Replication ................................. 141
Table 9.1. Areas under ROC curves in Figure 9.10 ............................................................. 158
Table 9.2. Empathizer suggested emotions .......................................................................... 160
Table 9.3. Distribution of empathetic responses .................................................................. 166
Table 9.4. Empathy model results by data type used for learning ....................................... 167
Table 10.1. Interpersonal Reactivity Index Results ............................................................... 182
Table 10.2. Analysis of Appropriateness Responses .............................................................. 182
Table 10.3. Breakdown of condition assignments by group and gender ............................... 185
Table 10.4. Presence results by group .................................................................................. 191
Table 10.5. Presence results by group continued ................................................................. 191
Table 10.6. Likelihoods for all transitions .......................................................................... 198
Table 10.7. Interesting likelihood for transitions ................................................................. 201
Table 10.8. Subject population by condition ...................................................................... 212
Chapter 1

Introduction

Affect\(^1\) plays a central role in human endeavors. It influences our interactions, our behavior and even our thinking. Affect is expressed by the words we choose, our facial expressions, our posture, and our actions. If computational models of affect can be designed that accurately reflect how emotion is recognized, understood, and expressed, we can design computational systems that more effectively support human-computer interaction. The field of affective computing addresses precisely these issues. Research in affective computing has found that emotion plays a role in decision making, perception, learning and rational thinking and can contribute to the successful deployment of a broad range of interactive systems including educational and training software, games, and virtual environments.

In recent years, there has been a growing recognition that emotion is rooted in perception, rational thinking, decision-making, and other cognitive processes (Cytowic, 1989; Damasio, 1994; Picard, 1997). Prior to 1990 it was believed that perception and decision-making were primarily cortical\(^2\) processes (Cytowic, 1989). Cytowic discovered that the limbic system of the brain (thought to lie beneath the cortex) was triggered during perception and decision-making processes in addition to the cortex. The limbic system is thought to be responsible for emotion, memory, and attention (Cytowic, 1989). This finding blurs the line between “thinking” and “feeling” and has led to the expansion of cognitive appraisal theories of emotion (discussed below).

---

\(^1\) This manuscript will use the terms emotion, affect, and affect state(s) interchangeably. Emotions, in general are thought to entail affect. That is, affect generally refers to states that are comprised of various cognitive-emotional constructs. The broader category of affect allows the discussion to include both emotions and mood. The approach to affect recognition and expression detailed in this work will also be used for recognizing a variety of affective and motivational constructs (i.e., self-efficacy, emotion, empathy, goal judgements, etc.)

\(^2\) The part of the brain nearest the surface.
In general, it is widely believed that customizing users’ interactions can significantly improve their effectiveness in performing tasks. User-adaptability can serve a broad range of functions in intelligent systems. For example, interactive learning applications can adapt their problem-solving challenge levels to better suit individual users. Such systems can help users focus their information-seeking tasks, provide pertinent information that takes into account users’ interests in specific topics, and present information in a manner that is appropriate for users’ visualization preferences (Jameson, 2003). To create effective customized interactions, interactive systems need to not only reason about users’ goals, but also how users feel. This requires recognizing user emotions.

Affect recognition is the task of inferring a user’s affective state from a sequence of observations of behavior. Recognizing affect can draw on a variety of information sources including user system behavior, facial expression, head and body posture, speech, and physiological responses (i.e., changes in heart rate or galvanic skin response) (Frijda, 1986). By providing knowledge about users’
affective state, interactive systems can dynamically modify their behavior to better support individual users, in some cases responding with expressions of emotion.

*Affect expression* is the task of determining how a system should communicate emotion. Expression of affect is often conveyed through both verbal and non-verbal channels including text or spoken dialog, animated agent behavior, color, or music. Affect recognition and expression are two components of the “affective loop” (Ståhl et al., 2005; Sundström et al., 2007) (Figure 1.1), which also includes affect understanding.

*Affect understanding* is the process of interpreting recognized user emotions, determining what it means for the user to feel the recognized emotion, and then formulating adaptation strategies based on how the user feels. Collectively, affect recognition, expression, and understanding have been the subject of growing attention and have been investigated in a broad range of applications relying on a variety of observation sources.

Affect has begun to play an increasingly important role in intelligent tutoring systems. Recent years have seen the emergence of work on affective student modeling (Conati and Mclaren, 2005), detecting frustration and stress (Burleson, 2006; Burleson and Picard, 2004; Prendinger and Ishizuka, 2005), modeling student uncertainty (Forbes-Riley and Litman, 2007), modeling agents’ emotional states (Andre and Mueller, 2003; Gratch and Marsella, 2004; Lester et al., 1999), devising affectively informed models of social interaction (Johnson and Rizzo, 2004; Paiva et al., 2005; Porayska-Pomsta and Pain, 2004), and detecting student motivation (de Vicente and Pain, 2002). All of this work seeks to increase the fidelity with which cognitive-emotional processes are modeled in intelligent tutoring systems in an effort to increase the effectiveness of tutorial interactions and, ultimately, learning.

There is a strong connection between affect and learning. Teachers and tutors alike motivate students to learn and craft educational experiences to increase student efficacy to support learning. Affect influences the cognitive, motivational, and behavioral processes of students (Linnenbrink and Pintrich, 2000), and educational psychologists believe affect impacts learning and cognition in at least four ways: memory, strategy use, attention, and motivation (Pekrun, 1992).

- **Memory.** Students experiencing positive emotions and moods are more likely and better able to retrieve positive information connected with similar positive feelings (Forgas, 2000).
Likewise, Linnenbrink and Pintrich found negative affect to have a negative effect on working memory (2000). Pekrun postulated that material which is learned in the presence of affect, and thus stored with memories of the affective state, is better remembered if the learner is experiencing the same affective state at the time of retrieval (1992).

- **Strategy use.** Positive affect is believed to elicit top-down, schema-based approaches to problem-solving and creative processes (Fiedler, 2000; Forgas, 2000). For example, Csikszentmihalyi’s *state of flow* (1990) is a positive state that is known to elicit these processes. On the other hand, negative affect is more likely to elicit the processing of information in a detail-focused, analytical way (Pekrun, 1992). Negative affect also influences avoidance orientations by shifting student focus to external, less relevant information (Fiedler, 2000). Further, negative affect has been shown to correlate with a reduction in the use of cognitive strategies known to result in deeper learning (Linnenbrink and Pintrich, 2000). Such effects of emotion on strategy selection and use are hypothesized to occur at a lower level than typical strategy selection (i.e., meta-cognitive level) (Pekrun, 1992).

- **Attention.** Because affect can occupy the limited resources of working memory, it can negatively impact the cognitive resources available to students in learning tasks (Linnenbrink and Pintrich, 2000; Pekrun, 1992). In particular, negative affective states such as anxiety, fear, or frustration can result in cognitive resources being utilized to focus on the source of the emotion, such as the worry inducing anxiety or the object causing frustration, rather than on the problem at hand. This finding has been repeatedly demonstrated in studies of test anxiety (Zeidner, 1998). Attentional resources occupied with any emotion (positive or negative) leave fewer resources available for students to achieve. However, in most cases positive emotions are signals of favorable situations, while negative affect is more likely to signal unfavorable situations requiring cognitive resources to cope with the emotion (Pekrun, 1992).

- **Motivation.** Lastly, emotion can affect student learning through motivation. First, affect influences intrinsic motivation, which inspires an individual to complete a task or perform an action solely for its own sake (Schunk et al., 2007). Positive emotions are likely to increase
task enjoyment and peak interest in similar, future tasks (Pekrun, 1992). While negative affect is likely to reduce enjoyment of tasks, thereby impeding intrinsic motivation, it is also likely to encourage negative intrinsic motivation. Negative intrinsic motivation promotes disengaged activity to avoid the activity altogether (Pekrun, 1992). Secondly, affect influences extrinsic motivation, which prompts an individual to engage in tasks and actions as a means to an end, such as a tangible reward or verbal praise (Schunk et al., 2007). Positive affect is likely to promote motivation to obtain positive outcomes, while negative affect (i.e., anxiety) may trigger motivation to avoid outcomes (Pekrun, 1992). For instance, the fear of failure is known to motivate students to exert greater effort to avoid negative outcomes.

Affect can be mediated by memory, strategy use, attentional resources, and intrinsic and extrinsic motivation to effect learning achievement (Pekrun, 1992). Thus, being able to recognize, understand, and respond to student affect could inform adaptable instruction methodologies resulting in valuable impacts on learning.

A common approach to obtaining information about a student’s affect is directly posing questions to her throughout a learning episode. However, periodic self-reports can be disruptive. Another approach is to understand affective experiences through post-hoc reports and analysis. While this approach can be effective in an experimental setting, it does not result in a system’s ability to recognize, understand, and respond through expression of affect. This calls for models of affect that can accurately recognize users’ affect during interaction, understand their affective states, and appropriately respond to their situations. One can distinguish two fundamental approaches to modeling affect: analytical and empirical. In the analytical approach, models of affective constructs (e.g., emotion, self-efficacy, empathy) can be constructed by analyzing the findings in the psychology literature. Although affective phenomena can frequently be explained with psychological models, it is challenging to use these findings as the basis for computational models. In addition, many affective constructs are not well understood. For example, it is only in the past two decades—this is very recent in the history of psychology—that empathy has become a focus of study for social psychologists (Davis, 1994).
An alternative to analytically devising models of affect is the *empirical* approach. If we could create models of affect that were derived directly from observations of “affect in action,” we could create empirically grounded models based on human-computer interactions and human-human behaviors exhibited during the performance of a specific task within a given domain. It is not apparent that this approach could produce universal models of affective constructs. (Universal models may not even be achievable, at least in the near term.) However, the empirical approach could nonetheless generate models of affect that significantly extend the capabilities of a system to adapt itself to its users to support of effective interaction and learning.

1.1 Motivation

Affective computing is beginning to make inroads into HCI research (Hudlicka, 2003; McNeese, 2003; Picard, 1997). Models of affect recognition and affect expression are now being incorporated into at least four types of applications:

*User-adaptive systems*

User-adaptive systems depend on the information about users’ goals, plans, beliefs, and preferences (Jameson, 2003). Recently, adaptive systems have begun to explore how to take user affective states into account to create more effective interactive experiences. For example, SPECTER (Kleinbauer et al., 2003) is a personal assistant that senses the user’s action and affective states based on data from a ubiquitous computing environment to provide personalized recommendations that are the most appropriate for the current context. Another example is the Sensual Evaluation Instrument (Isbister et al., 2006; Isbister et al., 2007) which has been introduced to understand users’ affective response to interactive systems.

*Human language technologies*

Language is emerging as useful resource for identifying user affect (D’Mello et al., 2008; Forbes-Riley et al., 2008; Graesser et al., 2006). Likewise, language is one of the many channels for expressing affect. Language as a communication medium has a long history in human-computer interaction.
Recently, there have been efforts to exploit affect to disambiguate dialogue acts (Bosma and André, 2004) by considering emotional features such as arousal and valence to estimate the probabilities of dialogue acts. The estimated probabilities are used to determine which interpretation is more likely to be correct. Another project, the MARY text-to-speech system (Schröder and Trouvain, 2003), is exploring the generation of affective speech for synthetic agents.

**Entertainment, story, and drama applications**

Recently, entertainment applications such as interactive narrative environments and games have begun to account for affect in both research laboratories and in commercially released products. The Oz Project (Bates et al., 1994; Bates, 1994; Reilly and Bates, 1992) employed emotions in characters and found positive effects on the perceived believability of the characters. Carmen’s Bright IDEAS (Marsella et al., 2000) also utilized emotional characters in a health intervention application designed for mothers of pediatric cancer patients. The characters’ emotions influenced their interactions with users. By incorporating personality and emotional dimensions into character control, the SIMS, a popular PC game from Electronic Arts leverage affect to increase user satisfaction and improve skill acquisition (Brown, 2006).

**Intelligent tutoring systems**

Intelligent tutoring systems (ITSs) comprise an important category of user-adaptive systems. ITSs employ student models, which are used to provide customized hints, generate explanations tailored to individual users, select problems, and determine affective responses for pedagogical agent expression. Student modeling research has investigated a broad spectrum of issues in supporting customized pedagogical decisions. Central problems of student modeling include attempts to recognize students’ intentions, their focus of attention, their current plan, their current goal, and their affective state as they solve problems.

An important form of learning supported by intelligent tutoring systems is exploratory learning, in which students have great freedom to explore the physical environment and the problem-solving
space (de Jong and van Joolingen, 1998). However, it is precisely this freedom that poses a significant challenge to student modeling, in general, and to affect modeling in particular. The openness of the environment and the complexity of student goals, plans, and problem-solving strategies compound the possible situations which might arise in student learning episodes. Together, these factors make it difficult to determine how a student feels and how the system should emotionally respond. When multiple affective states (and/or multiple affective responses) are in play, an affective student model must determine which is most likely given the current context (and which response would be most appropriate or effective).

The assessment of student affective states is an important aspect of student modeling. With information about the student’s affective state, ITSs can interact more effectively with the student. Detecting frustration can enable ITSs to make intelligent intervention decisions, and knowledge of affective states may lead to more accurate and earlier detection of student difficulties and help determine appropriate problem-solving challenge levels. Pedagogical planners can take further advantage of recognized affective states to formulate appropriate tutorial strategy and generate appropriate natural language advice and explanations.

1.2 Affect Recognition and Expression Challenges

Picard defines affect recognition as, “inferring a user’s emotional states from observations of emotional expressions and behavior, and through reasoning about an emotion-generation situation” (Picard, 1997). In other words, because emotions are internal to the observed users, the recognizer has to infer the underlying affective states based on observable indicators. For example, we can infer someone’s emotional states from her gestures or the way she speaks. We can “wire” a person to measure her heart rate and skin conductivity, which can be indicators of her affective states, or at least useful for disambiguating between two candidate affective states.

In contrast, affect expression is the problem of determining which affective state to convey to the user and how to embody the determined affective state. Affect can be expressed through a number of channels including color, music, speech, body language, gestures, posture, and facial

---

3 Intelligent tutoring systems that support exploratory learning are often referred to as interactive learning environments.
expression. The problem of affect expression is compounded in intelligent tutoring systems where we are not only interested in selecting the most appropriate affective state, but further choosing the affective state that is most effective for learning in the given situation.

Affect recognition and expression are difficult tasks. For example, drawing inferences about users’ affective states inherently involves reasoning under uncertainty because users can feel a variety of emotions and users in the same or similar situations may experience very different affective states. Much useful information about the user is not directly observable, as users frequently do not or cannot explicitly express their feelings. Further, even with concrete information about user affective state, there are a multitude of ways to effectively respond, which again exhibits variability among individual users as to what responses are most effective. In general, affect recognizers and expressers should address the following requirements: the selection of pertinent appropriate affective states, scalability and speed, early prediction, partial observability and noisy environments.

Selecting Affective States

There are several open questions in emotion research. Fundamental issues such as how many emotions exist, which emotions are pertinent in which types of applications (i.e., intelligent tutoring systems), and how do such emotions differ across demographics (i.e., age, gender) and cultures are all open research problems. Emotion research and affective computing have explored a broad range of affective states. These range from Elliott’s 26 emotions (1992), based on the OCC model’s 22 emotions (Ortony et al., 1988) to Ekman’s 6 basic emotions expressed in the face (Ekman, 2003; Ekman and Friesen, 1978) and D’Mello and Graesser’s sets of 8 and 7 emotions investigated in their intelligent tutoring system, AutoTutor (D’Mello et al., 2008). Recent research has even begun to isolate single affective states in efforts to explore recognition, understanding, and expression issues in the context of specific states such as uncertainty (Forbes-Riley et al., 2008), confusion (Forbes-Riley and Litman, 2007), frustration (McQuiggan et al., 2007), and flow (Burleson, 2006). To date, it is unclear which affective states comprise a comprehensive set of emotions, especially in learning experiences. D’Mello and Graesser have been the first to explore which emotions occur in ITS learning interactions (2008), but their set of explored emotions has evolved from study to study.
Additionally, we suspect that emotions in traditional ITSs may be very different from the affective states reported in interactions with narrative learning environments, such as CRYSTAL ISLAND, the test bed learning environment in which the majority of our experiments are conducted.

**Scalability and Speed**

Affect recognition and expression involve narrowing down possible hypotheses that best explain, characterize, or respond appropriately to observed user behavior. Classic approaches to affect recognition employ post-hoc analyses to understand the affective experiences of users. However, while such analyses are informative, they do not recognize user affect during runtime. Runtime affect recognizers and expressers need to make predictions that satisfy the realtime requirements of interactive systems. While probabilistic approaches to affect recognition and expression are attractive, they also present scalability challenges. For example, Dempster-Shafer Theory (Bauer, 1996a; Wilson, 2000) is known to be exponential in general and exact inference in Bayesian networks is an $NP$-hard problem. Because most interactive applications require affect recognition and expression results quickly (i.e., before the user succumbs to negative affective state), ideally the recognition should be done quickly (on the order of a few milliseconds) to appropriately serve the user.

**Early Prediction**

For interactive environments, affect recognizers should make early accurate predictions and they should converge as quickly as possible on the most likely affective state. These requirements are often problematic in plan and goal recognition (Albrecht et al., 1998; Blaylock & Allen, 2003; Mott et al., 2006; Yin et al., 2004). Early prediction of user affect affords narrative and tutorial planners sufficient time in the event that replanning is necessary. In addition, even if affect recognizers and expressers computation is fast enough to address scalability issues, if they are not able to predict affective states until they observe the affective state or the user confirms the affective state (i.e., via self-report), they would be impractical for many interactive environments.
Noisy Environments

User’s affective states may be generated in response to any number of factors (both external and internal) to interactive systems. The case may therefore arise where affect recognizers are unable to observe the entire situation resulting in particular user emotions. Furthermore, the individual differences of users may explain or characterize conditioned responses to various situations arising in interactive systems. Thus, affect recognizers and expressers must account for individual differences (i.e., personality, goal and learning orientations, domain efficacy, etc.) to better model affect for each individual user. Additionally affect recognizers and expressers that rely on physiological input must deal with noisy signals and misread biofeedback data that may not be consistent with observations of user actions in the system to effectively model affect. Thus, an effective solution to the affect recognition problem must cope with the uncertainty, real-time requirements, scalability issues, and absence of direct access to users’ ongoing cognitive, motivational, and affective processing.

1.3 The CARE Approach

To effectively reason about affect, affect models should satisfy three requirements. First, they should be realized in a computational mechanism that operates at runtime. Affective states are likely to vary throughout a learning episode, so pre-learning instrumentation may or may not be predictive of student affect at specific junctures during an interactive session. Second, affect diagnosis should be efficient. It should satisfy the real-time demands of interactive learning. Third, affect diagnosis should avoid interrupting the learning process. A common approach to obtaining information about a student’s affective-cognitive state is directly posing questions to them throughout a learning episode. However, periodic self-reports are disruptive.

To satisfy these requirements, we propose an empirical approach to devising computational models of affect that are accurate, runtime efficient, and non-disruptive to learning experiences. The proposed approach has been implemented in an inductive framework, called CARE (Computational Affect Recognition and Expression). CARE is a data-driven affective architecture and
methodology for learning models of affect from observation of student interactive behaviors. It has been modified and evaluated in a variety of affect recognition and affect expression experiments designed to investigate the inductive approach to generating models of affect. For example, we have used CARE to recognize a number of user emotional states (e.g., excitement, fear, frustration, sadness) (Lee et al., 2007; McQuiggan et al., 2006; McQuiggan et al., 2007) and student self-efficacy levels (McQuiggan and Lester, 2006; McQuiggan et al., to appear). For expressing affect, we have also used CARE to learn empirical models of empathy for companion and pedagogical agents (McQuiggan and Lester, 2006; McQuiggan and Lester, 2007; McQuiggan, Robison et al., 2008). Finally we have conducted several experiments to evaluate the perceived accuracy of empathy models and the relative impact of affect models on user and student experience (McQuiggan and Lester, 2007; McQuiggan, Rowe et al., 2008; McQuiggan, Robison et al., 2008). Collectively, these studies investigate the affective loop in the areas of affect recognition, affect expression, and user and student outcomes.

1.4 Contributions

This research primarily considers three points of the affective loop (Figure 1.2).

This work includes the design, implementation, and evaluation of an inductive framework for generating affect recognition and affect expression models. The implications and conclusions of this research are supported by a number of experiments investigating the merit of employing affect in interactive systems, with a focus on interactive learning environments. This includes user and student-focused experimentation. The work described in this dissertation yields the following contributions:

CARE framework. We have devised a framework for developing computational affect recognition and expression models. We have used the framework and associated methodologies to recognize user affect and to model affective expression in synthetic agents.

Inductive approach. The inductive approach to affect recognition and affect expression for learning models from training data is a novel bottom-up approach. The framework learns affect recognizers
and expressers from observation sequences of user actions, interactive system locations, physiological data, and world states.

**Affect Recognizers.**

- **Emotion recognition.** The CARE framework has been used to induce models of emotion. We consider various sets of emotion as we explore the relative emotional space. Induced models will consist of affect recognizers that predict user emotion from a set of candidate emotions. Other affect recognizers will focus on a single affective state (i.e., frustration and flow). These affect recognizers will make Boolean predictions of whether the user is experiencing the particular emotion or not.

---

**Figure 1.2.** Affective loop contributions. A – Affect recognition; B – Affect expression; C – Affective outcomes and individual differences.
• **Self-efficacy recognition.** We also use the CARE framework to induce models of self-efficacy. We consider various levels of granularity at which to model efficacy such as LOW vs. HIGH. These models determine whether students are confident in their abilities in relation to current learning tasks.

**Affect Expressers.**

• **Empathy models.** The CARE framework has been used to induce models of empathy. Empathy modeling is comprised of two complementary models: one for empathetic assessment (when to be empathetic), and one for empathetic interpretation (when empathy is called for, what affective state should be expressed). Empathy models can be used to drive affective expression modules for characters, pedagogical agents, companion agents, and various other synthetic characters of interactive systems.

• **Parallel and reactive empathy modeling.** We further used the CARE framework to investigate modeling of advanced empathy models. These models are able to distinguish between parallel and reactive empathy determining when each is most appropriate for various situations that arise in interactive systems, such as intelligent tutoring systems.

**Real-time and early methods for predicting affect.** Affect recognizers and expressers function in real-time. Two types of models are investigated, those that make predictions based on current situational contexts and those that make early predictions for future affective experiences.

**Empirical studies.** We conducted a series of empirical studies to assess the CARE framework and the effects of employing affect in interactive systems. The framework was primarily evaluated in an implemented intelligent tutoring system with college level (graduate and undergraduate), high school, and middle school students participating as subjects.

**Affective and motivational outcomes.** We explored the affective and motivational impact on student users of the implemented environments considered in this research. For example, we investigated how empathetic characters affect students’ sense of involvement in the learning environment. Considering such research questions evaluates the merit of implementing the work described her
1.5 Summary of Results

Our work has yielded three categories of results from studies conducted with more than 700 subjects in the areas of affect recognition, affect expression, and affective and motivational outcomes. Table 1.1 summaries completed studies. Studies 2, 3, 5, 6, 7, and 8 fall under the category of affect recognition. In each of these studies, the signified construct was the task of recognition. Studies 1 and 10 focus on modeling affect expression. Lastly, Studies 4, 8, and 10 focused, either in part (Study 10) or entirely (Studies 4 and 8), on evaluation. Below is a brief summary of each study. Table 1.2 summarizes the participants and materials of each study.

1. **Empathy modeling (completed October 2005).** This work investigated the CARE inductive approach to modeling empathy, both empathetic assessment (when empathy is called for) and empathetic interpretation (when empathy is called for, which emotion should be expressed). A Wizard-of-Oz study design was utilized allowing wizards to observe subjects’ interactions in the Treasure Hunt virtual environment. When the wizard believed empathetic expression from an accompanying companion agent was appropriate, she selected the emotion to be expressed by the companion resulting in a short spoken remark and body gesture by the companion agent. Models were constructed from representations of ongoing Treasure Hunt interactions monitoring the subject’s character location, actions, and progression. Empathetic assessment was best modeled by decision trees resulting in 89% accuracy. Empathetic interpretation was modeled with an accuracy of 80% by naïve Bayes models. Details of this study are reported in Chapter 9.

2. **Self-efficacy modeling (completed November 2005).** We conducted this foundational study to investigate the prospect of using the inductive approach to model self-efficacy. The online tutorial system consisted of an online interaction with genetic materials followed by a problem solving session in which subjects were wired and rated their self-efficacy. Models were constructed from representations of ongoing situations in the online tutorial system accounting for. The best performing models were able to classify student self-efficacy as HIGH EFFICACY or LOW EFFICACY with 87% accuracy.
<table>
<thead>
<tr>
<th>Construct</th>
<th>Environment</th>
<th>Description of Study</th>
<th>N</th>
<th>Publication(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 Self-efficacy</td>
<td>Online Tutorial</td>
<td>Subjects interacted with an online tutorial in the domain of genetics. Subjects then solved a series of related problems. Subjects were wired and self-reported self-efficacy.</td>
<td>33 G</td>
<td>McQuiggan and Lester, 2006 McQuiggan et al, 2008</td>
</tr>
<tr>
<td>3 Physiological Response</td>
<td>Treasure Hunt</td>
<td>Subjects interacted in the environment collecting treasures. Subjects were wired.</td>
<td>20 U &amp; G</td>
<td>McQuiggan et al, 2006</td>
</tr>
<tr>
<td>4 Empathy</td>
<td>Treasure Hunt</td>
<td>Subjects were presented a series of video clips of empathetic expressions in the environment from 3 sources: CARE-induced, human-controlled, &amp; reversed-human.</td>
<td>31 U</td>
<td>McQuiggan and Lester, 2007</td>
</tr>
<tr>
<td>5 Self-efficacy</td>
<td>Crystal Island</td>
<td>Subjects solved the scientific mystery while self-reporting on self-efficacy.</td>
<td>42 U</td>
<td>McQuiggan et al, 2008</td>
</tr>
<tr>
<td>6 Emotion Set</td>
<td>Crystal Island</td>
<td>Subjects solved the scientific mystery while self-reporting on emotions from a set of six emotions (excitement, fear, frustration, happiness, relaxation, and sadness).</td>
<td>36 G</td>
<td>Lee et al., 2007 McQuiggan et al, 2007</td>
</tr>
</tbody>
</table>
### Table 1 (cont.). Chronological summary of completed studies.
(Subject classifications: MS—middle school; HS—high school; U—undergraduate; G—graduate).

<table>
<thead>
<tr>
<th>Construct</th>
<th>Environment</th>
<th>Description of Study</th>
<th>N</th>
<th>Publication(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Frustration Threshold Crystal Island</td>
<td>Subjects solved the scientific mystery (comprised of insolvable goals). Subjects periodically self-reported emotion. When emotion was frustration subjects offered opportunity to continue or quit current goal.</td>
<td>46 U</td>
<td>n/a</td>
</tr>
<tr>
<td>8</td>
<td>Empathy Presence Crystal Island</td>
<td>Subjects solved the scientific mystery. Condition A consisted of 3 empathetic characters who responded to self-reported emotion. Condition B had no empathetic characters.</td>
<td>55 MS &amp; 35 HS</td>
<td>McQuiggan et al., 2008</td>
</tr>
<tr>
<td>9</td>
<td>Goal Monitoring Crystal Island</td>
<td>Subjects solved the scientific mystery while self-reporting on goal monitoring. Condition A presented the environment as a mastery task. Condition B presented the environment as a performance task.</td>
<td>59 MS &amp; 36 U</td>
<td>Nietfeld et al., 2008</td>
</tr>
<tr>
<td>10</td>
<td>Empathy       Crystal Island</td>
<td>Subjects solved the scientific mystery. Characters empathetically responded to reported emotions (either in parallel or reactivity). Subjects evaluated responses.</td>
<td>33 G</td>
<td>McQuiggan et al., 2008</td>
</tr>
<tr>
<td>11</td>
<td>Presence Learning Crystal Island</td>
<td>Subjects were randomly assigned to 3 conditions: Crystal Island narrative, Crystal Island non-narrative, or PowerPoint. The study assesses learning and presence against the effects of story-based learning.</td>
<td>179 MS</td>
<td>McQuiggan et al., 2008</td>
</tr>
<tr>
<td>12</td>
<td>Goal Monitoring Crystal Island</td>
<td>Replicating Study 9 using the Crystal Island non-narrative condition from Study 11 subjects solved the scientific mystery while self-reporting on goal monitoring in a mastery or performance instructed condition.</td>
<td>68 MS</td>
<td>n/a</td>
</tr>
</tbody>
</table>
3. **Physiological response change modeling** (*completed December 2005*). In this study we investigate the inductive approach to modeling changes in physiological response based on observations of user behavior in the Treasure Hunt environment. This study was motivated by the desire to predict changes in user physiology without requiring users to be tethered to intrusive biofeedback hardware during runtime usage. We explored naïve Bayes, Bayesian networks, and decision tree models. Decision tree models were able to accurately model changes in heart rate with 92% accuracy and skin conductivity with 93% accuracy. Models predicted changes of UP, DOWN, or SAME for both heart rate and skin conductivity.

4. **Empathy evaluation** (*completed March 2006*). A focus group study was conducted to evaluate the perceived accuracy of empathetic interpretation and assessment of companion agents. Subjects were presented a series of video clips depicting companion agents empathetically responded to various situations. Subjects evaluated companion agents whose empathetic interpretation was controlled by either a CARE-induced model of empathy, a human, or a reversed human assessment. CARE model empathetic expressions were evaluated to be comparable to human assessments and interpretations (no statistically significant differences).

5. **Self-efficacy modeling II** (*completed April 2006*). Subjects interacted with the CRYSTAL ISLAND narrative-centered learning environment, periodically self-reporting on self-efficacy. Models of self-efficacy were induced from observations of student behavior in the environment including representations of subject actions, locations, and other world state information. Decision tree models were able to correctly classify efficacy levels (HIGH or LOW) with 87% accuracy.

6. **Emotion modeling** (*completed November 2006*). Subjects interacting with the CRYSTAL ISLAND environment solved the science mystery while periodically reporting on their affective states, choosing from a set of six emotions (excitement, fear, frustration, happiness, relaxation, and sadness). Various models of emotion have been induced from observations of student behavior in the environment to predict student affective states. The first set of models predicted affective state using naïve Bayes and decision trees resulting in the best
performing model with 95% accuracy. We then investigated approaches to early prediction of student frustration by collapsing the dataset to two states: FRUSTRATION and NOT-FRUSTRATION. Induced models were able to predict student frustration up to 30 seconds before confirmation of student frustration (self-reported frustration) with 89% accuracy. Details of this study are reported in Chapter 7.

7. **Frustration threshold modeling (completed March 2007).** This study was designed to investigate the inductive approach to model frustration thresholds (the point at which students quit in the face of frustration). Subjects solved the science mystery of CRYSTAL ISLAND while periodically self-reporting on affective state, from a set of six emotions (excitement, fear, frustration, happiness, relaxation, and sadness). If subjects reported frustration, they were offered an opportunity to quit working on the current goal, or continue attempting to solve the current sub-problem. The environment included 3 tasks that were impossible to complete. Thus, subjects had to quit the impossible tasks at some point to continue solving the mystery. Collected data has not yet been analyzed and models have not been induced. See Chapter 7 for addition details of this study and the research plan for scheduled data analysis.

8. **Empathy evaluation II (completed May 2007; June 2007).** This study was designed to investigate the impact of empathetic characters on students’ sense of presence, particularly their sense of involvement in the educational narrative. Students’ participating in the condition whose environment was populated with empathetic characters reported significantly higher senses of presence and involvement then students participating in the condition with characters who were not empathetic. See Chapter 9 for more details on this study.

9. **Metacognitive monitoring (completed May 2007).** This study design involved two conditions. The first presented the environment as a mastery task in which students are expected to learn as much as possible. The second condition presented the environment as a performance task in which students were to obtain the highest scores possible (points were awarded for correctly solving the scientific mystery of CRYSTAL ISLAND in as few steps as
Table 2. Constructs, instrumentation, environments, and subject populations by study. (Subject classifications: MS = middle school; HS = high school; U = undergraduate; G = graduate).

<table>
<thead>
<tr>
<th>Constructs</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Emotion</td>
<td>x</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>x</td>
</tr>
<tr>
<td>Empathy</td>
<td>x</td>
</tr>
<tr>
<td>Presence</td>
<td>x</td>
</tr>
<tr>
<td>Metacog Monitoring</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Demographics</td>
<td>x</td>
</tr>
<tr>
<td>IRI</td>
<td>x</td>
</tr>
<tr>
<td>Goal Orientation</td>
<td>x</td>
</tr>
<tr>
<td>Achievement Goals</td>
<td>x</td>
</tr>
<tr>
<td>Domain Self-efficacy</td>
<td>x</td>
</tr>
<tr>
<td>Self-efficacy for SRL</td>
<td></td>
</tr>
<tr>
<td>Hope</td>
<td>x</td>
</tr>
<tr>
<td>Learning Beliefs</td>
<td>x</td>
</tr>
<tr>
<td>Personality</td>
<td>x</td>
</tr>
<tr>
<td>Immersion</td>
<td>x</td>
</tr>
<tr>
<td>Presence</td>
<td>x</td>
</tr>
<tr>
<td>Learning Assess.</td>
<td>x</td>
</tr>
<tr>
<td>Computer Usage</td>
<td></td>
</tr>
<tr>
<td>Character Eval.</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Environments</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Online Tutorial System</td>
<td></td>
</tr>
<tr>
<td>Treasure Hunt</td>
<td>x</td>
</tr>
<tr>
<td>Crystal Island (1.0)</td>
<td></td>
</tr>
<tr>
<td>Crystal Island (2.0)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Student Subjects</th>
<th>G</th>
<th>G</th>
<th>U</th>
<th>G</th>
<th>U</th>
<th>U</th>
<th>G</th>
<th>U</th>
<th>MS</th>
<th>HS</th>
<th>MS</th>
<th>U</th>
</tr>
</thead>
</table>
Throughout the interaction students were asked to monitor their goal progression. In Chapter 6 we detail induced models of student metacognitive monitoring.

10. **Empathy modeling II** (*completed October 2007*). This study has been designed to extend the previous approach of empathy modeling to account for both reactive and parallel empathy. Subjects evaluated the utility of character empathetic expressions to their situation allowing models to account for what works best in particular situations for different individuals. Further details of this study can be found in Chapter 9.

11. **Learning and presence** (*completed November 2007*). This study design controls for the role of narrative in learning environments. Participants were randomly assigned to four conditions: CRYSTAL ISLAND narrative, CRYSTAL ISLAND non-narrative, PowerPoint, and a holdout condition. The CRYSTAL ISLAND narrative condition includes the full curriculum and full storyline of Crystal Island 2.0. The CRYSTAL ISLAND non-narrative condition removes the full storyline, leaving only the story necessary to support the curriculum. The PowerPoint condition includes exactly the same curriculum content presented through a media different from CRYSTAL ISLAND, with no story. The participants in the holdout group serve as a control group for this study and planned longitudinal studies. A number of dependent variables are considered in this study design including learning, self-efficacy, and presence.

12. **Modeling metacognitive monitoring** (*completed December 2007*). This study replicates Study 9 with a new population. The design again consists of two conditions a performance and mastery condition in which the instructions for each condition are intended to guide students to approach CRYSTAL ISLAND with performance goals or mastery-oriented goals, respectively. The goals of study were to evaluate the extent to which goal-orientations can be manipulated and how a game-based learning environment affects performance-oriented students to take on mastery-oriented goals. We also investigated the use of CARE to model student metacognitive monitoring.
1.6 Dissertation Organization

This dissertation is organized as follows: Chapter 2 provides background on affect recognition and expression. Chapters 3 and 4 present the inductive approach to affect recognition and expression. The interactive systems utilized in this work are (1) an online tutorial system in the domain of genetics, (2) TREASURE HUNT, an interactive environment, and (3) CRYSTAL ISLAND, a narrative-centered learning environment in domains of genetics and microbiology. Each environment is described in Chapter 5. Chapter 6 presents the design, implementation, and evaluation of CARE for modeling self-efficacy. Chapter 7 presents modifications, implementations, and evaluations of CARE for modeling affect, including studies specifically focused on frustration and flow. Chapter 8 details the use of CARE to model student goal monitoring judgments. Chapter 9 describes studies which use CARE to model empathy (affect expression). Chapter 10 presents a series of studies and experiments designed to assess the affective experience of students interacting with interactive learning environments. Chapter 11 reflects on the studies discussed in the preceding chapters drawing implications from the findings and concludes, describing future directions of this research.
Chapter 2

Background and Related Work

Affective reasoning has been the subject of increasing attention in recent years. Collaborations between computer scientists, cognitive scientists, and psychologists are giving rise to computational models of affect. The CARE framework, presented in the next chapter, is used to generate computational models of affect for the tasks of recognizing and expressing emotion. This chapter begins with an overview of appraisal theory, which serves as the theoretical basis for the CARE inductive framework. The following section then presents related work, summarizing recent research at the intersection of affective computing and intelligent tutoring systems.

2.1 Appraisal Theory

Emotions comprise an intricate system in humans that influence our interactions, our behavior and our cognition. Emotions are almost always being expressed by the words we say, our facial expressions, our posture, and our actions. Recent findings have found that emotions play a role in decision making, perception, learning and problem solving and impact the activity and achievement of a broad range of tasks. In recent years, there has been a shift in research on models of emotion, based on the discovery that emotion is found in perception and decision-making processes. Prior to 1990 it was believed that perception and decision-making were primarily cortical (relating to the part of the brain nearest the surface) processes (Cytowic, 1989). Cytowic discovered that the limbic systems of the brain (thought to lie beneath the cortex) were triggered during perception and decision-making processes in addition to the cortex. The limbic systems are thought to be
responsible for emotion, memory and attention (Cytowic, 1989). This finding blurs the line between “thinking" and “feeling" and has led to the development of cognitive appraisal theories of emotion (Ellsworth and Scherer, 2003; Frijda, 1986; Ortony et al., 1988; Lazarus, 1991; Scherer, 1984; Smith and Lazarus, 1990; Roseman, 1991; Roseman et al., 1996; Roseman and Smith, 2001).

2.1.1 Overview of Appraisal Theory
Appraisal theory seeks to uncover the underlying cognitive processes that give rise to emotion. To understand appraisal theorists’ approach to emotion we consider the problems Roseman and Smith (2001) identified that appraisal theories seek to explain. These problems have caused complications in alternative models such as stimulus-response theories (Watson, 1919) and behavioral theories (James, 1890). Here we take several of the problems in turn and briefly summarize appraisal theory’s resolution.

1. How can we account for the differentiated nature of emotion response? Appraisal theorists’ seek to define appraisal configurations, that is, patterns of interpretation of person-environment relationship, that give rise to emotional states. This view differs from the behaviorist conceptualization of emotion as a dimension of behavior stemming from physiology.

2. How can we explain individual and temporal differences in emotional response to the same event? Appraisal theorists contend that it is not the event that evokes the emotion. Instead, it is the interpretation of the event and its relevance that gives rise to emotion. It is widely believed that few, if any, patterns exist tying situations to emotions. Different emotions will stem from the same event if there are different interpretations of events.

3. How can we account for the range of situations that evoke the same emotion? Individuals that interpret events in the same way, using the same appraisal configuration, should feel the same emotion. This question again addresses the lack of one-to-one relationships between situations and emotions, highlighting the observation that it is the appraisals of situations that give rise to emotion.

4. What initiates the process of emotional response? Roseman claimed that few emotion models have identified what initiates the emotion process (1984). Appraisal theorists
contend that appraisal, the evaluation of person-environment relationship and associated judgments, precede affective outcomes. This hypothesis will be highlighted in the next section of this chapter in the discussion of Smith and Lazarus’ model.

5. How can we explain the appropriateness of emotional responses to the situations in which they occur? While prior research viewed emotion as disorganized responses, appraisal theorists are identifying patterns that consistently give rise to emotions. Further, appraisal theorists view emotion as having utility in the appraisal process, particularly as a way to cope with situations. Such appraisal configurations suggest the rational aspects of emotional outcomes as they can be expected given certain patterns of appraisal and have utility in managing the person-environment relationship in a desirable fashion.

These issues highlight the components of appraisal theory that make it distinct as an emotional theory. Instead of viewing emotion as a stimulus-response process or a behavioral dimension, appraisal theorists define an evaluative process in which judgments are made regarding the events of unfolding situations and the personal relevance of such events. To understand appraisal theory we describe several appraisal models of emotion beginning with Smith and Lazarus’ process-oriented model, followed by structural models by Ortony et al (1988) and Roseman (1996).

**Smith and Lazarus’ Model**

The Smith and Lazarus appraisal theory of emotion (Figure 2-1) characterizes emotion as a two-staged cognitively informed process consisting of appraisal and coping (Lazarus, 1991; Smith and Lazarus, 1990). **Appraisal** refers to one’s interpreted relationship with her surrounding physical and social environment. Appraisal is a cognitively-constructed representation of events and how these events relate to internal goals. **Coping** is the process by which one considers actions that repair, maintain, or manipulate the existing relationship (both external and internal) with the environment based on affective behavioral tendencies, current affective state, desired affective state, and physiological factors (Lazarus, 1991). Coping determines the response to appraised situations based on past, present and future events.
Appraisal theorists have identified several components that collectively serve to judge the person-environment relationship. These, *appraisal variables*, include relevance, desirability, likelihood, expectedness, and coping potential (see Gratch and Marsella, 2004b for a more extensive
set). These judgments determine the meaning of person-environment relationship, accounting for the events surrounding the current situation and existing goals, beliefs, knowledge, and intentions (Lazarus, 1991). The attributes of the person-environment relationship are referred to as antecedent variables. The appraisal process is an evaluation of antecedent variables (e.g., environment conditions) which are judged according the appraisal variables (e.g., relevance) leading to emotional outcomes that reflect what has happened in relation to oneself, i.e., whether expectations have been met or goals realized, and the significance of the events (Lazarus, 1991). The heart of appraisal theory seeks to describe this process that leads from the situation construal and antecedent variables of the person-environment relationship to its interpretation and meaning (appraisal variables), resulting in emotion. In addition to emotional states, primary appraisal leads to action tendencies and physiological responses. Action tendencies refer to biologically wired coping strategies (e.g., the nature of humans to avoid or escape objects evoking fear) providing links between affective states and their associated physiological responses (Lazarus, 1991). Primary appraisal and emotional outcomes, possibly in conjunction with action tendencies, lead to changes in physiology from increases in heart rate, to decreases in skin conductance, to facial expressions (Frijda, 1986; Lazarus, 1991). Thus, the outcomes of primary appraisal are action tendencies, affective states, and physiological changes.1

The coping processes, or secondary appraisals, refer to cognitive and behavioral alterations made to the person-environment relationship to either maintain desirable ones or repair less-desirable situations. Coping is used to manage the conditions which led to an emotion, directing changes in subsequent appraisals (Lazarus, 1991). The coping process is comprised of two types of coping strategies: those that motivate change in the environment, which are known as problem-focused coping strategies, and those that change the interpretation of the person-environment relationship, which are known as emotion-focused coping strategies. Problem-focused coping acts to change the actual external relationship with the environment such as removing objects believed to cause negative emotion or seeking external assistance (Gratch and Marsella, 2004b; Lazarus, 1991). Emotion-focused coping, or cognitive coping, seeks to change the way one views the

---

1 Several appraisal theorists have begun to identify appraisal patterns, combinations of appraisal variables, which lead to specific affective states (Ortony et al., 1988; Roseman,). While Lazarus has ventured into identifying such patterns, the Smith and Lazarus appraisal model of emotions is more concerned with the process.
environment in an effort to modify the next appraisal. This type of internal coping effects how the person-environment relationship is interpreted and what meaning it will have. Such emotion-focused strategies consist of refocusing attention, changing goals, and disengaging in cognitively disengaging in situational events (see Gratch and Marsella, 2004b for a list of coping strategies). Coping behavior arises from the collection of cognitive appraisal, emotional outcomes, and coping strategies. Note that this theory focuses on the underlying cognitive processes that produce emotions and not a theory seeking to define affective states, like that of the OCC model (Ortony et al., 1988).

2.1.3 Other Structured Appraisal Theory Models
Largely, appraisal models to date have been concerned the contents of appraisal, that is, the variables, and their respective values, that lead to specified affective states. Smith and Lazarus’ model, presented in the previous section, is focused on the cognitive process of appraisal, rather than defining specific patterns that give rise to specific emotions. Many structured models have received attention in the affective computing literature, including the OCC model (Ortony et al., 1988) and Roseman’s model (Roseman, 1996). Both of these models identify structured patterns of appraisal variables that give rise to specific emotional states.

The OCC model (Figure 2-2) (Ortony, Clore, and Collins, 1998) is based on the assumption that human behavior is inherently goal-directed (Omdahl, 1995) and attention is focused on one of three entities: events, agents, or objects (Ortony et al., 1988). According to the OCC model emotion arises from valenced (a dimension ranging from positive to negative) reactions to events, agents, or objects.

- **Valenced reactions to consequences of events** (pleased/displeased). If one is focused on the consequences of events for another, then she must determine whether the consequences are desirable or undesirable. If the consequences of events are desirable for the other person, then she will feel happy-for or resentment toward the other person, depending on her own goals. If the consequences of the events are undesirable for the other person, then she will gloat or feel pity for the other person, again depending on her
Figure 2.2. OCC Model from Ortony et al., 1988.
own goals (Ortony et al., 1988). If, on the other hand, she is focused on the consequences of events for herself, then she will appraise the extent to which the consequences are relevant to her. If she determines the consequences to be relevant, she will feel hope or fear, depending on her goals. If the hope (or fear) is confirmed she will feel satisfaction (or fears-confirmed). If, however, hope (or fear) is disconfirmed she will feel disappointment (or relief). If she determines the consequences to be irrelevant, she will be joyful or distressed depending on her goal state (Ortony et al., 1988).

- **Valenced reactions to actions of agents** (approving/disapproving). If one is focusing on herself as an agent, she will feel pride or shame. On the other hand, if one is focused on another, she will feel admiration or reproach, depending on her goals (Ortony et al., 1988).

- **Valenced reactions to aspects of objects** (liking/disliking). If one likes aspects of objects she will feel love. She will feel hate if she dislikes aspects of the objects (Ortony et al., 1988).

Lastly, the OCC model accounts for interactions between well-being and attribution appraisals. Thus, if one feels joy that the consequences are irrelevant and attributes that irrelevance to another agent, she will feel gratitude. If she attributes the irrelevance to herself, she will feel gratification. On the other hand, if one feels distress that the consequences are irrelevant and attributes that irrelevance to another agent, she will feel anger. However, if she attributes the irrelevance to herself she will feel remorse.

The OCC model has received much attention in affective computing literature. The next section of this chapter will detail several projects that have utilized the appraisal configurations of the OCC model. Largely, the OCC model has been used for affect expression, starting with Elliott’s Affective Reasoner (1992). However, the OCC appraisal configurations were originally designed for the task of recognizing affect.

Roseman’s model (Figure 2-3) identifies a set of appraisal configurations and resulting emotional states. This model consists of five cognitive dimensions (Omdahl, 1995; Roseman, 1996):

1. Motivation towards a desired or undesired state,
2. Situational state - either motive-consistent (positive) or motive-inconsistent (negative),
3. Certainty (weak, strong, or unknown),
4. Legitimacy (deserved or undeserved), and
### Figure 2.3. Roseman’s model (1996) from Picard, 1997.

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Motive-Consistent</td>
<td>Motive-Inconsistent</td>
</tr>
<tr>
<td></td>
<td>Appetitive</td>
<td>Aversive</td>
</tr>
</tbody>
</table>

A. Caused by circumstances

1. Unknown

2. Weak
   - a. Uncertain
     - Hope
   - b. Certain
     - Joy
     - Relief

3. Strong
   - a. Uncertain
     - Hope
   - b. Certain
     - Joy
     - Relief

B. Caused by others

1. Weak
   - a. Uncertain
   - b. Certain

2. Strong
   - a. Uncertain
   - b. Certain

C. Caused by self

1. Weak
   - a. Uncertain
   - b. Certain

2. Strong
   - a. Uncertain
   - b. Certain

---

5. Agency, what or who caused the event.

Roseman assigns each appraisal configuration one of 14 unique emotional states. Roseman’s emotional states differ from the 22 emotions defined in the OCC model and include: anger, disgust, disliking, fear, frustration, guilt (shame), hope, joy, liking, pride, regret, relief, sorrow, and surprise.
(Roseman, 1996). Figure 2-3 contains the appraisal patterns; aligning each of these affective states with its associated appraisal configuration. To date, Roseman’s model has received attention from the affective computing community for its alternatively defined appraisal configurations. However, we are unaware of any current computational implementations in the literature.

Because there is variability in the structured models being developed by appraisal theorists’ and because the appraisal patterns are still very much the focus of much needed empirical work, we seek to exploit the theoretical basis Smith and Lazarus have defined in their process model. This point will be discussed further in the next chapter, describing our inductive approach to modeling affect.

2.2 Affective Computing
Affect influences our interaction with others, our behavior, our perception, and our decision making. Because affect plays a central role in human cognition, it is widely believed that affect modeling can contribute to a broad range of computational tasks (Picard, 1997). Affective computing investigates techniques for giving computers the ability to effectively recognize, understand, and express emotion. Incorporating affective computing into interactive applications holds much appeal. For example, affect-informed educational, training, and entertainment systems may create more effective, interesting, customized experiences for users.

2.2.1 Affect Recognition
Affect recognition (Picard, 1997) is the task of identifying the affective state of an individual from a variety of physical cues, which are produced in response to affective changes in the individual. These include visually observable cues such as body and head posture, facial expressions, and posture, and changes in physiological signals such as heart rate, skin conductivity, temperature, and respiration (Allanson and Fairclough, 2004; Frijda, 1986). Psychologists have used electroencephalograms (EEG) to monitor users’ brain activity for detection of task engagement (Pope et al., 1995) and user attention (Mekeig and Inlow, 1993), electromyograms (EMG) to detect electrical activity in muscles to obtain measurements of users’ sense of presence in virtual
environments (Weiderhold et al., 2003), and eye tracking devices to measure pupil responses to emotional stimulations (Partala and Surakka, 2003). Heart rate measurements have been used to adapt challenge levels in computer games (Gilleade and Allanson, 2003), detect frustration and stress (Prendinger et al., 2005), and monitor anxiety and stress (Healey, 2000). Galvanic skin response (GSR) has been correlated with cognitive load (Verwey and Veltman, 1996) and used to sense user affective states, such as stress (Healey, 2000), student frustration for learning companion adaptation (Burleson, 2006), frustration for life-like character adaptation in a mathematical game (Prendinger et al., 2005), and multiple user emotions in an educational game (Conati, 2002). Heart rate and GSR have jointly been used to determine user affect (Prendinger and Ishizuka, 2005) based on the model of Lang (1995), which characterizes emotions in a two-dimensional space of valence (positive to negative) and arousal (low to high).

Affect recognition work has explored emotion classification from self reports (Beal and Lee, 2005), post-hoc reports (de Vicente and Pain, 2002), self-reports, peer reports, and judges’ reports trained to recognize emotion in the face based on the work of Ekman and Friesen (1978) (Graesser et al., 2006), posture (Mota and Picard, 2003), and multimodal classifications including combinations of visual cues and physiological signals (Burleson, 2006; Burleson and Picard, 2004; Picard et al., 2001), and facial and head gestures, posture, and task information (Kapoor and Picard, 2005). Recent work has also begun to investigate linguistic features for prediction of affective states (Litman and Forbes-Riley, 2006) and comprehensive world models for predicting user physiological response to reduce the need for biofeedback apparatus in runtime environments (McQuiggan et al., 2006). Collectively, this body of work serves as a springboard for research proposed here, which, in part, reports on the use of measurements of user physiological response as a predictor of self-efficacy levels. Because users’ physiological responses follow directly from their affective states, which are known to be correlated with levels of self-efficacy (Zimmerman, 2000), accurate measurements of physiological response could be used to enable interactive environments to effectively predict user levels of self-efficacy in order to guide customized interactions.
Affect Recognizers

The Affective Learning Companion project at MIT (Burleson and Picard 2004) is focused on monitoring user physiological signals to determine frustration, which is used to trigger learning support. This approach allows learners to enter states of impasse with the system only intervening at points where the learner becomes “stuck” (Burleson et al. 2004; Burleson and Picard 2004; Picard 1997). The affective learning companion makes use of input from a variety of affective and physiological sensors including cameras for facial feature expression and eye gaze detection, seat pressure pads to detect posture, galvanic skin response, pressure mouse and game state (Burleson and Picard 2004). An affective agent in the learning environment is able to sense the user’s emotion and respond by expressing its own emotion. The Affective Learning Companion project focuses on motivating students through states of failure and frustration, required phases for becoming an expert, and in particular focuses on the characteristics of the affective agent that influence sustaining a high-level of motivation to continue through failure or a state of “stuck” (Burleson and Picard 2004). This is accomplished through two approaches: (1) manipulating the environment to maintain an optimal experience for the user, and (2) promoting self-awareness to users to empower them to self-regulate their own motivation (Burleson et al. 2004). Results suggest that the detection mechanisms provide enough detail to determine user affect and that affective agents can positively influence users in maintaining their motivational state and self-awareness.

The Prime Climb educational game (Conati 2002; Conati and McLaren 2005; Conati and Zhao 2002; Conati and Zhao 2004) addresses affective reasoning with probabilistic approaches. Prime Climb investigates using affective user models derived from probabilistic approaches in educational games (Conati 2002). It makes use of dynamic Bayesian networks to detect a variety of affective states based on the OCC model (Ortony et al. 1988). Prime Climb links detected personality traits (extraversion, agreeableness, and conscientiousness) directly to goals associated with the math game (Conati 2002). Bodily expression of emotion is measured, including eyebrow tracking, skin conductivity and heartbeat, while probabilities determine which trait is represented by a given set of measurement values (Conati 2002). While such techniques model internal affective states and, in this case, personality traits, building deployable systems cannot rely on the “wiredness” required by this implementation.
Craig, D’Mello, and Graesser (with others) have begun to explore the emotional experiences of students in a traditional intelligent tutoring system, AutoTutor (Craig et al., 2004; D’Mello et al., 2008; Graesser et al., 2006). Their work has begun with a frequency analysis to identify the most common emotions experienced during learning. The set of emotions considered in their work has changed over time resulting in the most current set of deep learning emotions: boredom, confusion, delight, flow, frustration, neutral, and surprise. Recent work has begun to use conversation cues to predict student affect (D’Mello et al., 2008). The attributes of conversations included temporal information (i.e., time since session began, number of conversation turn), student response information (i.e., number of words), quality assessment of student answers (as compared to expectations and student history), tutor directness (i.e., hint, correction of student misconception), and tutor feedback (i.e., positive or negative feedback) (D’Mello et al., 2008). To date, affect predictions are able to model boredom, confusion, flow and frustration with reasonable accuracies. It has proven to be more difficult to accurately predict delight and surprise from conversational cues.

de Vicente and Pain have investigated motivation as distinct from emotions. They have primarily investigated user motivation detection and management in their ITS development (de Vicente and Pain, 1998; de Vicente and Pain, 2002). Their system is based on data collected through interviews and surveys. The methodology used to collect data for the motivational model by de Vicente and Pain involve participants’ reviews of motivational characteristics of particular students, reviews of video clips of student interactions with the ITS and comments on the motivation of the student as well as techniques they were using to identify the students motivational state and what they would do as a teacher to encourage the student (de Vicente and Pain, 2002). Ten postgraduate students participated in this portion of the research; each had a minimum requirement of some previous teaching. What de Vicente and Pain constructed was a set of 85 inference rules to determine the motivational state of the student (de Vicente and Pain, 2002). When the participants made comments about the motivational state of the student, performance and contextual characteristics such as speed, whether the student gave up, and quality of the answer as well as difficulty of the question were recorded and contrasted with the resulting inference rule system (de Vicente and Pain, 2002).
Beal and Lee present an approach to assessing student’s motivation and mood from students’ self reports for an ITS for learning secondary school mathematics. Before students use the system, they complete a brief mood report by answering questions such as, “I am having a <great day, OK day, bad day>. From students’ self-reports of mood, the ITS makes pedagogical decisions about the difficulty of the problems, types of the problems, and types of multimedia help. For example, if a student reports that she is having a bad day, the system might suggest a tutorial of the material that the student has already mastered. Instead of recognizing motivational states, their work focused on determining pedagogical actions based on the student’s motivational state and mood which were self-reported by the students.

Recent efforts have explored the use of physiological response to detect users’ emotional states [Prendinger et al., 2003]. Prendinger and his team utilized bio-signals to investigate the effects of a like-like synthetic agent with affective behavior on users’ emotional states which were derived from user physiological data. They reported the result of an empirical study showing that affective feedback of a synthetic agent may reduce user stress. Conati et al. present a model of user affect for educational games [Conati, 2002; Conati and Maclaren, 2005]. The techniques are illustrated with PRIME CLIMB, an educational game in which a pedagogical agent uses the assessment of student affect to direct its intervention. Recent work on this project is exploring techniques for refining the previous user affect model based by utilizing user physiological data such as galvanic skin response [Conati and Maclaren, 2005]. In a similar vein, the Affective Learning Companion project at MIT senses students’ affective states through various channels (e.g., facial expression analysis, pressure mouse, and skin conductivity) to assess students’ progress, which could be used to determine appropriate interactions and interventions of the affective learning companion [Burleson and Picard, 2004].

2.2.2 Affect Expression through Agent Interaction

Believable Characters
The Oz architecture (Bates et al., 1994; Bates, 1994; Reilly and Bates, 1992) includes and generates a simulated physical world, multiple characters, an interactor, a theory of presentation, and a drama
manager. Oz also presents an emotion modeling system (Reilly and Bates, 1992) which is based on OCC (Ortony et al., 1988). The emotional model takes a clustering approach where emotions are grouped into clusters of emotion types with other emotions that share similar causes (Reilly and Bates, 1992). For example, the distress emotional type includes the emotions of sad, distraught and lovesick which differ on intensity and situation types but share similar causes. The emotional model is based on three types of subject appraisals suggested in (Ortony et al., 1988): (1) the pleasingness of events in respect to desired goals, (2) approval of actions with respect to a set of standard behaviors, and (3) the appraisal of liking certain objects with respect to attitudes. In addition (Ortony et al., 1988) proposed another set of emotions which are the result of other emotional combinations (Reilly and Bates, 1992). The emotions resulting from this model that Reilly and Bates (1992) explores include: joy, distress, hope, fear, pride, admiration, reproach, anger, gratitude, gratification, and remorse. The emotional model bases its decisions on the success or failure of goals and the events that may have led to the success or failure. The emotional model creates an affective state for an agent which is used to influence the agent’s behavior. This work is one of the first attempts to apply emotional modeling schemes from outside disciplines to influence the behavior of agents in a computer system. The emotion models explored in the Oz project have led to increased believability in agents and a more intensive immersion in the virtual environments. It is hoped that introducing empirically-grounded models of empathy, such as CARE, into Oz-like environments would increase the believability of agents.

EMA is an emotional model based on Smith and Lazarus’ Appraisal Theory of emotion (Gratch and Marsella, 2001; Gratch and Marsella, 2004a; Gratch and Marsella, 2004b; Marsella and Gratch, 2003). Smith and Lazarus’ theory centers on the cycle of event appraisal and selection of coping strategies. EMA has been implemented in Mission Rehearsal Exercise (MRE), a system that is designed to train military officer candidates in decision-making in highly volatile situations (Gratch and Marsella, 2001; Gratch and Marsella, 2004a; Gratch and Marsella, 2004b; Marsella and Gratch, 2003). The user, who is immersed in a virtual learning environment experiences typical sights, sounds, and events found in mission circumstances (Gratch and Marsella, 2001). She acts as the commanding officer in the situation with all other soldiers, enemies and civilians controlled by agents. Each agent is controlled by the EMA emotional model, which increases the believability and
realism of the characters, the environment and the situation (Gratch and Marsella, 2001; Gratch and Marsella, 2004b). The purpose of the exercise is to compel the user to assess the situation and make an effective decision in a short amount of time. Adding emotional factors, such as a crying citizen upset by her son’s (also a citizen) injury in the virtual Bosnian village, makes the training more life-like and as educational as possible without requiring the commanding officer to actually experience the mission physically.

**Affective Pedagogical Agents**

The Soar Training Expert for Virtual Environments (STEVE) is a pedagogical agent cohabiting virtual environments with students learning to perform physical, procedural tasks such as operating complex devices (Johnson and Rickel, 1998; Rickel and Johnson, 1999). Steve was developed in collaboration between the Center for Advanced Research in Technology for Education, Lockheed AI Center and USC Behavior Technology Laboratories. Steve’s goals for a particular situation may include engaging the user, and teaching the user so that the user performs better over time (Johnson and Rickel, 1998). From such goals Steve exhibits emotions of caring about the user and desiring to ensure successful learning. Steve is happy-for the student when the student successfully completes a task and is disappointed when the student fails. A goal of the user exhibiting caution is an indication that possible actions would include fear of possible future actions and/or hope of the user performing possible actions (Johnson and Rickel, 1998). Steve not only is able to simulate the expression of motion, but through its procedural tasks, it invokes emotions in the user.

Herman-the-Bug inhabits Design-A-Plant, a knowledge-based learning environment used to explore interactive problem solving in the domain of botanical anatomy and physiology, with an animated pedagogical agent (Elliott et al., 1999; Lester et al., 2000). Herman-the-Bug employs an affective user modeling architecture (Lester et al., 2000). Herman-the-Bug appraises the world allowing him to interpret situations that arise which may invoke emotional responses (Lester et al., 2000). The agent maintains emotional models about the fortune of its users, including enjoyment and happiness when goals are met (Lester et al., 2000). Herman-the-Bug maintains structures which monitor user goals, principles and preferences and supports inferences concerning motivations and
pleasure which can be used by Herman-the-Bug to produce emotional expressions to encourage a user (Lester et al., 2000).

Several studies have evaluated the impact of affective interface agents on both affective and motivational outcomes based a number of factors including gender, ethnicity and realism of the agent (Baylor, 2005). The first study asked 183 participants (undergraduates) to select the desired instructor from eight pedagogical agents. Participants were immediately asked to explain their selection. The responses cited perceived demeanor, gender, instructor-looking characteristics and ethnicity. Regression results discovered that the participants tended to select an instructor who was the same ethnicity. The second test randomly assigned agents and examined motivational outcomes such as self-regulation and self-efficacy. Other studies focused on expert agents vs. motivating agents and investigated tradeoffs between genders and ethnicity. The final two studies focused on non-human-like agents categorized by color, shape aliveness and complexity. One looked into associated meanings of agent images by asking participants what they thought the displayed figure was. Responses were categorized as either an emotion-related (sad, happy) or word association (spoon, dinosaur) response. The more complex the agent the more likely participants were to match the agents with their intended meanings. Another study focused on the role of color and shape. Color was determined to have a very low impact if any while the more complex the shape (fish and dog, vs. a geometric shape) the more the likely the agent was perceived as instructor-like. All six of Baylor et al.’s studies indicate that characteristics such as gender, ethnicity, and realism are influential agent characteristics affecting student learning experiences, in particular student affect and motivation (Baylor, 2005).

**Socio-emotional Agents**

Focused on modeling multimodal interactions in conversation, several projects led by Justine Cassell, have incorporated affect. Gestures, gaze, body posture and vocal intonation have been used in human-computer conversation to emulate the interactions of human-human conversation. Cassell along with Bickmore investigated relational agents (Bickmore, 2003). This work sought to design computational relational agents to interact with users and build social-emotional relationships with them (Cassell and Bickmore, 2003). (Cassell and Bickmore, 2003) targets caring
and empathetic behaviors because they play an instrumental role in relational strategies and need to be considered in their model of social relationships (Cassell and Bickmore, 2003).

Johnson and colleagues explore techniques for creating socially-skilled, polite pedagogical agents (Johnson et al., 2005; Wang et al., 2005; Wang et al., 2008). Their work emphasizes agents that are expressive in both emotion and attitude, sympathetic and sensitive to student’s motivational states, and are polite, knowing how to appropriately interact in different social contexts. Johnson uses such agents to exploit Reeves and Nass’ findings (Reeves and Nass, 1996) that people tend to relate to and other media as if they were human. For example, the Adele agent (Johnson et al., 2005) was used in a medical setting to teach domain knowledge and to serve as an assistant to physicians in clinical workups. Making use of a variety of emotive gestures and head movements, Adele is able to communicate in social-normal ways.

Johnson and Rizzo report on a Wizard-of-Oz experiment to evaluate the effects of politeness strategies on the self-confidence, interest and motivation of the learner (2004). Politeness is a technique used when a socially rich situation assessment calls for reactive empathy. In this case, understanding the need to use positively valenced emotions to affect the learner’s motivation and self-confidence is deemed necessary. It is not the case that all instances of politeness can be handled by empathetic assessment and interpretation, but some can. Regardless, Johnson et al. (Johnson and Rizzo, 2004; Johnson et al., 2005) clearly identify the type of socially intelligent opportunities that are now available based on rich models of affect (Lazarus, 1991; Ortony et al., 1988; Smith and Lazarus, 1990).

Elliott’s Affective Reasoner (Elliott, 1992) was based on the OCC model (Ortony et al., 1988), which he expanded to include 26 emotions, as opposed to the 22 emotions described in the original OCC model. Elliott also included jealousy, envy, like, and dislike in his Affective Reasoner implementation. The Affective Reasoner incorporates these emotions into a rule-based system to model multiple software agent personalities and the existing social relationships between the agents. Elliott modeled three types of social relationships: friendship, animosity and empathy. A friendship among agents prompts similar valenced emotions. For example, if an agent is satisfied with a good grade one could expect the agent’s friends to be happy-for the agent; both are positive emotions, while happy-for is empathetic in nature. Animosity typically provokes oppositely
valenced emotions between two agents. The emphasis of his work is reasoning about emotions for a social context. The Affective Reasoner does this by using a rule-based rule system to appraise events to determine its own emotions while modeling concern for other agents using a backward, case-based system (Elliott, 1992). In the Affective Reasoner, emotional recognition is not solely based on the expression of other agents; rather, the significance of situational context is considered in conjunction with expression to determine the other agent’s emotions. This is an important consideration because humans derive their emotions from a number of factors such as past emotion, mood, the emotions of those nearby, and the mood of the environment.

The Empathic Companion, an animated interface agent, resides in a simulated interview environment providing empathic feedback based on users’ affective states (Prendinger and Ishizuka, 2005). Interestingly, the user’s affective state is derived from only two physiological sensors measuring skin conductivity and electromyography in real time. Using a Bayesian network, signals’ distance measurements from a baseline determine user emotion (Prendinger and Ishizuka, 2003). Six emotions are associated with a job interview scenario: sad, frustrated and fear, relaxed, joyful and excited. The agent makes the user aware of negatively valenced emotions in response to interviewer questions. For example, if the interviewer asks the interviewee if they would mind working unpaid overtime the interviewee may become frustrated and cause the Empathic Companion to intervene and make the user aware of their affective state so that they may cope with such a situation differently if actually presented with such a scenario. Preliminary evidence suggests that there may be a positive effect for the interviewee’s management of stress (Prendinger and Ishizuka, 2005).

The FearNot! Project has explored techniques for eliciting empathetic responses in users responding to bullying scenarios depicted in on-screen interactions (Paiva et al., 2004; Paiva et al., 2005). The application proposed in (Paiva et al., 2005) aims to elicit empathetic interaction from children. Their work explores the hypothesis that creating emotionally charged situations which evoke empathy from users increases agent believability and immersion (Paiva et al., 2005). Empathy in users accomplishes this objective because, by its very nature, empathy requires being immersed in another’s perspective. If users are able to allow themselves to feel empathetic towards agents in the world, then they have achieved a level of immersion to be empathetic and increased.
the believability of the target agent. Using empirical approaches to model empathy for the synthetic agents inhabiting the FearNot! environment, the approach would allow a greater reciprocation of empathy and further develop believability because users would be able to recognize other agents as responding to their own affective state and situation.
Chapter 3

An Inductive Approach to Affect Recognition and Expression

In this chapter, we first define the tasks of affect recognition and affect expression. We propose an inductive approach embedded in an architecture consisting of both an affect recognizer and affect expresser. Finally, candidate machine learning algorithms for the proposed inductive framework are introduced.

3.1 Task Definitions: Affect Recognition and Affect Expression

Affect Recognition
The task of affect recognition (Figure 3.1) is to identify the affective state $S^*$ from a set of candidate affective states $s_1, s_2, ..., s_m$ given an observation sequence $O_1, \ O_2, \ ..., \ O_n$. Each $O_i$ represents an instance of the observable attribute vector that candidate affective states may be inferred by observing attributes at time slice $i$. The observational attribute vector includes a number of observable attributes including user goals, actions, locations, intentions, and physiological response data such as heart rate, skin conductivity, eye movement, voice intonation, and bodily movement.

Affect Expression
The task of affect expression (Figure 3.2) is comprised of two sub-tasks: 1) identify when affect expression is warranted (affective assessment), either true or false given an observation sequence $O_1, \ O_2, \ ..., \ O_n$; and 2) identify the affect state $I^*$, when affective assessment is true (affect expression is
warranted), from a set of candidate affective states $i_1, i_2, ..., i_m$ given the same observation sequence $O_1, O_2, ..., O_n$. Each $O_i$ represents an instance of the observable attribute vector from which candidate affective states may be inferred by observing attributes at time slice $i$. The observational attribute vector includes a number of observable attributes including user goals, actions, locations, intentions, and physiological response data such as heart rate, skin conductivity, eye movement, voice intonation, and bodily movement.

3.2 Appraisal Theory as the Theoretical Basis for CARE

Appraisal theory serves as the theoretical basis from which the CARE inductive framework is derived. In this section we briefly explore the components of the appraisal process that translate to elements of the CARE framework.
Lazarus identifies two sets of variables of emotion: observable and non-observable variables (1991). Observable variables relevant to emotions consist of the following four classes:

1. **Actions.** Action variables include facial expression, posture, and gesture in addition to movement toward or away from objects. CARE will account for student movements in virtual environments and will, at times, collect video recordings of subjects’ facial expressions during learning experiences.

2. **Physiological reactions.** Advancements in physiology tracking hardware allow researchers to measure many physiological changes in subjects, including heart rate, galvanic skin response, respiration, temperature. Further advancements are allowing researchers to monitor activity in regions of brain and measurements of muscle tension. Physiological responses are an appraisal outcome in the Smith and Lazarus model (Lazarus, 1991). CARE will account for physiological responses, at times, using apparati to measure changes in blood volume pulse (used to measure heart rate) and galvanic skin response (used to measure skin conductivity).

3. **Speech.** Emotion is frequently expressed in what people say and how they say it. People can concretely report emotions being felt (i.e., “I am angry,” or “This task is becoming frustrating”). Further, accounts of emotion can often reveal goals or uncover expectations and beliefs that motivate emotional experiences (Lazarus, 1991). None of the experimental environments in which we later investigate the CARE framework utilize speech functionalities. However, at times, experimental conditions will employ virtual characters to engage students in conversation by querying for student affective state experience. Scientists continue to debate the validity of speech variables of emotion. Self-reports of affect can fail to accurately capture subjects’ affect states because of social influences, desires to represent oneself in particular (often positive) states, and misinterpretation of unfolding events (Lazarus, 1991).

4. **Environmental events and contexts.** Environmental variables include the social, cultural, and physical events surrounding emotional experiences (Lazarus, 1991). The CARE framework utilizes a rich representation of the environment and unfolding situations surrounding student actions in virtual environments. For instance, CARE will track the
objects and agents with which a student interacts, as well as the locations the student visits. Further, CARE will monitor temporal attributes of environment variables to effectively represent a more extensive context.

The variables described above can directly be observed. Lazarus also defines a series of non-observable variables relevant to emotion (1991). These non-observable variables consist of the following five classes:

1. **Action tendencies.** Action tendencies refer to personal impulses that are non-observable because they may never be realized or acted upon (Lazarus, 1991). Clues to action tendencies may be found in observable variables such as actions or physiological changes (e.g., increased muscle tension) (Lazarus, 1991).

2. **Subjective emotional experience.** Affective states themselves are not directly observable. However, we are able to observe clues that allow us to infer which emotion(s) are being experienced. A common mechanism to obtain such information is to collect verbal reports of affect by simply prompting subjects to report how they feel. CARE will frequently employ just such a self-reporting mechanism to collect information regarding student affective experiences.

3. **Person-environment relationship.** The relationship that exists between the observable environment and the person is not directly observable (Lazarus, 1991). However, the relationship can be inferred through independent observation of each entity (environment and person). Using psychological instruments and surveys may allow us to better observe personalities, goal-orientations, and beliefs, thereby permitting more accurate inferences to be drawn about the person-environment relationship. We will utilize psychological instrumentation in many of the CARE experiments to acquire static information pertaining to the person while using dynamic observations of the environment to infer the relationship between the two entities and, ultimately, affective state experiences.

4. **Coping processes.** The coping process results in problem-focused strategies, whose actions can be observed in the environment, and emotion-focused strategies, which can be observed through prompted reporting (Lazarus, 1991). However, the deliberation of the coping process can only be inferred from such observations and not directly observed itself.
5. **Appraisal processes.** Likewise, inferences can be drawn about appraisals, but cannot directly be observed (Lazarus, 1991). It is through observation of the person-environment relationship combined with reports and inferences surrounding judgments of this relationship that we can infer the appraisal. Lazarus’ observable and non-observable variables relevant to emotion collectively constitute the emotion system and are identifiable in the Smith and Lazarus model of emotion. Many of these variables are also incorporated into the CARE inductive framework.

There are multiple approaches to developing computational models of affect. Wehrle and Scherer (2001) identify two approaches: a black box approach and a process approach. Black box approaches make predictions or decisions that are accurately similar to the operations of the appraisal process. Black box approaches do not provide granular, structural representations but instead are intended to provide functional models that yield accurate affective decisions using abstract mappings from an input space to correct outcomes (Wehrle and Scherer, 2001). Process modeling approaches seek to develop a system that exemplifies the processes and subsystems accounting for each component and their interdependencies of appraisal theory (Wehrle and Scherer, 2001). Process models are those useful for research aimed at understanding the appraisal process, while black box models are unable to lead to claims regarding the structure of appraisal models. Gratch and Marsella’s EMA is a process model approach incorporating appraisal processes into the architecture of their virtual agents (2004b). The CARE inductive framework will utilize a black box approach because we are fundamentally interested in the aptitude of the approach to produce models sufficient for informing future pedagogy in interactive learning environments.

We distinguish between two black box approaches: an analytical approach and an empirical approach. Both approaches are useful for indentifying rules, structures, and other relationships between an input space and affective outcomes. The analytical approach is exemplified by taking appraisal configurations identified in the relevant literature and incorporating them into a black box model. This approach often requires domain expertise and a significant amount of authoring-time. The empirical approach instead identifies the mappings, rules, and relationships between the input space and outcomes by observing the outcomes in action. CARE uses a black box, empirical approach to induce models of affect. CARE therefore requires corpus collection studies in which
observations of “affect in action” are observed in the presence of relevant emotion variables. Section 3.4 will present the black box algorithms we consider to effectively model the relationship between our input space and affective outcomes. The next section will present the CARE framework.

3.3 The Inductive Approach

The prospect of creating an “affect learner” that can induce empirically grounded models of affect and other psychological constructs (i.e., self-efficacy) to recognize or express affect from observations of student interactions holds much appeal. To this end, we propose CARE, Computational Affect Recognition and Expression, an affective data-driven paradigm that learns models of affect. CARE consists of a trainable architecture and a two-phase methodology of training and learning.

The CARE framework operates in two modes: affect model induction in which the architecture interacts with a student trainer to gather data and runtime operation, in which it monitors user affect based on observations of student interaction.

- **Affect Model Induction.** During model induction (depicted in Figure 3.3 with dashed arcs), CARE acquires training data and learns models of affect from training users interacting with the interactive environment. The training user is outfitted with biofeedback equipment which monitors her heart rate and galvanic skin response. Biofeedback signals are recorded in training logs via the interactive environment, which also records an event stream produced by the training users’ behaviors in the environment. Generally, this event stream includes user actions, user locations, world states, physiological responses, and temporal features of these events (i.e., time spent in locations, average heart rate, etc.). Together, the biofeedback signals and the corresponding elements in the event stream are assembled in temporal order into the observational attribute vector. After training sessions (typically involving multiple training users) are complete, the affect learner induces models from the observed situational data and physiological data. The users’ self-reported affective states serve as class labels for the training instances. During interaction users are presented opportunities to self-report on their affective state using any number of mechanisms from multiple choice dialog boxes to a “self-efficacy slider” with a scale ranging from 0 (low) to
100 (high) to dialogue choices in exchanges with characters. Users report their perceived affect. Additional labels may be obtained through trained-judge analysis of video of training users’ facial expressions (Ekman and Friesen, 1978; D’Mello et al., 2008).

Figure 3.3. CARE Inductive Framework.
• **Runtime Operation.** During runtime operation (represented in Figure 3.3 with solid arcs), which is the mode employed when users interact with fielded interactive environments, the induced models inform decision making runtime components by predicting end-users’ affective states. The interactive environment again tracks all activities in the world and monitors the same observable attributes reported to the affect learner during affect model induction. The induced model is used by the affect diagnosis controller to (1) assess the situation to determine what affective state the user is experiencing, and (2) determine which interactive environment modules need to be informed of the changes (if any changes exist) in the users’ affective state. In runtime operation mode users may don biofeedback equipment if the model being used is equipped to handle physiological response data, in which case the observational attribute vector expects to have a continuous feed of heart rate and skin conductance data.

Recall, the task of affect recognition is to identify the affective state $S^*$ given an observation sequence $O_1, O_2, ..., O_n$. $O_i$ representing an attribute vector comprised of emotion relevant variables. The choice of input attributes depends on the application. We will consider the following general input attribute vector:

- **User Actions** $A_1, A_2, ..., A_n$: An action sequence that a user has performed so far. An affect recognizer can observe users’ actions in the world; recognizers also have access to auxiliary information about the interactions, e.g., any artifacts manipulated such as which objects have been picked up or which doors have been opened, as well as the characters with which users have interacted.

- **User Locations** $L_1, L_2, ..., L_n$: A sequence of locations in which a user has performed actions. Affect recognizers can bring to bear a broad range of knowledge about the location in which users’ actions are performed in virtual environments. In contrast to activity recognition in physical environments where recognizers must cope with noise and errors in sensors and perception (e.g., vision and speech), affect recognition has access to precise locational information.

- **Physiological Data** $D_1, D_2, ..., D_n$: A sequence of physiological response data (e.g., heart rate, galvanic skin response, etc.) from which user affective states can be inferred. Affect
recognizers have access to sensor data. In contrast to user location information, there can be noise in sensor data, so affect recognizers must cope with noisy sensor data streams.

- **The State of the World** $W_1, W_2, \ldots, W_n$: Other miscellaneous information such as information about virtual agents or objects in the environments. Affect recognizers track all information in the world such as other characters’ actions, goals, and narrative states. Affect recognizers may further need to account for individual psychological differences such as goal orientation, domain efficacy, personality, immersion tendencies, etc.

Figure 3.3 depicts the CARE architecture. In runtime operation, the user interacts with a virtual environment via a user interface. The user navigates the virtual 3D world to accomplish problem-solving tasks. Throughout the interaction, the interactive environment tracks all activities in the world and monitors the observational data such as user actions, user locations, states of the world, and physiological data. User actions include navigation actions (e.g., move to the location), manipulation actions (e.g., pick up objects, stack objects, open a door, test an object), communication actions (e.g., talk to a person), and information-seeking actions (e.g., read a book). User locations represent the location in which user actions are performed. World states represent the current state of the world (e.g., the state of other character, narrative states, and the focus of user attention). Various sensor data can be used to detect user emotions such as users’ physiological state changes, user body posture, eye movements, and voice signal changes. Note that the task we propose to explore includes both the “tethered” and “untethered” versions of the problem, i.e., at runtime, sensor data may or may not be available.

### 3.4 Inductive Modeling Techniques

Because it is unclear which observable features will provide the greatest utility in disambiguating candidate affective states, we consider several inductive approaches for generating affect recognition models: naïve Bayes, decision trees, support vector machines, Bayesian networks, and $N$-gram models.

Naïve Bayes and decision tree classifiers are effective machine learning techniques for generating preliminary predictive models. Naïve Bayes classification approaches produce probability tables that can be incorporated into runtime systems and used to continually update
probabilities for assessing student self-efficacy levels. Decision trees provide interpretable rules that support runtime decision making. The runtime system monitors the condition of the attributes in the rules to determine when conditions are met for assigning particular values of student self-efficacy. Both the naïve Bayes and decision tree machine learning classification techniques are useful for preliminary predictive model induction for large multidimensional data, such as the attribute vectors used in experiments described in sections 6 and 7. Because it is unclear precisely which runtime variables are likely to be the most predictive, naïve Bayes and decision tree modeling provide useful analyses that can inform more expressive machine learning techniques (e.g., Bayesian networks) that also leverage domain experts’ knowledge.

**Naïve Bayes**

Naïve Bayes classification approaches produce probability tables that can be incorporated into runtime systems and used to continually update probabilities for predicting self-efficacy. Naïve Bayes classifiers make an unsupported assumption (referred to as the “naïve assumption”) that the attributes of the observational attribute vector are conditionally independent. Thus, the probability of two conditionally independent events, A and B both occurring is \( P(A \text{ and } B \mid C) = P(A \mid C)P(B \mid C) \), where C is an observed event. Under the naïve assumption, gaining knowledge of event A occurring, given that we already know C, has no effect on the probability of event B occurring, and vice versa (Russell and Norvig, 2003). This assumption does not hold in the interactive environments we consider in this work (described in Chapter 5). For example, in the interactive learning environment there are many actions that are dependent on the location of the student’s character (i.e., experiments can only be run in the laboratory). Despite the inaccurate assumption that all observable attributes are conditionally independent, it has been found that naïve Bayes classifiers can nevertheless perform well and often with performance comparable to other classification methods (Han and Kamber, 2005).

**Decision trees**

Decision trees provide interpretable rules that support runtime pedagogical decision making. The decision trees induced in this work make use of the well known C4.5 software extension of the ID3
decision tree induction algorithm (Quinlan, 1986), which has been incorporated in the WEKA machine learning toolkit as the J48 algorithm (Witten and Frank, 2005). The decision tree induction algorithm makes use of a top-down, divide-and-conquer approach. At each node, an information gain analysis is used to select the attribute with the highest information gain, thus reducing the amount of information needed, to a minimum, to make classifications in the node’s sub-tree (Han and Kamber, 2005).

**Support vector machines**

Support vector machines (SVMs) are also particularly effective at handling high-dimensional data. SVMs search for hyperplanes that linearly separate data into classes (affective states). The SVM algorithm searches for maximal margin hyperplane(s) creating decision boundaries with the largest possible margin between classes (Tan et al., 2006). SVMs make the assumption that greater margins translate to classifiers with better generalized error rates.

**N-gram models**

An N-gram model is a Markov model of order n-1 (Murphy, 2002). N-gram models make the assumption that the current state depends only on n-1 previous states. When n = 1, we have a unigram model and when n = 2, we have a bigram model. N-gram models can be thought of as hidden Markov models without hidden variables. One disadvantage of N-gram models is that they cannot model partially observable worlds because they cannot have hidden variables. When n is small, N-gram models are efficient because of the Markov assumption. However, when n is large, it suffers from the issues of space and expressivity.

In this chapter, we have defined the tasks of affect recognition and affect expression as they will be explored in the following chapters. In the next chapter we will examine how the inductive approach is adapted to our learning environments and how affect models are induced utilizing the machine learning techniques discussed above.
Chapter 4

Learning Models of Affect Recognition and Expression

Manually constructing affect recognition and affect expression models is labor intensive and for many domains impractical. We therefore propose the adoption of an inductive approach in which the acquisition of affect in action is used to model the affective phenomena. After describing the task of model induction, we discuss issues of training corpus acquisition and then turn to learning issues.

4.1 Defining Affect Model Learning

The task of affect model learning is to induce a model from observations of training data such as the interaction traces between a user and an environment. The central goal of model induction is to generate a model that makes accurate predictions for new (unseen) data. Ideally, the training data collected should be representative of all possible situations that users may encounter at runtime. However, this assumption rarely holds in practice, so the induction process must generalize over the training data to unseen data. The induced model should therefore represent the underlying systematic characteristics of the data rather capturing specific details of the particular training data.

Two fundamental approaches to learning can be adopted: supervised learning or unsupervised learning. In supervised learning, for each data item, the value of the target output is specified. Thus, supervised learning is the task of learning an input-output mapping (the target function). In unsupervised learning, for each input data, the value of target output is not specified. Instead of learning a target function, unsupervised learning may model the probability distribution of the input
data or discover clusters or other structure in the input data. We adopt a supervised learning approach in which labeled classifications, such as affective states, are provided by a trainer. The task of model learning will therefore occur in two phases: a training phase, which is followed by a learning phase. In the training phase, training data will be collected by recording detailed logs of the interactions. In the learning phase, the model will be induced from the training data.

4.2 Affect Modeling Corpora

To accurately model affect, an instrument needs to be devised that can provide a metric for the construct and that can be used by the induced models for prediction. Recall from Chapter 2 that a growing body of work reports on efforts to detect and recognize user affect from a variety of information sources including self-reports, peer reports, judges’ reports, physiological response, body posture, eye tracking, and linguistic features of interactions. While sophisticated techniques have been developed for third-party detection of affect, e.g., analyzing recordings of facial expression (Ekman and Friesen, 1978), and a multitude of validated instruments have been devised for a broad range of affective phenomena, analogous techniques and instruments have not yet emerged for affective constructs such as self-efficacy and empathy. Further, much affective analysis is conducted to report on user experience instead of modeling affect in real-time as users complete various tasks. To date, self-reports have been the most widely used method for obtaining quantitative affective state measurements (Baylor and Kim, 2004; Beal and Lee, 2005; Kim, 2005).  

We therefore employ periodic self-report mechanisms, as well as instruments embedded in character dialog exchanges. Emotion and empathy modeling utilize dialog boxes with available affective states. Self-efficacy reporting tools have utilized a slider, allowing subjects to report self-efficacy in the range of 0 to 100.

In addition, accurately modeling affect requires a representation of the situational context that satisfies two requirements: it must be sufficiently rich to support assessment of changing affective states, and it must be encoded with features that are readily observable at runtime. Because affect

---

1 One approach to validating self-reports of efficacy is the test-retest method and the subsequent analysis to determine the reliability between self-reports for like questions. While this method is common in survey instruments for obtaining self-efficacy measurements, similar methodologies have yet to be devised for validating self-reports of efficacy gathered in real-time environments.
is fundamentally a cognitive process in which the user appraises the relationship between herself and her environment (Gratch and Marsella, 2004; Smith and Lazarus, 1990) and similarly, because self-efficacy beliefs draw heavily on a student’s appraisal of the situation at hand, affect recognition models (and models of self-efficacy and empathy) should take into account both physiological and environmental information. For task-oriented learning environments, affect models can leverage knowledge of task structure and the state of the student in the learning environment to effectively reason about students’ affective states. In particular, for such learning environments, affect models can employ concepts from appraisal theory (Lazarus, 1991) to recognize student emotions generated from their appraisals of the environment. Likewise, self-efficacy stems from users’ assessment of how their abilities relate to the current learning objective and task. Thus, affect models can leverage representations of the information observable in the learning environment—note that this refers to the same information that students may use in their own appraisals—to predict student affective states. The CARE framework therefore employs an expressive representation of all activities in the learning environment, including those controlled by users and the interactive system, by encoding them in an observational attribute vector, which is used in both the model induction mode and model usage mode of operation. During model induction, the observational attribute vector is passed to the affect learner for model generation; during runtime operation, the attribute vector is monitored by a CARE-enhanced runtime component that utilizes knowledge of user affective states to inform effective pedagogical decisions. The observable attribute vector represents four interrelated categories of features for making decisions: actions, locations, physiological data, and world states. These features have been implemented in a number of ways. To illustrate the observational attribute vector we consider models of self-efficacy induced from observations in the Online Tutorial System and CRYSTAL ISLAND. We describe these features at a high level below, followed by concrete examples in Table 4.1.

• **Temporal Features:** In the online tutorial system, CARE monitors the amount of time students spend on each question and how long the cursor resides in particular locations of the interface, since users tend to move their mouse according to the focus of their attention (Chen et al., 2001). In the interactive learning environment, CARE continuously tracks the amount of time that has elapsed since the student arrived at the current location, since the
Table 4.1. Representative observational attributes monitored in the online tutorial system (OTS) and interactive learning environments (ILE), including temporal, locational, intentional and physiological features.

<table>
<thead>
<tr>
<th>Temporal Features</th>
<th>Attribute Description</th>
<th>Possible Values</th>
<th>Applicable Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Question Time</td>
<td>The amount of time that has elapsed since the question was first displayed to the student</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Difference from Average Question Time</td>
<td>How does the amount of time the student has spent on the current question compare to the average time spent on previous questions (less or more)</td>
<td>Positive and negative real values</td>
<td>✓</td>
</tr>
<tr>
<td>Time in Current Location</td>
<td>The amount of time that the student has spent in a defined location of the interface</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Time on Current Learning Goal</td>
<td>The amount of time that the student has spent on current learning goal being attempted</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Comprehensive Learning Time</td>
<td>The amount of time that has passed since the student began interacting</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Locational Features</th>
<th>Attribute Description</th>
<th>Possible Values</th>
<th>Applicable Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current Location</td>
<td>The defined area in which the student’s cursor is located (OTS) or the area in which the student’s embodied character is located (ILE)</td>
<td>OTS areas: Question, Answer, Self-efficacy Slider, Submit ILE areas: Dining Hall, Waterfall, Lab Testing Area, Lab Reading Room, etc.</td>
<td>✓</td>
</tr>
<tr>
<td>Previous Location</td>
<td>The defined area in which the student’s cursor was located (OTS) or the area in which the student’s embodied character was located (ILE) immediately before the Current Location.</td>
<td>Same as “Current Location” Observational Attribute above</td>
<td>✓</td>
</tr>
<tr>
<td>Goal Achievable in Current Location</td>
<td>Whether or not the learning goal is achievable in the student’s current location</td>
<td>True or False</td>
<td>✓</td>
</tr>
<tr>
<td>Visited Location L</td>
<td>Whether or not the student has visited the particular location, L, for all locations, as designated by cursor location (OTS) and embodied character location (ILE)</td>
<td>True or False</td>
<td>✓</td>
</tr>
<tr>
<td>Number of visits to Location L</td>
<td>The number of times the student has visited the particular location, L, for all locations, as designated by cursor location (OTS) and embodied character location (ILE)</td>
<td>Positive integer values (values reset to 0 after each problem/goal)</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intentional Features</th>
<th>Attribute Description</th>
<th>Possible Values</th>
<th>Applicable Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem/Goal</td>
<td>Identifier corresponding to individual problems (OTS) and learning goals (ILE)</td>
<td>OTS: Problem number (1-20) ILE: Goal name (test-milk, talk-to-Jin, locate-ill-characters, etc.)</td>
<td>✓</td>
</tr>
<tr>
<td>Progression</td>
<td>Number of problems/ goals solved</td>
<td>Positive integer values</td>
<td>✓</td>
</tr>
<tr>
<td>Progression Rate</td>
<td>Average amount of time required to solve problems and achieve goals</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Number of Locations Visited in Goal Pursuit</td>
<td>Average amount of time required to solve problems and achieve goals</td>
<td>Positive integer values</td>
<td>✓</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physiological Features</th>
<th>Attribute Description</th>
<th>Possible Values</th>
<th>Applicable Environments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Rate</td>
<td>The student’s beats per minute as measured by the interval between the last two heart beats</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Galvanic Skin Response</td>
<td>The electrical resistance of the student’s skin as measured by the biofeedback apparatus</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Average HR and GSR</td>
<td>The student’s average heart rate and galvanic skin response measured from the start of interaction</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Problem/Goal HR and GSR</td>
<td>The student’s average heart rate and galvanic skin response measured from the start of the current problem/goal</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Sliding Window HR and GSR Averages</td>
<td>The student’s average heart rate and galvanic skin response measured across multiple intervals of 5, 10, 15, 20, 30, 45 and 60 seconds</td>
<td>Positive real values</td>
<td>✓</td>
</tr>
<tr>
<td>Sliding Window HR and GSR Differences</td>
<td>The change in student’s average heart rate and galvanic skin response measured across multiple intervals of 5, 10, 15, 20, 30, 45 and 60 seconds from the previous interval’s window</td>
<td>Increasing, Decreasing, Same</td>
<td>✓</td>
</tr>
</tbody>
</table>

57
student achieved a goal, and since the student was last presented with an opportunity to achieve a goal. Temporal features are useful for measuring the persistence of the student on the current and past tasks.

- **Locational Features**: CARE tracks the location of the student’s cursor in the online tutorial system. In the interactive learning environment, CARE continuously monitors the location of the student’s character. It monitors locations visited in the past, locations recently visited, locations not visited, and locations being approached. There are 45 designated locations in the interactive learning environment (e.g., the laboratory, the living room of the men’s quarters, and the area surrounding the waterfall). Locational features are useful for tracking whether students are in locations where learning tasks and current goals are achievable. When a student arrives in a location where a learning objective can be completed combined temporal attributes and locational features can aid in the prediction of the student exhibiting command of the learning task and associated levels of efficacy.

- **Intentional Features**: In the interactive learning environment, CARE continuously tracks goals being attempted (as inferred from locational and temporal features, e.g., approaching a location where a goal can be achieved), goals achieved, the rate of goal achievement, and the effort expended to achieve a goal (as inferred from recent exploratory activities and locational features). These features enable models to incorporate knowledge of potential and student-perceived valence (positive and negative perceptions) of a given situation. Intentional features, such as goal progression, are useful for measuring how a student’s abilities match the demands of the learning tasks. For example, a student that is rapidly achieving goals is more likely to be confident in their abilities to drive themselves towards success.

- **Physiological Response**: CARE continuously tracks readings from a biofeedback apparatus attached to the student’s hand. Blood volume pulse and galvanic skin response readings are monitored at a rate of approximately 30 readings/second to accurately track changes in the student’s physiological response. Blood volume pulse readings are used to compute student’s heart rate and changes in their heart rate. CARE monitors trends in both student heart rate and galvanic skin response over a variety of fixed and sliding windows in addition
to moment-to-moment readings. For instance, several fixed width averages of HR and GSR are monitored over the entire learning episode, for individual questions in the online tutoring system, fixed by the time the student takes to complete the question), and across individual learning objectives in the interactive learning environment, fixed by the time the student takes to complete the learning objective. CARE monitors HR and GSR trends in several sliding window frames of 5, 10, 15, 20, 30, 45, and 60 seconds. These sliding windows allow self-efficacy models to isolate changes in physiological response in the smaller windows that have little or no impact to the trends tracked in the longer windows. Other physiological response features include comparison attributes that monitor the change between current and past windows; summarizing the transition between the windows, i.e., whether HR and GSR are going up or down, and determining the rate of change between the windows.

In the CARE implementation for the online tutorial system, the observational attribute vector encodes nearly 150 features while in the interactive learning environment, CRYSTAL ISLAND, the observational attribute vector encodes 283 features. During model induction, a continuous stream of physiological data is collected and logged approximately 30 times per second. In addition, an instance of the observational attribute vector is logged every time a significant event occurs, yielding, on average, hundreds of vector instances each minute. We define a significant event to be a manipulation of the environment that causes one or more features of the observational attribute vector to take on new values. At runtime, the same features are continuously monitored by the respective environment. This may or may not include physiological response data depending on the incorporated model type, static or dynamic.

4.3 Designing Corpus Acquisition Tasks

To successfully construct models of affect for recognition and expression tasks corpus collection studies must be conducted to obtain scientifically-controlled observation of “affect in action.” The work described in the latter chapters of the dissertation will present details of a number of corpus collection studies that have been conducted to obtain instances of the observational attribute vector, described above, around the elicitation of subjective affective experiences. While each
Each study will present subjects with a well-defined task. It is important that the task be sufficiently long for subjects to make significant progress on the task and that the task be designed in such a way that the psychological construct of interest can be observed. Thus in all corpus collection studies subjects receive a task description, controls sheet, map of the virtual environment, and a directory of characters including images, names, and roles (see Appendix A). The task description clearly defined any self-reporting of affect, along with definitions of affective states, if the experiment was to include reporting of affective experiences. In some studies, we designed a practice task for subjects to familiarize themselves with the controls. However, this was only

Figure 4.1. Corpus Acquisition Methodology.
included in the study design when the subject population was perceived to have little to no experience using controls similar to the ones needed in the environments described in the following chapter. Prior to the corpus collection subjects completed psychological instrumentation deemed to be relevant to the goals of the study. A comprehensive list of instrumentation can be found in Appendix A. Lastly, if changes physiological response were to be measured in the corpus collection than subjects entered a resting period prior to the start of the task to obtain baseline measurements of biofeedback.

The collection of affect reports was controlled throughout subject interactions within virtual environments. Either a constant time interval was used to query subjects or a constant method, such as character conversation prompts, were used. Observations were continuously recorded with each logged instance corresponding to a change in a variable of the observational attribute vector. Upon completion of the data collection task subjects were often asked to complete a series of post-task surveys and questionnaires (see Appendix A). However, this information was collected to obtain subjective responses reflecting on the subject’s experience in the virtual environment task. This post-data is not folded into the corpus. The final corpus consists of relevant student psychological instrument data, demographic data, dynamic observational attribute vector data, and observations of affect (e.g., obtained from self-reports, video analysis, etc.) which serve as class labels in the corpus.

4.4 Model Induction

With naïve Bayes, decision tree, SVMs, Bayesian networks, or n-gram classifiers, CARE-enhanced runtime tutorial control components can monitor the state of the attributes to determine when conditions are met for predicting particular affective states. Naïve Bayes, decision tree, and SVM classification techniques are useful for preliminary predictive model induction for large multidimensional data, such as the 278-attributes taken from the 283-observed attribute vector used for learning self-efficacy models in the interactive learning environment, CRYSTAL ISLAND. Two approaches can be distinguished in learning techniques: those that are completely automated, and those that require the knowledge provided by a domain expert. CARE experiments reported below
focus on fully automated learning approaches. CARE model induction proceeds in three phases (exemplified with modeling self-efficacy).

4.4.1 Data Construction
Each training log is first translated into a full observational attribute vector. For example, blood volume pulse (BVP) and galvanic skin response (GSR) readings were taken nearly 30 times every second reflecting changes in both heart rate and skin conductivity. The 278 attributes observed directly in the environment were combined with the selected self-reported levels of self-efficacy class labels, since only one class label can be used. Thus, 4 datasets are constructed; one for each level of granularity. Consider observable attributes $a_1, a_2, ..., a_{278}$, and class labels $c_{279}, c_{280}, c_{281}, c_{282}, c_{283}$ ($c_{279}$ corresponds to the raw self-efficacy reports, $c_{280}$ corresponds to two-level self-efficacy self-reports, $c_{281}$ corresponds to three-level, $c_{282}$ corresponds to four-level, and $c_{283}$ corresponds to five-level self-efficacy self-reports). Each constructed dataset consists of all observable attributes, $a_1, ..., a_{278}$, and one non-raw self-efficacy self-report class label.

4.4.2 Data Cleansing
After data are converted into an attribute vector format, a dataset is generated that contains only instances in which the biofeedback equipment was able to successfully monitor BVP and GSR throughout the entire learning session. For example, in the foundational evaluation described below, data from two sessions had to be discarded for this reason: BVP (used for monitoring heart rate) readings were difficult to obtain from this participant. Two sessions did not satisfy these requirements and were subsequently removed from the interactive learning environment evaluation.

4.4.3 Model Learning

*Naïve Bayes Classifier, Decision Tree, and SVM Learning*

Once the dataset is prepared, it is passed to the learning systems. Each dataset is loaded into the *WEKA* machine learning toolkit (Witten and Frank, 2005), where naïve Bayes classifiers, decision
trees, and SVMs can be learned. Tenfold cross-validation analyses are run on the resulting models to obtain accurate estimates of the classifier’s error.

**Bayesian networks and n-Gram Learning**

Preliminary techniques, such as naïve Bayes, decision trees, and SVMs, are useful for informing the construction of advanced, more expressive machine learning techniques that also require domain expert knowledge. Because of computational complexities, Bayesian networks and n-gram model learning cannot handle the same multi-dimensional data as the preliminary techniques. For this reason, datasets must be reconstructed utilizing only the most informative attributes (determined through information analysis) and often require slightly less expressive attributes (i.e., reducing possible values of attributes by collapsing data). Newly constructed datasets are passed to learning software, developed in our lab, and tenfold cross-validation is used again on resulting models.

### 4.5 Evaluating Induced Models

We will investigate competing machine learning approaches to affect modeling. Candidate machine learning techniques include naïve Bayes, decision trees, support vector machines, Bayesian networks, and N-grams. In addition to exploring this set of machine learning algorithms we will often consider two conditions: the “wired” case, in which sensors provide physiological and other data, and the “unwired” case, in which no sensor data is available. Because affect models are learned from training data, the models may be able to learn the relationship between observational data and affective states without the added boost from physiological data. Further, it may be possible to infer physiological response changes at an abstract level, thereby rendering the use of biofeedback apparatus in runtime environments unnecessary (McQuiggan et al., 2006). If users are wired in runtime, we may obtain more accurate prediction of affective states. To determine the efficacy of various affect models we will use the following evaluation techniques to compare, contrast, and effectively assess CARE-induced affect models.

All CARE-induced models will be evaluated using a \( k \)-fold cross-validation scheme (Han and Kamber, 2005; Witten and Frank, 2005). In this scheme, data is decomposed into \( k \) equal partitions, known as folds, \( F_1, F_2, \ldots, F_k \). Each iteration uses one part, \( F_i \), for testing and the remaining \( k-1 \) folds
for training. After k iterations, each fold will have been held out for testing. The estimate of the classifier accuracy averages the error rate stemming from each iteration. Tenfold cross-validation, $k = 10$, is widely used for obtaining the best estimate of error (Witten and Frank, 2005). In this scheme data is divided into 10 equal subsets, $F_1, F_2, ..., F_{10}$. In the first iteration $F_1$ is held out for testing and the remaining folds, $F_2, F_3, ..., F_{10}$, are used to train a classifier that will be testing on $F_1$. Likewise the second iteration will holdout $F_2$ for testing and use $F_1, F_3, F_4, ..., F_{10}$ for training. After 10 iterations each fold will have been held out for testing. The error calculated from each iteration is then averaged to arrive at the accuracy estimate for the classifier. Another popular $k$-fold cross-validation method is leave-one-out. In this method $k$ is equal to the number of samples. This method has received some attention from the educational data mining community because the estimate of error describes the expected error for a given student since each sample corresponds to data collected from 1 student. Thus, each iteration uses one student’s data for testing and the other $k$-1 students for training until each student’s data has been held out for testing. In comparison, the estimate of error obtained from a tenfold cross-validation describes a more general error rate given any sample, which may include data from more than one student.

In several of the studies presented in chapters 6, 7, 8, and 9 we will use cross-validated ROC curves for presenting the performance of classification algorithms for two reasons. First, they represent positive classifications, included in a sample, as a percentage of the total number of positives, against negative classifications as a percentage of the total number of negatives (Witten and Frank, 2005). Second, the area under ROC curves is widely accepted as a generalization of the measure of the probability of correctly classifying an instance (Hanley and McNeil, 1982). Thus, ROC curves are useful for visually assessing classifier performance.

Lastly, we will evaluation CARE-induced models by comparing the results to baseline models. Such comparisons will justify the inductive approach and associated machine learning algorithm by demonstrating statistically significant improvements in accuracy over baseline models. Our baseline models will be composed of selecting the most frequent class label and measuring the error rate for a classifier that always predicts this most frequent class. For example, if we are conducting an affect recognition study concerning two emotions, happy and sad, then each instance in the corpus will be labeled with either happy or sad. Now, suppose that the happy label occurs 60% of the time and
sad only 40% of the time. In this case, the baseline model would constantly predict happy as the affective state. This means that the baseline model would make inaccurate predictions in 40% of the instances. Given this example, we would hope that CARE-induced models are able to accurately recognize whether a subject is happy or sad better than 60% of the time.

Other Metrics
We will compare baseline models and competing CARE-induced affect models in terms of accuracy and efficiency.

Accuracy measures the predictive power of the system and is defined as follows. Accuracy can be reported in terms of the true positive rate (the rate of positive cases correctly classified as such), which measures the sensitivity of a prediction system, and the true negative rate (the rate of negative cases correctly classified as such), which measures the specificity of the prediction system2 (Tan et al., 2005):

- Accuracy = (a + d) / (a + b + c + d) (i.e., the number of correct predictions divided by the total number of observations).
- Sensitivity (Recall, true positives of actual positives) = d / (c + d)
- Specificity (true negative rate) = a / (a + b)
- Precision (true positives of predicted positives) = d / (b + d)

We need to measure how well the model can make accurate early predictions and converge as quickly as possible on the most likely interpretation. To measure “early prediction” ability, we will use converged and convergence point (Blaylock and Allen, 2003). We will be particularly concerned with these efficiency metrics when we utilize N-grams for modeling affect (e.g., section 7.3).

- Converged: Percentage of observation sequences in which the affect model’s final prediction is correct.

---

2 Precision and Recall are also widely used matrices to report accuracy. Note that recall is the same as sensitivity. Precision represent the ratio of records that are actually classified to positive class in the group that classifier has declared as a positive class (i.e., Precision = d/(b+d)). Ideally, we want to build the classifier that maximizes both recall and precision (Tan et al., 2005).
• **Convergence Point**: For observation sequences which converged, the point within the sequence when the affect model started making the correct prediction and continued to make the correct prediction for the remainder of the sequence.

• **Average Observations of Converged**: Average number of observations contained in observation sequences in sequences which converged.

### 4.6 Model Usage

After the models are analyzed, the best performing models can then be deployed in runtime environments to recognize affect or determine when and how to express affect, depending on the task for which the model was constructed. For example, a CARE-induced model of affect for the task of recognizing student emotion can, during runtime operation, inform decision making components by predicting end-users’ affective states. Such components may include narrative and tutorial planners, student models, dialogue managers, and camera planners. To make runtime decisions the induced model tracks all activities in the world and monitors the same observable attributes reported to the affect learner during affect model induction. However, the model is now able to make predictions regarding affect instead of needing to observe “affect in action” as required in corpus collection studies.
Chapter 5

Implementations

The CARE paradigm has been used to train models of affect (emotion, self-efficacy, and empathy) to inform runtime control components of interactive systems. CARE has been tested and evaluated in several interactive systems. These systems include 1) an Online Tutorial System in the domain of genetics, 2) TREASURE HUNT, a virtual environment test bed in which users are instructed to collect treasures in the allotted time, and 3) CRYSTAL ISLAND, a narrative-centered learning environment in the domains of genetics and microbiology. Each interactive system is introduced below.

5.1 Online Tutorial System

The Online Tutorial System consisted of an online genetics tutorial and a genetics problem-solving system. The online genetics tutorial consisted of an illustrated 15-page web document (Figure 5.1) which included some animation and whose content was drawn primarily from a middle-school biology textbook (Padilla et al., 2000). The online genetics problem-solving system consisted of 20 questions, which covered material in the online genetics tutorial. The problem-solving system presented each multiple-choice question individually and required participants to rate their confidence, using a “self-efficacy slider,” in their answer before proceeding to the next question (Figure 5.2).
Mendel was the first scientist to recognize that principles of probability can be used to predict the results of genetic crosses.

A tool that can help you understand how the laws of probability apply to genetics is called a Punnett square. A Punnett square is a chart that shows all the possible combinations of alleles that can result from a genetic cross.

The Punnett square below shows a cross between two hybrid tall pea plants (Tt). Each parent can pass either of its alleles T or t, to its offspring. The possible alleles that one parent can pass are written across the top of the Punnett square. The possibilities of the other parent are written down the left side. The boxes in the Punnett square represent the possible combinations of alleles that the offspring can inherit. The boxes are filled in like a multiplication problem, with one allele contributed by each parent.

**Figure 5.1.** Genetics Online Tutorial System.

**Figure 5.2.** Genetics Online Tutorial System problem solving environment.
Treasure Hunt Virtual Environment

TREASURE HUNT (Figure 5.3) is a prototype virtual environment featuring a synthetic agent controlled by the user and a companion agent whose empathetic behaviors are controlled by CARE. The user navigated the 3-D virtual world in search of hidden (and some not-so-hidden) treasures. Each treasure box was labeled with the value of its contents, representing points obtained by collecting the associated treasure. Throughout the user’s quest for treasure, the companion agent followed along and expressed empathetic behaviors as appropriate situations arose. The following description of TREASURE HUNT (Figure 5.3) was presented to target trainers in an effort to establish a controlled backstory. Each target trainer received a copy of the same backstory.

You are about to find yourself on what appears to be an abandoned island with your companion, Alyx. On the island is an old warehouse, formerly used by pirates. The pirates have since left the island leaving some of their treasures behind! Scattered throughout the
island and particularly in the warehouse you may find boxes labeled by the value of their contents. Break open the boxes to collect the treasure within using your crowbar. Beware, some boxes may be unusually marked and have unknown contents. Collect such treasure at your own risk. You will have 7 minutes to explore the environment and collect treasure. The treasure you have collected, the number of remaining treasure boxes and the time left will be displayed in the bottom left corner of your display. Those that have ventured to the island before have left with treasure valued over 3,500!!!

TREASURE HUNT was implemented with Valve’s Source™ engine and the accompanying 3-D game platform for Half-Life 2. As described in Section 3, CARE models tracked observations of user behaviors in the interactive environment. Examples of those features monitored in the TREASURE HUNT environment include the following:

- **10-second score window.** A temporal feature which tracked whether the user had scored in the last 10 seconds of the interaction. Many time-window attributes were monitored because empathetic behaviors such as “excitement” often happened within a 10 second window of time after a significant score, while empathetic emotions such as “frustration” most frequently occurred after collecting a treasure worth a minuscule value took longer than expected. Time-windows for monitoring the last score were monitored at 5, 10, 15, 20, 30, 50, 75 and 100 second intervals.

- **Been to rocks on the beach.** A locational feature, the “been to the rocks on beach” attribute monitored whether the user has visited a specific location where a known high-valued treasure was hidden behind a group of rocks. Other locational features tracked the user’s navigation trends over time. For example, there were attributes to monitor where the user has been in a series of time-window attributes like those described above.

- **Moving towards high-valued treasure in sight.** This is an example of an intentional attribute that monitored whether or not there was a treasure in the user’s view, if there was a treasure box, the value of the treasure and whether the user was navigating in the direction of the treasure. This represented the user’s intent to approach and perhaps collect the treasure in sight.
• *Time left and total score.* The time remaining and the total score were displayed to both the target trainers and the empathizers. As time began to expire certain empathetic selections were made based on the performance of the target trainer. For instance, if the score was well below the expected value and the potential that the target would reach the goal based on the amount of time left empathizers chose different empathetic affective states (e.g., *frustration* and *sadness*). In contrast if, as time expired, the target had surpassed the expected goal then empathizers made empathetic selections of *joy* and *excitement*, based on the target trainers overall success.

• *Last emotion.* Tracking the last emotion could prove to be quite predictive of the next empathetic interpretation. Especially in the male empathizers it was uncommon to move from highly-aroused, positive-valenced emotion (*excitement*) to a low-aroused, negative-valenced emotion (*sadness*). This exact example, however, was witnessed several times with female empathizers. Tracking the last emotion is also useful for determining how far in the two-dimensional affective space empathizers moved between empathetic interpretations.

As users interacted with *TREASURE HUNT*, the observational attribute vector monitored variables such as those noted above to update when and how to be empathetic. Figures 5.4 and 5.5 depict similar
situations that have been assessed as a situation calling for empathy but have been interpreted differently. Notice that a treasure box was in the user’s view in Figure 5.4 where the companion is excited, in an “easy” environment, but not in Figure 5.5 the companion is relaxed, in a challenging environment.

To illustrate the empathetic behavior control posed by CARE, consider the following scenario in TREASURE HUNT, which repeatedly played out in CARE training sessions. As we catch up with the user, she has navigated her synthetic agent throughout the virtual environment as she struggles to find significant, high-valued treasure. The user and empathizer are aware that the user has not yet met her expected treasure collection quota (as specified in the graphical HUD representation) and is quickly running out of time. Only 30 seconds remain.

Now, the user has found her agent’s way into a location on the beach, a location visited by the user’s agent in the early moments when the session began. The empathizer realizes that this particular location has been previously visited and was already determined to be an area without
any treasure boxes. It has now been over one minute since the user last discovered any treasure at all.

Assessing the situation, the empathizer selects the frustrated affective state, igniting a behavioral sequence in which the companion agent announces her frustration, “This is becoming quite frustrating,” and uses gestures and posture similar to Figure 5.6. (The agent’s speech segments are stored in high quality pre-rendered audio clips.) CARE’s empathy learner has monitored a variety of environmental characteristics, including those described above, during its training sessions. The resulting instances aid the empathetic models in inferring that the same response is suitable (“when” and “how”) to similar situations when time is running out, the user’s agent is in a previously visited location known to be without treasure, the user’s intended treasure collection goal is likely to fail, etc. Thus, given the same situation with CARE driving the empathetic behaviors of the companion agent at runtime, empathetic assessment and interpreter models are likely to make the appropriate empathetic decisions. The next section discusses how effectively the models learned by the agent are able to predict empathizer actions.

5.3 Crystal Island Learning Environment

5.3.1 Narrative-centered Learning

It has long been recognized that discovery is a key element of the scientific enterprise, and recent years have seen a growing focus on discovery in education. For more than a decade, science education reform efforts by organizations such as the US National Research Council and the National Academy of Sciences have set forth standards promoting a greater emphasis on discovery learning (American Association for the Advancement of Science, 1993; National Research Counsel, 1996). In discovery learning (or inquiry-based learning), students approach a new topic via learning-by-doing. Instead of being presented problems and solutions in an expository fashion, students are given problems to solve, a rich environment in which to explore the problems, and a set of tools and techniques for constructing solutions. While early accounts of discovery learning focused on concept discovery (Bruner, 1961), contemporary work views discovery learning as scientific investigation. Thus, the process of discovery learning is analogous to the scientific method: students design and perform experiments, collect data, and evaluate hypotheses (de Jong and van
that Joolingen, 1998). First and foremost, discovery learning is active learning. As stated in the National Science Education Standards (National Research Counsel, 1996), discovery learning is “something that students do, not something that is done to them.”

Discovery learning offers several advantages over more didactic approaches. It tends to increase students’ ability to remember what they have learned, to apply their new knowledge, and to transfer it to new tasks more effectively than with more passive approaches that might emphasize activities such as reading textbooks (Blumenfeld et al., 2000; de Jong and van Joolingen, 1998). In addition to the cognitive benefits of discovery learning, it also offers potential motivational benefits. It enables students to become more active science learners (rather than passive consumers of information), it increases students’ beliefs that scientific theories change as new evidence becomes available (rather than being seen as unchangeable entities), and perhaps most importantly, it makes science more concretely meaningful (rather than seeming too abstract) (White and Fredricksen, 1998).

Despite the potential benefits of discovery learning, in the absence of appropriate scaffolding, discovery learning can be ineffective. Early findings suggested that discovery learning augmented with guidance can be more effective than pure discovery learning in enabling students to apply their knowledge to new problems (Shulman and Keisler, 1966). Furthermore, students may sometimes learn incorrect concepts through discovery learning, and discovery learning may be inefficient (Hammer, 1997). A recent analysis of thirty years of studies on discovery learning suggests that discovery learning accompanied by guidance in the form of feedback and coaching is more effective than unguided discovery learning (Mayer, 2004). Thus, guided discovery appears to be a promising alternative to didactic instruction and pure discovery learning.

Narrative could serve as the foundation for guided discovery learning. It is becoming apparent that narrative can be used as an effective tool for exploring the structure and process of “meaning making.” Narrative analysis is being adopted by those seeking to extend the foundations of psychology (Bruner, 1990), and one can imagine narrative-centered curricula that leverage students’ innate metacognitive apparatus for understanding and crafting stories. This insight has led educators to recognize the potential of contextualizing all learning within narrative (Wells, 1986). Narrative is central to human cognition. Because of the motivational force of narrative, it has long
been believed that story-based education can be both engaging and effective. Much educational software has been devised for story-based learning. These systems include both research prototypes and a long line of commercially available software. However, this software relied on scripted forms of narrative: they employed either predefined linear plot structures or simple branching storylines. In contrast, one can imagine a much richer form of narrative learning environment that dynamically crafts customized stories for individual students at runtime. Recent years have seen the emergence of a growing body of work on dynamic narrative generation (Cavazza et al., 2002; Riedl and Young, 2004; Si et al., 2005), and narrative has begun to play an increasingly important role in intelligent tutoring systems (Machado et al., 2001; Mott and Lester, 2006b; Riedl et al., 2005).

Narrative experiences are powerful. In his work on cognitive processes in narrative comprehension, Gerrig identifies two properties of reader’s narrative experiences (1993). First, readers are transported, i.e., they are somehow taken to another place and time in a manner that is so compelling it seems real. Second, they perform the narrative. Like actors in a play, they actively draw inferences and experience emotions as if their experiences were somehow real. It is becoming apparent that narrative can be used as an effective tool for exploring the structure and process of “meaning making.” For example, narrative analysis is being adopted by those seeking to extend the foundations of psychology (Bruner, 1990) and film theory (Branigan, 1992).

Learning environments may utilize narrative to their advantage. One can imagine narrative-centered curricula that leverage a student’s innate metacognitive apparatus for understanding and crafting stories. This insight has led educators to recognize the potential of contextualizing all learning within narrative (Wells, 1986). Because of the active nature of narrative, by immersing learners in a captivating world populated by intriguing characters, narrative-centered learning environments can enable learners to participate in the construction of narratives, to engage in active problem solving, and to reflect on narrative experiences (Mott et al., 1999). These activities are particularly relevant to inquiry-based learning. Inquiry-based learning emphasizes the student’s role in the learning process via concept building (Zachos et al., 2000) and hypothesis formation, data collection, and testing (Glaser et al., 1992). For example, a narrative-centered inquiry-based learning environment for science education could foster an in-depth understanding of how real-
world science plays out by featuring science mysteries whose plots are dynamically created for individual students.

5.3.2 Student Motivation in Narrative Learning

Narrative-centered inquiry-based learning environments may also offer motivational benefits. Motivation is critical in learning environments, for it is clear that from a practical perspective, educational software that fails to engage students will go unused.

Motivation is a powerful force: it drives humans to act (Schunk, Pntrich and Meese, 2007). There are two types of motivation that have been studied extensively, extrinsic and intrinsic motivation. *Extrinsic motivation* refers to engaging in a behavior because of external influences such as tangible rewards or pressures (Ryan and Deci, 2000). Extrinsic motivation does not stem from one’s internal interests. Instead, extrinsically motivated behavior can often be attributed to acting for the reward of pleasure or security manifested by something other than the task itself. *Intrinsic motivation* refers to engaging in a behavior because it is inherently interesting (Malone 1981; Malone and Lepper, 1987; Ryan and Deci, 2000). The behavior is undertaken solely for the challenge it poses, the enjoyment it yields, or the curiosity it satisfies; the act has some internal utility. Intrinsic motivation is favored because it has been associated with quality learning and creativity (Ryan and Deci, 2000). Further, it is believed that pedagogy that cultivates interest in a subject matter is more likely to lead to self-initiated learning beyond instructional experiences (Bandura, 1997).

Malone and Lepper’s taxonomy of intrinsic motivations (Malone and Lepper, 1987) consists of both individual and interpersonal factors. We focus on the four individual intrinsic motivators: challenge, control, curiosity, and fantasy.

- **Challenge.** Tasks that are too easy or impossibly difficult will foster little or no intrinsic interest and may lead to student boredom or frustration, respectively. Designing optimally challenging tasks will maximize student motivation.

- **Curiosity.** Student interest can be maintained by controlling for an optimal level of discrepancy between the student’s current knowledge and skills and the expected knowledge and skills following engagement in particular activities.
• **Control.** Humans have a basic tendency to want to have a hand in their own fate. Providing mechanisms that allow students to manipulate the learning experience results in a sense of power and choice.

• **Fantasy.** Playing to students’ abilities to develop mental models of situations that are not present contributes to motivation. Fantasies can evoke each of the other intrinsic motivators in ways that otherwise are unavailable to the student in reality.

It is has been determined that supporting the perception of student autonomy and devising tasks with optimal challenge levels is critical in student motivation (Lepper and Henderlong, 2000). Computational models of narrative may serve as the basis for narrative-centered learning that supports student autonomy with optimal task levels.

Game playing experiences and educational experiences that are extrinsically motivating can be distinguished from those that are intrinsically motivating (Malone, 1981). Narrative-centered discovery learning could provide the four key intrinsic motivators identified in the classic work on motivation in computer games and educational software (Malone and Lepper, 1987): challenge, curiosity, control, and fantasy.

Narrative-centered inquiry-based learning should feature challenging tasks of intermediate levels of difficulty, i.e., tasks that are not too easy and not too difficult, targeting desirable levels of student intrinsic motivation. Dynamically created narratives can feature problem-solving episodes whose level of difficulty is customized for individual students. In inquiry-based approaches, learning is inherently presented as a challenge, a series of problem-solving goals, that once achieved provide a deeper understanding of the domain. Devising narratives and providing tutorial feedback that both maintain a delicate level of uncertainty about the possibility of attaining each goal and sufficient reporting of student performance and progress is critical to sustaining effective levels of challenge.

Curiosity plays a central role in successful learning in narrative-centered inquiry-based learning environments. Since inquiry-based learning compels students to obtain knowledge throughout learning episodes on their own (materials are not provided explicitly prior to interaction) students are likely to question the completeness of their acquired knowledge as they progress, searching for new answers, stimulating their curiosity.
Narrative-centered inquiry-based learning environments can empower students to take control of their learning experiences; students can choose their own paths, both figuratively (through the solution space) and literally (through the storyworld), while being afforded significant guidance crafted specifically for them. The narrative structure of inquiry-based learning can provide unobtrusive direction by indirectly highlighting a subset of possible goals (i.e., blinking lights in a particular room in the environment, or a character audibly coughing in the student’s right audio channel) for the student’s next action consideration, maintaining the student’s perception of control.

Narrative-centered inquiry-based learning is innately fantasy-based. Fantasy refers to a student’s identification with characters in the interactive narrative and the imaginative situations created internally and off-screen by the student. All narrative elements ranging from plot and characters to suspense and pacing can contribute to vivid imaginative experiences. The openness of discovery learning provides scaffolding to support all levels of student imagination, increasing motivation and engagement. Effective narrative tutorials will engage characters in the storyworld that either the individual students perceive as possessing some cognitive, emotional, or physical similarities with themselves, or that the individual student admires, expresses feelings of compassion towards, or for which the student conveys empathetic feelings. In short, narrative can provide the guidance essential for effective inquiry-based learning and the “affective scaffolding” for achieving high levels of motivation and engagement.

5.3.3 Crystal Island – Version 1.0

In our laboratory we are developing a narrative-centered inquiry-based learning environment. Some components are fully designed and implemented while others are in the early stages of design. The prototype learning environment, CRYSTAL ISLAND (Mott et al., 2006), is being created in the domains of microbiology and genetics for middle school students (Figure 5.7).
CRYSTAL ISLAND features a science mystery set on a recently discovered volcanic island where a research station has been established to study the unique flora and fauna. The user plays the protagonist attempting to discover the genetic makeup of the chickens whose eggs are carrying an unidentified infectious disease at the research station. The story opens by introducing her to the island and the members of the research team for which her father serves as the lead scientist. As members of the research team fall ill, it is her task to discover the cause of the specific source of the outbreak. She is free to explore the world and interact with other characters while forming questions, generating hypotheses, collecting data, and testing her hypotheses. Throughout the mystery, she can walk around the island and visit the infirmary, the lab, the dining hall, and the living quarters of each member of the team. She can pick up and manipulate objects, and she can talk with characters to gather clues about the source of the disease. In the course of her adventure she must gather enough evidence to correctly choose which breeds of chickens need to be banned from the island.

Figure 5.7. CRYSTAL ISLAND.
The virtual world of CRYSTAL ISLAND, the semi-autonomous characters that inhabit it, and the user interface were implemented with Valve Software’s Source™ engine, the 3-D game platform for Half-Life 2. The Source engine also provides much of the low-level (reactive) character behavior control. The character behaviors and artifacts in the storyworld are the subject of continued work. The narrative planner of CRYSTAL ISLAND has been implemented with an HTN planner that is based on the SHOP2 planner (Nau et al., 2001). For efficiency, the planner was designed as an embeddable C++ library to facilitate its integration into high-performance 3-D gaming engines. A decision-theoretic “director” agent based on dynamic decision networks has been implemented to guide the narrative in the face of uncertain user actions (Mott and Lester, 2006a), and the method and operator libraries for the genetics and microbiology domains are currently being constructed.

To illustrate the behavior of the CRYSTAL ISLAND learning environment, consider the following situation in the domain of genetics. Suppose a student has been interacting within the story world and learning about infectious diseases, genetic crosses and related topics. In the course of having members of her research team become ill, she has learned that an infectious disease is an illness that can be transmitted from one organism to another. As she concludes her introduction to infectious diseases, she learns from the camp nurse that the mystery illness seems to be coming from eggs laid by certain chickens and that the source or sources of the disease must be identified. The student is introduced to several characters. Some characters are able to help identify which eggs come from which chickens while other characters, with a scientific background, are able to provide helpful genetics information (Figure 5.8). The student discovers through a series of tests that the bad eggs seem to be coming from chickens with white-feathers. The student then learns that this is a co-dominant trait and determines that any chicken containing the allele for white-feathers must be banned from the island immediately to stop the spread of the disease. The student reports her findings back to the camp nurse.

5.3.4 Crystal Island – Version 2.0

Version 2.0 of CRYSTAL ISLAND has seen significant extensions to the curriculum, story, and the virtual environment itself. Here we summarize these major changes to distinguish it from version 1.0. In
the following chapters experiments conducted in the CRYSTAL ISLAND environment will signify whether it is version 1.0 or 2.0.

CRYSTAL ISLAND is a narrative-centered learning environment that features a science mystery set on a recently discovered volcanic island. The curriculum underlying CRYSTAL ISLAND’s science mystery is derived from the North Carolina state standard course of study for eighth-grade microbiology. Students play the role of the protagonist, Alyx, who is attempting to discover the identity and source of an unidentified infectious disease plaguing a newly established research station. The story opens by introducing the student to the island and members of the research team for which the protagonist’s father serves as the lead scientist. Several of the team’s members have fallen gravely ill, including Alyx’s father. Tensions have run high on the island, and one of the team members suddenly accuses another of having poisoned the other researchers. It is the student’s task to discover the outbreak’s cause and source, and either acquit or incriminate the accused team member.
CRYSTAL ISLAND’s expansive setting (Figure 5.9) includes a beach area with docks (1), a large outdoor field laboratory (2), underground caves, and a research camp with an infirmary (8), lab (5), dining hall (6), and living quarters for each member of the team (3, 4, and 9). There are twelve characters in the Crystal Island 2.0 storyworld (Figure 5.10): Al (camp foreman), Alex (student user), Agatha (field scientist), Audrey (research scientist), Bryce (lead scientist), Elise (lab manager), Ford (lab scientist), Jin (camp nurse), Quentin (camp cook and maintenance engineer), Robert (research scientist), Sebastian (expedition financier), and Teresa (senior scientist). Throughout the mystery, the student is free to explore the world and interact with other characters while forming questions, generating hypotheses, collecting data, and testing hypotheses. The student can pick up and manipulate objects, take notes, view posters, operate lab equipment, and talk with non-player characters to gather clues about the source of the disease. During the course of solving the mystery, the student is minimally guided through a five problem curriculum (see Appendix B for a detailed description of the curriculum). The first two problems deal with gaining knowledge of pathogens, including viruses, bacteria, fungi, and parasites. Students gather information from interacting with in-game pathogen “experts”, viewing books, and posters in the environment. In the third problem, students are asked to compare and contrast their knowledge of four types of
Figure 5.10. CRYSTAL ISLAND 2.0 Cast of Characters.
pathogens (bacteria, fungi, parasites, and viruses). Problem four guides the student through an inquiry-based hypothesis-test-and-retest problem. In this problem students must complete a “fact sheet” (figure 5.11) with information pertaining to the disease inflicting members of the CRYSTAL ISLAND research team. Once the “fact sheet” is completed and verified by the camp nurse, the student completes the final problem regarding the treatment of viruses, bacteria, fungi, and parasites, and selects the appropriate treatment plan for sickened CRYSTAL ISLAND researchers. The story and curriculum are interwoven throughout the student experience. As the curriculum unfolds, so does the story unveiling new information pertaining to the poisoning scenario. As the student concludes the fourth problem, solving the mystery, the student is informed of Quentin’s acquittal in light of the student’s recent discovery of the mystery illness.

Figure 5.11. Crystal Island 2.0 Factsheet.
In addition, note-taking functionalities were recently added to the CRYSTAL ISLAND environment. Students access their notes using the ‘N’ key, which launches an in-game dialog where students can review their entered notes and supply additional comments if they so choose. Note-taking capabilities were a feature request stemming from focus group studies conducted with middle school students.¹

The virtual world of CRYSTAL ISLAND, the semi-autonomous characters inhabiting it, and the user interface were implemented with Valve Software’s Source™ engine, the 3D game platform for Half-Life 2. The Source engine also provides much of the low-level (reactive) character behavior control. The Source™ SDK includes several components that have made the development of CRYSTAL ISLAND. The SDK includes the C++ source code for the multi-player Half-Life 2: Deathmatch on which CRYSTAL ISLAND has been developed. Many of the 3D models and sound effects are included in the SDK and are a part of the single player Half-Life 2 or Deathmatch game. The SDK includes several tools that

¹ For an analysis of student note-taking in CRYSTAL ISLAND see McQuiggan, Goth, et al., 2008.
have assisted in the realization of CRYSTAL ISLAND. These include the Hammer Level Editor allowing the creation of customized 3D game maps (Figure 5.12). Another tool utilized in this work is Face Poser tool for creating character behaviors.

The CRYSTAL ISLAND C++ game code includes the initialized source code from the multi-player Half-Life 2: Deathmatch. The multi-player version does not offer any particular advantages over the single player version for the current incarnation of CRYSTAL ISLAND, but it does provide support for future CRYSTAL ISLAND extensions such as Wizard-of-Oz experiments (e.g., the experiment reported in Chapter 9) and multiplayer versions of CRYSTAL ISLAND. In addition to the removal of weaponry—Half-Life 2 is a first-person shooter game—and logging functionality to support the observation of student actions, character actions, and visited locations several extensions were implemented to support curriculum and narrative directives. These extensions include a set of client information panels for presenting textual information. These custom client-side GUI components are used to display character dialog, present the student with dialog options, allow students to read books (Figure 5.13), and support the functionality of the factsheet and the note-taking utility. An in-game web browser was implemented to support interactive posters that depict information relevant to the curriculum (Figure 5.14). Lastly, hooks were provided in the source to recognize and communicate with biofeedback apparatus such that changes in student physiological response can be logged along side of other observations in CRYSTAL ISLAND. Collectively, these extensions along with those described by Mott (2006) comprise Version 2.0 of CRYSTAL ISLAND.

Figure 5.13. Virtual books in CRYSTAL ISLAND. Figure 5.14. Web-based posters in CRYSTAL ISLAND.
Chapter 6

Modeling Self-Efficacy

We have explored CARE in several controlled studies. In this chapter, we investigate the CARE paradigm for modeling self-efficacy. Section 6.1, will first present background, related work, and motivation for examining self-efficacy. Section 6.2 describes a foundational study utilizing a modified version of CARE for inducing models of student efficacy in an online tutorial system in the domain of genetics. In Section 6.3 a follow-up study is detailed which explores CARE to again model student efficacy in CRYSTAL ISLAND, a narrative-centered learning environment. Finally, Section 6.4 summarizes the studies completed with CARE for modeling self-efficacy and discusses several design implications stemming from the results of this Chapter.

6.1 Self-Efficacy Background, Related Work, and Motivations

Self-efficacy\(^1\) influences students’ reasoning, their level of effort, their persistence, and how they feel; it shapes how they make choices, how much resilience they exhibit when confronted with failure, and what level of success they are likely to achieve (Bandura, 1995; Schunk and Pajares, 2002; Zimmerman, 2000). While it has not been conclusively demonstrated, many conjecture that given two students of equal abilities, the one with higher self-efficacy is more likely to perform better than the other over time. Highly efficacious students exhibit more control over their future

\(^{\text{1}}\) Self-efficacy is closely related to the popular notion of confidence. To distinguish them, consider the situation in which a student is very confident that she will fail at a given task. This represents high confidence but low self-efficacy, i.e., she is exhibiting a strong belief in her inability (Bandura, 1997).
through their actions, thinking, and feelings than ineffectual students (Bandura, 1986). Self-efficacy is intimately related to motivation, which controls the effort and persistence with which a student approaches a task (Lepper et al., 1993). Effort and persistence are themselves influenced by the belief the student has that she will be able to achieve a desired outcome (Bandura, 1997). Students with low self-efficacy perceive tasks to be more challenging than they actually are, often leading to feelings of anxiety, frustration and stress (Bandura, 1986). In contrast, students with high self-efficacy view challenge as a motivator (Bandura, 1986; Malone and Lepper, 1987). Self-efficacy has been studied in many domains with significant work having been done in computer literacy (Delcourt and Kinzie, 1993) and mathematics education (Pajares and Kranzler, 1995). It is widely believed that self-efficacy is domain-specific; whether it crosses domains remains an open question. For instance, students with high self-efficacy in mathematics may be ineffectual in science, or a highly efficacious student in geometry may experience low efficacy in algebra.

A student’s self-efficacy is influenced by four types of effectors (Bandura, 1997; Zimmerman, 2000). First, in enactive mastery experiences, the student performs actions and experiences outcomes directly. These are typically considered the most influential category as successful experiences typically raise self-efficacy, while unsuccessful experiences tend to lower self-efficacy. However, easy successes often encourage expectations of quick successes leading to a reduction in student resilience when faced with challenges. Second, in vicarious experiences, the student models her beliefs based on comparisons with others. These can include peers, tutors, and teachers, especially those with similarly perceived capabilities. Thus, seeing a perceived parallel peer succeed at a task typically increases self-efficacy. Vicarious experiences are particularly useful when the only way to gauge adequacy is to relate personal results with the performance of others. For instance, a student who completes a timed math test in 53 seconds has to gauge her performance by comparing completion times of her peers. Third, in verbal persuasion, the student experiences an outcome via a persuader’s description. For example, she may be encouraged by the persuader, who may praise the student for performing well or comment on the difficulty of a problem. Her interpretation will be affected by the credibility she ascribes to the persuader. Thus, it is pedagogically constructive to suggest a student has the capabilities to succeed at a given task verbally, likely raising the student’s self-efficacy. Verbal persuasion is particularly useful in enabling
students to overcome self-doubt. Although verbal persuasion does not have a large impact on lasting student persistence it can encourage immediate action and effort. Fourth, with *physiological and emotional effects*, the student responds affectively to situations and their anticipation. These experiences, which often induce stress and anxiety, are manifested in physiological responses, such as increased heart rate and sweaty palms, call for emotional support and motivational feedback since they can be detrimental to success.

Student self-efficacy beliefs regulate human behavior through four major processes central to human performance (Bandura, 1997):

- **Cognitive Processes.** Self-efficacy affects student reasoning and problem-solving (Bandura, 1995; Schunk and Pajares, 2002; Zimmerman, 2000) to the point that performance can be elevated or impaired. High self-efficacy affords students the abilities to set ambitious future goals and a rigid commitment to achieve them. Furthermore, self-efficacious students are better able to select favorable problem-solving strategies and more quickly disregard inadequate approaches. On the other hand, low self-efficacy reduces the payoff of achieving weakly structured goals and elicits an inability to select optimal problem-solving strategies.

- **Motivational Processes.** Students with high self-efficacy are more likely to visualize successful outcomes. Setting challenging goals in turn yields elevated levels of motivation (Lepper et al., 1993), another construct affected by self-efficacy. Low self-efficacy interferes with visualizing, thereby reducing resilience and persistence abilities.

- **Selective Processes.** The activities that students choose to engage in significantly affects their potential to achieve. Students with high self-efficacy select challenging activities and environments that regularly present opportunities to exhibit persistence. Students with low self-efficacy tend to select activities and environments that present little or no challenge and can often be detrimental to the development of cognitive and social skills.

- **Affective Processes.** Self-efficacy influences students’ abilities to regulate their own affective states. There are three fundamental ways in which self-efficacy influences affective state: self-control over thought, action, and affect (Bandura, 1997). First, *thought-oriented mode* refers to cognitive processes that are emotionally arousing and the ability to
self-regulate such thoughts. Self-efficacy beliefs about one’s ability to overcome risks and to persist through or avoid emotionally disturbing thoughts have great influence on behavior. Second, action-oriented mode refers to taking courses of action that effect change in the environment so that there is an increased potential for desirable affective outcomes. Third, affect-oriented mode refers to one’s abilities to conceive adverse affective states when faced with adverse-emotion-invoking situations. Self-relaxation, calming inner monologue and controlled breathing are techniques often used to reduce undesirable emotional arousal.

Predicting self-efficacy holds great promise for intelligent tutoring systems and educational software in general. Self-efficacy beliefs have a stronger correlation with desired behavioral outcomes than many other motivational constructs (Graham and Weiner, 1996), and it has been recognized in multiple educational settings that self-efficacy can predict both motivation and learning effectiveness (Zimmerman, 2000). Thus, if it were possible to enable ITSs to accurately model self-efficacy, they might be able to leverage it to increase students’ academic performance. Two recent efforts have explored the role of self-efficacy in ITSs. One introduced techniques for incorporating knowledge of self-efficacy in pedagogical decision making (Beal and Lee, 2005). Using a pre-test instrument and knowledge of problem-solving success and failure, instruction is adapted based on changes in motivational and cognitive factors. The second explored the effects of pedagogical agent design on students’ traits, which included self-efficacy (Baylor and Kim, 2004; Kim, 2005). The focus of the experiments reported in this article is on the automated induction of self-efficacy models for runtime use by ITSs.

One can distinguish two fundamental approaches to modeling self-efficacy: analytical and empirical. In the analytical approach, models of self-efficacy can be constructed by analyzing the findings of the educational psychology literature. However, self-efficacy is not well understood computationally, i.e., the literature has not produced a set of rules defining precise characteristics of particular levels of self-efficacy. While we do have expressive computational models of affect, e.g., the OCC model (Ortony et al., 1988) and EMA (Gratch and Marsella, 2004) based on the Smith and Lazarus’ appraisal theory of emotion (Lazarus, 1991), we do not have similarly rich, comprehensive models of self-efficacy. Moreover, because self-efficacy reasoning requires drawing inferences
about a student’s past experiences, her beliefs, her intentions, her affective state, her current situational context, and her capabilities, devising a complete and universal model of self-efficacy seems to be well beyond our grasp at the current juncture.

An alternative to analytically devising models of self-efficacy for intelligent tutoring systems is the empirical approach. If somehow we could create models of self-efficacy that were derived directly from observations of “efficacy in action,” we could create empirically grounded models based on student behaviors exhibited during the performance of a specific task within a given domain. While it is not apparent that this approach could produce a universal model of self-efficacy—a universal model may not even be achievable, at least in the near term—the empirical approach could nonetheless generate models of self-efficacy that significantly extend the pedagogical capabilities present in current educational software and intelligent tutoring systems.

6.2 Self-efficacy Modeling in an Online Tutorial System

In this experiment, two families of self-efficacy models were induced: the model learner constructed (1) static models, which are based on demographic data and a validated problem-solving self-efficacy instrument (Bandura, 2006), and (2) dynamic models, which extend static models by also incorporating real-time physiological data. Both families of resulting models operate at runtime, are efficient, and do not interrupt the learning process.

6.2.1 Method

Participants and Design

In a formal evaluation, data was gathered from thirty-three subjects in an Institutional Review Board (IRB) of North Carolina State University approved user study. There were 6 female and 27 male participants varying in age, race, and marital status. Approximately 12 (36%) of the participants were Asian, 20 (60%) were Caucasian, and 1 (3%) was Black or African-American. 27% of the participants were married. Participants average age was 26.15 (SD=5.32).
Materials and Apparatus

The pre-experiment paper-and-pencil materials for each participant consisted of a demographic survey, tutorial instructions, Bandura’s Problem-solving Self-Efficacy Scale (Bandura, 2006), and the problem-solving system directions. Post-experiment paper-and-pencil materials consisted of a general survey. The demographic survey collected basic information such as gender, age, ethnicity, and marital status. The tutorial instructions explained to participants the details of the task, such as how to navigate through the tutorial and an explanation of the target domain. Bandura’s validated Problem-solving Self-Efficacy Scale (Bandura, 2006), which was administered after participants completed a tutorial in the domain of genetics, asked them to rate how certain they were in their ability to successfully complete the upcoming problems (which they had not yet seen). The problem-solving system directions supplied detailed task direction to participants, as well as screenshots highlighting important features of the system display, such as the “self-efficacy slider.”

The computer-based materials consisted of an online genetics tutorial and an online genetics problem-solving system. The online genetics tutorial consisted of an illustrated 15-page web document which included some animation and whose content was drawn primarily from a middle school biology textbook (Padilla et al., 2000). The online genetics problem-solving system consisted of 20 questions, which covered material in the online genetics tutorial. The problem-solving system presented each multiple-choice question individually and required participants to rate their confidence, using the “self-efficacy slider,” in their answer before proceeding to the next question.

Apparati consisted of a Gateway 7510GX laptop with a 2.4 GHz processor, 1.0 GB of RAM, 15-in. monitor and biofeedback equipment for monitoring blood volume pulse (one sensor on the left middle finger) and galvanic skin response (two sensors on the left first and third fingers). Participants’ right hands were free from equipment so they could make effective use of the mouse in problem-solving activities.

6.2.2 Procedure

Each participant entered the experimental environment (a conference room) and was seated in front of the laptop computer. First, participants completed the demographic survey at their own rate. Next, participants read over the online genetics tutorial directions before proceeding to the
online tutorial. On average, participants took 17.67 (SD = 2.91) minutes to read through the genetics online tutorial. Following the tutorial, participants were asked to complete the Problem-Solving Self-Efficacy Scale considering their experience with the material encountered in the genetics tutorial. The instrument asked participants to rate their level of confidence in their ability to successfully complete certain percentages of the upcoming problems in the problem-solving system. Participants did not have any additional information about the type of questions or the domain of the questions contained in forthcoming problems. Participants were then outfitted with biofeedback equipment on their left hand while the problem-solving system was loaded. Once the system was loaded, participants entered the calibration period in which they read through the problem-solving system directions. This allowed the system to obtain initial readings on the temporal attributes being monitored, in effect establishing a baseline for HR and GSR.

The problem-solving system presented randomly selected, multiple-choice questions to each participant. The participants selected an answer and then manipulated the self-efficacy slider representing the strength of their belief in their answer being correct. Participants completed 20 questions. They averaged 8.15 minutes (SD = 2.37) to complete the problem-solving system. Finally, they were asked to complete the post-experiment survey at their own rate before concluding the session.

After all participants’ sessions were completed, the procedure (described in Section 4) was used to induce models of self-efficacy ratings from the training sessions (Figure 6.1). Each session log, containing on average 14,645.42 (SD = 4,010.57) observation changes (e.g., a change in location, student heart beat detected, or changes in selected answer), was first translated into a full observational attribute vector. For example, BVP and GSR readings were taken nearly 30 times every second reflecting changes in both heart rate and skin conductivity. Blood volume pulse (used for monitoring HR) readings were difficult to obtain from two participants resulting in the elimination of that data. The entire dataset was used to generate several types of self-efficacy models, each predicting self-efficacy with varying degrees of granularity. These included two-level models (Low, High), three-level models (Low, Medium, High), four-level models (Very Low, Low, High, Very High), and five-level models (Very Low, Low, Medium, High, Very High).
6.2.3 Results

Below we present the results of the naïve Bayes and decision tree classification models and provide analyses of the collected data. Various ANOVA statistics are presented for results that are statistically significant. Because the tests reported here were performed on discrete data, we report Chi-square test statistics ($\chi^2$), including both likelihood ratio Chi-square and the Pearson Chi-square values. Fisher’s Exact Test is used to find significant p-values at the 95% confidence level ($p < .05$).

Model Results

Naïve Bayes and decision tree classifiers are effective machine learning techniques for generating preliminary predictive models. Naïve Bayes classification approaches produce probability tables that can be implemented into runtime systems and used to continually update probabilities for assessing student self-efficacy levels. Decision trees provide interpretable rules that support
runtime decision making. The runtime system monitors the condition of the attributes in the rules to determine when conditions are met for assigning particular values of student self-efficacy. Both the naïve Bayes and decision tree machine learning classification techniques are useful for preliminary predictive model induction for large multidimensional data, such as the 144-attribute vector used in this experiment. Because it is unclear precisely which runtime variables are likely to be the most predictive, naïve Bayes and decision tree modeling provide useful analyses that can inform more expressive machine learning techniques (e.g., Bayesian networks) that also leverage domain experts’ knowledge. Both static and dynamic models of self-efficacy were induced using naïve Bayes and decision tree classification techniques. Dynamic models were induced from all observable attributes, while static models excluded physiological response data.

All models were constructed using a tenfold cross-validation scheme. In this scheme, data is decomposed into ten equal partitions, nine of which are used for training and one used for testing. The equal parts are swapped between training and testing sets until each partition has been used for both training and testing. Tenfold cross-validation is widely used for obtaining the best estimate of error (Witten and Frank, 2005). Cross-validated ROC curves are useful for presenting the performance of classification algorithms for two reasons. First, they represent positive classifications, included in a sample, as a percentage of the total number of positives, against

![ROC curves](image)

**Figure 6.2.** ROC curves for naïve Bayes (a) and decision tree (b) three-level models of self-efficacy. Overall the naïve Bayes model correctly classified 72% of the instances while the decision tree was able to correctly classify 83%.

95
negative classifications as a percentage of the total number of negatives (Witten and Frank, 2005). Second, the area under ROC curves is widely accepted as a generalization of the measure of the probability of correctly classifying an instance (Hanley and McNeil, 1982).

The ROC curves depicted in Figure 6.2 show the results of both a naïve Bayes and decision tree three-level model. Low-confidence was noted by a student self-efficacy rating lower than 33 (on a 0 to 100 scale). Medium-confidence was determined by rating between 33 and 67, while High-confidence was represented all ratings greater than 67. The smoothness of the curve in Figure 6.2(a) indicates that the data collected seems to have sufficiently covered the multidimensional space for inducing naïve Bayes models. The jaggedness of the curves in Figure 6.2(b) indicates that training data did not cover the entirety of the instance space. While sufficient data was collected for the induction process and modeling the phenomena of self-efficacy, further training may be useful to obtain complete coverage of the multidimensional space. In particular, further investigation will be required to gather data from situations in which there are more opportunities for students to experience low self-efficacy. Although training data did not cover all possible instances in the multidimensional space (notice how the ROC curves for induced decision tree models do not extend to the axis in Figure 6.2b), the decision tree model performed significantly better than the naïve Bayes model (likelihood ratio, χ² = 21.64, Pearson, χ² = 21.47, p < .05). The highest performing model induced from all data was the two-level decision-tree based dynamic model, which performed significantly better than the highest performing static model, which was a two-level decision tree model (likelihood ratio, χ² = 3.99, Pearson, χ² = 3.97, p < .05). The three-level dynamic decision tree model was also significantly better than the static three-level decision tree (likelihood ratio, χ² = 18.26, Pearson, χ² = 18.13, p < .05). All model results are shown in Table 6.1. The performance of two dynamic naïve Bayes models proved to be significantly better than baseline models. Both of the dynamic two-level model (likelihood ratio, χ² = 4.272, p = 3.87 × 10⁻², and Pearson, χ² = 4.26, p = 3.9 × 10⁻², df = 1) and the dynamic four-level model (likelihood ratio, χ² = 10.647, p = 1.1 × 10⁻³, and Pearson, χ² = 10.615, p = 1.1 × 10⁻³, df = 1) yielded significant improvements over the baseline models. No static naïve Bayes models’ performance was significantly better than baseline models. The performance of static decision tree models also did
not produce significant results over baseline performance. However, all dynamic decision tree models did perform significantly better than baseline models (Table 6.2).

Table 6.1. Model accuracy results (area under ROC curves). Static models were induced from non-intrusive demographic and Problem-Solving Self-Efficacy data. Dynamic models were also based on physiological data. Baseline models report the portion of the distribution pertaining to the most reported efficacy level (i.e., 80.6% of self-efficacy reports for the two-level models were High). * Value represents model performance statistically significant from baseline performance.

<table>
<thead>
<tr>
<th>Granularity</th>
<th>Baseline Accuracy</th>
<th>Decision Tree Accuracy</th>
<th>Likelihood Ratio</th>
<th>Pearson</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level Models</td>
<td>80.6%</td>
<td>82.2%</td>
<td>83.4%</td>
<td>86.9%</td>
<td>86.9%</td>
</tr>
<tr>
<td>Three-level Models</td>
<td>69.8%</td>
<td>70.1%</td>
<td>73.4%</td>
<td>83.4%</td>
<td>83.4%</td>
</tr>
<tr>
<td>Four-level Models</td>
<td>65.4%</td>
<td>68.8%</td>
<td>69.0%</td>
<td>74.7%</td>
<td>74.7%</td>
</tr>
<tr>
<td>Five-level Models</td>
<td>60.9%</td>
<td>63.4%</td>
<td>63.9%</td>
<td>75.3%</td>
<td>75.3%</td>
</tr>
</tbody>
</table>

Table 6.2. Dynamic decision tree model improvements were statistically significant over baseline model accuracies.
Model Attribute Effects on Self-efficacy

Heart rate and galvanic skin response had significant effects on self-efficacy predictions (Table 6.3). Participants’ age was the only demographic attribute to have a significant effect on all levels of self-efficacy models. Table 6.3 presents several effects of physiological response and pre-experiment survey data, including demographic information and Bandura’s problem-solving self-efficacy scale, on self-efficacy predictions. These results suggest that when modeling self-efficacy at higher-granularity it becomes more important to account for student demographics. Two-level self-efficacy models have the least significant effectors. This is likely due to the results of the two-level baseline model, in which 80.6% of the efficacy self-reports are classified with the label, “High”.

Table 6.3. Chi-squared values representing the significance effects of physiological signals, demographics, and Bandura’s problem-solving self-efficacy scale instrument on dynamic self-efficacy models (p < 0.5).

<table>
<thead>
<tr>
<th>Effect</th>
<th>Self-efficacy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two-level</td>
</tr>
<tr>
<td>Physiological Signals</td>
<td>Heart rate</td>
</tr>
<tr>
<td></td>
<td>Galvanic Skin Response</td>
</tr>
<tr>
<td>Demographics</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Race &amp; Ethnicity</td>
</tr>
<tr>
<td>Bandura’s Problem-Solving Self-efficacy Scale</td>
<td>Collective Responses</td>
</tr>
<tr>
<td></td>
<td>10% Response</td>
</tr>
<tr>
<td></td>
<td>20% Response</td>
</tr>
<tr>
<td></td>
<td>30% Response</td>
</tr>
<tr>
<td></td>
<td>40% Response</td>
</tr>
<tr>
<td></td>
<td>50% Response</td>
</tr>
<tr>
<td></td>
<td>60% Response</td>
</tr>
<tr>
<td></td>
<td>70% Response</td>
</tr>
<tr>
<td></td>
<td>80% Response</td>
</tr>
<tr>
<td></td>
<td>90% Response</td>
</tr>
<tr>
<td></td>
<td>100% Response</td>
</tr>
</tbody>
</table>
Self-efficacy is closely associated with motivational and affective constructs that both influence (and are influenced by) a student's physiological state. It is therefore not unexpected that a student's physiological state can be used to more accurately predict her self-efficacy. For example, Figure 6.3 shows the heart rates for one participant in the study over the course of solving two problems. In Figure 6.3, in the upper left image, the participant reported high levels of self-efficacy for a particular question, while the same participant whose heart rate progression is also shown in the upper right image of Figure 6.3 reported a low level of self-efficacy for another question. The heart rate for the student reporting high self-efficacy gradually drops as they encounter a new question, presumably because of their confidence in their ability to successfully solve the problem. In
contrast, the heart rate for the same student reporting low self-efficacy spikes dramatically, an increase of 5 beats per minute in less than 2 seconds, when the student selects an incorrect answer. This phenomenon is noteworthy since the students were in fact not given feedback about whether or not their responses were correct. Instead the student’s self-appraisal seems to lead to the determination of low efficacy, an inability to successfully achieve at the current task, without a requirement of confirmation of their assessment. It appears that through some combination of cognitive and affective processes the student’s uneasiness with her response, even in the absence of direct feedback, was enough to bring about a significant physiologically manifested reaction. Curiously, there is a subsequent drop in heart rate after the student reports her low level of self-efficacy. In this instance, it seems that providing an opportunity to acknowledge a lack of ability and knowledge to perform may itself reduce anxiety.

The experiment has two important implications for the design of runtime self-efficacy modeling. First, even without access to physiological data, induced decision-tree models can make predictions about students’ self-efficacy that are more accurate than baseline models. Sometimes physiological data is unavailable or it would be too intrusive to obtain the data. In these situations, decision-tree models that learn from demographic data and data gathered with a validated self-efficacy instrument administered prior to problem solving and learning episodes, can model self-efficacy. Second, if runtime physiological data are available, they can significantly enhance self-efficacy modeling. Given access to HR and GSR, self-efficacy can be predicted more accurately.

In summary, the static models are able to predict students’ real-time levels of self-efficacy with 73% accuracy, and the resulting dynamic models are able to achieve 83% predictive accuracy. Thus, non-intrusive static models achieve a statistically significant improvement over baseline performance, and their predictive power can be increased by further enriching them with physiological data at varying levels of granularity.

6.2.5 Study Limitations

The foundational evaluation reported above is limited in several ways. First, the subject population was mainly comprised of males, with only 6 female participants. The role of gender cannot extensively be evaluated with such a skewed population. However, while the population is the best
representation of potential users the results are nonetheless encouraging and motivate further investigation of modeling self-efficacy using CARE-like frameworks. Second, the study was limited by the perceived challenge of the problem-solving activities. Participants reported much higher levels of confidence than low levels of confidence. Their confidence might be affected by the online tutorial system itself, promoting student confidence in the curriculum, or it may be an effect of the challenge levels of the problems in the online tutorial system. Future study designs should carefully account for levels of challenge. It is necessary to collect instances of both high and low student efficacy during training problem solving sessions. Thus, challenge should be tunable to induce a range of efficacious students. Note that while it is not the goal of intelligent tutoring systems, to promote low efficacy they should be able to accurately model it.

6.3 Self-efficacy Modeling in an Interactive Learning Environment

The results of the foundational evaluation reported in Section 6.1 indicated that an inductive approach offered potential as a method for modeling self-efficacy and called for further investigation of the techniques. The design of the second experiment was motivated by three factors: explicitly controlling the challenge levels of the learning environment; exploiting task structure to study self-efficacy with an appraisal-theoretic (Lazarus, 1991) framework; and increasing the complexity of the learning environment and, therefore the dimensionality of the induction problem.

- Explicitly controlling the level of challenge of learning tasks in an effort to increase the frequencies of reported low self-efficacy. In the first evaluation the majority of reported levels of self-efficacy were classified as “high” (see baseline model results in the preceding section, Table 5). The dynamic nature of an interactive learning environment would allow for the design of tasks of varying degrees of difficulty, presenting a variety of challenge levels to study participants. Individual tasks could be designed to be more complicated, require more actions to complete, and elicit student persistence to reach achievement.
- Exploiting task structure and notions from appraisal theory (Lazarus, 1991) to model self-efficacy. An immersive, visually-rich interactive learning environment would offer an ideal testbed in which to study the interaction between student self-appraisals and self-efficacy.

101
Recall that self-efficacy beliefs arise from one’s appraisal of the environment and the current situation in conjunction with appraisals of one’s abilities to achieve goals given the current and possible future states of the surrounding environment. Thus, it is likely that a rich learning environment would elicit patterns of self-efficacy in response to student appraisals of unfolding events in learning episodes. In turn, the representation of the environment should then enable induced models to accurately predict student self-efficacy.

- **Automatically inducing models of self-efficacy from observations in an increasingly complex interactive narrative-centered learning environment.** The induction task becomes increasingly difficult as more dimensions are added to represent more complex learning environments. The second empirical study was designed to investigate the potential and the value of creating models of self-efficacy in more complex interactive learning environments, and to “stress-test” the induction approach in higher dimensions.

Together, these factors motivated the second experiment investigating CARE model induction in a rich interactive learning environment.

### 6.3.1 Method

**Participants and Design**

In a formal evaluation, data was gathered from 42 subjects in an Institutional Review Board (IRB) of North Carolina State University approved user study. There were 5 female and 37 male participants. Participants average age was 21.2 (SD = 1.96).

**Materials and Apparatus**

In The pre-experiment materials for each participant consisted of an online demographic survey and Bandura’s Self-Efficacy Scale (Bandura, 2006). The experiment materials consisted of the following: tutorial directions, the online genetics tutorial, the CRYSTAL ISLAND backstory and directions, the CRYSTAL ISLAND interactive environment control sheet, the CRYSTAL ISLAND character profiles and world map, the genetics problem-solving self-efficacy questionnaire (Bandura, 2006), the genetics problem-solving system directions, the online problem-solving system, and a post-experiment survey. The demographic survey collected participant information such as age, gender, race and
ethnicity. Bandura’s Self-Efficacy Scale rates the participants’ self-efficacy in a variety of more general domains. The tutorial directions described the simple navigation controls and lack of time constraints for reading through the genetics tutorial. The CRYSTAL ISLAND backstory and directions presented the participant’s task and some background information about their character. The controls reference sheet described which keys and mouse movements would be needed to manipulate their agent in the training task. The character profiles provided pictures with associated names and job descriptions of characters the participant might meet on the island. The CRYSTAL ISLAND map was a tool to help the participants maintain orientation within the environment and provide navigational assistance. The genetics problem-solving self-efficacy questionnaire was administered to gauge the participants’ self-efficacy with respect to solving genetics problems after completing both the tutorial and CRYSTAL ISLAND interaction. The post survey was used to determine how participants would feel about using similar software in educational settings and their thoughts on affect and self-efficacy uses in videogames and educational software.

Apparati consisted of a Gateway 7510GX laptop with a 2.4 GHz processor, 1.0 GB of RAM, 15-in. monitor and biofeedback equipment for monitoring blood volume pulse (one sensor on the right ring finger) and galvanic skin response (two sensors on the right middle and little fingers).

6.3.2 Procedure
First participants completed the online demographic survey and the online general self-efficacy questionnaire (Bandura, 2006). Participants then completed the genetics tutorial which took anywhere from 5 minutes to 25 minutes. Next, participants were wired with biofeedback sensors similar to those worn by the user in Figure 6.4. The practice task was then completed allowing participants to become familiar with the controls for CRYSTAL ISLAND. Participants were then presented with the CRYSTAL ISLAND materials (backstory, controls, map, and character profiles) while the virtual environment was loaded. Once participants indicated they were prepared and had any questions answered by the research assistant, they began their interaction in CRYSTAL ISLAND. As participants solved the genetics mystery on CRYSTAL ISLAND, they were periodically asked to rate their current level of self-efficacy, i.e., their current belief in their abilities to solve the science mystery. Upon completion of interacting with CRYSTAL ISLAND, participants completed the genetics self-efficacy
questionnaire (Bandura, 2006) prior to receiving the problem-solving system directions. Once participants indicated they were prepared and physiological response measurements had been calibrated, they began solving 20 randomly displayed genetics problems. Each question was presented together with 4 multiple-choice answers and a “self-efficacy slider” which participants adjusted indicating their belief in their ability to correctly solve the given problem. Finally, participants completed the post-experiment questionnaire before the experiment session concluded.

After all participants’ sessions were completed, the same procedure as the one described in Section 6.2 was used to induce models of self-efficacy ratings from the training sessions (Figure 6.5). Training sessions lasted at least eight minutes, and each session log contained at least 15,000 (32,487 at most) observation changes (e.g., a change in location, completing a goal, manipulating an object, or detected heart beat). These changes were first translated into a full observational attribute vector. For example, BVP and GSR readings were taken approximately 30 times every second reflecting changes in both heart rate and skin conductivity. After data were converted into
an attribute vector format a dataset was generated that contained only records in which the biofeedback equipment was able to successfully monitor BVP and GSR throughout the entire training session and in which participants actively participated in the experiment by providing self-reports. Two training sessions from male participants did not satisfy these requirements.

Self-efficacy models were again produced at varying levels of granularity. These included two-level models (Low, High), three-level models (Low, Medium, High), four-level models (Very Low, Low, High, Very High), and five-level models (Very Low, Low, Medium, High, Very High).

6.3.3 Results

All models were evaluated using a tenfold cross-validation scheme for producing training and testing datasets. The ROC curves (Figure 6.6) show the results of decision tree and naïve Bayes modeling for predicting student levels of self-efficacy. The lack of smoothness of the curves indicates that training data did cover the entirety of the multidimensional space. However, collected training data was sufficient for inducing SELF models of self-efficacy. The highest performing model induced from
interactive learning environment training data was the two-level decision tree model, correctly predicting more than 87% of reported levels of self-efficacy. Table 6.4 reports the results of the self-efficacy model induction mode of SELF. Decision tree models’ prediction improvements over naïve Bayes models were significant at the two-level models (likelihood ratio, $\chi^2 = 7.321$, $p = 6.8 \times 10^{-3}$, and Pearson, $\chi^2 = 7.291$, $p = 6.9 \times 10^{-3}$, df = 1) and four-level models (likelihood ratio, $\chi^2 = 24.085$, $p = 9.218 \times 10^{-7}$, and Pearson, $\chi^2 = 23.96$, $p = 9.835 \times 10^{-7}$, df = 1). Furthermore, decision tree models performed significantly better than baseline models: two-level models (likelihood ratio, $\chi^2 = 29.319$, $p = 6.139 \times 10^{-8}$, and Pearson, $\chi^2 = 28.929$, $p = 7.506 \times 10^{-8}$, df = 1), three-level models (likelihood ratio, $\chi^2 = 62.443$, $p = 2.74 \times 10^{-15}$, and Pearson, $\chi^2 = 61.56$, $p = 4.29 \times 10^{-15}$, df = 1), and four-level models (likelihood ratio, $\chi^2 = 25.759$, $p = 3.869 \times 10^{-7}$, and Pearson, $\chi^2 = 25.617$, $p = 4.163 \times 10^{-7}$, df = 1). Naïve Bayes models performance was significantly better than baseline models for two-level models (likelihood ratio, $\chi^2 = 7.433$, $p = 6.4 \times 10^{-3}$, and Pearson, $\chi^2 = 7.412$, $p = 6.5 \times 10^{-3}$, df = 1) and three-level models (likelihood ratio, $\chi^2 = 43.494$, $p = 4.25 \times 10^{-11}$, and Pearson, $\chi^2 = 43.099$, $p = 5.2 \times 10^{-11}$, df = 1). Table 6.5 reports the results of self-efficacy models induced in both the online tutorial system and the interactive learning environment.

In the online tutorial system evaluation, the majority of self-efficacy self-reports were classified as being high efficacy, as indicated by the baseline models (the portion of the distribution belonging

![ROC curves for CARE three-level models induced from CRYSTAL ISLAND interactions.](image)
### Table 6.4. Model results – area under ROC curves for dynamic self-efficacy models. * Value represents model performance statistically significant from baseline performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline (High)</th>
<th>Naive Bayes</th>
<th>Decision Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level Models</td>
<td>76.1%</td>
<td>82.1%*</td>
<td>87.3%*</td>
</tr>
<tr>
<td>Three-level Models</td>
<td>61.2%</td>
<td>77.6%*</td>
<td>80.4%*</td>
</tr>
<tr>
<td>Four-level Models</td>
<td>63.5%</td>
<td>64.0%</td>
<td>72.9%*</td>
</tr>
<tr>
<td>Five-level Models</td>
<td>62.4%</td>
<td>59.2%</td>
<td>63.2%</td>
</tr>
</tbody>
</table>

### Table 6.5. Model results – area under ROC curves for online tutorial system static and dynamic self-efficacy models, and interactive learning environment dynamic models. * Value represents model performance statistically significant from baseline performance.

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline (High)</th>
<th>Naive Bayes</th>
<th>Decision Tree</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two-level Models</td>
<td>80.6%</td>
<td>82.2%</td>
<td>82.9%</td>
</tr>
<tr>
<td>Three-level Models</td>
<td>69.8%</td>
<td>70.1%</td>
<td>73.4%</td>
</tr>
<tr>
<td>Four-level Models</td>
<td>65.4%</td>
<td>68.8%</td>
<td>69.0%</td>
</tr>
<tr>
<td>Five-level Models</td>
<td>60.9%</td>
<td>63.4%</td>
<td>63.9%</td>
</tr>
<tr>
<td>Online Tutorial System</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static Model Accuracy</td>
<td>80.6%</td>
<td>85.2%*</td>
<td>86.9%*</td>
</tr>
<tr>
<td>Dynamic Model Accuracy</td>
<td>76.1%</td>
<td>82.1%*</td>
<td>87.3%*</td>
</tr>
<tr>
<td>Interactive Learning Environment</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic Model Accuracy</td>
<td>65.4%</td>
<td>74.7%*</td>
<td>78.9%*</td>
</tr>
<tr>
<td>Dynamic Model Accuracy</td>
<td></td>
<td>63.5%</td>
<td>64.0%</td>
</tr>
<tr>
<td>Dynamic Model Accuracy</td>
<td></td>
<td>62.4%</td>
<td>59.2%</td>
</tr>
</tbody>
</table>
Thus, in the interactive learning environment development, some tasks were designed to present more challenging scenarios to students than were presented in the online tutorial system in an effort to elicit a higher percentage of low efficacy self-reports. While the baseline results indicate that the majority of self-efficacy self-reports in the interactive learning environment evaluation were also classified as high and very high efficacy, we obtained significantly more instances of students reporting low efficacy. Table 6.6 reports the baseline dynamic models from both evaluations and likelihood ratio and Pearson’s statistics indicating the reduced accuracy in the interactive learning environment dynamic baseline models to be statistically significant. Since baseline models are composed of high self-efficacy report instances, a drop in baseline models (drop in the count of high self-efficacy reports) corresponds directly to an increase in counts of low self-efficacy report instances. This observation of a reduction in the quantity of high self-efficacy reports indicates a significant gain in the quantity of low self-efficacy reports. This fact is supported by the results presented in Table 6.6.

### Table 6.6

Baseline comparisons between the online tutorial system and the interactive learning environment evaluations. The percentage increase in the number of instances in which students reported low levels of self-efficacy from the online tutorial system to the interactive learning environment evaluation was statistically significant. * For the five-level dynamic baseline model, comparison p-values are slightly above .05 indicating weak significance.

<table>
<thead>
<tr>
<th>Models</th>
<th>Online Tutorial System</th>
<th>Interactive Learning Environment</th>
<th>Likelihood Ratio</th>
<th>Pearson</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80.6%</td>
<td>76.1%</td>
<td>17.264</td>
<td>0.05</td>
</tr>
<tr>
<td>Two-level Dynamic</td>
<td>80.6%</td>
<td>69.8%</td>
<td>9.476</td>
<td>0.05</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>80.6%</td>
<td>65.4%</td>
<td>5.004</td>
<td>0.05</td>
</tr>
<tr>
<td>Three-level Dynamic</td>
<td>80.6%</td>
<td>60.9%</td>
<td>3.470</td>
<td>0.05</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>80.6%</td>
<td>62.4%</td>
<td>6.250</td>
<td>0.05</td>
</tr>
<tr>
<td>Four-level Dynamic</td>
<td>80.6%</td>
<td>59.9%</td>
<td>6.290</td>
<td>0.05</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>80.6%</td>
<td>62.4%</td>
<td>6.290</td>
<td>0.05</td>
</tr>
<tr>
<td>Five-level Dynamic</td>
<td>80.6%</td>
<td>60.9%</td>
<td>6.250</td>
<td>0.05</td>
</tr>
<tr>
<td>Baseline Model</td>
<td>80.6%</td>
<td>62.4%</td>
<td>6.290</td>
<td>0.05</td>
</tr>
</tbody>
</table>

6.3.4 Discussion

A notable difference between the online tutorial system evaluation and the interactive learning environment evaluation is the dimensionality of the observational attribute vector. Recall that 150
features were observed in the online tutorial system, while in the interactive learning environment over 275 features were continuously monitored. This added dimensionality called for a larger dataset covering a larger space to improve the predictive accuracy of self-efficacy modeling. The training data obtained from the 40 usable sessions appears to have been sufficient for modeling self-efficacy in CRYSTAL ISLAND. The design of CRYSTAL ISLAND learning tasks, and particularly the varying challenge levels of the tasks, led to an increase in reports of low-efficacy in the interactive learning environment evaluation. This observation may explain why CARE-induced models of self-efficacy obtained similar levels of accuracy among comparable models in the interactive learning environment as they did in the online tutoring system. It is noteworthy considering the increased dimensionality and complexity constraints placed on the induction process for learning self-efficacy models in the CRYSTAL ISLAND learning environment.

One of the challenging tasks in the design of CARE for the interactive learning environment evaluation was selecting observable attributes to monitor throughout student interactions that would also be used in student appraisal and self-efficacy determination. Because of the difficult nature of identifying attributes used by most students in appraisal, we elected to monitor the large 283-dimensional space designed for CRYSTAL ISLAND. The performance of induced models suggests that there is overlap between the features contained in the observable attribute vector and the attributes of the learning environment used by students in appraisal and realized in reports of self-efficacy.

We have considered a variety of models in the online tutorial system and the interactive learning environment along three dimensions: static vs. dynamic data, classification technique, and granularity. The online tutorial system evaluation found that dynamic models (inclusion of physiological data) performed significantly better, i.e., they correctly classified student self-efficacy more accurately, than static models (exclusion of physiological data) of self-efficacy. This result motivated the focus of investigating only dynamic models in the interactive learning environment evaluation. We hypothesize the performance improvements of dynamic models stems from the relationship between self-efficacy and physiological response. Because physiological responses follow from emotional reactions to situation appraisals (Frijda 1986; Picard, 1997) and self-efficacy beliefs arise from a similar cognitive appraisal process (Bandura, 1997), it seems appropriate to infer
that changes in physiology are perhaps generated in response to a combination of interacting affective factors, such as emotional state, self-efficacy beliefs, and motivational states. Following research that has demonstrated the ability to recognize affective state from classification of physiological data (Burleson, 2006; Conati, 2002; Healey, 2000; Picard et al., 2001; Prendinger et al., 2005), it seems reasonable to infer that physiological response data may also be useful in predictions of self-efficacy.

Both evaluations investigated two families of classification techniques: rule-based models (decision trees) and probabilistic models (naïve Bayes). In the online tutorial system and the interactive learning environment, decision tree models outperformed naïve Bayes models. We hypothesize that this is likely due to the naïve Bayes assumption that all observable attributes are conditionally independent. As noted above, this is clearly not the case in CRYSTAL ISLAND where particular events can only occur in particular locations, such as running an experiment on an artifact, which requires the use of stationary machinery that can only be found in the laboratory on CRYSTAL ISLAND.

Induced models of self-efficacy also vary in the levels of granularity in which they predicted student efficacy. There is a noticeable decay in model performance as the granularity is increased in both evaluations. For instance, the performance of dynamic decision tree models from the interactive learning environment evaluation were 87.3%, 80.4%, 72.9%, and 63.2% for two, three, four, and five-level models respectively. Despite the trend of decreasing accuracy with increasing levels of granularity there are several instances worth noting, such as the performance of the two-level dynamic decision tree model from the interactive learning environment which accurately predicted 80.4% of instances, outperforming the associated baseline by 19.2% (the baseline model achieved 61.2% accuracy). However, it remains clear that as granularity is increased the multidimensionality of the observation attribute vector hinders the ability to accurately predict student efficacy levels. For runtime environments, this decay effect raises the question of which level of granularity should implemented models use to predict self-efficacy? The answer to this question must consider the tradeoffs between models. This calls for analyzing the increasing number of misclassifications associated with each additional level of granularity and how misclassifications affect system performance. For instance, consider the two-level dynamic decision
tree model from the interactive learning environment which was able to predict 87.3% instances correctly. The 12.7% of instances that were incorrectly classified were predicted to be in the other class of the two-level model, i.e., instances of high self-efficacy were misclassified as low, and low self-efficacy instances were misclassified as high in 12.7% of all instances. After introducing another level of granularity, yielding a three-level model, we notice performance slips to 80.4% with an increase in misclassifications accounting for 19.6% of all instances. While higher granularity models do indeed provide more information than low levels of granularity, misclassifications can increase. This highlights the tradeoff question: when should models with higher levels of granularity (and therefore more precision) but with lower predictive accuracy be preferred to models with lower levels of granularity (and therefore less precision) but with higher predictive accuracy? In the future, it will be important to consider the tradeoff question in evaluations of runtime self-efficacy models.

6.3.5 Study Limitations

It is unclear how well induced models will perform with diverse user populations. The subjects of this study were mainly comprised of male undergraduate students majoring in computer science. The models induced with the collected corpus described above may not be effective for other populations. Additional studies are required to account for a greater variance in demographics and behaviors in the learning environment.

6.4 Summary of Self-Efficacy Modeling

Both the foundational evaluation with an online tutorial system and the follow-up evaluation with an interactive learning environment suggest that it is possible to model self-efficacy from observable attributes with induced models achieving statistically significant improvements in performance over baseline models. The two experiments suggest that it may be possible to devise empirically based models that can then be used to support learning in interactive settings. Recall from Section 6.1 that Bandura distinguishes four types of self-efficacy effectors: enactive mastery experiences, vicarious experiences, verbal persuasion, and physiological and affective state (Bandura, 1997).
Here, for each type of effector, we consider how ITSs may employ tutorial strategies to enhance and maintain ideal levels of student self-efficacy in conjunction with a CARE-like self-efficacy diagnostic framework.

ITSs can facilitate mastery learning (Bloom, 1984) by creating experiences in which the difficulty of the task or specific problems is adapted to the individual student. Diagnosing self-efficacy can better inform the pedagogical decisions bearing on the selection of problem difficulty by ensuring that the student has not only mastered the concept but believes in her abilities to use acquired knowledge in the domain. When self-efficacy models determine that a student has low efficacy beliefs during particular problem-solving tasks, an ITS can redirect the student’s tasks to prior concepts or sub-problems that will help the student gain confidence in the skills required to solve the problem eliciting low self-efficacy. Self-efficacy models could contribute to improved pedagogical planning by informing the planner when replanning is necessary for individual students. Self-efficacy models could also contribute to error correction decision making, and they could play a role in determining when to intervene to provide tutorial guidance. Since efficacious students are likely to persist longer than students with low self-efficacy, pedagogical monitoring components might permit efficacious students to work through their own mistakes and consider intervening when mistakes are made by ineffectual students. Challenge is an intrinsic motivator that is often employed by human tutors (Lepper et al., 1993). Self-efficacy models could inform decisions about the appropriate challenge level of tasks to create adaptive learning experiences that sustain ideal levels of self-efficacy and motivation, which in turn support effective learning. The amount of learning that takes place relates to the amount of mental effort students exhibit which has an “inverted U” relationship to self-efficacy (Clark, 1999). Thus, the difference between low self-efficacy and high self-efficacy needs to be handled delicately by ITSs. Just as too low self-efficacy can constrain learning, so too can too high self-efficacy.

The adaptability of ITSs may enable them to create vicarious experiences, which are sometimes difficult to elicit in a classroom setting. In particular, peer learning companions (Aimeur et al., 2000; Burleson and Picard, 2004; Chan and Baskin, 1990; Chou et al., 2003; Goodman et al., 1998; Kim, 2004) can create adaptable vicarious experiences for students. Student observation of similar peers succeeding may enhance the observing student’s self-efficacy if she believes she can also succeed at
the same or similar tasks (Schunk, 1987). Consider an ITS in which a peer companion agent fails or struggles at a task. Witnessing this event may enable less efficacious students to exert more effort if they believe their abilities to be greater than the companion agent’s abilities. Likewise, highly efficacious students may persist as a companion agent begins to succeed at similar tasks and problems. This form of competition with a learning companion could contribute to increases in student efficacy. It has been determined that student perception of a companion agent’s knowledge level can have a material effect on student self-efficacy (Baylor and Kim, 2004). Monitoring such perceptions could support the orchestration of agent and environment behaviors, and it could inform the adaptive selection of agent personae that most effectively support interactions with individual students. Enabling an ITS to adaptively control the perception of peers in the learning environment through personae selection, agent task completion, and interactive dialogue to demonstrate agent knowledge (or lack thereof) are promising techniques for enhancing and maintaining student self-efficacy.

Verbal persuasion is a common motivational tool used by tutors (Lepper et al., 1993), both human and automated. Tutors who express confidence in a student’s abilities can have a profound effect on the student’s own self-efficacy beliefs. The impact is determined by the value the student places on the persuader, so an established relationship between a tutor and the student makes verbal persuasion all the more powerful. ITS research has considered several approaches to providing feedback (Aleven et al., 2004; Corbett and Anderson, 2001; Moreno, 2004), but feedback that improves self-efficacy can also be less performance-driven. In a study that targeted students with academic problems, direct feedback on success did not affect self-efficacy; rather, feedback on the selected cognitive strategies to develop a solution substantially enhanced student self-efficacy beliefs (Schunk and Rice, 1987). This is not to discount the potential effects of rewarding performance, especially verbally. Verbal performance feedback ensures that students are aware of goal progression, immersed in challenging tasks, and may contribute to student task persistence. Verbal persuasion is not as powerful as enactive mastery or vicarious experiences, particularly for inducing lasting effects on student efficacy beliefs (Bandura, 1997). Verbal persuasion is a technique that learning companions might employ if students are closing in on learning goals and self-efficacy.
models are beginning to detect declining student efficacy. In short, verbal persuasion can quickly elicit short bursts of efficacy to motivate students at critical junctures in learning episodes.

The final effector Bandura considers is physiological and affective state. This calls for self-efficacy modeling and affect recognition to operate in tandem. Changes in affective state and the subsequent changes in student physiology will impact self-appraisals of efficacy. Thus, devising strategies to guide students toward affective states with lower arousal levels will diminish the adverse effects of high-arousal physiological responses on student efficacy. For example, stress elicits aroused responses, such as increased heart rate and sweaty palms. Such responses may cause adverse self-appraisals of efficacy. Employing affect recognition combined with self-efficacy models can inform interactive pedagogical components to take action when situations of arousal and low self-efficacy co-occur. One approach to addressing student affect is to respond appropriately, given the social interactive context of an ongoing learning episode, through empathetic companion agents (Kim, 2005; McQuiggan and Lester 2006a, Paiva et al., 2005; Prendinger and Ishizuka, 2005). The empathetic nature of such agents may help students better self-regulate their own affective state leading to stronger senses of efficacy. Recognizing that physiological and affective states influence self-efficacy beliefs and in turn, that self-efficacy affects affective processes (Bandura, 1997), self-efficacy modeling can play an important role in the affective-loop of ITSs.
Chapter 7

Modeling Emotion

In this chapter we investigate the CARE inductive framework for the task of modeling emotion. First Section 7.1 will present relevant background, associated related work, and motivation for examining emotion (detailed related work can be found in Chapter 2). Then we present a series of corpus collection studies designed for the induction of various affect recognition models. Section 7.2 describes a study aimed at modeling the emotional states excitement, fear, frustration, happiness, relaxation, and sadness. One of the challenges of this work, as noted in Chapter 1, is identifying the set of emotions that need to be considered. This particular set of emotions was chosen for its coverage of Lang’s two dimensional-space of emotion (1995). In Chapter 10 we further investigate the challenges surrounding emotions that are relevant to narrative-centered learning. Section 7.3 presents an isolated corpus collection study aimed at modeling frustration. This work will also investigate the merit of early prediction methods. Section 7.4 then summarizes the affect modeling research presented in this chapter.

7.1 Emotion Background, Related Work, and Motivation

In this chapter we will focus our discussion on affect recognition. Of primary interest will be the capability to model the emotions of frustration and flow. Both of these states are relevant to learning (Craig et al., 2004; D’Mello et al., 2006) and have occurred with high frequencies in CRYSTAL ISLAND evaluations (McQuiggan, Robison, et al., 2008).
Frustration
Frustration occurs when something or someone impedes an individual’s progress towards a particular goal. As an emotional response, frustration is not fundamentally different from another negative affective response common to a variety of situations, anxiety. Anxiety is often more than merely an emotional response; it also consists of behavioral, cognitive, and physiological responses (Seligman et al., 2001). However, because our work focuses on interactive task-oriented narrative learning environments, where the construction and achievement of goals is critical to student learning episodes, we primarily focus on frustration. Both anxiety and frustration can lead students to fixate on the impeding source of frustration, diverting attention from, and in some cases causing students to ignore, the task at hand. Anxiety particularly arises when students affectively respond to their focus on planning contingencies for potential future events. Detecting situations that will likely lead to student anxiety or frustration that in turn may eventually lead to student impasses would allow learning environments to intervene early, i.e., before the emotion is fully realized as the student approaches her threshold for the particular emotion.

Several strategies can be employed to identify levels of anxiety and frustration that are not detrimental to learning. Setting realistic expectations based on a student’s abilities and observed past performance can contribute to student successes. Encouragement, and specific feedback directed at particular behaviors, not merely global performance assessments, may help motivate students and provide them with guidance so that they can improve their self-assessment and help them cope with frustration and anxiety (Ormrod, 2002). The central questions that must be answered are, “How can we detect and monitor anxiety and frustration levels so that our learning environments have sufficient time to plan and execute appropriate scaffolding?” and, “With what computational mechanisms can we draw inferences about the student, the task, and the environment to accurately predict student frustration?”

Flow
For nearly four decades Csikszentmihalyi has studied flow and its relevance in a broad range of activities ranging from working crossword puzzles and playing chess to rock climbing and performing surgery (Csikszentmihalyi, 1975; Csikszentmihalyi, 1990; Csikszentmihalyi, 1996; Csikszentmihalyi
The primary function of these flow activities is that the player or performer finds them to be enjoyable, autotelic experiences (experiences that are self-motivated and worth doing for the sole purpose of doing the activity). Many creative activities such as those performed by composers, dancers, chess players, surgeons, medical researchers, and mathematicians elicit flow. Csikszentmihalyi observes that all flow experiences share one commonality: they “provide a sense of discovery, a creative feeling of transporting the person into a new reality.” (Csikszentmihalyi, 1990, p.74). Csikszentmihalyi has identified nine characteristics of flow (1996).

Because flow leads to growth, discovery, and higher performance (Csikszentmihalyi, 1990), applications in business, education, training, and a variety of other domains may benefit from the ability to automatically recognize states of flow. Accurate models of flow and affect will enable adaptable systems to design and support tasks to maintain flow (for users already in flow) and to scaffold experiences for users not in flow (i.e., users who may be bored, anxious, or frustrated).
When one veers from the flow channel they must find a way back, perhaps by increases in challenge or skill (less likely are decreases along these dimensions). Following Csikszentmihalyi’s model of flow (Figure 7.1) a person in a state of boredom should enter or increase challenge in their activity, while an anxious or frustrated person might focus on skill development. Adaptive systems might consider practice tasks, or compartmentalization of goals and tasks, to reduce challenge and align with current skill levels.

7.2 Affect Modeling in a Narrative-centered Environment

In this study, we investigate the CARE architecture for inducing models of student affect. In this case we focus on the recognition of six emotions: excitement, fear, frustration, happiness, relaxation, and sadness in CRYSTAL ISLAND.

7.2.1 Method

Participants and Design

In a formal evaluation, data was gathered from thirty-six subjects in an Institutional Review Board (IRB) of North Carolina State University approved user study. There were 5 female and 31 male participants varying in age, race, and marriage status. Approximately 44% of the participants were Asian, 50% were Caucasian, and 6% were of other ethnicities. Participants’ average age was 26.0 (SD=5.4).

7.2.2 Procedure

After filling out a consent form and demographic survey, participants began training sessions by first completing a practice task. The practice task allowed them to become familiar with the keyboard and mouse controls as well as interacting in a 3D virtual environment. Following the practice task, participants were presented a controlled backstory for CRYSTAL ISLAND situating them on the island and providing details about their task. For reference, participants had access to a cast of agents found in the CRYSTAL ISLAND environment as well as an overview map. Participants then interacted with the environment to solve the science mystery. The training testbed provided them with
specific goals to focus on, guiding them through the solution to the mystery. Periodically a “self-report emotion dialog” box would appear requesting input from them regarding their affective state. Participants were asked to select the emotion, from a set of six emotions (excitement, fear, frustration, happiness, relaxation, and sadness), that was most closely related to their own feelings at that particular juncture. This set of emotions was chosen to effectively cover the affect space so that most subjects would easily be able to relate their feelings during interaction to one of the six affective states. In addition to periodic reports, participants had the ability to trigger the self-report emotion dialog if they felt compelled to report a change in their affective state. In practice, this functionality was used sparingly. After solving the science mystery, participants completed a post-experiment survey before exiting the training session.

7.2.3 Results

Both naïve Bayes and decision tree models were induced from data collected in the training sessions described above. Models were evaluated using tenfold cross-validation (Witten and Frank, 2005). Table 7.1 below reports the overall results of naïve Bayes and decision tree affect recognition models. The percentages refer to correctly classified instances. The highest performing model is a decision tree affect recognition model induced from representations of user actions, locations, task structure, and temporal information. Since participants choose from a selection of six affective states chance is 16.7%. An additional baseline to consider is selecting the most common affective state, frustration, which appeared in 34.4% of self-reported affective states.

Table 7.1. Classification results for decision tree and naïve Bayes models with specified datasets.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Physiological Data</th>
<th>Goals, Actions, Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>56.72%</td>
<td>62.94%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>71.34%</td>
<td>95.23%</td>
</tr>
</tbody>
</table>
Table 7.2. Precision and recall analysis for the decision tree affect recognition model induced from non-physiological data.

<table>
<thead>
<tr>
<th>Affective State</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excitement</td>
<td>0.949</td>
<td>0.950</td>
</tr>
<tr>
<td>Fear</td>
<td>0.947</td>
<td>0.948</td>
</tr>
<tr>
<td>Frustration</td>
<td>0.962</td>
<td>0.972</td>
</tr>
<tr>
<td>Happiness</td>
<td>0.965</td>
<td>0.934</td>
</tr>
<tr>
<td>Relaxation</td>
<td>0.961</td>
<td>0.966</td>
</tr>
<tr>
<td>Sadness</td>
<td>0.957</td>
<td>0.937</td>
</tr>
</tbody>
</table>

Table 7.2 drills down to further analyze the performance of the best affect recognition model reporting precision and recall analysis of individual affective states. Precision refers to the percentage of instances correctly classified as particular value of all instances classified as the same value. For instance, more than 96% of the instances recognized as “happiness” were actually instances in which the model predicted an affective state of “happiness.” Recall refers to the percentage of instances correctly classified as a particular value of all instances recorded to be the same value (e.g., the percentage of instances correctly classified as “frustration” that were actually reported as “frustration” in self-reports during training sessions.

7.2.4 Discussion

The results suggest that an approach to affect recognition based on appraisal theory can be effective in task-oriented environments and that representation of user action, location, task structure and temporal information can be used to realize it in a computational model. The performance of decision tree models likely stems from an apparently broad and deep relationship between self-reported affective states and appraisal factors that were explicitly represented in the model. The affect recognition models reported on here seem to be able to capture the relationship between user actions and goals that are assessed during users’ appraisal periods.
The CRYSTAL ISLAND task-oriented environment was specifically designed to feature goals with many possible action sequences to achieve the goal. Particular tasks were designed to be challenging, i.e., certain artifacts were difficult to find in the environment and even when found required problem-solving skills to access the object (e.g., having to climb on boxes to reach a book on the top of a high shelf). A variety of tasks and goals ranging from easy to difficult to complete were presented to each participant. The varying degree of difficulty seemed to elicit a wider range of emotions. Many participants began difficult tasks in positive states (e.g., happy, relaxed) and found themselves feeling frustrated or sad after spending a significant amount of time on particular goal without making progress.

7.2.5 Study Limitations
The approach taken in this study was to cover the two-dimensional affective space (Lang, 1995). However, as research begins to narrow in on a set of emotions relevant to learning it will be necessary to revisit the results presented in this section to determine if such a level of induced model accuracy is achievable with unique sets of emotions. The implications of the results may also limited by the demographics of the participants in this study. The participants were largely male and average 26 years of age. It is unclear if these results would be replicable given more females as well as older and younger participants.

7.3 Frustration Modeling in an Interactive Learning Environment
To create models that make accurate predictions of student frustration as early as possible, we first collect training data by observing students interacting with an intelligent tutoring system. From this training data, we then induce n-gram models to make early predictions of student frustration. Models based on n-grams are useful for early prediction because they are induced from sequences of observations, making predictions with each new observation until they arrive at the final observation of the sequence. In the final observation, concrete evidence of student affect (used as the class label) is obtained. Each prediction from an n-gram model is attempting to determine the affective state of the student recorded in this final observation of the sequence. In many cases, n-gram model predictions will converge on the correct affective state early in a sequence of
observations. The point at which an n-gram model first begins making the correct prediction and then continues to make a correct prediction for the remainder of the sequence is known as the convergence point.

While sequential models such as n-grams allow us to make early predictions, they are not computationally well suited to large multidimensional data. To address this issue, we investigate three non-sequential modeling techniques: naïve Bayes, decision trees, and support vector machines. To enable non-sequential models to make early predictions, we exploit the results of the n-gram models. Utilizing convergence point information, we construct data sets that contain only those observations preceding the convergence point. Thus, models induced from the newly constructed datasets are able to make predictions of student affect long before the observation in which the student’s affective state was recorded.

7.3.1 Early Prediction Methods

n-Gram Models for Early Prediction of Frustration

Given an observation sequence $O_1, O_2, \ldots, O_n$, the objective of affect recognition is to identify the student’s most likely affective state $E^*$ (i.e., frustrated or not frustrated) such that:

$$E^* = \arg \max P(E | O_1, O_2, \ldots, O_n)$$

$$= \arg \max P(E | O_{1:n})$$

where each $O_i$ is an observation encoding the user’s goals, user’s action, the location at which action was performed, and physiological responses such as heart rate and galvanic skin responses. The observation sequence $O_1, O_2, \ldots, O_n$ is denoted by $O_{1:n}$. Applying Bayes rule and the Chain Rule, the equation becomes:

$$E^* = \arg \max P(O_n | O_{1:n-1}, E)P(O_{n-1} | O_{1:n-2}, E) \cdot P(O_{n-2} | O_{1:n-3}, E) \ldots P(O_1 | E)P(E)$$

However, estimating these conditional probabilities is impractical – it would require exponentially large training data sets – so we make a Markov assumption that an observation $O_i$ depends only on the affective state $E$ and a limited window of the preceding observations. We explore two n-gram affect recognition models for detecting student frustration, a unigram model and a bigram model. The unigram model is based on the assumption that, given the affective state $E$, $O_i$ is conditionally
independent of all other observations. Thus, the affect recognition formula for the unigram model can be simplified to:

\[ E^* = \arg \max \ P(E) \prod_{i=1}^{n} P(O_i \mid E) \]

The bigram model is based on the assumption that, given the affective state \( E \) and the preceding observation \( O_{i-1} \), \( O_i \) is conditionally independent of all other observations. Thus, the affect recognition formula for the bigram model can be simplified to:

\[ E^* = \arg \max \ P(E) \prod_{i=1}^{n} P(O_i \mid O_{i-1}, E) \]

The resulting formulae for the unigram and bigram models are very efficient because updating the affect prediction for each new observation only requires computing the product of the probability returned by the previous prediction and the current conditional probability.

During training, we estimate \( P(E) \), \( P(O_i \mid E) \), and \( P(O_i \mid O_{i-1}, E) \) using training data acquired with an interactive learning environment as described below. Because training data is necessarily sparse, i.e., we are unlikely to observe all possible combinations of actions, locations, goals, and physiological response levels, the unigram and bigram models employ a standard smoothing technique (a flattening constant and simple Good-Turing frequency estimation [8] to re-evaluate zero-probability and low-probability n-grams.

**Naïve Bayes, Decision Tree, and SVMs for Early Prediction of Frustration**

We also consider naïve Bayes, decision tree, and support vector machine classifiers for modeling frustration. Chapter 3 provided details regarding these machine learning techniques. We have used the WEKA machine learning toolkit (Witten and Frank, 2005) to analyze naïve Bayes, decision tree, and SVM approaches for generating models of student affect to predict student frustration as early as possible.
7.3.2 Study Method

Participants and Design
There were 5 female and 31 male participants varying in age, race, and marriage status. Approximately 44% of the participants were Asian, 50% were Caucasian, and 6% were of other ethnicities. Participants’ average age was 26.0 (SD=5.4).

7.3.3 Procedure
After filling out a consent form and demographic survey, participants began training sessions by first completing a practice task. The practice task allowed them to become familiar with the keyboard and mouse controls as well as interacting in a 3D virtual environment. Following the practice task, participants were presented a controlled backstory for CRYSTAL ISLAND situating them on the island and providing details about their task. For reference, participants had access to a cast of agents found in the CRYSTAL ISLAND environment as well as an overview map. Participants then interacted with the environment to solve the science mystery. The training test bed provided them with specific goals to focus on, guiding them through the solution to the mystery.

Self-reporting mechanisms are frequently used in studies to obtain information about a subject’s affective state (Conati and Mclaren, 2005; de Vicente and Pain, 2002). In the study reported here, periodically (every 75 seconds) a “self-report emotion dialog” box would appear requesting input from participants about their affective state. They were asked to select the emotion, from a set of six emotions (excitement, fear, frustration, happiness, relaxation, and sadness), that best summarized their own feelings since they were previously asked to assess their emotional state. This set of emotions was chosen to cover the affect space (Lang, 1995) so that most subjects would easily be able to relate their feelings during interaction to one of the six affective states. In addition to periodic reports, participants had the ability to trigger the self-report emotion dialog if they felt compelled to report a change in their affective state. This functionality proved in practice to be used sparingly. After solving the science mystery, participants completed a post-experiment survey before exiting the training session.
7.3.4 Results

Unigram, bigram, naïve Bayes, decision tree, and SVM affect recognition models for detecting student frustration were learned from the collected datasets. The n-gram models were evaluated using the following the criteria (Blaylock and Allen, 2003):

- **Accuracy**: Ratio of correct predictions to the total number of observations.
- **Converged**: Percentage of observation sequences in which the goal recognizer’s final prediction is correct.
- **Convergence Point**: For observation sequences which converged, the point within the sequence when the affect recognizer started making the correct prediction and continued to make the correct prediction for the remainder of the sequence.
- **Average Observations of Converged**: Average number of observations contained in observation sequences in sequences which converged.

The induced n-gram models were tested using the standard k-fold cross validation evaluation methodology (Mitchell, 1997), with k=10. (In each fold, nine segments are used for training and one, which is held out of training, is used for testing.) The results of n-gram affect recognition models are presented in Table 7.3. Figure 7.2 shows a bigram convergence graph depicting the amount of data (actions) required by the model to converge on the correct affective state and the associated probability of that emotion classification. Note that the bigram model utilizing a flattening constant converged after consuming 6.5% of the records leading up to the student self-reported affective state (the class label). In instances where n-gram models converged the models

<table>
<thead>
<tr>
<th></th>
<th>Unigram Flattening Constant</th>
<th>Unigram Good-Turing</th>
<th>Bigram Flattening Constant</th>
<th>Bigram Good-Turing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>68.5%</td>
<td>73.4%</td>
<td>73.6%</td>
<td>73.5%</td>
</tr>
<tr>
<td>Converged</td>
<td>39.7%</td>
<td>67.1%</td>
<td>67.8%</td>
<td>67.2%</td>
</tr>
<tr>
<td>Converged Point</td>
<td>22.6%</td>
<td>7.1%</td>
<td>6.5%</td>
<td>6.9%</td>
</tr>
<tr>
<td>Average Observations of Converged</td>
<td>54.3</td>
<td>51.7</td>
<td>51.8</td>
<td>51.8</td>
</tr>
</tbody>
</table>
were able to correctly classify whether the student was frustrated, on average, 35 seconds prior to the self-report.

The results of the naïve Bayes, decision tree, and support vector machine (SVM) affect recognition models are presented in Table 7.4. Using the results of the n-gram analysis in which bigram models converged after 6.5% of observations contained in a defined window, we
constructed datasets containing only these records for naïve Bayes, decision tree, and SVM model induction. Let us consider the observations \( O_1, O_2, \ldots, O_m \) as the observations leading up to a student self-reported affective state used by n-gram model induction. The data set consisting of observations \( O_1, O_2, \ldots, O_m \), where \( m \) equals \( n \times 0.065 \), leading to each self-report are then used to induce naïve Bayes, decision tree, and SVM affect models. Thus, all induced models are able to predict student affective states (i.e., whether students are frustrated or not) early, i.e., long before we receive confirmation of the student’s affective state from self-reports. These early prediction datasets contain the same data n-gram models consumed up to the convergence point. The constructed dataset allows induced models to make the same early predictions as the n-gram models, approximately 35 seconds prior to the self-reported affective state.

Below ANOVA statistics are presented for results that are statistically significant. Because the tests reported here were performed on discrete data, we report Chi-square test statistics \( (\chi^2) \), including both likelihood ratio Chi-square and the Pearson Chi-square values. To analyze the performance of induced models we first establish a baseline level. Because six affective states were reduced to a two-class predictive classifier (frustrated vs. not frustrated), we consider chance as a baseline measure of performance. If our baseline model were to predict the most frequent classifier (not frustrated, \( n=3859 \)), then the model would correctly predict a student’s frustration state 65\% of the time. Using this model as a baseline, we observe that all induced models outperform the baseline model. The lowest performing induced model, a unigram model using a flattening constant accurately predicted 68.5\% of instances correctly in testing. This performance is statistically significantly better than the baseline (likelihood ratio, \( \chi^2 = 16.075 \), \( p = 6.089 \times 10^{-5} \), and Pearson, \( \chi^2 = 16.067 \), \( p = 6.1 \times 10^{-5} \), \( df = 1 \)). Thus, the performance of all induced models is a statistically significant improvement over the baseline. On the high performing end, the induced decision tree model performed best, accurately predicting 88.8\% of all test instances. This is statistically significant compared to the baseline (likelihood ratio, \( \chi^2 = 980.87 \), \( p = 2.6 \times 10^{-215} \), and Pearson, \( \chi^2 = 943.92 \), \( p = 2.8 \times 10^{-207} \), \( df = 1 \)), and is also statistically significant compared to the next highest performing model, the induced SVM model (likelihood ratio, \( \chi^2 = 105.28 \), \( p = 1.06 \times 10^{-24} \), and Pearson, \( \chi^2 = 104.49 \), \( p = 1.58 \times 10^{-24} \), \( df = 1 \)).
7.3.5 Discussion

The experiment has two important implications for the design of runtime student frustration modeling. First, by monitoring student physiological response, the student’s learning task, and events unfolding in the learning environment, induced models can make early, accurate predictions of forthcoming student frustration. Second, using models that can make early predictions of student frustration creates a significant window of opportunity for the learning environment to take corrective action; early-prediction models offer an improvement over traditional approaches that predict affective states and self-reports on a moment-by-moment basis.

7.3.6 Study Limitations

One study limitation is the design of the problem solving tasks ability to induce particular emotions. Frustration accounted for 35% of reported emotions. It is unclear whether frustration would have such a high frequency of occurrence if the set of emotions was expanded or even tailored to the individual task or participant. This limitation is along the lines of a constant source of criticism of self-reporting, “How does one know the participant is reporting reliably?” Thus, it is uncertain how affect reporting might have been different given a different set of candidate emotions.

7.4 Summary of Affect Modeling Studies

In this chapter the Care framework has been used to induce models of emotion with predictions selecting from six candidate emotions (excitement, fear, frustration, happiness, relaxation, and sadness). The framework has also been used to induce models capable of predicting emotion early, before the emotion report and perhaps before its onset. These early prediction models have been applied to frustration modeling.

The studies reported in this chapter represent a first step towards inductive modeling of affect. The results highlight several important directions of future work. First, as research settles on sets of relevant emotions for learning it will be necessary to conduct new corpus acquisition studies to induce new models of emotion accounting for current research. Second, it is not clear how these models scale. It is unclear how induced models will perform in light of many more affective states,
such as the 22 emotions included in the OCC Model (Ortony et al., 1988), or given emotions more closely related in interpretation, such as happiness and joy.
Chapter 8
Modeling Metacognitive Monitoring

The previous sections reported on experiments focused on the induction and evaluation of affect recognizers and expressers. In this chapter, we turn to a discussion to modeling metacognitive monitoring. Here we utilize the CARE inductive framework to construct models of metacognitive monitoring. We begin this chapter with a discussion of relevant background followed by two studies designed to evaluate the inductive approach to modeling metacognitive monitoring.

8.1 Background

The field of educational psychology has yet to define a universal model of self-regulated learning. However, many agree that self-regulated learning consists of phases including forethought, monitoring and control, and reflection\(^1\) (Pintrich, 2000; Winnie and Hadwin, 1998; Zimmerman, 2000b). Figure 8.1 depicts a high-level overview of the cognitive components of self-regulated learning. Forethought refers to processes that precede acting and often include planning, setting goals and expectations of outcomes, organizing knowledge, adopting goal orientations, and judging efficacy and interest (Pintrich, 2000; Zimmerman, 2000b). Monitoring and control refer to processes occurring during task effort and may include monitoring motivation and affect, monitoring cognition and effort, strategy selection, and focusing of attention (Pintrich, 2000; Zimmerman, 2000b). Lastly, self-reflection refers to processes occurring post effort and after task completion. Self-reflection may include attribution of outcomes, evaluation of self, and evaluation of task (Pintrich, 2000; Zimmerman, 2000b).

\(^1\) These phases are included at varying degrees of granularity and under varying terminology.
Figure 8.1. Self-regulated learning model modified from Winne and Hadwin, 1998; Pintrich, 2000; Zimmerman, 2000b.

Understanding how learners develop and improve skills of self-regulation in computer-based environments is becoming increasingly important. Self-regulated learning (SRL) refers to learning that results from students’ self-generated thoughts and behaviors that are systematically oriented toward the attainment of their learning goals (Schunk and Zimmerman, 2003). Research in SRL has made significant advancements as models have been developed and refined (Butler and Winne, 1995; Greene and Azevedo, 2007; Pintrich, 2000; Winne and Hadwin, 1998; Zimmerman, 2000b). However, there is now a shift in efforts to understand self-regulation not only in traditional learning environments but also in computer-based environments capable of providing intelligent tutoring (Azevedo, 2005; Graesser, McNamara, and VanLehn, 2005; Roll et al., 2007). This poses significant challenges because such environments are very complex and may require additional processing demands from the user (Lajoie and Azevedo, 2006; Schraw, 2007). In order to provide sophisticated models of the development of SRL in computer-based environments (i.e., narrative-centered learning environments) and subsequently build intelligent tutoring systems that support the
development of SRL we must delve into the models and start to examine the relationships between the key variables within the models. Models of SRL, at the broadest levels, are composed of strategic, metacognitive, and motivational components (Zimmerman, 2000b). In this work we begin to examine the function of variables within these components by students working within a rich NLE.

In this chapter we focus our attention on the process of monitoring (Figure 8.2) and their associated metacognitive judgments. Monitoring is a crucial component of self-regulated learning. Self-regulated learners continuously monitor cognition, motivation, and behavior guided by goals and the constraints of the environment (Pintrich, 2000). Metacognitive monitoring skills and the regulation of strategies and tactics are core components within information processing models of self-regulation (Butler and Winne, 1995) and the development of human expertise in general (Glaser and Chi, 1988). More accurate monitoring has been shown to lead to improved self-regulation that, in turn, translates into improved performance (Thiede et al., 2003). Thus, intelligent tutoring systems would be able to further tailor instruction if they were able to effectively and accurately
diagnosis student monitoring. Traditionally this is accomplished through student reports of monitoring judgments. However, this chapter reports on the results of two studies that investigate an inductive approach to modeling metacognitive monitoring judgments in a narrative-centered learning environment, CRYSTAL ISLAND. The results of these experiments indicate that we can build accurate computational models of student monitoring that can operate efficiently at runtime to diagnose student monitoring without the use of interruptive self-reporting. These results will serve as a springboard for future research investigating intelligent technologies for scaffolding self-regulated learning.

8.2 Modeling Metacognitive Monitoring Study

By incorporating learning into narrative-based, virtual environments, investigators hope to tap into students’ innate facilities for crafting and understanding stories. Contextualized learning experiences are known to encourage regulated learning behavior (Perry, 1998). Narrative is also well suited to alternative learning paradigms such as guided discovery and inquiry-based learning. Leveraging stories’ ability to draw audiences into plots and settings, narrative-centered learning environments can introduce novel perceptual, emotional, and motivational experiences, as well as establish connections between narrative content and pedagogical subject matter in young learners (Schraw, 1997). Further, narrative-centered learning environments can effectively support the factors shown to contribute to student levels of motivation (Malone and Lepper, 1987). Such contextual experiences influence student learning and motivation (Linnenbrink and Pintrich, 2001) by providing an environment whereby students are self-directed in their goal-based behaviors and yet orientation toward the goals themselves can be guided by the organization and structure of the learning environment (Schunk et al., 2007). Narrative-centered learning environments allow students a degree of autonomy and control within a structured context that is particularly fertile for the development of self-regulation. Understanding how learners develop and improve skills of self-regulation in computer-based environments is becoming increasingly important. This study takes a pivotal first step to investigate the prospect of using the CARE framework for modeling metacognitive monitoring; a critical component of self-regulated learning.
8.2.1 Method

Participants and Design
The participants included 59 eighth grade students from a highly diverse (e.g., 46% minority; 32% free/reduced lunch) magnet school in Raleigh, North Carolina. Thirty-two were boys and 27 were girls. The students ranged in age from 13 to 15 ($M = 13.73$, $SD = 0.59$). Approximately 56% of the student participants were Caucasian ($n = 33$), 39% were African-American ($n = 23$), and 5% were Hispanic or Latino ($n = 3$). The students participated as part of their science class.

Materials and Apparatus
Materials consisted of the following computer-based instruments:

- **AGQ (Achievement Goals Questionnaire):** This is a 12-item scale that measures achievement goal orientation in the form of four factors: mastery approach, mastery avoidance, performance approach, and performance avoidance (Elliot and McGregor, 2001). The four factors have been shown to have strong reliability and validity through a validation study (Finney et al., 2004). The AGQ was given both as a pretest and posttest.

- **Goals Inventory** (Roedel et al., 1994). This 17-item inventory measures one's tendency to adopt mastery and performance goals and is based upon Dweck and Leggett's (1988) seminal paper on this subject. The items were answered on a five-point Likert scale. Scores for the mastery goals variable could range from 12 to 60 and scores for the performance goals variable could range from five to 25. The Goals Inventory was given both as a pretest and posttest.

- **Gaming Survey.** The Gaming Survey was created for this study and asked questions about experience with video games and computer use. For this study we were particularly interested in two 5-point Likert scale items that included, “Do you play videogames,” that measured the frequency of game play and, “How skilled are you when playing video games,” that measured perceived skill in video game play. The Gaming Survey was given only as a pretest.

- **Situational Interest rating.** At the end of CRYSTAL ISLAND interaction participants were asked to rate their interest in the game (“Did you enjoy this game”) on a 10 point Likert scale with
one being “Not at all true of me” and 10 being “Very true of me.” Participants rated their interest a second time following feedback from the Scoreboard.

8.2.2 Procedure

Before playing, all students were given background information on CRYSTAL ISLAND in addition to a sheet listing the characters of CRYSTAL ISLAND and a map of the island. The character handout includes images of the characters, their names, and their narrative roles (i.e., virus expert, or camp nurse). The map identifies the layout of the virtual environment and the spatial relationship between areas of interest, such as the dining hall, the research lab, and the infirmary. Students completed pretests and then played CRYSTAL ISLAND for 35 minutes. At an interval of 90 seconds during interaction with CRYSTAL ISLAND, participants were asked, “How well are you accomplishing your overall goal for the game?” They responded by entering a number from zero to 100 where zero corresponded to “not confident at all” and 100 corresponded to “very confident.” These monitoring judgments were collected through an in-game dialog box and recorded in the same log tracking student behavior in the CRYSTAL ISLAND environment. Following game play the participants completed the AGQ and Goals Inventory.

8.2.3 Results

Nietfeld et al. (2008) examined what variables impacted performance in CRYSTAL ISLAND. Specifically, they investigated the relationship between several SRL variables of interest (goal orientation, monitoring, and situational interest) and CRYSTAL ISLAND outcome measures (number of actions completed, number of goals completed, number of mystery solution guesses, and score). Monitoring was clearly the most prominent SRL variable, it showed significant correlations with actions completed ($r = .33$), goals completed ($r = .59$), and score ($r = .74$). Monitoring also showed a significant negative correlation ($r = -.45$) with number of guesses indicating that students who were confident that they were attaining their goals tended to see less of a need to risk making a guess without having all of the necessary information. This decision appears to be a metacognitively accurate one in that the payoff was the tendency for higher game scores in the end. Neither the mastery or performance facets from the Goals Inventory nor the mastery approach or performance
approach variables from the AGC showed significant relationships with the CRYSTAL ISLAND outcome variables. A similar lack of relationships held for the situational interest variable. The mastery avoid facet of the AGC did show negative relationships with goals completed ($r = -.26$) and score ($r = -.39$). In sum, these findings reveal the importance and centrality of monitoring with performance.

To accurately model metacognitive monitoring judgments we utilize procedures which have been used to model self-efficacy (McQuiggan and Lester, 2006; McQuiggan, Mott, and Lester, 2008) and affect (Lee et al., 2007; McQuiggan, Lee, and Lester, 2007) using the WEKA machine learning toolkit (Witten and Frank, 2005). Both a naïve Bayes classifier and decision tree model were learned. Naïve Bayes and decision tree classifiers are effective machine learning techniques for generating preliminary predictive models. Naïve Bayes classification approaches produce probability tables that can be incorporated into runtime systems and used to continually update probabilities for assessing student self-efficacy levels. Decision trees provide interpretable rules that support runtime decision making. Runtime systems with decision trees monitor the condition of the attributes in the rules to determine when conditions are met for assigning particular values of student self-efficacy. Both the naïve Bayes and decision tree machine learning classification techniques are useful for preliminary predictive model induction for large multidimensional data. Because it is unclear precisely which runtime variables are likely to be the most predictive, naïve Bayes and decision tree modeling provide useful analyses that can inform more expressive machine learning techniques (e.g., Bayesian networks) that also leverage domain experts’ knowledge.

A tenfold cross-validation analysis was utilized to obtain an estimate of model error. In this method the data is broken into ten equal partitions. In each of the ten iterations (folds), nine partitions are used to construct the model and one partition is held out for testing. Each fold uses a unique partition for testing. Tenfold cross-validation is widely used for obtaining a sufficient estimate of error (Witten and Frank, 2005).

Data consisted of traces of student behavior in the CRYSTAL ISLAND learning environment, including all actions, visited locations, goals accomplished, and character interactions. Values (0-100) of student monitoring judgments served as class labels.

The results (Table 8.1) indicate that the decision tree model is able to accurately predict student metacognitive monitoring judgment values with 93% accuracy. The decision tree model result is
Table 8.1. Modeling metacognitive monitoring results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>20.76%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>14.16%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>93.36%</td>
</tr>
</tbody>
</table>

Statistically significant from the performance of the naïve Bayes model ($\chi^2(1, N = 2752) = 1740.2, p < 0.0001$) and a baseline model ($\chi^2(1, N = 2752) = 1486.2, p < 0.0001$). Here, we define a baseline model which constantly predicts the most frequently occurring monitoring judgment value (e.g., 100). Surprisingly, the baseline model significantly outperforms the naïve Bayes model ($\chi^2(1, N = 2752) = 20.4, p < 0.01$).

8.2.4 Discussion

This study represents a first step in examining SRL variables in a rich and interactive NLE. A major implication of this study is the centrality that monitoring judgments have within SRL (Butler and Winne, 1995). Future work should follow up not only on the monitoring judgments themselves in narrative-centered learning environments but also the level of calibration accuracy of such judgments, which have been shown in other contexts to be significant predictors of performance (Nietfeld et al., 2006).

Further, we find that through observation of student behavior in the CRYSTAL ISLAND learning environment that we can sufficiently predict student monitoring values, utilizing a constructed decision tree model, at runtime. Future work should also examine whether similar approaches can be used to model levels of calibration accuracy in addition to other SRL variables. By combining models of student metacognitive monitoring with other models, such as self-efficacy (McQuiggan and Lester, 2006; McQuiggan, Mott, and Lester, 2008), we may come to better understand student self-regulation in real-time affording adaptive pedagogical strategies tailored to individual students.

Knowledge of how, when, and where students self-regulate via metacognitive and motivational...
monitoring judgments within a NLE may in turn aid teachers in developing instructional approaches that support self-regulated learning in both traditional and technology-oriented classrooms. If we are able to extend models of student metacognitive monitoring to understand student levels of calibration, we believe we would be able to further tailor pedagogical interventions for students who are over-confident and increase efficacy of under-confident students. In fact, pedagogical feedback on problem-solving performance has the greatest impact when students are confident but the solution is wrong (Hattie and Timperley, 2007). On the other hand, similar feedback would be less useful for students with low confidence. Instead, when students are in low-confident situations, pedagogical strategies should include instruction and information (Hattie and Timperley, 2007). Thus, the utility in modeling self-regulatory components, such as metacognitive monitoring, is the ability to better construct the content of pedagogical feedback.

8.2.5 Study Limitations

With regard to limitations, the experiment was designed to control for time on task, allowing 35 minutes for the intervention. As a result of this constraint and the amount of content in CRYSTAL ISLAND, not all participants had finished solving the CRYSTAL ISLAND mystery at the end of the 35 minute session. An alternative design, which will be adopted in future work, will control for task completion rather than time on task.

8.3 Modeling Metacognitive Monitoring Replication

The study described above was replicated with a new population of eighth-grade students at a different middle school in North Carolina. One notable difference is that a different version of Crystal Island was used, version 2.0. Recall from chapter 5 that Crystal Island 2.0 included a completely revised curriculum (developed with a cross-disciplinary team and checked against standards by two eighth-grade science teachers) and story. The result is a longer learning experience exposing the student to a variety of problem-solving activities including a hypothesis-testing problem where the student must rely on the scientific method to complete the problem.
The analyzes reported below aim to understand if the results found in the study described in the earlier section could be replicated in the new version of Crystal Island.

8.3.1 Method

Participants and Design
The participants of the follow up study included 66 eighth grade students from a middle school in Burnsville, North Carolina. Forty-seven were girls and 19 were girls. The students ranged in age from 12 to 15 (\(M = 13.42, SD = 0.63\)). The students participated as part of their science class.

Materials and Apparatus
The Pre-experiment questionnaires collected data on participant demographics.

Materials consisted of the following computer-based instruments:

- **AGQ (Achievement Goals Questionnaire)**: This is a 12-item scale that measures achievement goal orientation in the form of four factors: mastery approach, mastery avoidance, performance approach, and performance avoidance (Elliot and McGregor, 2001). The four factors have been shown to have strong reliability and validity through a validation study (Finney et al., 2004). The AGQ was given both as a pretest and posttest.

- **Goals Inventory** (Roedel et al., 1994). This 17-item inventory measures one's tendency to adopt mastery and performance goals and is based upon Dweck and Leggett's (1988) seminal paper on this subject. The items were answered on a five-point Likert scale. Scores for the mastery goals variable could range from 12 to 60 and scores for the performance goals variable could range from five to 25. The Goals Inventory was given both as a pretest and posttest.

- **Gaming Survey**. The Gaming Survey was created for this study and asked questions about experience with video games and computer use. For this study we were particularly interested in two 5-point Likert scale items that included, “Do you play videogames,” that measured the frequency of game play and, “How skilled are you when playing video games,” that measured perceived skill in video game play. The Gaming Survey was given only as a pretest.
• **Situational Interest rating.** At the end of CRYSTAL ISLAND interaction participants were asked to rate their interest in the game ("Did you enjoy this game") on a 10 point Likert scale with one being "Not at all true of me" and 10 being "Very true of me." Participants rated their interest a second time following feedback from the Scoreboard.

### 8.3.2 Procedure

Participants entered the experiment room with completed informed consent documentation. Before playing, all students were given background information on CRYSTAL ISLAND in addition to a sheet listing the characters of CRYSTAL ISLAND and a map of the island. The character handout includes images of the characters, their names, and their narrative roles (i.e., virus expert, or camp nurse). The map identifies the layout of the virtual environment and the spatial relationship between areas of interest, such as the dining hall, the research lab, and the infirmary. Students completed pretests and then played CRYSTAL ISLAND for 35 minutes. At an interval of 90 seconds during interaction with CRYSTAL ISLAND, participants were asked, “How well are you accomplishing your overall goal for the game?” They responded by entering a number from zero to 100 where zero corresponded to “not confident at all” and 100 corresponded to “very confident.” These monitoring judgments were collected through an in-game dialog box and recorded in the same log tracking student behavior in the CRYSTAL ISLAND environment. Following game play the participants completed the AGQ and Goals Inventory.

### 8.3.3 Results

The results (Table 8.2) indicate that the decision tree model is able to accurately predict student metacognitive monitoring judgment values with 95% accuracy. The decision tree model result is statistically significant from the performance of the naïve Bayes model ($\chi^2(1, N = 3715) = 2379.2, p < 0.0001$) and a baseline model ($\chi^2(1, N = 3715) = 2717.3, p < 0.0001$). Here, we define a baseline model which constantly predicts the most frequently occurring monitoring judgment value (e.g., 100). The baseline model was significantly outperformed by the naïve Bayes model ($\chi^2(1, N = 3715) = 26.6, p < 0.01$).
Table 8.2. Modeling metacognitive monitoring results – replication study.

<table>
<thead>
<tr>
<th>Model</th>
<th>Correctly Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>12.31%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>14.15%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>95.55%</td>
</tr>
</tbody>
</table>

Figure 8.3. Distribution of monitoring values for studies 1 and 2.
8.3.4 Discussion
There was noticeable change in the baseline model for the replicated study. In the first study the baseline model, selecting the most frequently occurring monitoring value, was 100. In the replicated study, the baseline model utilized the monitoring value 50. The study seems to be slightly skewed with nearly 21% of reported monitoring values being 100. 9.9% of reported monitoring values were 100 in the replicated study. A monitoring value of 50 was reported most frequently in the replicated study, 12.3% of self-reports. The difference in distribution (Figure 8.3) of metacognitive monitoring values is interesting considering the first study (skewed towards values of 100) used version 1.0 of Crystal Island, while the replicated study was conducted with version 2.0. Recall that the notable difference between the two versions is the refinement of content, both curriculum and narrative.

Due to the more even distribution of metacognitive monitoring values in the replicated study we see a notable decline in the performance of the baseline model (12.31% accuracy) compared to the baseline model of the skewed distribution of monitoring values collected in the first study (20.76% accuracy). There is relatively little difference between the naïve bayes models of both studies. The decision tree model from the replicated study (95.55% accuracy) is able to more frequently, accurately predict the correct monitoring value compared to the decision tree of the first study (94.36% accuracy) ($\chi^2(1, N = 3570) = 82.9, p < 0.001$). The shift in the distribution if monitoring values can be attributed to the level of challenge and difficulty found in version 2.0 of Crystal Island. This version of Crystal Island requires the application of microbiology domain knowledge and the scientific method, whereas version 1.0 did not require such depth of knowledge to be applied for success.

8.3.5 Study Limitations
With regard to limitations, the experiment was again designed to control for time on task, allowing 35 minutes for the intervention. As in the first study, 35 minutes was not sufficient for students to solve the mystery and complete the content of Crystal Island. The alternative design, of controlling for task completion, as mentioned in the limitations of the first study is worth investigation.
8.4 Summary of Modeling Metacognitive Monitoring

Narrative is receiving increasing attention in the ITS community as a medium for contextualizing learning in meaningful ways while creating rich, engaging experiences for learners. NLEs provide a framework for learning engagement and self-regulation by providing a supportive context for student control and autonomy, necessary for the growth of self-regulated learning. Thus, it will be required by future intelligent narrative-centered learning environments to diagnose, understand, and respond to student self-regulation. One approach to beginning the exploration of developing intelligent SRL detection models is the use data-driven methodologies that can enrich runtime SRL models. The results of this study indicate that an inductive approach leads to models of metacognitive monitoring that are capable of accurate runtime diagnosis.

The results highlight several important directions for future work. First, analysis of student calibration in metacognitive monitoring accuracy should be investigated. This is a necessary extension to further customize pedagogical feedback on student metacognitive monitoring. Second, it is important to understand the misclassifications of induced models and how the misclassifications may inadvertently affect pedagogical strategies. Third, it is necessary to begin to develop and systematically evaluate how to scaffold student learning experiences in light of information obtained through induced SRL models, such as the models of metacognitive monitoring described in this work. Lastly, the inductive approach holds much promise for modeling a variety of SRL components (e.g., efficacy, affect, interest). Future work should investigate the merit of combining data-driven SRL models for scaffolding student learning experiences in narrative-centered learning environments and identifying key features of student behavior useful for diagnosing self-regulated learning.
Chapter 9

Modeling Empathy

We have explored CARE in a controlled studies design to collect corpora for modeling empathy. First, Section 9.1 provides relevant background exploring the psychological basis on which the studies are designed and implemented. Section 9.2 describes a foundational Wizard-of-Oz study utilizing a modified version of CARE, for inducing models of empathy. Section 9.3 describes a study designed to investigate an extension to empathy models accounting for parallel and reactive empathy.

9.1 Empathy Background, Related Work, and Motivations

Empathy is receiving more attention from social psychologists now than ever before (Davis, 1994; Hoffman, 2000). Empathy is not comprehensively addressed in computational models of affect, perhaps because of the recentness with which it has been a focus in social psychology. Although recent developments in computational models of affect have reported success, their limited empathetic abilities suggest a need for enhanced computational models focused on empathy to drive social interactions in synthetic agents.

Devising computational models of empathy contributes to the broader enterprise of modeling affective reasoning (Picard 1997). Beginning with Elliott’s implementation (Elliott, 1992) of the OCC model (Ortony et al., 1988), advances in affective reasoning have accelerated in the past few years, including the appearance of a sophisticated theory of appraisal (Gratch and Marsella, 2004b) based on the Smith and Lazarus appraisal theory (Lazarus, 1991; Smith and Lazarus, 1990). We have also begun to see probabilistic approaches to assessing users’ affective state in educational games (Conati, 2002) and investigations of the role of affect and social factors in pedagogical agents.
Figure 9.1. Empathy construct from Davis, 1994.

(Baylor, 2005; Burleson and Picard, 2004; Elliot et al., 1999; Johnson and Rizzo, 2004; Lester et al., 1999; Porayska-Pomsta and Pain, 2004). Recent work on empathy in synthetic agents has explored their affective responsiveness to biofeedback information and the communicative context (Prendinger et al., 2003). It has also yielded agents that interact with one another and with the user in a virtual learning environment to elicit empathetic behaviors from its users (Paiva et al., 2005).

Empathy is a complex socio-psychological construct (Figure 9.1). Defined as “the cognitive awareness of another person’s internal states, that is, his thoughts, feelings, perceptions, and intentions” (Ickes, 1997), empathy enables us to vicariously respond to another via “psychological processes that make a person have feelings that are more congruent with another’s situation than with his own situation” (Hoffman, 2000).

The social-psychological study of empathy is a relatively recent development (Davis, 1994). Defined as an awareness of another's affective state that generates emotions in the empathizer that reflect more than their own situation, empathy is formalized in a tripartite tableau: an antecedent, an assessment and an empathetic outcome (Davis, 1994). The antecedent captures the affective and situational context of the target (the recipient of empathetic expression) that is then assessed by the empathizer. This assessment yields an empathetic outcome that can be cognitive (e.g., greater awareness of the target's situation) or affective (e.g., flow, frustration, delight, etc.). As noted above, two types of empathetic outcomes can be distinguished: parallel outcomes and reactive outcomes. In parallel outcomes, the empathizer mimics the affective state of the target. For example, the empathizer may become fearful when the target is afraid. In reactive outcomes, empathizers exhibit a higher cognitive awareness of the situation and react with empathetic
behaviors that do not necessarily match those of the target’s affective state. For example, empathizers may become encouraging when the target is frustrated with the problem-solving tasks.

- **In parallel outcomes**, the empathizer mimics the affective state of the target; an individual exhibits an emotion similar to that of the target (Davis, 1994). This is typically based on an understanding of the target’s situation and shows the empathizer’s ability to identify with the target. For example, the empathizer may become fearful when the target is afraid.

- **In reactive outcomes**, empathizers exhibit a higher cognitive awareness of the situation to react with empathetic behaviors that do not necessarily match those of the target’s affective state. The focus is on the target’s affective state, in addition to her situation (Davis, 1994). Reactive empathizers will display emotions that are different from the target’s, often in order to alter or enhance the target’s own affective state. This type of empathy is focused on the target whereas parallel empathy is more self-oriented. As such, reactive empathy can be viewed as a higher level of empathetic behavior. For example, empathizers may become frustrated when the target does not meet with success in her task, even if the target herself may not be frustrated.

Recent years have seen a growing interest in empathetic reasoning in virtual agents. Bickmore (2003) showed how embodied agents can form social relationships with users by employing empathy and thereby improving the users’ motivation. The Empathic Companion (Prendinger and Ishizuka, 2005) tracks a user’s bio-signals (e.g., GSR and heart rate) in order to assess the effect of empathetic interventions within stressful job interview scenarios. It was found that empathetic feedback successfully reduced the user's arousal. Burleson (2006) studied empathetic interventions in frustrating learning environments and explored their effect on meta-affective-strategy learning. In a similar vein, Paiva et al. (2005) studied the requirements for eliciting user empathy and showed that psychological proximity (e.g., gender, shared qualities) is important for generating empathy. While significant advances have been made in modeling empathy, previous work has not addressed the problem of parallel and reactive empathetic reasoning, which will be addressed in Section 9.3.
9.2 Modeling Empathy in an Interactive Environment

This section presents a discussion of a user study, training sessions and the experiment conducted to create, implement, and evaluate CARE models of empathy for companion agents. The purpose of the experiment is to develop models of human-human social interaction by observing situational data changes in the Treasure Hunt virtual environment.

9.2.1 Method

A brief description of participants and the design of their participation are presented, followed by a presentation of materials and apparatus, first for empathizers and second for trainers. Section 9.2.2 discusses the procedures, which is followed by the details of the experiments results. Finally, a discussion of the experiment and evaluation responses concludes the chapter.

Participants and Design

In a formal evaluation, more than two hours of data were gathered from thirty-one subjects in an Institutional Review Board (IRB) of North Carolina State University approved user study. The subjects were divided into 25 targets and 6 empathizers. There were 20 male subjects serving as target trainers and 5 female subjects serving as target trainers varying in race, ethnicity, age and marital status who participated as training targets. There were 3 male and 3 female subjects participating as training empathizers. On average, empathizers completed 4 training sessions, each with a unique training target participant.

Materials and Apparatus – Training Target

For each target trainer pre-experiment paper-and-pencil materials consisted of a demographic survey, Half-Life 2 controls reference sheet, and a controlled backstory in preparation for interacting within the environment. The post-experiment paper-and-pencil materials consisted of a general survey about the training target’s experience and opinions on affect in applications such as games. The demographic survey collected basic information such as gender, age, ethnicity, marital status, and number of children. The Half-Life controls reference sheet described which keys and mouse movements would be needed to manipulate the agent in both the practice task and the training
task. The controlled backstory for the interactive environment was constructed in such a way that each participant would be given the same preparatory information.

The computerized materials for the target trainers consisted of three 3D Treasure Hunt virtual environments, each of varying degrees of difficulty, and the practice task drawn directly from the game Half-Life 2. The easiest version of Treasure Hunt offered many opportunities to find treasures and meet the expectations that were set in the backstory. The most challenging version of Treasure Hunt made it difficult to find treasures; there were fewer treasures worth less value and more occluded treasure boxes making it difficult to meet backstory expectations. The practice task from the game Half-Life 2 was an opportunity for target trainers to become familiar with the required controls. The practice task required completing activities such things as climbing a ladder, stacking boxes, and jumping.

The target training apparatus consisted of a Gateway 7510GX laptop with a 2.4 GHz processor, 1.0 GB of RAM, 15-in. monitor and built-in speakers.

**Materials and Apparatus - Empathizer**

For each empathizer pre-experiment paper-and-pencil materials consisted of a demographic survey, Davis’ Interpersonal Reactivity Index questionnaire, a two-paged background on emotions and empathy, and an empathizer controls reference sheet. Post-experiment paper-and-pencil materials consisted of a survey inquiring about the emotions used/unused, other emotions that could have been useful, and general opinions regarding affect in applications, such as games. The demographic survey collected basic information such as gender, age, ethnicity, marital status, and number of children. Before empathizers began training, they completed Davis’s Interpersonal Reactivity Index (IRI) to gain a measure of their empathy (Davis, 1983).

The IRI consists of 28 statements in which respondents are instructed to rate the degree to which each item describes them on a Likert scale of 0 to 4. The result is a set of 4 subscale values pertaining to the following qualities of empathy: fantasy, perspective taking, empathetic concern and personal distress (Davis, 1994). These empathetic qualities are described below:

- **Fantasy scale.** The fantasy scale refers to tendency one has to immerse themselves into fictional situations.
- **Perspective taking.** Perspective taking measure indicates tendency one has to adapt to the psychological point-of-view of others.

- **Empathetic concern.** Empathetic concern reflects tendencies to have feelings of sympathy and compassion when others experience unfortunate circumstances.

- **Personal Distress.** Person Distress describes the general discomfort and distress one experiences in response another’s distress.

The computerized materials consisted of a spectator view (third person point of view) of the 3D virtual environment, Treasure Hunt, that target trainers would be interacting in. Empathizers did not view target trainer practice tasks and they were not informed of the degree of difficulty.

The empathizer apparatus consisted of a Gateway 7510GX laptop with a 2.4 GHz processor, 1.0 GB of RAM, 15-in. monitor and built-in speakers.

### 9.2.2 Procedure

Each training target participant entered a conference room and was seated in front of a laptop computer. First, target participants completed the demographic survey at their own rate. Concurrently, empathizers entered a second room and were seated in from of another laptop computer. Training targets were unaware of the empathizer’s participation at this point. Empathizers were only aware that a target training participant was in the next room. There was no contact between the participants at any point disabling the empathizers’ ability to distinguish any characteristics of the target trainer other than those assumed from the interaction portrayed on their monitor. Empathizers also first completed the same demographic survey as the targets, also at their own pace. Next, empathizers completed Davis’ IRI questionnaire at their own rate while targets where given the Half-Life 2 controls reference sheet to read until the practice task was loaded on the laptop in front of the target. Once loaded target trainers were able to complete the practice task at their own rate until the task was accomplished. At this point empathizers were given the emotion and empathy reference sheet and instructed to read over the definitions and empathizer controls. Next, one of the degrees of difficulty was randomly selected and that Treasure Hunt training environment was loaded on the target machine while the spectator view application was concurrently loaded on the empathizer machine.
Figure 9.2. Evaluation Data Flow.
Once the training environment was loaded target trainers had 7 minutes to explore the environment and collect treasure. Empathizers viewed the interaction and made empathetic behavior decisions by selecting the appropriate control for the affective state they desired the companion agent to have. When empathetic behaviors were selected by the empathizer, both participants had the opportunity to hear the companion agent’s spoken language and see the associated gestural behaviors and posture. Upon completion of the 7 minute training session, both training targets and empathizers were given post-session surveys and were interviewed. Finally, target trainers were offered information about the details of the experiment and informed about the presence of the empathizer during the training session.

The following procedural steps were used to generate models of empathy from the training sessions (Figure 9.2 presents the evaluation data flow):

- **Data Construction.** Each session log, containing 6,000 – 9,000 observation changes, was first translated into a full observational attribute vector. For example, if a treasure box came into view (and all other observable attributes remained constant) then the observational attribute vector would modify the previous vector to account for the noted change.

- **Data cleansing.** After data was converted into the observational attribute vector format the data was ready to be cleaned. This step included generating the dataset containing only records in which the empathizer selected an empathetic emotion.

- **Naïve Bayes classifier and Decision Tree analysis.** Once the dataset was ready it was loaded into the Weka machine learning package (Witten and Frank, 2005), a naïve Bayes classifier and decision tree were learned, and tenfold cross-validation analyses were run on the resulting models. The entire dataset was used to generate models for empathetic assessment (when to be empathetic) and empathetic interpretation (how to be empathetic). Empathetic assessment is determined using the entire dataset, while empathetic interpretation is determined from a transformed dataset containing only empathetic records.

The following section presents the results of the naïve Bayes and decision tree classification models and describes statistical analyses of the training sessions.
9.2.3 Results

The following sections report results from the training sessions and on the performance of constructed models of empathetic assessment and empathetic interpretation.

**IRI Empathy Instrument Results**

Female empathizers scored higher than male empathizers on the pre-experiment Davis Interpersonal Reactivity Index, in each quality except for perspective taking. Males averaged one-half point higher than the female empathizers for perspective taking. Subjects were found to be representative of the general population in empathetic characteristics (Davis, 1983). Figure 9.3 reports the IRI results for each empathizer. IRI results are also reported by subscale (Figure 9.4), and by subscale and gender (Figure 9.5). Empathizer 2 was deemed the most likely to be empathetic based on Davis’ index. This empathizer chose to be empathetic 93 times over 4 training sessions, significantly more than the 72 empathetic selections of Empathizer 6 over 4 training sessions. Furthermore, Empathizer 2 chose to be “excited” 55 of 93 empathetic selections. In all, “excited” was chosen 119 times across the 25 training sessions, one-half stemming from Empathizer 2. All
emotion frequencies are reported in Figure 9.6. Figures 9.7 and 9.8 examine affect frequencies by gender, and average emotion frequencies per training session by gender. From the figures on the following pages it is evident that female empathizers were more likely to be empathetic and more likely to use the extreme emotions: excited, representative of positive valence and high arousal, and sad, representative of negative valence and low arousal. Male empathizers were less likely to venture too far from the origin of the two-dimensional affective space, using the emotions fear, frustrated, relaxed and joyful much more than excited and sad.

**Model Results**
Models were induced from data collected in the training sessions described above. As noted earlier, 192 observational attributes were used to define the feature vectors. Figure 9.9 shows the ROC curves for Care’s naïve Bayes and decision tree approaches for modeling empathetic assessment. Figures 9.10 shows a ROC curve for Care’s naïve Bayes and decision tree approaches for modeling empathetic interpretation. Associated areas under the curve can be found in the Table 9.1.
Figure 9.4. Individual Empathizer IRI Subscale Results.

Figure 9.5. Average Empathizer IRI Subscale Results by Gender.
Figure 9.6. Affective State Frequencies from 25 Training Sessions.

Figure 9.7. Affective State Frequencies by Gender.
**Figure 9.8.** Average Affective State Frequencies by Gender.

**Figure 9.9.** ROC Curves for Empathetic Assessment. The ROC curves for each model predicting assessed empathetic behavior triggers in a ten second interval. The area under the Naïve Bayes curve is 0.72 and the area under the Decision Tree curve is 0.89
Figure 9.10. ROC Curves for Empathetic Interpretation (Emotions).
Table 9.1. Areas under ROC curves in Figure 9.10.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Area under naïve Bayes ROC curve</th>
<th>Area under decision tree ROC curve</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Fearful</td>
<td>0.74</td>
<td>0.66</td>
</tr>
<tr>
<td>(b) Excited</td>
<td>0.80</td>
<td>0.56</td>
</tr>
<tr>
<td>(c) Frustrated</td>
<td>0.78</td>
<td>0.56</td>
</tr>
<tr>
<td>(d) Joyful</td>
<td>0.69</td>
<td>0.56</td>
</tr>
<tr>
<td>(e) Sad</td>
<td>0.69</td>
<td>0.50</td>
</tr>
<tr>
<td>(f) Relaxed</td>
<td>0.57</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Cross-validated ROC curves are useful for presenting the performance of classification algorithms for two reasons. First, the curve represents the positive classifications (true positives), included in a sample, as a percentage of the total number of positives along the vertical axis, against the negative classifications (false positives) as a percentage of the total number of negatives (Witten and Frank, 2005). Second, the area under the ROC curve has widely been accepted as a generalization of the measure of the probability of correctly classifying an instance (Hanley and McNeil, 1982).

Two categories of functionality can be distinguished. First, the decision tree classifier was best suited for modeling empathy assessment, i.e., it was better able to determine when to be empathetic (Figure 9.9). The depth of the decision tree constructed for empathetic assessment surpasses 1,000-levels. However, many of the decision trees constructed for empathetic interpretation were much shallower. For instance, empathizers tended to select the affective state fearful most often when the target trainer appeared in the dark hallway of the Treasure Hunt virtual environment. These locational features coupled with several other observational attributes were used to construct shallower decision trees compared to the decision tree for determining when to be empathetic.

Second, the naïve Bayes classifier was best suited to modeling empathy interpretation, i.e., it was better able to determine how to be empathetic. The smoothness of the curve in Figure 9.9 indicates that sufficient data seems to have been used for training empathy assessment, while the jaggedness of the curve in the empathetic interpretation ROC graphs (Figure 9.10) indicates that more data covering a larger space of situations is called for in training empathy interpretation.
Many empathizers only rarely used particular emotions, e.g., sad, and some trainers suggested that having more affective states available would have been helpful. In general, however, it appears that effective classifiers can indeed be learned for both empathy assessment and empathy interpretation.

All six emotions were evaluated and the naïve Bayes classifier bested the decision tree classifier in every case (Figures 9.10a through 9.10f). Areas under the curve can be found in Table 9.1.

Of course, there are other ways to classify empathetic interpretation other than into individual the emotions. One such classification is to determine the valence (positive or negative) of the interpretation in conjunction with a separate model determining the arousal (High, Medium, or Low) (Results are reported in McQuiggan, 2005). The naïve Bayes classifier also outperformed the decision tree for both determining valence and arousal, also likely due to the amount of data available. Yet another classification of emotions that can be used to determine how to be empathetic is by determining the quadrant of the two-dimensional affective space (Lang, 1995) the affective state should come from (see McQuiggan, 2005 for further detail).

9.2.4 Discussion
The results clearly demonstrate that a decision tree classification approach is sufficient for modeling empathetic assessment and that significantly more training is needed to produce large quantities of empathetic instances for the same approach to have such compelling results for empathetic interpretation. Although, naïve Bayes makes the assumption that all attributes of the observational attribute vector are independent – this assumption is false – it nonetheless induces a model sufficient for controlling empathetic interpretation in companion agents.

Only 388 instances were available for modeling empathetic interpretation. Collecting more data would likely improve the predictability of the decision tree classifier for interpreting how to be empathetic. We speculate that for this reason, the decision tree classifier was outperformed by the naïve Bayes classifier for modeling empathetic interpretation. Although more data would likely improve decision tree classification it is unclear whether more data would increase the true positive rate of the decision tree model such that it would surpass that of the naïve Bayes classifier.
The empathetic assessment models had over 10,000 instances from which to learn from; a significantly larger dataset. There was a sufficient amount of collected data to generate naïve Bayes and decision tree models of empathetic assessment while the transformed dataset of empathetic records seems to have been insufficient for decision tree models of empathetic interpretation.

Part of the post-experiment experience for target trainers was to answer several oral questions posed by the researcher. One such question concerned the target trainer’s expectations of the companion agent’s role contrasted with the role the companion agent portrayed in actual training. An overwhelming majority of target trainers expected the companion to play the role of a guide, or at least be able to make constructive comments concerning exploration by providing some form of assistance or advice. Some target trainers expected the companion agent to explore the world on her own, dividing the task between the user and the companion agent. Most target trainers described the companion agent has providing comic relief based on the current situation and still others commented that the companion agent provided emotional commentary throughout the interaction.

In post-interviews with empathizers it was discovered that up to an additional set of 4 emotions, for a set of set of 10 emotions, would have been preferred. Only one empathizer responded that they would not have preferred any more empathetic emotions be available. Table 9.2 lists the additional empathetic affective states that empathizers expressed interest in having at their disposal. The number in parentheses represents the number of empathizers suggesting the emotion as an additional empathetic affective state choice.

**Table 9.2.** Empathizer suggested emotions.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Number of Suggestors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angry (2)</td>
<td>Curious (2)</td>
</tr>
<tr>
<td>Annoyed (2)</td>
<td>Encouraging (1)</td>
</tr>
<tr>
<td>Bored (1)</td>
<td>Relieved (1)</td>
</tr>
<tr>
<td>Confused (3)</td>
<td>Worried (1)</td>
</tr>
<tr>
<td>Congratulatory (1)</td>
<td></td>
</tr>
</tbody>
</table>
Three empathizers expressed, in post-experiment interviews, that they wished they could have experienced the target trainers’ response. Empathizers expressed that from solely watching the interaction, assessing only situational context, it was difficult to judge the impact of their empathetic selections.

One empathizer expressed an interesting view in the process they used to determine which empathetic emotion to select. They described their thinking as follows, “I tried to think of what the user’s (target trainer’s) character was feeling and how my character (the companion agent) felt and should respond.” This perspective is an interesting one because it points to immersion of not only the empathizer, but the target trainer as well, and that empathetic assessment and interpretation were embedded completely within the virtual environment. Another direction for future work is to investigate the sense of immersion created when a user, in this case the empathizer, is thinking empathetically. It seems that eliciting users to be empathetic may draw them deeper into the virtual environment.

9.2.5 Study Limitations

The study utilized six “wizards” playing the role of empathizers, as the companion agent, while interacting with many subjects acting as “targets”. Future study designs might consider increasing the number of empathizers. The empathy models are based on the empathetic assessment and interpretation of a small number of empathizers which may not be representative of a larger population. Another limitation of this study is the lack of evaluation of the impact of empathetic behaviors. It is unclear if the empathetic behaviors were motivating, situationally appropriate, or distractive to the targets. This limitation is addressed in the study described in Section 9.3 as well as in studies detailed in Chapter 10.

9.3 Modeling Parallel and Reactive Empathy

This work investigates extending models of empathy to more closely reflect the current socio-psychological understandings of empathy. As mentioned in Section 9.1 that there are two forms of empathy. Parallel empathy refers to empathizers that express the same affective state as the
target. *Reactive empathy* considers slightly more cognitive awareness, where the empathizer accounts for the full antecedent of empathy (target emotion, target situation, and his own affective state) before arriving at an emotional response they deem to be more appropriate for the situation than his own emotion.

**9.3.1 Method**

*Participants and Design*

To empirically investigate “empathy in action,” a study was conducted with subjects interacting with virtual agents. The subjects of the study consisted of 35 college students ranging in age from 21 to 60 (*M* = 24.4, *SD* = 6.41) including 9 females and 26 males. Among these students, 60% were Asian (*n* = 21), approximately 37% were Caucasian (*n* = 13) and one participant chose not to respond.

**9.3.2 Procedure**

Participants entered the experiment room where they completed informed consent documentation. They were randomly assigned to either the control condition or the empathy condition and were seated in front of a laptop computer. They were then given an overview of the experiment agenda, and they completed the pre-experiment questionnaires including the demographics survey, the interpersonal reactivity index survey (Davis, 1983), and the achievement goals questionnaire (Elliot and McGregor, 2001).

The interpersonal reactivity index (Davis, 1983) includes 28 items that measure subjects’ empathetic nature by asking them to rate the degree to which each statement describes them. These items are assessed on a 5-point Likert scale (0 - does not describe me well to 4 - describes me very well). The IRI is divided into four subscale measurements quantifying the following components of empathy: fantasy, perspective taking, empathetic concern, and personal distress (Davis, 1983). The achievement goals questionnaire (Elliot and McGregor, 2001) measures four achievement goal constructs (mastery-approach, performance-approach, mastery-avoidance, and performance-avoidance goals). Subjects indicate the extent to which each statement is true of them on a 7-point Likert scale (1 - not at all true of me to 7 - very true of me).
Upon completing the pre-experiment questionnaires, participants were instructed to review CRYSTAL ISLAND instruction materials. These materials consisted of the backstory and task description, the character overviews, the map of the island, the control sheet, and definition sheet of the self-report emotions. Participants were then further briefed on the controls via a presentation summarizing the task and explaining each control in detail.

Participants were given 35 minutes to solve the mystery. Solving the mystery consisted of completing 15 goals including learning about various diseases, compiling the symptoms of the sickened researchers, testing a variety of possible sources, and reporting the solution (cause and source) back to the camp nurse.

Six characters (Audrey, Elise, Jin, Quentin, Robert, and Teresa), each play distinct roles in the CRYSTAL ISLAND environment. When subjects decided to interact with the characters, the following schema was used to direct subject-character interactions and virtual character empathetic responses (the schema is also depicted in Figure 9.11):

1. The virtual character queries the subject for a self-reported affective state. The subject is presented with a dialog box asking the question, “Hi Alex, how are you feeling?” The subject may respond by selecting one of the 10 available emotions (anger, anxiety, boredom, confusion, delight, excitement, fear, flow, frustration, and sadness).

2. The virtual character then responds to the subject’s reported affective state with an empathetic response. The empathetic response is determined from the subject-reported emotion and the character’s empathizer type, i.e., whether the character is a reactive empathizer or a parallel empathizer. Empathetic responses are short, consisting of 1 to 2 sentences. Parallel responses consist of the character expressing the same emotion as the user through text responses (i.e., I feel frustrated by this as well); alternatively, reactive responses demonstrate advanced cognitive processing on the character’s part by providing responses designed to be more motivating, revealing the character’s desire for the user to be in a positive emotional state (i.e., I really feel that you can get it!).

3. A follow-up dialog box is then presented to the subject asking her to respond with the prompt, “...and you respond”. The subject is able to choose from 4 Likert-scaled responses designed to evaluate the appropriateness and effectiveness of the virtual character’s
empathetic response. Subjects can issue responses ranging from (1) “That does not help me at all.” to (4) “Thanks, I feel a lot better!”

4. The virtual character responds with a one-word quip (e.g., “Thanks”, or “Great!”) directed towards the subject’s evaluation response (Step 3). In addition, the virtual character provides narrative and problem-solving information.

5. The virtual character then asks the subject how she feels one final time before concluding the interaction. The subject is presented a dialog box similar to the one described in Step 1 without the character greeting. Here, the character prompts the subject with, “How are you feeling now?”

6. Finally, the virtual character again empathetically responds to subject-reported affective states in the same manner as described in Step 2.

Immediately after solving the science mystery of CRYSTAL ISLAND (or after 35 minutes of elapsed interaction time for subjects who had not solved the mystery), subjects completed the post-experiment questionnaire. This researcher-designed questionnaire assessed perceptions of individual CRYSTAL ISLAND characters. The results of this instrument are outside the scope of this discussion.
Figure 9.11. Empathetic character schema.
9.3.3 Results

The In Step 3 of the user-agent interaction schema presented above, subjects evaluated virtual character responses as part of the “conversation” with the character. The distribution of empathetic responses (parallel and reactive) and the associated ratings are detailed in Table 9.3. The first two rows show the number of empathetic responses that were found to be appropriate by subjects (i.e., the instances of parallel and reactive empathetic behaviors that were given an evaluative rating of 3 or 4), while the next two rows indicate empathetic responses that were found to be inappropriate by subjects (i.e., the instances of parallel and reactive empathetic behaviors that were given an evaluative rating of 1 or 2).

<table>
<thead>
<tr>
<th>Evaluative Rating</th>
<th>Empathetic Responses</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parallel</td>
<td>Reactive</td>
</tr>
<tr>
<td>4</td>
<td>1200</td>
<td>1305</td>
</tr>
<tr>
<td>3</td>
<td>1663</td>
<td>1600</td>
</tr>
<tr>
<td>2</td>
<td>1558</td>
<td>1498</td>
</tr>
<tr>
<td>1</td>
<td>1033</td>
<td>1033</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>5454</td>
<td>5436</td>
</tr>
</tbody>
</table>

Naïve Bayes, decision tree, and support vector machine (SVM) models were induced from data collected in the training sessions described above using the WEKA machine learning toolkit (Witten and Frank, 2005). All models were constructed using a tenfold cross-validation scheme for producing training and testing datasets. In this scheme, data is decomposed into ten equal partitions, nine of which are used for training and one used for testing. The equal parts are swapped between training and testing sets until each partition has been used for both training and
testing. Tenfold cross-validation is widely used for obtaining an acceptable estimate of error (Witten and Frank, 2005).

Two distinct datasets were used. The first dataset was comprised only of empathetic responses receiving high ratings of 4 \((n = 2505)\) from subjects during conversations with virtual characters. The

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Induced Empathy Models</th>
<th>Highest-Rated</th>
<th>Favorably-Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (Reactive)</td>
<td>0.52</td>
<td>0.50</td>
<td></td>
</tr>
<tr>
<td>Situation Attributes Only</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.73</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.75</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.75</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Situation + Affect</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.74</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.87</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.83</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>Situation + Affect + Bio</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.75</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.95</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.84</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Situation + Affect + Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.74</td>
<td>0.70</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.98</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.97</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Situation + Affect + Bio + Characteristics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>0.75</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.98</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>0.97</td>
<td>0.92</td>
<td></td>
</tr>
</tbody>
</table>
second dataset was comprised of empathetic responses receiving a rating of either a 3 or 4 \( (n = 5768) \). Within each dataset, several versions of models were learned from the various types of observational attributes (situation data, user affect reports, user characteristics, and physiological response data). Table 9.4 provides results of all induced models and baselines.

A baseline measure determines the most frequent class label (in this case reactive empathy) and predicts all empathetic responses to call for reactive responses. For the dataset containing only empathetic responses rated level 4 (Highest-Rated), reactive empathy accounted for 52% of the instances. Reactive empathy occurred in 50% of the instances in the dataset containing empathetic responses rated as level 3 or 4 by subjects (Favorably-Rated).

All induced models outperformed baseline models. The improvement of induced models over baselines is statistically significant. For instance, the least accurate induced model from the Highest-Rated dataset is the naïve Bayes model (73% accuracy), which was learned from situation attributes only. The accuracy of this naïve Bayes model was statistically significant from baseline accuracy, \( \chi^2(1, N = 2505) = 238.94, p < 0.001 \). Also, in the Highest-Rated dataset, decision tree performance was statistically significant from naïve Bayes performance, except for the models induced from situation attributes only. For instance, in the models learned from situation attributes and user affective states, decision tree accuracy (87%) was statistically significant from naïve Bayes accuracy (74%), \( \chi^2(1, N = 2505) = 136.42, p < 0.001 \).

Similar results appear in models learned from the Favorably-Rated dataset. All induced models again outperformed the baseline (reactive empathy) in which 50% of instances were reactive responses. The worst performing induced model’s accuracy (64%) was the naïve Bayes model learned from situation attributes only. This model was statistically significant from baseline accuracy, \( \chi^2(1, N = 3263) = 131.04, p < 0.001 \). Decision tree model accuracies were statistically significant from their corresponding naïve Bayes models for each category of data in the Favorably-Rated dataset. For instance, the accuracy of the decision tree model learned from situation data only (70%) was statistically significant from the naïve Bayes model (64%) induced from the same data, \( \chi^2(1, N = 2505) = 26.65, p < 0.001 \).
9.3.4 Discussion

The study found that models of empathy induced from knowledge of the user’s situation and the user’s affective state can effectively determine which type of empathy is most appropriate for interactions requiring empathetic expression. The results suggest that designers of empathetic virtual agents should consider both parallel and reactive empathy in virtual agent architectures. The study has the following implications, each of which is discussed below:

- Empathetic response modeling should be integrated with other virtual agent response functionalities.
- Individual differences of users should be accounted for when determining empathetic responses.
- Empathy models induced from “good” examples (examples rated highly by subjects) should improve the quality of interaction with virtual agents.

Approximately one-half of the empathetic responses appearing in the user study received high marks from subjects (ratings of 3 or 4) leaving the other half as “bad” examples of empathetic responses. While we have not fully explored which attributes of the interaction correspond to low ratings, it seems reasonable to conclude that in instances where a parallel response received a low rating, it would not be improved by replacing the response with an instance of reactive empathy. To deal with this, virtual agent architectures should combine empathy models with other response strategies. For example, cases in which subjects reported feelings of frustration or confusion and subsequently rated empathetic responses poorly might be better addressed not by emotional responses, but by directed content feedback, e.g., by addressing the obstacle that is the source of user frustration or providing hints that may relieve user confusion. Certainly, emotional empathetic responses are not always appropriate and should be combined with a broad range of agent response strategies.

The models of empathy described in the previous section account not only for knowledge of user situation and user emotion, but also other demographic information such as gender, age, user empathetic nature, and goal orientation. This information is typically accounted for subconsciously in human-human interaction but often discounted in human-agent interaction. Agents that
understand who users are, male or female, young or old, mastery- or performance-oriented, may be able to more effectively determine how to empathetically respond to user emotional situations. Finally, models of empathy that account for both parallel and reactive empathy should lead to more effective human-agent interactions. Since subjects rated each empathetic response on a Likert scale, models were able to be induced solely from “good” examples, instances receiving ratings of 3 or 4. While future investigations will consider the effects of empathy models on learning and task performance, the addition more flexible empathetic abilities yields an immediate improvement in an agents’ abilities to respond appropriately to social situations.

9.3.5 Study Limitations
The results of the study are affected by the virtual characters that interacted empathetically with participants. First, it is possible that the gender, narrative role, and pedagogical role of the empathetic characters may not generalize to other characters and across domains. Further investigation is required to assess the effect of character persona on perception of character empathy. Second, the population participating in this study is a small group of college students studying computer science. Different empathy models may be appropriate for different demographic segments. To determine the generalizability of empathy models, additional user studies are required. The procedure used in the study utilized only empathetic responses. It may be the case, particularly for instances of empathetic behavior that were not favorably evaluated, that empathy was inappropriate and that other types of responses, such as those providing pedagogical assistance, or other content would be more effective.

9.4 Summary of Empathy Modeling
Recent advances in affective reasoning demonstrate the important role that emotion plays in cognitive accounts of social interaction and suggest that it should therefore play an equally important role in virtual agents. Because empathy is a natural extension of the appraisal process and appears continuously in human-human interaction, it is important to endow virtual agents with the capability to respond empathetically. Empathy modeling requires accurately assessing a social
situation context in order to determine (1) if an empathetic reaction is warranted, and (2) if so, what sort of empathetic behavior should be performed, including whether parallel or reactive empathetic expression would be most appropriate for the user, her situation and her affective state\(^1\).

The inductive framework of CARE has been used in this Chapter to learn empathy models that account for both parallel and reactive empathetic expression and determine when to deliver the expressions. The data-driven approach centers on the observation of “empathy in action” and acknowledges the psychological understanding of empathetic assessment and appraisal processes by including appropriate information in model induction. Such data include situational contexts (e.g., user actions, visited locations, goals), user affect and affective responses (as measured through physiological changes), and user demographics. While previous work has either focused solely on parallel empathy or not distinguished between the two forms, the inductive approach proposed here was evaluated in a user study designed to examine the effects of parallel and reactive empathy upon the recipient. By allowing virtual agents to employ parallel and reactive empathy that is appropriate for the social situation, it is hoped they will become more effective and engaging.

---

\(^1\) The effect of parallel and reactive empathetic expression on user affect is further explored in Chapter 10.
Chapter 10

Empirical Studies of Affect-based Interventions

The previous chapters have reported on experiments focused on the induction and evaluation of affect recognizers and expressers, models of self-efficacy and models of metacognitive monitoring. In this chapter, we turn to a discussion of an examination of users’ affective experiences when interacting with narrative-centered learning environments and the virtual characters that inhabit them. This corresponds to a student-focused shift in research that is concerned with understanding the effects and impact the work described in earlier chapters has on students. Here, in Section 10.1, we review the background of psychological constructs of interest to the research reported in this chapter and introduce a new psychological construct, presence. We then begin to examine student affective experience with the CRYSTAL ISLAND learning environment and the pedagogical characters inhabiting the island. Much of this evaluation seeks to assess the merit of employing affect-sensitive modules in interactive learning environments. Through evaluation of factors engaging students and the impact of empathetic virtual characters we strive to motivate merit of employing induced models of affect (chapters 6-9) and to further understand the affective experience of students interacting in narrative-centered learning environments.

10.1 Background, Related Work, and Motivation

We have evaluated the accuracy of models or the ability of models to correctly predict user affect in the preceding chapters. In this chapter, we investigate the merit of employing affect-sensitive
interactive learning environments examining student engagement, as measured through student reports of presence, and observing student affective experience in narrative-centered learning episodes with empathetic characters. Here we provide the background relevant to presence and reiterate the psychological foundations of empathy.

10.1.1 Presence

Presence relates to quality of experience (Gaggioli et al., 2003). Although there has been substantial debate on formal definitions, there is a general consensus that presence describes a user’s sense of “being there” when interacting with a mediated environment (Schubert et al., 1999; Ijsselsteijn et al., 2000; Insko, 2003). Presence has been alternatively defined as “the perceptual illusion of nonmediation” (Lombard and Ditton, 1997), as well as “the subjective experience of being in one place or environment, even when one is physically situated in another” (Witmer and Singer, 1998). It is distinguished from related concepts such as immersion and involvement. Immersion refers to the extent and nature of technology‐provided sensory stimuli; it is often associated with the pervasiveness and fidelity of visual, audial, olfactory, and tactile inputs (Schubert et al., 1999). Involvement refers to the degree of attention and meaning devoted to some set of stimuli (Witmer and Singer, 1998). Numerous additional conceptualizations of presence have also been discussed; they can be divided into two types: physical presence, the sense of being physically located in a mediated space, and social presence, the sense of co‐location and social interaction with a virtual or remote partner (Lombard and Ditton, 1997; Ijsselsteijn et al., 2003).

The virtual reality and human factors communities use the notion of presence to describe users’ perception of transportation into a virtual environment (Garau et al., 2005; Ijsselsteijn et al., 2000; Meehan et al., 2002; Schubert et al., 1999; Witmer and Singer, 1998). This definition also points to presence as a useful construct for describing the types of experiences audiences report when feeling transported into a story. Within the presence community, there has been a large body of work examining the effects of technological and fidelity‐related factors on presence, but comparatively little work examining the influence of content. By examining interactive, narrative‐centered learning environments, a broad range of opportunities arises for investigating the effects of narrative content on student perceptions of presence. Strong narrative is often heavily character‐
driven (Egri, 1946), so by examining educational, story-related interactions with virtual characters, we can begin to understand the variables inherent in story, their impact on presence, and their relationship to learning experiences. Shifting focus from interface issues to content has long been advocated within the presence community. By exploring narrative content’s ability to affect presence, we focus on the meaning of an experience, the “deep structure for the virtual world,” the “cinema” rather than the “celuloid” (Bates, 1992).

Presence contributes to the goal of transparency in technology-mediated interactions Norman, 1998). Although there has been substantial debate on formal definitions, there is a general consensus that presence describes a user’s sense of “being there” when interacting with a mediated environment (Insko, 2003; Johnsen et al., 2006; Schubert et al., 1999). Presence has been alternatively defined as “the perceptual illusion of nonmediation” (Lombard and Ditton, 1997), as well as “the subjective experience of being in one place or environment, even when one is physically situated in another” (Witmer and Singer, 1998). It is distinguished from related concepts such as immersion and involvement. Immersion generally refers to the extent and nature of technology-provided sensory stimuli; it is often associated with the pervasiveness and fidelity of visual, audile, olfactory, and tactile inputs (Schubert et al., 1999). Involvement refers to the degree of attention and meaning devoted to some set of stimuli (Witmer and Singer, 1998).

Several groups of factors have been proposed as contributors to presence, including the extent and fidelity of sensory information; the mapping between actions and effects; content factors including characters, objects, and events; and user characteristics including perceptual and cognitive abilities (IJsselsteijn et al., 2000). Lee and Nass recently observed that social factors, such as the relationships between synthesized voice personality, user personality, and presented content can also have significant effects on users’ perceptions of social presence (Lee and Nass, 2003). Partially in response to these factor analyses, several separate conceptualizations of presence have been proposed. These are traditionally divided into two types: physical presence, the sense of being physically located in a mediated space, and social presence, the sense of co-location and social interaction with a virtual or remote partner (Lombard and Ditton, 1997). In this work, we do not distinguish between physical and social presence. We are interested in how the content of the experience, varied through the empathetic behavior of various characters, influences overall...
perceptions of presence in a narrative learning interaction. This relates to Lee and Nass’ work in investigating social phenomena’s relationship with presence using relatively low levels of immersion (2003). However, our current investigation does not focus on student perceptions of individual characters; rather, we seek to investigate easily manipulable content factors that affect student perceptions of presence within the narrative.

Presence is of particular concern because it can play a variety of important roles for supporting education in narrative-centered learning environments (Rowe, McQuiggan, and Lester, 2007). Pedagogical approaches that foster experiential learning, where students actively engage and learn from their experiences (Kolb et al., 2000), can benefit from enhanced levels of presence; students focus their cognitive resources on the experience rather than the interface or other peripheral features. Waterworth and Waterworth (2000) emphasize the use of presence and absence in virtual learning environments, thereby interweaving perceptual and conceptual learning, respectively. This distinction provides insight into the types of cognitive states desirable for concrete and abstract learning tasks. Presence also relates to positive affective states for learning, such as flow (Csikszentmihalyi, 1990). Although the relationship between presence and flow has not been clearly defined, the states share a number of common characteristics, including highly focused attention and lost sense of time. These connections suggest that presence serves an important role in understanding the characteristics of effective narrative-centered learning experiences. In this work we seek to investigate content-based features (i.e., the role of narrative and empathetic characters) hypothesized to contribute to presence, thereby enabling a deeper understanding of interactions with narrative-centered learning environments.

10.1.2 Empathy

Recall that the social-psychological study of empathy is a relatively recent development (Davis, 1994; Hoffman, 2000; Ickes 1997). Defined as an awareness of another’s affective state that generates emotions in the empathizer that reflect more than their own situation, empathy is formalized in a tripartite tableau: an antecedent, an assessment and an empathetic outcome (Davis, 1994). The antecedent captures the affective and situational context of the target (the recipient of empathetic expression) that is then assessed by the empathizer. This assessment yields an
empathetic *outcome* that can be cognitive (e.g., greater awareness of the target’s situation) or affective (e.g., flow, frustration, delight, etc.). As noted above, two types of empathetic outcomes can be distinguished: parallel outcomes and reactive outcomes. In parallel outcomes, the empathizer mimics the affective state of the target. For example, the empathizer may become fearful when the target is afraid. In reactive outcomes, empathizers exhibit a higher cognitive awareness of the situation and react with empathetic behaviors that do not necessarily match those of the target’s affective state. For example, empathizers may become encouraging when the target is frustrated with the problem-solving tasks.

Recent years have seen a growing interest in empathetic reasoning in virtual agents. Bickmore (2003) showed how embodied agents can form social relationships with users by employing empathy and thereby improving the users’ motivation. The Empathic Companion (Prendinger and Ishizuka, 2005) tracks a user’s bio-signals (e.g., GSR and heart rate) in order to assess the effect of empathetic interventions within stressful job interview scenarios. It was found that empathetic feedback successfully reduced the user’s arousal. Burleson (2006) studied empathetic interventions in frustrating learning environments and explored their effect on meta-affective-strategy learning. In a similar vein, Paiva et al. (2005) studied the requirements for eliciting user empathy and showed that psychological proximity (e.g., gender, shared qualities) is important for generating empathy. Finally, we have extracted empathetic behavior protocols mimicking human empathetic behavior, as described in Chapter 9 (McQuiggan and Lester, 2006; McQuiggan and Lester, 2007). While significant advances have been made in modeling empathy, previous work has not addressed the problem of parallel and reactive empathetic reasoning. In this chapter we will evaluate the impact of empathetic characters and the effects of parallel and reactive empathy on student affect.

### 10.1.3 Psychological Assessment Instrumentation

Below are a number of metrics and instruments that have been used in the studies and experiments reported in this body of work. Much of the evaluation reported in this chapter will make use of the results from these psychological assessments. Thus, we take time to detail each before proceeding with individual investigations in the following sections.
• **Demographics:** The demographic survey collects information regarding student gender, age, race and ethnicity, and in some pertinent cases, marriage status, number of children, native language, degree major, and level of education.

• **Situational interest:** To determine how situational interest relates to affective and learning experiences we analyze the relationship through questionnaire assessments. For measurement purposes, situational interest scales will be adapted from those used by Schraw (1997) to capture differences across groups for interest and to examine within-subject relationships with learning outcomes.

• **Self-efficacy:** Affective experiences have an impact on self-efficacy and visa-versa (Bandura, 1997). Therefore we measure efficacy before and after the task using self-efficacy scales modified from Bandura. Chapter 6 details Self-Efficacy at greater length. The primary instrument for assessing efficacy will be the Self-Efficacy for Self-Regulated Learning Scale (Bandura, 2006). The particular advantage of this scale is that it breaks down the various components of self-efficacy that exert influence on student efficacy (mastery experience, vicarious experience, persuasion, physiological response).

• **Goal orientation:** We assess the relationship of goal orientations to performance in affect-informed narrative-centered learning. Of particular interest will be exploration of how mastery-oriented students interact with characters (such as empathetic characters) differently from performance-oriented students, and how such results might influence pedagogy. To assess goal orientations we will consider two instruments from educational psychology literature. First, Roedel et al.’s Goal Inventory questionnaire (1994) which consists of 18 items divided into 2 subscales: learning and performance. Second, we will use Elliot and McGregor’s Achievement Goals questionnaire (2001), which consists of 12 items broken into 4 subscales (mastery approach, mastery avoidance, performance approach, and performance avoidance).

• **Empathetic Tendencies:** To understand subjects’ tendencies and abilities to be empathetic we will use Davis’ IRI (1983). The IRI measures the participants’ empathetic nature by asking them to rate the degree to which each statement describes them. These items are assessed on a 5-point Likert scale (0 – does not describe me well to 4 – describes me very well). The
IRI is divided into four subscale measurements quantifying the following qualities of empathy: fantasy, perspective taking, empathetic concern, and personal distress (Davis, 1983).

- **Personality:** For measurement purposes we utilize McCrae and Costa’s OCEAN instrument to assess what is commonly referred to as the “Big Five” dimensions of personality: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness (2003). It may prove particularly useful to design support strategies to target personality characteristics that correlate with flow experiences.

- **Presence:** The most widely used measurement of presence is Witmer and Singer’s Presence Questionnaire (1998). One subscale of the PQ measures involvement. We believe involvement will be an interesting aspect of our investigations in narrative learning environments.

- **Immersion Tendencies:** Witmer and Singer also developed the Immersion Tendencies Questionnaire (ITQ) (1998). This instrument measures one’s natural tendency to immerse in various media. Individual results from the ITQ have been shown to predict presence, measured by the PQ (Witmer and Singer, 1998). The instrument is divided into several subscales including tendency to become involved in activities, tendency to maintain focus, and tendency to play video games (Witmer and Singer, 1998).

- **Test of Knowledge:** To assess learning a collaboratively designed test of knowledge was created with education researchers and eighth-grade science teachers. The test is designed to assess students’ knowledge of microbiology content. Using the multiple-choice assessment as a pre-test and post-test instrument learning gain analyses can be conducted. The pre-test measures prior microbiology knowledge while the post-test measures prior knowledge plus information learned from the intervention.

10.2 Evaluating Perceived Empathetic Accuracy of Induced Empathy Models

Recall that *perceived accuracy* is the degree to which a model of empathy makes assessment and interpretation decisions that are perceived by humans to be situationally appropriate. This section
reports on an evaluation of CARE models to determine their perceived accuracy. Perceived accuracy tells us whether the behaviors generated by a model are actually perceived to be socially appropriate in practice. Perceived accuracy is an important aspect of empathetic accuracy because, ultimately, we seek to create models of empathy that will generate behaviors that are deemed to be appropriate for a given social context by human observers.

10.2.1 Method

Participants and Design

In a formal evaluation, thirty-one undergraduate students, in an Institutional Review Board (IRB) of NCSU approved user study, evaluated empathetic responses of the companion agent in video clips from interactions in Treasure Hunt. There were 29 male subjects and 2 female subjects varying in race, ethnicity, and age. 6.5% were aged 18-19, 87.0% were aged 20-24, and 6.5% were aged 25-29.

Materials and Apparatus

For each participant the pre-experiment materials consisted of a consent form, demographic survey, Davis’ Interpersonal Reactivity Index questionnaire, Chapin’s Social Insight Test (Gough, 1993), and a one-page summary of the construct of empathy. Experiment paper-and-pencil materials consisted of response worksheets for each video clip. The experiment’s computerized materials consisted of ten clips of interactions captured from the Treasure Hunt virtual environment. The post-experiment paper-and-pencil materials consisted of a general survey about the participant’s experience and opinions on affect in interactive applications, such as games. The demographic survey collected basic information such as gender, age group, ethnicity, marital status, and number of children. Participants completed Davis’s Interpersonal Reactivity Index (IRI) to obtain a measure of their empathy (Davis, 1983). Chapin’s Social Insight Test quantifies a person’s ability to appraise another person by assessing her ability to predict future events involving the other person in interpersonal and social situations. Chapin’s Social Insight Test asks subjects to assess twenty-five social dilemmas.

---

1 There was no overlap in the 31 participants in the predictive accuracy evaluation (Chapter 9) with the 31 subjects participating in the perceived accuracy evaluation described here.
by selecting the best resolution from the four presented possibilities (Gough 1993). The background document provided the definitions and explanations of empathy from Davis (1994) and Hoffman (2000). Each response worksheet asked subjects the same series of questions about each video clip. Using the response worksheets, subjects evaluated the appropriateness and accuracy of the empathetic emotion, behavior, and timing viewed in the clip, and identified a more appropriate empathetic response, if the participant felt one was applicable.

Each video clip depicted a companion agent exhibiting an empathetic behavior in response to a situation involving another character in the Treasure Island environment. Three types of behaviors were depicted:

- **CARE-generated behaviors:** One set of video clips depicted empathetic responses that were exhibited by companion agents with CARE-induced decision tree models of empathy.

- **Inverse empathetic behaviors:** One set of video clips depicted empathetic responses that were, in effect, the opposite of what CARE recommended. These were determined by identifying the valence of the CARE-generated behavior and then selecting an “opposing” behavior from the classic two-dimensional affective space (Lang, 1995) that had an opposing valence. The inverse empathetic behavior for excited was sad, the inverse empathetic behavior for frustrated was relaxed, and the inverse empathetic behavior for joyful was fearful.

- **Human-generated behaviors:** One set of video clips depicted captures of empathizer-target trainer interactions following the procedure discussed in Section 7.1, i.e., the behaviors were in fact produced by humans (training empathizers) with the empathizer controls.

Video clips averaged approximately 90 seconds. So that viewers could assess the social context in which an empathetic behavior played out, each clip included several events in the virtual environment leading up to the empathetic behavior, as well as the empathetic behavior itself.

10.2.2 Procedure

Participants entered a conference room where they were first presented the details of the study and a consent form. They then completed the demographic survey, Davis’s IRI questionnaire, and Chapin’s Social Insight questionnaire. Next, they read the background on empathy and task
directions. Research assistants then fielded any questions from participants regarding empathy and their prescribed task. Participants were then presented, in random order, a series of ten video clips of captured user-interactions in the Treasure Hunt virtual world. There were four clips of CARE-generated behaviors, three clips of inverse empathetic behaviors, and three clips of human-generated behaviors. After viewing each clip, participants completed the associated response worksheet at their own pace. Following the completion of reviewing and responding to all of the video clips, participants completed the post-experiment survey before the study session concluded.

10.2.3 Results
This section analyzes the study participants’ assessments of the empathetic response clips. A variety of ANOVA statistics are presented for results that are statistically significant. Because the tests reported here were performed on discrete data, we report Chi-square test statistics ($\chi^2$), including both likelihood ratio Chi-square and the Pearson Chi-square values. Fisher’s Exact Test is used to find significant p-values at the 95% confidence level (p < .05).

The IRI results of participating subjects are reported in Table 10.1. Participants averaged 16.55 (SD = 4.33) on the Social Insight Test. Gough (1993) reported the results of a study conducted with a similar population consisting of undergraduate engineering students, whom averaged 25.01 (SD = 4.83) on the Social Insight Test. The difference between the subjects in the two studies is not significant (p < 0.5).

Analysis of participant responses to video clips depicting CARE-generated behaviors yielded 90.3% of participant responses who agreed that the displayed empathetic emotion was appropriate for the situation, 10.8% agreed clips of inverse empathetic behaviors were appropriate, and 87.1% agreed clips of human-generated behaviors from training episodes were appropriate (Table 10.2a). Participants also assessed whether the displayed empathetic emotion was the best emotion for the situation. Three-fourths (75.8%) of the participants agreed the displayed empathetic emotion was the best emotion in clips of CARE-generated behaviors, while 10.8% agreed for clips of inverse empathetic behaviors, and 73.1% agreed for clips of human-generated behaviors from training episodes (Table 10.2b). The third response had participants assess the timing of the empathetic behavior (i.e., was there a more appropriate instance in which the companion agent should have
been empathetic, or not). Fully 87.9% agreed that the timing of the behavior was appropriate in clips of CARE-generated behaviors, while 37.6% agreed for clips of inverse empathetic behaviors, and 82.8% agreed for clips of human-generated behaviors from training episodes (Table 10.2c). Table 10.3 reports the significant distinctions that can be made between these categories of empathetic behavior clips for each of the above participant responses.

Table 10.1. Interpersonal Reactivity Index Results.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Mode</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fantasy</td>
<td>16.48</td>
<td>4.75</td>
<td>17</td>
<td>11</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>Perspective Taking</td>
<td>17.26</td>
<td>4.77</td>
<td>17</td>
<td>20</td>
<td>6</td>
<td>25</td>
</tr>
<tr>
<td>Empathetic Concern</td>
<td>17.87</td>
<td>4.42</td>
<td>18</td>
<td>18</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>Personal Distress</td>
<td>9.13</td>
<td>5.19</td>
<td>9</td>
<td>11</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>60.74</td>
<td>12.08</td>
<td>62</td>
<td>63</td>
<td>34</td>
<td>84</td>
</tr>
</tbody>
</table>

Table 10.2. Analysis of Appropriateness Responses. Grayed-cells indicate no significance (p < .05).

<table>
<thead>
<tr>
<th>Clip Comparison</th>
<th>Likelihood Ratio ($\chi^2$)</th>
<th>Pearson ($\chi^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) CARE vs. Inverse</td>
<td>155.13</td>
<td>136.70</td>
</tr>
<tr>
<td>CARE vs. Human</td>
<td>0.56</td>
<td>0.56</td>
</tr>
<tr>
<td>Human vs. Inverse</td>
<td>122.76</td>
<td>108.46</td>
</tr>
<tr>
<td>(b) CARE vs. Inverse</td>
<td>99.75</td>
<td>90.12</td>
</tr>
<tr>
<td>CARE vs. Human</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Human vs. Inverse</td>
<td>81.24</td>
<td>74.28</td>
</tr>
<tr>
<td>(c) CARE vs. Inverse</td>
<td>62.51</td>
<td>60.16</td>
</tr>
<tr>
<td>CARE vs. Human</td>
<td>1.12</td>
<td>1.13</td>
</tr>
<tr>
<td>Human vs. Inverse</td>
<td>41.46</td>
<td>39.59</td>
</tr>
</tbody>
</table>

Participants also assessed the accuracy of the displayed empathetic behaviors, using a Likert scale from 0 to 4, with respect to the accuracy of the emotion (Table 10.3a), the timing (Table 10.3b), and the overall response (Table 10.3c). Table 10.3 reports the results of the participants’ accuracy assessment.
These results suggest that participants perceived the empathetic behaviors controlled by CARE-induced empathy models as being as appropriate and as accurate as human empathizers were in similar situations.

10.2.4 Discussion

Participant responses to clips of CARE-generated behaviors cannot be statistically distinguished from the responses to clips of human-generated behaviors from training episodes. This result indicates that CARE models generate empathetic behaviors that are similar to those made by humans and are perceived to be situationally appropriate. The fact that participants were able to distinguish, with statistical significance, inverse empathetic behaviors from both CARE-generated behaviors and human-generated behaviors suggests that both CARE models and human models of empathy differ fundamentally from “inverse” empathetic models.

There was no significant effect of psychological instruments (IRI and Social Insight) on participant responses. While a larger study may reveal significant results, it may be the case that measures of one’s own empathy does not correspond directly to one’s ability to interpret another’s empathetic accuracy. A more diverse population that includes subjects beyond undergraduate engineering students may yield different results.

In post-interviews 90.1% of participants indicated that emotions play a valuable role in interactive systems, and 80.1% responded that they would like such systems to account for their own feelings. Most (80.1%) participants indicated that the six emotions displayed in the empathetic behavior clips need to be extended to incorporate additional emotions. Anger, disappointment (as distinct from frustration), and apathy were the emotions most frequently suggested for addition to the current model. Relaxed was the emotion most frequently suggested for removal from the current model.

10.3 Evaluating Student Experience with Empathetic Characters

Following the line of research of employing empathetic characters in education, training, and entertainment we investigate the impact of empathetic characters in a narrative-centered learning
environment, CRYSTAL ISLAND, on users’ sense of presence. We utilize a controlled study (control vs. empathy condition) design to assess the effects of interacting with empathetic characters on recorded experiences of presence.

10.3.1 Method

Participants and Design
A total of 90 students (38 females, 52 males, mean age = 14.78, SD = 1.47) from two populations participated in the studies designed to evaluate the effect of empathetic virtual humans on presence. The subjects in the first study consisted of 55 eighth-grade middle school students (32 females and 23 males). The students ranged in age from 13 to 15 (M = 13.73, SD = 0.59). Approximately 60% of the middle student participants were Caucasian (n = 33), 25% were African-American (n = 14), 5% were Hispanic or Latino (n = 3), 5% were of mixed race (n = 3), and 4% were Asian (n = 2). The students participated as part of their science class.

The study was replicated with a second population consisting of 35 high school students (6 females and 29 males). The high school students ranged in age from 14 to 17 (M = 16.43, SD = 0.74). Approximately, 74% were Caucasian (n = 26), 11% were African-American (n = 4), 9% were Asian (n = 3), and 6% were Hispanic or Latino (n = 2).

Students were randomly assigned to either the control condition or the empathy condition to measure the effect of empathetic characters on the dependent measure of presence. The empathy condition exposed participants to three empathetic characters: Jin, the camp nurse; Elise, the lab technician; and Audrey, a research assistant. Participants in both conditions were able to interact with these characters; however, only in the empathy condition did those characters ask how the student felt and then empathetically responded to the student emotion selections. Table 10.3 depicts the results of random assignment.
Table 10.3. Breakdown of condition assignments by group and gender. Note: MS = Middle School students, HS = High School students, F = Female, M = Male.

<table>
<thead>
<tr>
<th>Condition (n = 90)</th>
<th>Control (n = 47)</th>
<th>Empathy (n = 43)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MS (n = 29)</td>
<td>HS (n = 18)</td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>12</td>
</tr>
</tbody>
</table>

Materials
Pre-experiment questionnaires collected data on participant demographics (including age, gender, and race), game and computer usage, empathy, goal orientation, and immersion tendencies. The researchers designed, game and computer usage questionnaire measured how frequently participants play games (e.g., “How many hours per week do you spend playing video games?”), their perceived level of game playing skill on a Likert scale (1 – not at all to 5 – very skilled), what type of games they play (e.g., role-playing, strategy, sports, etc.), and how many hours per week they use a computer. The interpersonal reactivity index (Davis, 1983) includes 28 items that measure the participants’ empathetic nature by asking them to rate the degree to which each statement describes them. These items are assessed on a 5-point Likert scale (0 – does not describe me well to 4 – describes me very well). The IRI is divided into four subscale measurements quantifying the following qualities of empathy: fantasy, perspective taking, empathetic concern, and personal distress (Davis, 1983). Participant IRI scores were collected to evaluate the interaction between participant empathetic nature and presence. The achievement goals questionnaire\(^2\) (Elliot and McGregor, 2001) measures four achievement goal constructs (mastery-approach, performance-

---
\(^2\) Only the eighth-grade middle school population (n = 55) received Elliot and McGregor’s achievement goal questionnaire (Elliot and McGregor, 2001).
approach, mastery-avoidance, and performance-avoidance goals). Participants indicate the extent to which each statement is true of them on a 7-point Likert scale (1 – not at all true of me to 7 – very true of me). The results of the achievement goals questionnaire were used to determine if there is an effect of student goal orientation on presence. Witmer and Singer developed the Immersive Tendencies Questionnaire (ITQ) to measure individual predispositions towards presence experiences (1998). The ITQ consists of three subscales: activity involvement tendency, activity focus tendency, and video game playing tendency. Participants indicate their level of tendency for each item on a 7-point Likert scale. Witmer and Singer found individual tendencies, as recorded by the ITQ, to be predictive of presence (1998).

Prior to and during participant interactions with CRYSTAL ISLAND students had access to several materials. These materials consisted of a backstory and task description, character handout, map of the island, and a control sheet. The backstory and task description detail the relationship of the students’ character to CRYSTAL ISLAND (a visiting child, named Alyx, of the lead researcher, Bryce) who plays the role of “medical detective” seeking to discover the cause and source of illness causing members of the research team to fall ill. The character handout has a picture, name, and role of each character included in the CRYSTAL ISLAND narrative. The map depicts each of the buildings of the research camp, various foliage, and the waterfall. Finally, the control sheet layouts each of the keyboard and mouse controls with short descriptions used to maneuver the student’s character through the learning environment.

Post-experiment questionnaires collected data on participant presence experience and perceptions of the various virtual humans inhabiting CRYSTAL ISLAND. Participants’ presence experience was captured using the Presence Questionnaire (PQ) developed by Witmer and Singer (1998). The PQ contains several subscales including an involvement/control, naturalism of experience and quality of the interface scales. The PQ accounts for four categories of contributing factors of presence: control, sensory, distraction, and realism. The results of the PQ serve as the primary dependent variable investigated in this paper. In addition to the PQ participants also received a questionnaire to assess participant perceptions of the virtual humans inhabiting CRYSTAL ISLAND. Questions assessed the quality of the characters along several dimensions including friendliness, empathy, and trust.
**Apparatus**

Middle school participants completed CRYSTAL ISLAND interactions on Pentium M 1.73 GHz IBM notebooks with 512 MB of RAM and an ATI Mobility Radeon 9000 Graphics card. The high school student participants completed CRYSTAL ISLAND interactions on Pentium IV 3.0 GHz Dell PCs with 2 GB of RAM and an NVidia Quadro FX 1300/1400 Graphics card. All participants controlled their character (Alyx) through the various goals and character interactions utilizing an extended version of standard Half-Life 2 controls requiring both keyboard and mouse manipulations.

The virtual world of CRYSTAL ISLAND, the semi-autonomous characters that inhabit it, and the user interface were implemented with Valve Software’s Source™ engine, the 3D game platform for Half-Life 2. The Source engine also provides much of the low-level (reactive) character behavior control. The character behaviors and artifacts in the storyworld are the subject of continued work. Students direct their character through CRYSTAL ISLAND by using the keyboard controls (WASD) and mouse movements.

**10.3.2 Procedure**

Participants entered the experiment room with completed informed consent documentation. Participants were randomly assigned to either the empathy or the control condition and were seated in front of a laptop computer. They were given an overview the experiment agenda. Next, participants completed the pre-experiment questionnaires including the demographics, game and computer usage survey, IRI (Davis, 1983), goal orientation¹ (Elliot and McGregor, 2001), and ITQ (Witmer and Singer, 1998).

Upon completion of pre-experiment questionnaires, participants were instructed to review CRYSTAL ISLAND instruction materials. These materials consisted of the backstory and task description, the character handout, the map of the island, and the control sheet. Participants were then further directed on the controls via a presentation explaining each control in detail.

Participants, in both conditions, had 35 minutes to solve the mystery. Solving the mystery consisted of completing 15 goals including learning about various diseases, compiling the symptoms
of the sickened researchers, testing a variety of possible sources, and reporting the solution (cause and source) back to the camp nurse.

Immediately after solving the science mystery of CRYSTAL ISLAND, or 35 minutes of interaction, participants completed the post-experiment questionnaires. First to be completed was the PQ (Witmer and Singer, 1998) followed by the researcher designed questionnaire assessing perceptions of individual CRYSTAL ISLAND characters.

10.3.3 Results
In the first study (middle school participants), there was a significant effect of empathetic character exposure on presence, $F(1, 54) = 4.13, p = 0.047$. Participants reported a greater sense of presence (totaled PQ) in the empathy condition ($M = 147.04, SD = 23.13$) than in the control condition ($M = 132.03, SD = 30.62$). Figure 10.1 shows a box plot of the presence reports of the control and empathy condition for participants from the middle school population.

![Figure 10.1](image)

**Figure 10.1.** Box plots of presence results (total PQ) from each study for control and empathy conditions with middle school participants and high school participants, respectively.
In the replicated study (high school participants), there was a weak significant effect of empathetic character exposure on presence for the high school population $F(1, 34) = 3.11, p = 0.087$. High school participants reported a greater sense of presence (totaled PQ) in the empathy condition ($M = 140.29, SD = 23.14$) than in the control condition ($M = 127.89, SD = 18.31$). Figure 10.2 shows a box plot of the presence reports of the control and empathy conditions for participants from the high school population.

Within each condition there was no significant difference among reported presence between populations. There was no significant difference found in the empathy condition, $t(42) = 0.93, p = 0.36$, between middle school and high school participants. In addition, there was no significant difference in the mean presence reports found in the control condition between populations, $t(46) = 0.58, p = 0.56$. With these results in mind we consider both populations, middle school students ($n = 55$) and high school students ($n = 35$), as a whole.

Examining the population as a whole we find a strong significant effect of empathetic character exposure on presence, $F(1, 89) = 7.02, p < 0.01$. As a single population, participants reported a
greater sense of presence (total PQ) in the empathy condition ($M = 144.37, SD = 23.09$) than in the control condition ($M = 130.44, SD = 26.43$).

Similar results were found for the involvement/control subscale of the PQ. In the first study there was a significant effect of empathetic character exposure on reported involvement/control in the middle school population $F(1, 54) = 4.32, p = 0.042$. Middle school participants in the empathy condition reported more involvement and greater control ($M = 56.42, SD = 9.89$) than middle school participants in the control condition ($M = 50.06, SD = 12.46$). Likewise, in the replicated study, there was a significant effect of empathetic character exposure on reported involvement/control in the high school population $F(1, 34) = 3.97, p = 0.054$. High school participants in the empathy condition reported more involvement and greater control ($M = 57.35, SD = 8.84$) than high school participants in the control condition ($M = 51.0, SD = 9.95$).

We again find strong significance when we consider the population as a whole, i.e., the middle school students and the high school students together. Within conditions there is no significance between groups. There is no significant difference between the populations in reported involvement/control within the empathy condition, $t(42) = 0.32, p = 0.75$. There is also no significant difference between the populations in reported involvement/control within the control condition, $t(46) = 0.28, p = 0.78$. Considering both populations as a whole we find a significant effect of exposure to empathetic characters on reported involvement/control $F(1,89) = 8.21, p = 0.005$. Tables 10.4 and 10.5 summarize the presence results by groups.

In addition to the effects of participants’ interaction with empathetic characters on perceptions of presence, there were several other interesting and significant results. In the first study, with the middle school population, goal orientation was found to affect students’ reported presence. In particular there was a significant effect of mastery approach on presence within the empathy condition, $F(1, 25) = 5.34, p = 0.029$, and performance avoidance on presence within the empathy condition, $F(1, 25) = 6.22, p = 0.019$. Mastery oriented students reported greater levels of presence than performance oriented students.

Middle school participants IRI score also had a significant effect on reports of presence, $F(1, 25) = 6.38, p = 0.018$, and involvement, $F(1, 25) = 5.90, p = 0.022$, within the empathy condition only. Students with greater IRI scores, thus more empathetically natured, reported greater presence than
Table 10.4. Presence results by group (MS – middle school, HS – high school, CC – combined) – C refers to the control condition and – E refers to the empathy condition. INV = involvement/control, NAT = natural, IFQ = interface quality. Pop. sizes: n₁ = 29, n₂ = 26, n₃ = 18, n₄ = 17, n₅ = 47, n₆ = 43.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>INV</th>
<th>NAT</th>
<th>IFQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-C¹</td>
<td>M = 132.03</td>
<td>M = 50.06</td>
<td>M = 11.89</td>
<td>M = 11.72</td>
</tr>
<tr>
<td></td>
<td>SD = 30.62</td>
<td>SD = 12.46</td>
<td>SD = 4.27</td>
<td>SD = 3.75</td>
</tr>
<tr>
<td>MS-E²</td>
<td>M = 147.03</td>
<td>M = 56.42</td>
<td>M = 13.26</td>
<td>M = 12.46</td>
</tr>
<tr>
<td></td>
<td>SD = 23.13</td>
<td>SD = 9.89</td>
<td>SD = 3.50</td>
<td>SD = 2.92</td>
</tr>
<tr>
<td>HS-C³</td>
<td>M = 127.88</td>
<td>M = 51.0</td>
<td>M = 12.61</td>
<td>M = 9.88</td>
</tr>
<tr>
<td></td>
<td>SD = 18.30</td>
<td>SD = 9.95</td>
<td>SD = 3.23</td>
<td>SD = 2.16</td>
</tr>
<tr>
<td>HS-E⁴</td>
<td>M = 140.29</td>
<td>M = 57.35</td>
<td>M = 13.88</td>
<td>M = 10.0</td>
</tr>
<tr>
<td></td>
<td>SD = 23.13</td>
<td>SD = 8.84</td>
<td>SD = 3.29</td>
<td>SD = 1.65</td>
</tr>
<tr>
<td>CC-C⁵</td>
<td>M = 130.44</td>
<td>M = 50.42</td>
<td>M = 12.17</td>
<td>M = 11.02</td>
</tr>
<tr>
<td></td>
<td>SD = 26.43</td>
<td>SD = 11.46</td>
<td>SD = 3.89</td>
<td>SD = 3.33</td>
</tr>
<tr>
<td>CC-E⁶</td>
<td>M = 144.37</td>
<td>M = 56.79</td>
<td>M = 13.51</td>
<td>M = 11.48</td>
</tr>
<tr>
<td></td>
<td>SD = 23.09</td>
<td>SD = 9.39</td>
<td>SD = 3.39</td>
<td>SD = 2.76</td>
</tr>
</tbody>
</table>

Table 10.5. Presence results by group cont. (MS – middle school, HS – high school, CC – combined) – C refers to the control condition and – E refers to the empathy condition. RES = resolution, HAP = Haptic, AUD = auditory. Pop. sizes: n₁ = 29, n₂ = 26, n₃ = 18, n₄ = 17, n₅ = 47, n₆ = 43.

<table>
<thead>
<tr>
<th></th>
<th>RES</th>
<th>HAP</th>
<th>AUD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS-C¹</td>
<td>M = 8.75</td>
<td>M = 7.89</td>
<td>M = 14.06</td>
</tr>
<tr>
<td></td>
<td>SD = 3.23</td>
<td>SD = 2.90</td>
<td>SD = 4.68</td>
</tr>
<tr>
<td>MS-E²</td>
<td>M = 10.19</td>
<td>M = 9.07</td>
<td>M = 16.57</td>
</tr>
<tr>
<td></td>
<td>SD = 2.22</td>
<td>SD = 2.92</td>
<td>SD = 2.98</td>
</tr>
<tr>
<td>HS-C³</td>
<td>M = 9.11</td>
<td>M = 8.5</td>
<td>M = 14.16</td>
</tr>
<tr>
<td></td>
<td>SD = 2.74</td>
<td>SD = 3.16</td>
<td>SD = 3.86</td>
</tr>
<tr>
<td>HS-E⁴</td>
<td>M = 10.35</td>
<td>M = 9.17</td>
<td>M = 14.88</td>
</tr>
<tr>
<td></td>
<td>SD = 3.10</td>
<td>SD = 3.35</td>
<td>SD = 4.59</td>
</tr>
<tr>
<td>CC-C⁵</td>
<td>M = 8.89</td>
<td>M = 8.12</td>
<td>M = 14.10</td>
</tr>
<tr>
<td></td>
<td>SD = 3.03</td>
<td>SD = 2.98</td>
<td>SD = 4.35</td>
</tr>
<tr>
<td>CC-E⁶</td>
<td>M = 10.25</td>
<td>M = 9.11</td>
<td>M = 15.90</td>
</tr>
<tr>
<td></td>
<td>SD = 2.57</td>
<td>SD = 3.06</td>
<td>SD = 3.74</td>
</tr>
</tbody>
</table>

low-IRI students in the empathy condition. In the replicated study, a weak significant effect was also found in the high school population, F(1, 16) = 4.03, p = 0.063, within the empathy condition. In the combined population the effect of participant empathetic nature on presence was significant, F(1, 42) = 11.82, p = 0.0014.
10.3.4 Discussion
The studies found that participants’ sense of presence was increased by enabling characters to interact empathetically. Empathetic characters had a significant effect on measurements of overall presence (total PQ), involvement and control, naturalism of the experience, and resolution. The results suggest that designers of narrative-centered learning environments who seek to increase their students’ sense of presence should consider introducing empathetic characters. The studies also have the following implications, each of which is discussed below:

- Individual differences in student empathy should be considered in designing characters and their interactions with students.
- The design of characters and interactions should also account for differences in students’ achievement goal orientation (mastery vs. performance).
- Learning experiences that seek to involve students in content and motivate them with a sense of control should consider the deployment of empathetic characters.
- Empathetic characters constitute only one of the many narrative content factors to consider in designing effective narrative-centered learning environments. Future studies should consider varying other aspects of the narrative learning experience as well.

The studies found that participants’ empathetic nature, as gauged by Davis’ IRI (1983), affected the participants’ perceptions of presence. The transportation of students, who are empathetic, and perhaps socially intelligent, is supported by interactions with empathetic characters.

Also found in these studies was an effect of student goal orientation on perceptions of presence among the middle school participants. The gaming environment, on which CRYSTAL ISLAND is built, may have had an effect on performance-oriented students, encouraging them to solve the mystery quickly. Meanwhile, it seems that mastery oriented students, who tend to measure success by absorbing content and learning, reported a greater perception of presence. It is possible that mastery oriented students were more likely to take their time during character interactions, perhaps leading to recognition of the characters’ empathetic nature.

The strong effect of empathetic characters on involvement and control may have important implications for learning. There are known motivational benefits user and learner of control
Having characters that respond to student affect may give the student a greater sense of control over the environment by regulating how they feel. Thus, empathetic characters may be able to scaffold student experiences to support regulation of emotions that benefit learning. One such emotion shown to correlate positively with learning is flow (Craig et al., 2004). Not surprisingly, there is a relationship between the constructs of flow and presence, expressed in the overlap of oft-mentioned qualities of each, such as focused attention and the sense of losing time after the experience. Thus, deeper understanding of the relationship between presence and flow, as well as the content variations that affect presence in narrative-centered learning, may lead to effects that promote flow, and ultimately effective learning experiences.

The results of the studies should motivate further investigation into the effects various content factors can have on presence in virtual environments. Beyond empathetic exchanges, there are a variety of narrative-related factors that require further exploration, including student perceptions of narrative drama, participants’ assigned or chosen role within the story, virtual characters’ personality, narrative structure, and plot coherence. Such investigations will be critical to the design of future training and educational environments, particularly after establishing a better understanding of the processes supporting presence and learning.

10.3.5 Study Limitations

The results of the studies are affected by the three virtual characters that interacted empathetically with participants in the empathy condition. It is possible that the gender, narrative role, and pedagogical role of the empathetic characters may not generalize to other characters and across domains. Another shortcoming was that presence was solely assessed after interaction via a questionnaire. Although post-hoc surveys are the currently the most accepted means for gauging presence, it would nevertheless be useful to include additional methods for presence assessment in the future. Finally, to determine how broadly the results hold, the effect of empathetic characters on additional populations should also be studied.
10.4 Affective Transitions

Recent work seeking to characterize the affective experience of learners interacting with intelligent learning environments has considered student affective trajectories occurring during learning. D’Mello et al. (2006) studied the likelihood of affective transitions among six affective states (boredom, flow, confusion, frustration, delight, and surprise) that were found to be relevant to complex learning (Craig et al., 2004). By in large, learners are likely to persist in the same affective state (i.e., transitioning from a state of boredom to boredom is likely, and in some cases, significantly more likely than transitioning to another affective state). This analysis was completed in the AutoTutor learning environment (Craig et al., 2004; D’Mello et al., 2006). Baker et al. were able to replicate many of D’Mello et al.’s (2006) findings when they calculated the likelihood of affective transitions in the “Incredible Machine: Even More Contraptions,” a simulation-based learning environment (Baker et al., 2007). Baker et al. extend their analysis to investigate how usage choices (2004) affect emotion transitions. This work found that confused learners are likely to game the system. Further, it was found that students who game the system are significantly unlikely to transition into a confused state (Baker et al., 2007).

In this section we investigate the likelihood of affective transitions in a narrative-centered learning environment, CRYSTAL ISLAND. The CRYSTAL ISLAND environment utilizes narrative as a mechanism to contextualize learning, making the experience meaningful. Contextualized learning experiences are known to encourage regulated learning behavior (Perry, 1998) and influence student learning and motivation (Linnenbrink and Pintrich, 2001). Because CRYSTAL ISLAND incorporates an engaging storyline into the learning experience we supplement the known relevant emotions to learning used by D’Mello et al. (2006) and Baker et al. (2007) with affective states that may be relevant to the story (anger, anxiety, boredom, confusion, delight, excitement, fear, flow, frustration, and sadness). We extend our analysis of affective transitions to evaluate the impact of character empathetic responses (parallel vs. reactive empathy) to student affect and the relative impact on transitions.

In this section we first present findings regarding common affective transactions observed with our narrative-centered learning environment, CRYSTAL ISLAND. These findings are followed by an
analysis comparing and contrasting likely affective transitions stemming from parallel and reactive empathetic reactions by CRYSTAL ISLAND characters.

To compute transition likelihoods we adopt D’Mello et al.’s $L$ (2006), which is based on Cohen’s Kappa (1960), and has been used by Baker et al. for affective transition analysis in their simulation learning environment (Baker et al., 2007). $L$ computes the probability that a transition between two affective states ($\text{CURRENT} \rightarrow \text{NEXT}$) will occur, where CURRENT refers to a reported emotion at time $t$, while NEXT refers to the next reported emotion at time $t+1$. D’Mello et al.’s $L$ accounts for the base frequency of the NEXT affective state in assessing the likelihood of a particular transition. Formally, $L$’s numerator is divided by $1-P(\text{NEXT})$ to normalize scores between $-\infty$ and 1 (D’Mello et al., 2006). A result of $L$ equal to 1 translates to emotion NEXT always following the CURRENT emotion; an $L$ value equal to 0 means the likelihood of transitioning to emotion NEXT is equal to chance, i.e., the probability of experiencing NEXT (the base rate) regardless of the CURRENT emotion. An $L$ value less than 0 translates to the likelihood of transitioning to emotion NEXT being less than chance (the probability of experiencing NEXT regardless of the CURRENT emotion).

To characterize affective transitions we first compute $L$ for each transition ($\text{CURRENT} \rightarrow \text{NEXT}$), for each student. We then use mean $L$ values to determine the likelihood of transitioning from each emotion CURRENT to each emotion NEXT. The results of ANOVAs determine whether the differences in likelihoods of transitioning to each NEXT emotion are significantly different for particular CURRENT emotions.

10.4.1 Method

Participants and Design

The subjects of the study consisted of 35 graduate students ranging in age from 21 to 60 ($M = 24.4$, $SD = 6.41$) including 9 females and 26 males. Among these students, 60% were Asian ($n = 21$), approximately 37% were Caucasian ($n = 13$) and one participant chose not to respond.
10.4.2 Procedure

Participants entered the experiment room where they completed informed consent documentation. They were randomly assigned to either the control condition or the empathy condition and were seated in front of a laptop computer. They were then given an overview of the experiment agenda, and they completed the pre-experiment questionnaires including the demographics survey, the interpersonal reactivity index survey (Davis, 1994), and the goal orientation survey (Elliot and McGregor, 2001).

Upon completing the pre-experiment questionnaires, participants were instructed to review CRYSTAL ISLAND instruction materials. These materials consisted of the backstory and task description, the character overviews, the map of the island, the control sheet, and definition sheet of the self-report emotions. Participants were then further briefed on the controls via a presentation summarizing the task and explaining each control in detail. Participants had continued access to the materials, including the definition sheet of the self-report emotions, throughout their interaction.

Participants were given 35 minutes to solve the mystery. Solving the mystery consisted of completing 15 goals including learning about various diseases, compiling the symptoms of the sickened researchers, testing a variety of possible sources, and reporting the solution (cause and source) back to the camp nurse.

Six CRYSTAL ISLAND characters (Audrey, Elise, Jin, Quentin, Robert, and Teresa), each play distinct roles in the CRYSTAL ISLAND environment. When subjects decided to interact with these particular characters, they were greeted with empathetic reactions to their expressed affective state, which they expressed through self-reports via an in-game dialog. The self-report dialog asked participants to select the affective state that best described their feelings at that time from a set of 10 affective states (anger, anxiety, boredom, confusion, delight, excitement, fear, flow, frustration, and sadness). This set of emotions was comprised of emotions identified with learning (Craig et al., 2004; D’Mello et al., 2006; Kort et al., 2001) together with basic emotions (Ekman and Friesen, 1978) that may play a role in students’ experience of the CRYSTAL ISLAND narrative.

Immediately after solving the science mystery of CRYSTAL ISLAND (or after 35 minutes of elapsed interaction time for subjects who had not solved the mystery), subjects completed the post-
experiment questionnaire. This researcher-designed questionnaire assessed perceptions of individual CRYSTAL ISLAND characters. The results of this instrument are outside the scope of this discussion.

10.4.3 Results
Aggregating self-reported affective states across the 35 participants we find flow to be the most frequently reported state (42%), followed by excitement (14%), confused (13%), delight (11%), anxiety (8%), frustration (6%), boredom (3%), sadness (2%), anger (1%), and fear (1%)

ANOVAs indicated that six affective states had statistically significant differences among the likelihoods of transitions. Affective transitions were statistically significantly different transitioning from frustration ($F(9, 340) = 2.06, p = 0.03$), flow ($F(9, 340) = 18.3, p < 0.0001$), confused ($F(9, 340) = 1.79, p = 0.06$), delight ($F(9, 340) = 5.22, p < 0.0001$), anxiety ($F(9, 340) = 2.98, p = 0.002$), and excitement ($F(9, 340) = 2.62, p = 0.06$).

Frustrated learners are most likely to remain frustrated (Mean $L = .28$) followed by transitions to confusion (.10) and fear (.09). The remaining transitions were below chance levels (i.e., flow (-.19, $t(34) = -4.24, p < 0.0001$) and excitement (-.10)).

Learners in the state of flow were most likely to remain in flow (.19) followed by confusion (.04, $t(34) = -3.09, p = 0.003$), anxiety (.03), and delight (.02). Both frustration (-.04, $t(34) = -7.91, p < 0.0001$) and excitement (-.07) were below chance levels. The remaining transitions did not occur or occurred at chance levels.

Confused students were likely to remain in a confused state (.16) followed by excitement (.10), boredom (.05), frustration (.04), and flow (.04). The likelihood of these and all remaining conditions are summarized for the interested reader in Table 10.6.
Table 10.6. Likelihoods for all transitions CURRENT → NEXT for the affective states: Frustration, Flow, Confused, Delight, Boredom, Anxiety, Excitement, Anger, Sadness, and Fear.

Next

<table>
<thead>
<tr>
<th>Current</th>
<th>Fr</th>
<th>Fl</th>
<th>Co</th>
<th>De</th>
<th>Bo</th>
<th>Anx</th>
<th>Ex</th>
<th>Ang</th>
<th>Sa</th>
<th>Fe</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fr</td>
<td>0.28</td>
<td>-0.19</td>
<td>0.10</td>
<td>-0.05</td>
<td>-0.07</td>
<td>-0.15</td>
<td>-0.10</td>
<td>-0.02</td>
<td>-0.01</td>
<td>0.09</td>
</tr>
<tr>
<td>Fl</td>
<td>-0.04</td>
<td>0.19</td>
<td>0.04</td>
<td>0.02</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.07</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Co</td>
<td>0.04</td>
<td>0.04</td>
<td>0.16</td>
<td>-0.03</td>
<td>0.05</td>
<td>-0.04</td>
<td>0.10</td>
<td>-0.01</td>
<td>-0.01</td>
<td>-0.03</td>
</tr>
<tr>
<td>De</td>
<td>0.01</td>
<td>0.10</td>
<td>-0.13</td>
<td>0.21</td>
<td>-0.03</td>
<td>-0.05</td>
<td>-0.33</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Bo</td>
<td>0.13</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.08</td>
<td>0.13</td>
<td>-0.04</td>
<td>-0.04</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.04</td>
</tr>
<tr>
<td>Anx</td>
<td>-0.08</td>
<td>0.06</td>
<td>0.04</td>
<td>-0.07</td>
<td>-0.01</td>
<td>0.14</td>
<td>-0.19</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Ex</td>
<td>-0.05</td>
<td>-0.11</td>
<td>0.06</td>
<td>-0.03</td>
<td>-0.03</td>
<td>0.03</td>
<td>0.24</td>
<td>-0.01</td>
<td>0.01</td>
<td>-0.02</td>
</tr>
<tr>
<td>Ang</td>
<td>0.00</td>
<td>-0.07</td>
<td>0.09</td>
<td>-0.39</td>
<td>0.00</td>
<td>0.23</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sa</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Fe</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Affective Transitions across Empathy

Empathy is the expression of emotion based on another’s situation and not merely one’s own (Davis, 1994; Ickes, 1997). Its expression can demonstrate that the target’s (the recipient of empathetic expression) feelings are understood or shared. In the case of parallel empathy, an individual exhibits an emotion similar to that of the target (Davis, 1994). This is typically based on an understanding of the target’s situation and shows the empathizer’s ability to identify with the target. Reactive empathy, in contrast, focuses on the target’s affective state, in addition to her situation (Davis, 1994). Reactive empathizers will display emotions that are different from the target’s, often in order to alter or enhance the target’s own affective state. This type of empathy is focused on the target whereas parallel empathy is more self-oriented. As such, reactive empathy can be viewed as a higher level of empathetic behavior.
Recent research with the characters of CRYSTAL ISLAND has investigated the merit of providing characters with empathetic capabilities to effectively respond to unfolding student experiences. In CRYSTAL ISLAND, empathetic responses are short, text-based responses consisting of 1 to 2 sentences. Parallel responses consist of the character expressing the same emotion as the user through text responses; alternatively, reactive responses demonstrate advanced cognitive processing on the character’s part by providing responses designed to be more motivating revealing the character’s desire for the user to be in a positive emotional state. The results below compare the likelihood of affective transitions based on empathetic expressions from CRYSTAL ISLAND characters in response to student CURRENT emotions. The findings suggest that in certain situations parallel and reactive empathy have significant differences in the affective transitions (NEXT emotion) likely to occur.

While the relatively low frequencies of some transitions prevents much many of the visible differences from being statistically significant, interesting patterns do emerge. Figures 10.3 and 10.4 present the transitions from the state of flow and frustration by empathetic reaction type (parallel or reactive). Analyzing the transitions from the state of flow we find that, interestingly, parallel empathy (.11) is weakly significantly more likely to support students’ remaining in the state of flow than reactive empathy (-.05), t(12) = -2.08, p = 0.06. Similarly, we find that the likelihood of transitioning to frustration from a frustrated state is significantly more likely when characters empathetic reactions are parallel in nature (.57) and reactive (-.13), t(12) = -2.09, p = 0.059. Other patterns with visible differences emerging from this analysis of affective transitions are summarized in Table 10.7. Note that the transition frequencies were not sufficiently high for concluding the differences to be statistically significant, but merit discussion.
**Figure 10.3.** Transitions from flow to FRustration, FLow, COnfused, DElight, BOredom, ANxiety, EXcitement, ANger, SAdness, and FEar.

**Figure 10.4.** Transitions from frustration to FRustration, FLow, COnfused, DElight, BOredom, ANxiety, EXcitement, ANger, SAdness, and FEar.
Table 10.7. Interesting likelihood for transitions differences by empathetic response type
(parallel or reactive).

<table>
<thead>
<tr>
<th>CURRENT</th>
<th>Transition State (NEXT)</th>
<th>Parallel Empathy Likelihood</th>
<th>Reactive Empathy Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boredom</td>
<td>Boredom</td>
<td>.35</td>
<td>-.04</td>
</tr>
<tr>
<td></td>
<td>Confused</td>
<td>0</td>
<td>-.41</td>
</tr>
<tr>
<td></td>
<td>Flow</td>
<td>-.13</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>-.08</td>
<td>.26</td>
</tr>
<tr>
<td>Anxiety</td>
<td>Anxiety</td>
<td>.33</td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td>Frustration</td>
<td>-.20</td>
<td>.17</td>
</tr>
<tr>
<td>Frustration</td>
<td>Frustration</td>
<td>.57</td>
<td>-.13</td>
</tr>
<tr>
<td></td>
<td>Flow</td>
<td>.10</td>
<td>-.25</td>
</tr>
<tr>
<td></td>
<td>Confused</td>
<td>-.17</td>
<td>.15</td>
</tr>
<tr>
<td>Flow</td>
<td>Flow</td>
<td>.11</td>
<td>-.05</td>
</tr>
<tr>
<td></td>
<td>Confused</td>
<td>.04</td>
<td>.08</td>
</tr>
<tr>
<td>Delight</td>
<td>Delight</td>
<td>.21</td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>Flow</td>
<td>.07</td>
<td>.17</td>
</tr>
</tbody>
</table>

Individual Differences in Affective Transitions

There were significant differences in the frequencies with which males and females reported emotions of boredom. Females ($n = 9$) did not report feeling bored while the males did, leading to a marginally significant difference, $t(34) = 1.87, p = .07$. There were no other differences across gender.

Student personalities also affected the frequency with which certain affective states were reported, namely, anger, boredom, confusion, delight, and flow. There was a significant difference in the frequency of reported states of flow along the Extraversion dimension. Students who were more extroverted reported affective states of flow less frequently than less extraverted students, $t(34) = 2.14, p = .04$. Also along the Extraversion dimension were differences in the frequencies of delight and anger. Marginally significant was the frequency of which more extraverted students reported delight than less extraverted students, $t(34) = 1.82, p = .07$. The more extraverted students reported delight almost 5 times per interaction compared to just 2 times for the less extraverted.
students. Anger was reported more frequently by more extraverted students than less extraverted students, \( t(34) = 2.77, p = .009 \).

There were significant differences across the personality dimensions of Agreeableness, Conscientiousness, and Neuroticism in reports of confusion. Less agreeable students reported confusion more frequently (\( M = 6.06, SD = 1.5 \)) than more agreeable students (\( M = 2.36, SD = 1.4 \)), \( t(34) = 1.77, p = .08 \). Similarly, less conscientious students reported confusion more frequently (\( M = 6.0, SD = 1.43 \)) than more conscientious students (\( M = 2.0, SD = 1.47 \)), \( t(34) = 1.94, p = .06 \). Students with greater emotional stability (Neuroticism dimension) reported confusion more frequently (\( M = 7.93, SD = 1.48 \)) than less emotionally stable students (\( M = 1.47, SD = 1.2 \)), \( t(34) = 3.37, p = .001 \).

The final significant difference in emotion frequencies along personality dimensions is reports of boredom across student agreeableness. More agreeable students reported being bored less frequently (\( M = 0.1 \), \( SD = 0.4 \)) than less agreeable students (\( M = 2.2 \), \( SD = 0.44 \)), \( t(34) = 3.45, p = .001 \).

Student goal orientation also affected the frequency of which students reported anger, anxiety, and flow. Anger was reported more frequently by students scoring higher on the performance approach subscale than students scoring below the performance approach population mean, \( t(34) = 2.28, p = .03 \). Marginally significant was the frequency with which students who were dominantly performance-oriented reported feeling anxious (\( M = 3.62, SD = 0.89 \)) than students who were dominantly mastery-oriented (\( M = 1.2, SD = 1.1 \)), \( t(34) = 1.71, p = .09 \). Also significant was the frequency with which students scoring high on the performance avoidance subscale reported feeling anxious (\( M = 4.05, SD = 0.87 \)) than students scoring below the performance avoidance population mean (\( M = 0.8, SD = 1.01 \)), \( t(34) = 2.43, p = .02 \). Flow was more frequently reported by students who were dominantly mastery-oriented (\( M = 18.2, SD = 2.8 \)) than students who were dominantly performance-oriented (\( M = 10.04, SD = 2.2 \)), \( t(34) = 2.25, p = .03 \). The frequency of flow reports was impacted by students’ performance-orientations. Students scoring lower on the performance avoidance subscale reported more feelings of flow than students scoring above the performance avoidance population mean, \( t(34) = 2.13, p = .04 \). Comparatively, students scoring lower on the performance approach subscale reported more feelings of flow than students scoring above the population mean for performance approach, \( t(34) = 1.87, p = .07 \).
Lastly, there were differences in the frequencies of reports of frustration and anxiety across students’ reported sense of presence. Students scoring below the population mean of the presence questionnaire reported frustration with greater frequency than students reporting a greater sense of presence, marginally significantly, \( t(34) = 1.70, p = .09 \). Anxiety was reported more frequently by students scoring above the population mean on the presence questionnaire than students reporting lower levels of presence, \( t(34) = 2.23, p = .03 \).

There were few statistically significant differences in affective transitions across individual differences. This is likely due to a small population size \( (n = 35) \) resulting in small split population sizes. However, there are noticeable trends that may be concretely uncovered in a large scale study. We report on several of these findings below.

For instance, there are interesting differences in affective transitions when we consider student dominant goal-orientations. Mastery oriented students are not likely to stay confused (Figure 10.5) and are most likely to transition to a state of flow, a finding that suggests that mastery-oriented students are engaged or motivated by the cognitive disequilibrium associated with confusion. Being in a confused state is associated with a need to learn and the CRYSTAL ISLAND environment supports mastery-oriented students’ goal of acquiring knowledge. There is a chance that performance-oriented students may stay confused or transition to negative states such as frustration, boredom, or anxiety. Perhaps this is indicative of the fact that CRYSTAL ISLAND is guiding performance-oriented
students into situations where they must master content to proceed, thus slowing progress and inadvertently decreasing perceived performance. Also, in Figure 10.5, we notice that bored mastery-oriented students are not likely to remain bored and are more likely to transition to a state of flow or confusion. These affective states are thought to be preferred learning emotional states (Craig et al., 2004).

There are interesting differences in likely transitions when we consider reported student presence as well (Figure 10.6). The participant population was broken into two groups around the population mean for the involvement/control subscale: low and high. Here we notice that students reporting high levels of involvement are not likely to stay in a state of confusion and are most likely to transition to a state of flow. On the other hand, students reporting lower levels of involvement in their experience were likely to stay confused or transition to other affective states, such as frustration or boredom. We notice a similar trend in transitions from a state of boredom. Students reporting high levels of involvement are not likely to stay bored and are more likely to become confused, excited, or enter a flow state. Students reporting lower levels of involvement are somewhat likely to stay bored, but are surprisingly more likely to transition to flow or delight. However, the occurrences of the vicious boredom cycles may in part be the cause for lower levels of reported involvement and control due to student disengagement.

Lastly, there is a noticeable trend in affective transitions among student agreeableness (Figure 10.7). For instance, less agreeable students were likely to enter vicious frustration cycles, with some transitioning to confusion. While, on the other hand, more agreeable students were not likely to remain frustrated and were most likely to transition to a state of confusion from frustration. Anxious, less agreeable students were very likely to stay anxious, with a small chance of transitioning to flow. More agreeable students were not likely to stay anxious and were most likely to transition to a flow state, with some chance of transitioning to delight or anger.

10.4.4 Discussion
The analysis of affective state transitions in CRYSTAL ISLAND replicate findings by D’Mello et al. (2006) and Baker et al. (2007). For instance, the state of flow dominated self-reported affect. The dominance of the flow state has been reported in a number of affective studies with intelligent
learning environments (Baker et al., 2007; Craig et al., 2004; D’Mello et al., 2006). Frustration and boredom were reported notably less frequently than in D’Mello et al.’s study and was comparably reported to frequencies found in Baker et al. Perhaps surprisingly, emotions found to be relevant to learning (namely, boredom, confusion, delight, flow, and frustration) were more dominate than the affective states hypothesized to be relevant affective outcomes to the CRYSTAL ISLAND story.
Among the most likely transitions, CURRENT → NEXT, were transitions where NEXT = CURRENT. This was true for the affective states frustration, flow, confused, delight, boredom, anxiety, excitement, and anger. This result also replicates the findings of (D’Mello et al., 2006) and (Baker et al., 2007). D’Mello termed these cycles vicious cycles for negative affective states (similar to Burleson’s notion of “state of stuck” (Burleson, 2006)) and virtuous cycles when students are likely to stay in positive states (i.e., flow).

When we consider affective transitions where NEXT occurs at time, t+1, after an empathetic response from a CRISTAL ISLAND character, we notice differences in the likely affective outcomes. For instance, if a student is in a frustrated state, parallel empathy is likely to elicit a transition in which the student stays frustrated. In contrast, reactive empathy is less than chance likely to prompt the same vicious cycle. Instead reactive empathy likely promotes transitions to a confused state, which is known to have better correlations with learning (Craig et al., 2004).

When we consider likely transitions from the state of flow, we find that parallel empathy is likely to encourage students to enter a virtuous cycle and remain in the state of flow. Reactive empathy is less likely than chance to produce the flow state and is likely to promote an affective state transition to confusion. Since a flow state cannot really be motivated to be more positive, as it is considered an optimal state of experience (Csikszentmihalyi, 1990), we can understand why a student might experience adverse effects to reactive empathy when in a state of flow.

Analyzing transition patterns from the state of boredom, we find parallel empathy is likely to encourage a vicious cycle while reactive empathy is less than chance likely to produce the same cycle. Instead, reactive empathy is most likely to transition to flow, with frustration slightly less likely than flow. If we are able to drill-down into the student’s experience and eventually differentiate when reactive empathy is likely to encourage flow as opposed to when it is likely to promote frustration, this diagnostic information can inform pedagogical agents when a reactive empathy response might be appropriate for alleviating student boredom and promoting a state of flow.

Among the differences between personality traits, those relating to extroversion and conscientiousness are perhaps the most interesting. Highly extroverted individuals were more likely to report narrative-based emotions such as anger and delight, and less likely to focus on learning, or
flow. Perhaps these individuals were more focused on the narrative aspects of the environment such as interacting with characters, and consequently their attention was drawn away from learning tasks. Additionally, individuals who reported high levels of conscientiousness were less likely to report experiencing confusion. Conscientious individuals are more likely to regulate their own behavior and perhaps this leads them to focus on finding solutions to resolve their confusion. This notion is also supported by the increased likelihood of conscientious individuals to transition into flow and the very low likelihood that they will remain confused.

Overall, the trend among affective frequencies shows that increased levels of performance orientation leads to reduced levels of flow and increased levels of anxiety. This is true when examining student’s dominant orientation as well as their avoidance and approach subscales. This correlates well with understanding of the approaches used by these two categories. Individuals who are mastery oriented are focused strongly on learning and may therefore be more likely to immerse themselves in learning oriented activities in the environment. Similarly, as suggested by the rates of affective transitions, they may return more quickly to flow after experiences of other affective states. Conversely, performance dominant students are focused on their measures of success. The higher level of anxiety reported by these students may be a direct result of concerns of performance. Because, there is no objective measure of performance in the CRYSTAL ISLAND environment, performance dominant students may become nervous over supposed comparison to others and opinions of the researcher present.

Interestingly, differences were found based among individual reports of presence. Students reporting higher levels of presence were more likely to be anxious and less likely to have experienced frustration. Perhaps students who became frustrated disengaged themselves from the environment resulting in lower levels of presence. Also, students who were highly engaged may have felt more salient responses to the narrative aspects of the environment. They may have become more concerned over the wellbeing of the characters and anxious over the outcome of the events. These differences are especially significant as it suggests that anxiety might be used to indicate measures of presence. Similarly, it appears that given an objective of maintaining presence, it would be highly important to avoid frustrating users.
10.4.5 Study Limitations

The results of this study are affected by the virtual characters that interacted empathetically with participants. It is possible that the gender, narrative role, and pedagogical role of the characters may affect the likelihood of transitions in addition to the type of empathy. Another shortcoming is that affective states were solely collect from student self-reports. In contrast, both D’Mello et al. (2006) and Baker et al. (2007) used judged reports of affect in their transition analysis. In the study reported here, video recordings of participants’ faces were collected during their interactions with the learning environment to permit future work to consider judged reports of affect with this dataset. Finally, to determine how broadly the results hold, the transitions found likely with this subject population need to be validated with other populations, such as middle school students, and with larger populations participating in larger scale studies.

10.5 Effects of Narrative on Learning and Student Engagement

Much of the work on NLEs has focused on developing AI-based approaches that provide rich, adaptive narrative-based learning experiences and respond appropriately to student actions in the environment. FearNot! is a character-driven learning environment for the domain of anti-bullying social education (Aylett et al., 2005). The environment emphasizes autonomous, highly affective characters that foster empathetic relationships with students, who in turn offer coping suggestions to the victimized virtual character. FearNot! has been the subject of several small- and large-scale studies, although the subjective nature of the domain renders objective, learning-gain results impractical. Carmen’s Bright IDEAS seeks to teach health intervention skills to mothers of pediatric cancer patients (Marsella et al., 2003). The environment combines autonomous characters with director and cinematographic agents in order to provide a dramatic story that satisfies pedagogical goals. Students control the main character’s (Carmen’s) decisions as she copes with the stresses and problems inherent in caring for an ill child. Carmen’s Bright IDEAS has been the subject of clinical trials, but reported results have also been limited.

Intelligent NLEs have recently been developed for military soft-skills training, particularly in leadership and language learning scenarios. IN-TALE is an interactive narrative system that integrates autonomous character behaviors and an Automated Story Director to provide dramatic
simulation experiences for social and cultural leadership training (Riedl and Stern, 2006). The system draws upon previous work in narrative planning and believable agent behavior to balance narrative coherence and user-agency in the simulation environment. The Tactical Language and Culture Training System (TLCTS) is a story-centric, serious game designed for language learning (Johnson, 2007). TLCTS use a combination of interactive lessons and games to train students in spoken and non-verbal communication, as well relevant cultural knowledge. Over the course of the last several years, TLCTS has transitioned into widespread use by the US military and other groups. However, large-scale, summative empirical results for learning outcomes have not yet been presented for either IN-TALE or TLCTS (Johnson and Beal, 2005).

Despite the presence of several promising and successful examples of NLEs, empirical evaluation remains limited. We seek to extend preliminary results in narrative-centered learning by reporting on a controlled experiment assessing learning outcomes between several versions of a NLE and a more traditional, didactic format.

10.5.1 Method

Participants and Design
There were 88 female and 91 male participants varying in age and race. Approximately 2% of the participants were American Indian or Alaska Native, 5% were Asian, 29% were Black or African American, 58% were Caucasian, 6% were Hispanic or Latino, and 6% were of other races. Participants were all eighth-grade students ranging in age from 12 to 15 (M = 13.27, SD = 0.51). The students had recently completed the microbiology curriculum mandated by the North Carolina state standard course of study before receiving the instruments, tests, and interventions of this experiment.

Materials
The pre-experiment paper-and-pencil materials for each participant were completed one week prior to intervention. These materials consisted of a researcher generated CRYSTAL ISLAND curriculum test, demographic survey, Achievement Goals Questionnaire (Elliot and McGregor, 2001), Self-Efficacy for Self-Regulated Learning scale (SESRL) (Bandura, 2006), Science Self-Efficacy scale, modified from
(Nietfeld et al., 2006), and Immersion Tendencies Questionnaire (Witmer and Singer, 1998). The CRYSTAL ISLAND curriculum test consists of 23 questions created by an interdisciplinary team of researchers and was approved for language and content by the students’ eighth-grade science teachers. Elliot and McGregor’s Achievement Goals Questionnaire is a validated instrument which measures four achievement goal constructs (mastery-approach, performance-approach, mastery-avoidance, and performance-avoidance goals) (2001). Bandura’s Self-Efficacy for Self-Regulated Learning scale (2006) consists of 11 items rated by participants on a 7-point Likert scale. Witmer and Singer developed and validated the Immersive Tendencies Questionnaire (ITQ) to measure individual predispositions towards presence experiences (1998). The ITQ consists of three subscales: activity involvement tendency, activity focus tendency, and video game playing tendency. Participants indicate their level of tendency for each item on a 7-point Likert scale. Witmer and Singer found individual tendencies, as recorded by the ITQ, to be predictive of presence (discussed in Section 6.2) (1998).

Post-experiment materials were completed immediately following intervention. These materials consisted of the same CRYSTAL ISLAND curriculum test, Achievement Goals Questionnaire (Elliot and McGregor, 2001), Science Self-Efficacy scale, an interest scale (Schraw, 1997), and the Presence Questionnaire (Witmer and Singer, 1998). The interest scale was adapted from those used by Schraw to capture differences across groups and to examine within-subject relationships with learning outcomes (1997). Participants’ presence experience was captured with the Presence Questionnaire (PQ) developed and validated by Witmer and Singer (1998). The PQ contains several subscales including involvement/control, naturalism of experience and quality of the interface scales.

10.5.2 Procedure
Participants entered the experiment room having completed the pre-test and instrumentation one week prior to the intervention. Participants were first instructed to review CRYSTAL ISLAND instruction materials. These materials consisted of the CRYSTAL ISLAND back-story and task description, a character handout, a map of the island, and a control sheet. Participants were then further directed on the controls via a presentation explaining each control in detail.
Participants in the three intervention conditions (narrative, minimal-narrative, and PowerPoint) were given 50 minutes to work on solving the mystery. Solving the mystery consisted of completing a number of goals including learning about pathogens, viruses, bacteria, fungi, and parasites, compiling the symptoms of the researchers who had fell ill, recording features of hypothesized diseases causing the CRYSTAL ISLAND illness, testing a variety of possible sources, and reporting the solution (cause and source) to the camp nurse to design a treatment plan.

Immediately after solving the science mystery of CRYSTAL ISLAND, or 50 minutes of interaction, participants completed the post-experiment questionnaires. First to be completed was the CRYSTAL ISLAND curriculum test, followed by the remaining post-experiment questionnaires described above. Completion of post-experiment materials took no longer than 35 minutes for participants. In total, sessions lasted 90 minutes.

The experiment randomly assigned the entire eighth grade population of Centennial Campus Middle School in Raleigh, North Carolina to four groups: holdout, CRYSTAL ISLAND narrative condition, CRYSTAL ISLAND minimal-narrative condition, or PowerPoint condition (see Table 1 for condition breakdown). Participants in the holdout condition did not receive an intervention and served as the control group for this experiment and planned longitudinal studies. In the remaining three conditions, students received an intervention consisting of the CRYSTAL ISLAND microbiology curriculum delivered in one of three formats. The CRYSTAL ISLAND narrative condition supplemented the curriculum with the full CRYSTAL ISLAND narrative, including a poisoning scenario, character back-stories, and rich character personalities. The CRYSTAL ISLAND minimal-narrative condition supplemented the curriculum with the minimal story required to support the curriculum. In this condition, the story strictly consisted of research members falling ill and the request for the student to identify the mysterious illness. The minimal-narrative condition did not include the poisoning storyline, character back-stories, or explicit character personality. The PowerPoint condition consisted of a narrated PowerPoint presentation of the same curriculum that was used in CRYSTAL ISLAND. Much of the text and images of the slides actually appear in CRYSTAL ISLAND in the form of books, posters, and character dialogue. The PowerPoint condition did not contain a corresponding storyline.
Table 10.8. Subject population by condition.

<table>
<thead>
<tr>
<th>Condition</th>
<th>n = 179</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holdout</td>
<td>n = 29</td>
</tr>
<tr>
<td>Male</td>
<td>n = 11</td>
</tr>
<tr>
<td>Female</td>
<td>n = 18</td>
</tr>
<tr>
<td>Narrative</td>
<td>n = 60</td>
</tr>
<tr>
<td>Male</td>
<td>n = 30</td>
</tr>
<tr>
<td>Female</td>
<td>n = 30</td>
</tr>
<tr>
<td>Min-narrative</td>
<td>n = 56</td>
</tr>
<tr>
<td>Male</td>
<td>n = 32</td>
</tr>
<tr>
<td>Female</td>
<td>n = 24</td>
</tr>
<tr>
<td>PowerPoint</td>
<td>n = 33</td>
</tr>
<tr>
<td>Male</td>
<td>n = 17</td>
</tr>
<tr>
<td>Female</td>
<td>n = 16</td>
</tr>
</tbody>
</table>

10.5.3 Results

Learning Outcomes

Investigating learning in CRYSTAL ISLAND as measured by the difference of post-test and pre-test scores, we find that, overall, students exhibited learning gains (learning gain, $M = 0.07$, $SD = 0.14$). On average, students answered 1.6 ($SD = 3.3$) more questions correctly on the post-test than on the pre-test. Matched pairs t tests (comparing post-test to pre-test scores) indicate that these learning gains are significant overall, $t(149) = 5.51$, $p < 0.0001$, and significant (although weakly significant in the narrative condition) within each condition (narrative condition: $t(58) = 1.43$, $p = 0.07$, minimal-narrative condition: $t(55) = 2.97$, $p < 0.005$, and the PowerPoint condition: $t(34) = 5.74$, $p < 0.0001$). Further, the learning gains in each condition were significantly different, $F(2, 149) = 10.38$, $p < 0.0001$. There was no significant difference among pre-test scores between conditions, $F(4, 179) = 0.94$, $p = 0.44$. The largest learning gains occurred in the PowerPoint condition ($M = 0.15$, $SD = 0.15$), followed by learning gains in the minimal-narrative condition ($M = 0.06$, $SD = 0.14$), and the lowest learning gains in the narrative condition ($M = 0.02$, $SD = 0.11$). Students in the hold out condition did not take a post-test, and therefore no learning gain results are available for that condition. The CRYSTAL ISLAND curriculum test consisted of 23 items leading to a learning gain of 0.043, which equates to getting one additional question correct in the post-test compared to the pre-test. Thus, on average, students in the PowerPoint condition answered 3.5 more questions correctly ($SD = 3.6$) on the post-test, with participants in the minimal-narrative and narrative conditions answering 1.3
(SD = 3.2) and 0.5 (SD = 2.7) more questions correctly, respectively. Learning gains are depicted in Figure 10.8. If we consider only the students who completed the CRYSTAL ISLAND mystery in the narrative condition, we find no significant difference between post-test scores with the PowerPoint condition, $F(1, 48) = 0.32, p = 0.58$. However, the learning gains in the PowerPoint condition were somewhat significantly better than the students who finished the CRYSTAL ISLAND narrative experience, $F(1, 48) = 4.09, p = 0.05$.

Interestingly, there was an effect of gender on learning in CRYSTAL ISLAND. When we consider only the problems on the CRYSTAL ISLAND curriculum test for which students were exposed to (not all students solved the CRYSTAL ISLAND mystery and completed all problem-solving activities), we find gender played a significant role, $F(1, 114) = 4.44, p = 0.037$. In CRYSTAL ISLAND, on average, male students got an additional 1.3 problems correct (SD = 2.4) on post-tests compared to pre-tests, while female students got an additional 0.4 problems correct (SD = 1.7).
Presence Outcomes

Presence contributes to the goal of transparency in technology-mediated interactions. Although there has been substantial debate on formal definitions, there is a general consensus that presence describes a user’s sense of “being there” when interacting with a mediated environment (Insko, 2003; Schubert et al., 1999). Here we report on students’ reported sense of presence while interacting in the CRYSTAL ISLAND storyworld (narrative and minimal-narrative conditions only).

Narrative had a significant effect on student presence, $F(1, 115) = 4.23, p = 0.04$. Higher presence was reported in the narrative condition ($M = 147.35, SD = 30.6$) compared to the minimal-narrative condition ($M = 136.5, SD = 25.8$). Gender was also found to have a weakly significant effect on presence, $F(1, 115) = 2.87, p = 0.09$, with females reporting higher levels of presence ($M = 146.9, SD = 26.1$) than males ($M = 137.9, SD = 30.5$). Students reporting high-levels of interest (as gauged by the interest scale modified from (Schraw, 1997)) reported higher levels of presence than students with low-levels of interest. There was a significant correlation of interest with student presence, $r(114) = 0.36, p = 0.0001$, and several of the PQ’s subscales, including: involvement/control ($r(114) = 0.42, p < 0.0001$), naturalism of experience ($r(114) = 0.27, p = 0.003$), and resolution ($r(114) = 0.29, p = 0.002$). Self-efficacy and presence also had a significant interaction. Students with high science efficacy reported higher levels of presence than less efficacious students, $r(114) = 0.35, p = 0.0001$. Likewise, students reporting greater levels of involvement and control (a PQ subscale) also reported higher science efficacy, $r(114) = .28, p = 0.002$.

Student goal orientation was found to affect presence as well. In particular, there was a significant effect of mastery approach on presence in both CRYSTAL ISLAND conditions, $F(1, 114) = 8.65, p = 0.004$, and performance avoidance on presence, $F(1, 114) = 4.59, p = 0.034$. Mastery oriented students reported greater levels of presence than performance-oriented students. Students who sought to avoid negative performance outcomes also reported higher levels of presence than students who did not seek to avoid negative performance outcomes.
10.5.4 Discussion

The experiment found that students who interacted with the CRYSTAL ISLAND environment achieved significant learning gains. While pre- to post-test performance differences were greatest in the PowerPoint condition, the findings support the hypothesis that students received clear motivational benefits from interacting with CRYSTAL ISLAND. Further, student levels of presence had significant relationships with factors relevant to learning and motivation, including self-efficacy, interest, involvement/control, and goal orientation. While learning gains were higher in the minimal-narrative condition, students reported higher levels of presence in the narrative condition, carrying promising implications for motivation.

Drawing upon the experiment’s learning gain results, it is possible that the narrative condition’s additional story content overloaded cognition, enabling students to learn more without the supplemental storyline on proximal assessment. An important direction for future work is conducting longitudinal studies to determine the long-term effects of narrative on learning and inform scaffolding strategies for reducing cognitive load.

The study found significant effects of presence, with higher reported presence in the narrative condition. With the benefits of efficacious learners having been widely demonstrated (Bandura, 1997; Zimmerman, 2000a), it is important to note that higher presence levels also lead to higher levels of reported self-efficacy. If further study can identify the narrative factors that contribute to motivation and efficacy, we can enhance the ability of NLEs to support student problem solving, increase student effort, persistence, and resilience when confronted with failure, and raise the levels of success students are likely to achieve (Bandura, 1997; Schunk and Pajares, 2002; Zimmerman, 2000a).

When considering the Involvement/Control subscale of the Presence Questionnaire (Witmer and Singer, 1998), the findings indicated that high levels of Involvement/Control are correlated with higher reports of self-efficacy. Perception of control is known to have motivational benefits (Malone and Lepper, 1987). As a factor contributing to presence, involvement/control suggests probable relationships between presence and motivation. The findings of this study highlight the potential for positive connections between narrative and motivation that deserve further investigation. Further exploration of these relationships will contribute to a deepened understanding of the
narrative factors that relate story content, presence, learning, motivation, and self-efficacy, as well as our ability to regulate these factors in an effort to support pedagogical objectives.

The study also found an effect of student goal orientation on perceptions of presence among the middle school participants. The gaming environment, on which CRYSTAL ISLAND is built, may have had an effect on performance-oriented students, encouraging them to attempt to solve the mystery quickly. Meanwhile, it seems that mastery oriented students, who tend to measure accomplishments by learning successes, reported a greater perception of presence. It is probable that mastery oriented students were more likely to take their time throughout their interactions, focusing their attention on the content of learning environment so that their presence experience was heightened.

10.5.5 Study Limitations
The experiment was designed to control for time on task, allowing 50 minutes for the intervention. As a result of this constraint and the amount of content in CRYSTAL ISLAND, only 49 of the 116 CRYSTAL ISLAND participants finished or were working on the final problem at the end of the 50 minute session. An alternative design, which will be adopted in future work, would consider controlling for task completion. Another limitation is that this study, at the time of writing, does not include a longitudinal test to assess the hypothesized benefits of narrative.

10.6 Summary of Individual Differences of Outcomes in NLEs
Given the central role of affect and motivation in cognitive processes, it is becoming increasing more important for intelligent tutoring systems to consider the affective experiences of students. In this chapter we investigated the student in the affective loop. This chapter evaluated the perception of character empathy finding that the character empathetic responses driven by CARE-induced empathy models are equal to those driven by humans. Next, student presence was evaluated in a controlled experiment designed to assess the impact of empathetic characters. Students were randomly assigned to a condition with or without empathetic characters. Students who had interacted with empathetic characters reported greater levels of presence, including higher levels of involvement and control. Then, we investigated how empathetic responses from characters
effected student affect by analyzing affective transitions. This section also began to look at individual differences in affective transitions as well. The results indicated that parallel and reactive empathy have distinctive effects for certain emotions. Lastly, the role of narrative in the learning environment was investigated in terms of student learning and student presence. The results indicate that the motivational principles of narrative lead to increases in student presence. The work reported in this chapter is useful for evaluating the merit of building computational models of affect from observations of affect in action addressing the utility of effectively modeling constructs such as student emotion, engagement and devising models of empathy and other socio-psychological constructs for pedagogical agents and characters of interactive learning environments.
Chapter 11

Conclusions

Affect plays a central role in human cognition, behavior, and interaction. It is imperative that interactive systems begin to reason in the affect dimension. Capitalizing on affective information can allow systems to plan, react, and adapt to user emotion, similar in ways that humans interact. Affective reasoning requires recognition of affect, an understanding of recognized affect, and the ability to respond and express affect.

For the task of recognizing and expressing affect we have developed an inductive framework, CARE (Computational Affect Recognition and Expression) for learning models of affect, using appraisal theory as a psychological foundation. This data-driven approach learns empirically grounded models of affect from observations of “affect in action” while students interact in a learning environment. In this approach, a training phase is used to gather observations of affect during student problem solving activities. Training data strictly employs observable features of student interactions in the environment such as actions and locations visited. Thus, the same features are available at runtime for induced model control. During training sessions, CARE collects reports of student affect. To date, this has largely been accomplished with self-report mechanisms whereby the students express how they feel. Care then uses a variety of machine learning techniques (i.e., decision trees) to induce models of affect from the observable features, making predictions about student affect. Evaluations of induced models in three interactive environments demonstrate the CARE empirical paradigm offers a promising technique for modeling student emotion, self-efficacy, metacognitive monitoring, and character empathy. It consistently outperforms baseline models.
11.1 Implications

The work presented in the preceding chapters has contributed primarily to the affective reasoning community that is now only a decade old. Contributions are situated in a number of intersected disciplines including human-computer interaction, education, educational psychology, and artificial intelligence. Below we examine several contributions stemming from this body of work.

- **CARE Framework.** We have devised an inductive framework for developing computational affect recognition and expression models founded in the appraisal theory literature. This methodology has been used to construct models to recognize student emotion, self-efficacy, and metacognitive monitoring. The CARE framework has also been used to induce models of empathy for the synthetic pedagogical agents inhabiting our learning environments.

- **Inductive Approach.** In addition to the framework itself, the inductive approach is novel; using student situational data to learn models that predict student affect. The approach has been adapted for the tasks of affect recognition and affects expression utilizing data obtained from training sessions in which observations of the psychological construct of interest (i.e., emotion, self-efficacy, etc.) are collected in the act of student psychological experience.

- **Affect Recognizers.** The affect recognizers we have constructed are derived from observations of affect in action. Through training sessions we collect situational data as students solve problems and, commonly, self-report on the psychological construct of interest. From this situational data machine learning is utilized to construct models capable of predicting student affect for informing narrative and pedagogical control modules in our learning environments.
  - **Emotion recognition.** The CARE framework has been used to induce models of student emotion. Our experimental set of emotions changed over the course of empirical evaluations as dictated by ongoing in research, but at one time or another consisted of a subset of anger, anxiety, boredom, excitement, fear, flow, frustration, happiness, and sadness. The affect recognizers developed in Chapter 9 focused on
predicting an emotion from a set of candidate emotions and even focused on determining the presence of single emotions, i.e., frustration.

- **Self-efficacy recognition.** We have also used the CARE framework to induce models of self-efficacy. In these empirical investigations we obtain observations of student efficacy through self-reports. Using the CARE framework we then make predictions regarding levels of student efficacy (i.e., LOW vs. MEDIUM vs. HIGH). These models are able to determine whether students are confident in their abilities in relation to the demands of current learning tasks.

- **Metacognitive monitoring recognition.** The CARE framework has most recently been used to model student metacognitive monitoring. In much the same way that we obtained models of emotion and self-efficacy we too construct models of metacognitive monitoring affording our learning environments to diagnose metacognitive judgments from students in runtime.

- **Affect Expressers.** The affect expressers we have constructed are derived from observations of affect in action, or empathy in action for these cases of affect expression. In training sessions, similarly designed to sessions intended for affect recognition the CARE framework collects observations of empathy in action with unfolding student situations. From this situational data we use machine learning techniques to induce models for informing or driving character behavior modules.

  - **Empathy models.** The CARE framework has been used to induce models of empathy for synthetic characters in both roles as a companion and pedagogical agent. Our first approach to modeling empathy resulted in the construction of two complimentary models. One for empathetic assessment (when to be empathetic) and one for empathetic interpretation (when empathy is called for, what affective state should be expressed). Empathy models have been used to drive the affective expressions of characters inhabiting our learning environments and have been evaluated in a number of empirical investigations.

  - **Parallel and reactive empathy models.** Recently have investigated approaches to extend our models of empathy to account for both types, parallel and reactive,
empathy. These models determine when parallel or reactive empathy has a greater utility for responding to a student’s current situation. Through empirical investigation such models of been induced. Additional analyses found that there are significant differences in the impact parallel and reactive empathy may have on student affect, including instances where empathy may even be inappropriate signaling other pedagogical intervention strategies.

- **Real-time and early methods for predicting affect.** The affect recognizers and expressers described in chapters 6-9 are capable of functioning in real-time. We have investigated two modeling approaches: one that solely utilizes students’ current situational contexts for prediction and those that make early predictions for future affective experiences.

- **Empirical studies.** We conducted a number of empirical studies to investigate the CARE framework itself, assess the merit of deploying affect-sensitive learning environments, and to evaluate aspects of the learning environment itself. The studies have been conducted with over 700 participants ranging in age from 12-60, including middle school, high school, college undergraduates, and college graduate students.

### 11.2 Future Work

The work described in the preceding chapters has uncovered a number of directions for future investigations. We examine each below.

- **Generalized Modeling.** The observational attribute vector is comprised of a number of features that are directly observable in Care training and runtime scenarios. Such features include student actions (i.e., picking up an egg) and visited locations (i.e., the infirmary). The Care-induced models described in the preceding chapters all make use of domain-specific features. The low-level nature of the observational attribute vector makes it virtually impossible for induced models to be ported to other interactive systems. It is even difficult for induced models to be folded into new versions of interactive systems, due to the level of domain-specificity. While outside the scope of this work, future investigations should include evaluation of models induced from varying levels of domain-specific features.
Certainly, there will be a tradeoff consideration in model performance for varying levels. However, if successful, such models could be portable to new versions of interactive systems and perhaps across domains. Research in modeling affect with domain independent features should rely heavily on psychological literature surrounding appraisal patterns. In interesting avenue of investigation here is to explore how the psychological top-down approach appraisal patterns match those derived for a computational, bottom-up approach.

- **Extending the CARE Framework with further aspects of Appraisal Theory.** We have utilized a somewhat limited view of appraisal theory as a psychological foundation for the Care framework. Largely, Care relies on information obtained from the interactive learning environment, which coincides with appraisal theories, situation construal. Care is able to monitor changes in the situation construal, but lacks a full representation of the environment. Appraisal theorists account for one’s goals, beliefs, intentions, etc. as part of the environment. This information is the focus of ongoing research in the artificial intelligence community, exploring approaches to modeling beliefs, desires, and intentions, as well as recognizing user goals and plans. Incorporating mechanisms for gaining a full representation of the person-environment relationship is a promising direction of future work. Such research could examine the relationship of user coping strategies and their effects on the person-environment relationship enabling interesting applications, such as scaffolding the user appraisal and coping processes.

- **Extending physiological instrumentation and measurements.** In some studies we have utilized measurements from sensors monitoring galvanic skin response and blood volume pulse. This is a small subset of currently available physiological instruments. Other researcher teams are exploring the merit of employing such physiological instruments as eye tracking, facial feature tracking, seat and back posture, gesturing, temperature, EEG, etc. As these devices continue to become more readily available and more practical for fielded applications it will be essential to understand the impact each has on the performance of affect modeling.
• **Apply CARE paradigm and appraisal theory to modeling of other constructs.** Beyond modeling emotion we have used the Care inductive framework for modeling self-efficacy, metacognitive monitoring, and empathy. The success of Care induced models relies largely on the representation of the situational environment and the relation of changes in the situation to the construct of interest. In addition to the constructs modeled in the preceding chapters, there are numerous other psychological constructs that rely on student appraisal of the environment and their relationship to the unfolding situation. Thus, Care holds much potential for modeling those psychological constructs where appraisal processes are pertinent.

• **Investigate self-reporting mechanisms.** The Care framework requires observations of affect in action where, during learning episodes, we are able to obtain reports or readings of the construct of interest. Largely, we have relied on self-reports, whereby students answer an in-game question regarding how they feel. We have used self-reporting constraints where dialog boxes appear at set intervals querying students to select an emotion related to how they feel, or to enter a number between 0 and 100 to indicate how confident they are that they can achieve the goal they are attempting to achieve. In addition to the self-report observations, we have also collected physiological data and video recordings of student faces during learning episodes. Future work should consider how this data can be used as either a validation methodology to verify self-reported data or serve as class labels for the machine learning processes themselves. There may be a tradeoff in modeling a student’s self-reported affect compared to models of the true emotion in terms of how to deliver feedback. An interesting question is to investigate how pedagogical agents should respond in an instance perhaps were a model of self-reported emotion differs from a model of true emotion, learned from the face data. Lastly, future work should consider how self-reports of affect our obtained. A periodic report dialog box which appears out of context may receive very different reports from affect reports obtained directly from character inquiries.

• **CARE models for generating computational models of Self-Regulated Learning.** Above we noted the possibility of extending the Care inductive framework to model a variety of other
psychological constructs that may depend on appraisal processes. In Chapter 8 we focused on modeling metacognitive monitoring. This is merely one component of self-regulation. To effectively scaffold experiences designed to create regulated learners the environment will have to make accurate predictions of a number of self-regulation components, metacognitive monitoring being one of them. Future work should consider investigations to model other self-regulated components exploring the potential of accurately modeling self-regulated components and also determining which components are necessary to model such that environments can effectively adapt to learners.

- **Role of Individual differences and how to account for such differences.** In Chapter 10 we explored trends in affective transitions accounting for student differences along a number of dimensions including gender, personality, and goal-orientation. These results are but one example of the merit of accounting for student individual differences. Future work should investigate along what dimensions do students significantly differ, how can these differences be effectively noted and obtained (survey methods, self-reporting, prediction models, etc.), how might pedagogical strategies differ given student individual differences, and which, if any, individual differences merit attention for intelligent tutoring systems.

- **Persona effect investigations.** An interesting avenue for future work that was outside the scope of the preceding chapters is consideration of agent persona. Particularly, agent persona should be explored in future empathy studies. It seems plausible that the agents persona (i.e., appearance, pedagogical role, narrative role, language, body posture, etc) could impact the perception of empathy as well as student self-reports of affect obtained through character inquiry. Intelligent tutoring systems may be able to capitalize of persona if particular characters can be determined to be better facilitators for learning and are more accurately perceived along tutorial, narrative, and affective dimensions. Agent persona has not been neglected by the ITS community, but a significant extension to the work described here would be consideration of agent persona.

- **Wizard-of-Oz Interface extensions.** In one of the studies investigating the Care paradigm for modeling empathy, human wizards participated as training empathizers, selecting
empathetic responses for a companion agent to express. Care-induced models used observable features to make predictions of when these wizards chose to be empathetic and how (what emotion) they chose to have the companion agent respond to the student user. The wizards display was a view of the student user’s display from a slight distance behind the user’s character. The display was intentionally vague in hopes that the wizards would rely on many of the same features Care was tracking for model induction. An interesting direction of future work is to consider how and what additional information might merit displaying to the wizard. For instance, we have utilized biofeedback equipment in several studies and many other groups are investigating numerous other devices for monitoring changes physiology. How might this information be useful to wizards and how should it be displayed? Also, we could summarize some of the features being monitored by Care and make them explicit to the wizard. For instance, we could display how much time a student has been working on a particular problem or spent in a particular location. Using a Care-induced affect recognition model we may also be able to display predictions about how the student feels. Beyond representing and displaying features of the student and the environment that Care may use for inducing affect models future work should also consider additional wizard controls. In our wizard-of-oz style studies the controls available to the wizard have been limited to affective control of a companion agent; controlling only the timing and type of empathetic response of the companion agent. Advanced wizard interfaces might also consider giving control of agent communication to wizards. For instance, wizards may prefer to indicate subtle emotion expressions with simple, non-verbal gestures while other instances may call for explicit verbal communication including the likes of tone and word-choice. Expanding wizard controls will enable Care to induce far richer models of empathy enabling expansive control of character behavior.

- **Combine with goal and plan recognition systems.** We have already mentioned the role goal and plan recognizers in extending the Care framework to more closely resemble person-environment relationship of appraisal theory. The capability of recognizing user goals and plans has significant potential to improve the abilities of interactive systems to predict appraisal outcomes from pairings of how the situation construal matches the goals
and plans of the user. Likewise, there may be a role for affect recognition in the process of recognizing user goals and plans. It seems reasonable that an interactive system may be able to disambiguate between competing actions, goals, or plans through knowledge of user affect. Is one likely to engage in different actions, or pursue different goals, or even change her plan depending on whether she is frustrated, confident, or confused? Some believe the answer is “perhaps” making this in interesting direction for future exploration.

- **Combine with Dialogue systems.** In chapters 9 and 10 our research utilized empathetic characters. In chapter 9 we used Care to induce models of empathy while in chapter 10 we explored the merit of employing empathetic characters in a narrative learning environment. The first study in chapter 9 relied on pre-scripted, spoken empathetic responses, while the other studies delivered pre-scripted, text-based empathetic responses. The opportunity to support dialogue systems by providing affective information such as student emotion and guidance of how characters might respond holds much potential. Future work should consider the utilization of induced models for such purposes are delivering affective context to dialogue systems informing the management, understanding, and synthesis of dialog between pedagogical characters and students.

- **Extension to other domains and applications.** This work has been explored in a limited set of problem-solving activities leaving one to question how the approach might be applied and the results contrasted in other settings across domains and applications. Because the Care framework utilizes a training phase the utility of the approach drops in applications of risk to the user, where requesting trail experiences in which observations of affect can be obtained may be too risky and even not probable. In such situations it is apparent that the Care approach will not work. Future work should consider other feasible domains and applications in which a training phase is practical and low-cost.

- **Extensive laboratory and in vivo experimentation.** While the studies reported in the preceding chapters have enlisted the participation of X subjects much research is required to understand the limits of the results. Exploring various other populations may reveal contrasted patterns in affect models necessitating investigations to account for population
and individual differences. Additionally, much of this work has not been concerned with the deployment of fielded applications and the issues surrounding systems installed in classroom settings. Future work should investigate whether the results obtained in controlled experimental settings can be replicated in vivo experimentation.

- **Privacy.** As the socio-psychological domain continues to merge with technical domains, like computer science, increased consideration of privacy is warranted. What if I don’t want a computer or application, such as a learning environment, to know how I feel or what I believe about myself? As technologies continue to evolve into affect-sensitive products, the technologies may be forced to open up themselves and reveal what information about the user is being utilized and how. This issue not only focuses on the rights of the person and the technology but the relationship between the two entities including trust, rapport, and other social constructs.

### 11.3 Concluding Remarks

The capability to recognize, understand, respond to, and express affect holds significant potential for improving the quality of interaction in interactive systems. In interactive learning environments, there is potential to create effective learning experiences through adaptations that account for the likes of student emotion and efficacy. The CARE framework is a promising methodology for constructing models of affect. The approach can be applied to interactive systems in a variety of domains and has been successfully demonstrated in interactive learning environments. This body of work offers a foundation for further exploring computational models of affect in interactive learning.
References


Beal, C., and H. Lee (2005) Creating a Pedagogical Model that Uses Student Self Reports of Motivation and Mood to Adapt ITS Instruction. Workshop on Motivation and Affect in Educational Software. 12th International Conference on Artificial Intelligence in Education.


Fleiss, J. L., Measuring nominal scale agreement among many raters, Psychological Bulletin, 76, 5, pp 378-382 (1971)


243


Appendix A

Modeling Physiological Response

Because physiological responses are directly triggered by changes in affect, biofeedback data such as heart rate and galvanic skin response can be used to infer affective changes. However, biofeedback hardware is intrusive and cumbersome in deployed applications. To this end we have also investigated the merit of utilizing the CARE inductive framework for automatically learning models of users’ physiological response from observations of user behaviors in interactive environments. These models can be used at runtime without biofeedback hardware to continuously predict users’ physiological state directly from situational context in the interactive environment. Empirical studies with induced decision tree, naïve Bayes, and Bayesian Network physiological response models suggest that they may be sufficiently accurate for practical use.

Affective reasoning offers much promise for interactive technologies. One can imagine an “affective barometer” (Picard, 1997) that could be incorporated into interactive environments to create customized experiences that are optimally stimulating and maintain ideal levels of engagement for individual users. Of course, the barometer would need to detect changes in affective state since users’ emotions vary widely. For example, a player can be surprised, relaxed, fearful, frustrated, bored, and excited within a single gaming session. Changes in affective state are accompanied by physiological responses such as changes in heart rate, respiration, temperature, and perspiration (Frijda, 1986). Thus, detecting emotions is typically accomplished by “wiring” users, i.e., attaching biofeedback devices that monitor their physiological state changes. Although biofeedback has been demonstrated to accurately predict users’ affective state (Conati, 2002; Picard et al., 2001), it is intrusive and cumbersome to use in practice. The limited feasibility of employing a biofeedback apparatus in a deployable system calls for an alternate approach to modeling users’ physiological states that leaves them untethered. Because physiological state is so closely
associated with affective state, an accurate model of physiological response could enable interactive environments to effectively reason about users’ affective states, their level of stress, and their level of interest to craft customized interactions that are appropriately stimulating and engaging.

This appendix presents an inductive approach to modeling users’ physiological response in interactive environments. The CARE framework learns empirically informed models of physiological response from observations of user interactions in an interactive virtual environment. During training sessions, users are outfitted with biofeedback sensors. CARE monitors both situational data, including locational, intentional, and temporal information, and physiological data, including heart rate and galvanic skin response as users direct their characters to perform a sequence of tasks. CARE then induces models of physiological response from situational data. These models can be used without biofeedback hardware to continuously predict a user’s physiological state directly from her situational context in the virtual environment. Empirical studies of CARE indicate that it can induce models of physiological response that appear to be sufficiently accurate for practical use.

A.1 Modeling Physiological Response

Because users’ physiological responses follow directly from their affective states, accurate models of physiological response could be used to enable interactive environments to effectively determine users’ level of interest, stress, and emotion in order to guide customized interactions. However, for practical purposes, these determinations should ideally be made without resorting to the use of invasive biofeedback devices. We propose using the CARE framework that first acquires situational and physiological data from gameplay and biofeedback hardware and then learns models of physiological response from this training. CARE’s induced models can then be used at runtime to predict users’ physiological response directly from situational data without requiring biofeedback monitoring.

Recall that the CARE framework operates in two modes, model induction and model usage. During model induction CARE acquires training data and learns models of physiological response from training users interacting with the interactive environment. The training user is outfitted with biofeedback equipment which monitors her heart rate and galvanic skin response. Biofeedback
signals are recorded in training logs via the interactive environment, which also records an event stream produced by the training users’ behaviors in the environment. Together, the biofeedback signals and the corresponding elements in the event stream are assembled in temporal order into the observational attribute vector. After training sessions (typically involving multiple training users) are complete, the physiological response learner induces models from the observed situational data and physiological data. The physiological data serves as class labels for the training instances. During model usage, which is the mode used in runtime interactive environments to be deployed, the induced models inform the decision making of CARE-enhanced runtime components by predicting end users’ physiological responses. Examples of candidate CARE-enhanced runtime components include NPC behavior controllers for games, narrative planners for interactive story worlds, and tutorial planners for intelligent tutoring systems.

A.2 Evaluation

A.2.1 Method

In a formal evaluation, data was gathered from 20 subjects in an Institutional Review Board (IRB) of NCSU approved user study. There were 11 female and 9 male participants. Participants average age was 21.6 (SD = 2.96).

A.2.2 Procedure

First participants completed and reviewed pre-experiment materials consisting of a demographic survey, Half-Life 2 controls reference sheet, and a controlled backstory in preparation for interacting within the environment. The pre-experiment phase also contained a practice task from the game Half-Life 2 presenting an opportunity for training users to become familiar with the controls. Participants were then outfitted with a biofeedback apparatus for the experiment. The experiment consisted of two 3D Treasure Hunt virtual environments, each of varying degrees of difficulty. The easiest version of Treasure Hunt offered many opportunities to find treasures and meet the expectations that were set in the backstory. The most challenging version of Treasure Hunt made it difficult to find treasures; there were fewer treasures worth less value and more occluded treasure
boxes making it difficult to meet backstory expectations. Participants collected as many treasures as possible in the allotted 7 minutes. The post-experiment materials consisted of a survey about the training user’s experience and opinions on affect in applications such as games.

A.2.3 Results

All models were evaluated using a tenfold cross-validation scheme for producing training and testing datasets. In this scheme, data is decomposed into ten equal partitions, nine of which are used for training and one used for testing. The equal parts are swapped between training and testing sets until each partition has been used for both training and testing. Tenfold cross-validation is widely used for obtaining a sufficient estimate of error (Witten and Frank, 2005).

Cross-validated ROC (Receiver Operating Characteristic) curves are useful for presenting the performance of classification algorithms for two reasons. First, they represent the positive classifications (true positives), included in a sample, as a percentage of the total number of positives along the vertical axis, against the negative classifications (false positives) as a percentage of the total number of negatives (Witten and Frank, 2005). Second, the area under ROC curves is widely been accepted as a generalization of the measure of the probability of correctly classifying an instance (Hanley and McNeil, 1982).

The ROC curves (Figure A.1) show the results of Bayesian networks, decision tree and naïve Bayes model results for predicting physiological response. The smoothness of the curves indicates that sufficient quantities of data were used to induce PRP models and that there was adequate coverage of possible instances. The highest performing models induced from Treasure Hunt training data was the decision tree models in both classification of heart rate and galvanic skin response, accurately predicting more than 90% of changes in physiological response.
A.2.4 Discussion

The results of the experiment suggest that the CARE framework can support the automated induction of accurate physiological response prediction models. It is interesting that the accuracy levels were surprisingly high, both for heart rate and GSR (changes in skin conductivity) prediction. The naïve Bayes and Bayesian network models both performed reasonably well, and the decision tree model performed particularly well. It seems that the high performance of the decision tree classifier was perhaps influenced by the fact that the data available for learning was voluminous. Decision trees seem to perform well on tasks that can furnish very large data sets, and physiological response modeling is such a task. Drawing strong conclusions such as, “Decision trees are superior to naïve Bayes or Bayesian networks for physiological response prediction,” is not supported by a single experiment such as the one reported here. In the Treasure Hunt experiment, both the conditional probabilities and the structure of the Bayesian network were learned automatically from the data. The results were similar to those for naïve Bayes, but it seems possible with the incorporation of additional domain knowledge, a Bayesian network enhanced by a domain expert might achieve stronger results.1

---

1 The authors also investigated Bayesian networks enhanced with hidden variables suggested by a domain expert. Because of technical limitations associated with learning conditional probabilities for Bayes nets with real-valued variables and hidden variables, success with “authored” Bayesian networks for physiological response prediction has to date been limited.
In the study, binary labels (down and up) were induced. Thus, given an environmental context, the model predicts whether the user’s heart rate will increase or decrease and whether the user’s skin conductivity will increase or decrease within a predefined temporal window. Binary labels were chosen because of their simplicity and because of the potential increases in prediction accuracy that might arise from reduced granularity. One can imagine fine-grained labels, e.g., (down, stable, up) or (large decrease, small decrease, no change, small increase, large increase).\(^2\) It could be argued that fine-grained physiological reaction models could be beneficial. However, it appears that a CARE-enhanced runtime component can function effectively with coarse grained class labels because predictions will be made continuously and it is the trends in physiological change rather than an isolated prediction that seem to be the most informative for decision making. For example, a sequence of 35 up values in heart rate clearly suggests that the user is experiencing an increase in stimulation.

Inspection of the data reveals that there were many repeated instances in the data, i.e., multiple users reacted similarly in identical situations. In the experiment reported here, although the interaction data were gathered from 20 different training users, and they were permitted flexibility in exploring the environment, there is significant regularity in the training data. This training data regularity stems from regularity in the environment itself, regularity in how users explored and interacted with the environment, and regularity with how users reacted physiologically to events in the environment.

One possible explanation for the results is that physiological response patterns are common across users, i.e., there are inherent physiological similarities between people. If this is true, then physiological response learning approaches such as that embodied by the CARE framework are promising. An alternative (and complementary) explanation is that lower levels of predictability might hold for a more diverse set of training users with virtual environments that are more expansive, complex and dynamic. This could well be the case and is an interesting direction for future work. Nonetheless, it seems likely that to a great extent, people react in predictable patterns that can be automatically learned from observation.

\(^2\) In fact, another set of experiments were run in which ternary rather than binary class labels were used. Accuracy rates were only slightly less than those for binary class labels, e.g., 88% for the decision tree GSR model.
A.3 Conclusion

Dynamically crafting interactive experiences that are highly customized for individual users is a long-term goal of digital entertainment, education, and training. If interactive environments can be given the ability to accurately reason about users’ affective characteristics, they can tailor user’s experiences to moment-by-moment changes in their levels of engagement, interest, stress, motivation, and emotional state. Physiological responses follow directly from changes in affect and thus can be used as key predictors of affective state. Although biofeedback devices can be used to obtain actual physiological signals, it may be impractical to require users to don biofeedback equipment and deploy additional hardware with applications. The CARE framework can automatically induce models of users’ physiological responses that can predict physiological changes from observable events in interactive environments. It appears that the accuracy of CARE induced models may be sufficiently high to facilitate the control of users’ experience in the virtual worlds that increasingly dominate gaming, education, and training.

This works represents a first step toward deployable affective reasoning for interactive environments. It will be interesting to explore the predictive capabilities of induced models in more complex, dynamic environments. Will different types of models be required? Can complementary affective constructs such as immersion, confidence, and fantasy be induced from observable elements in these environments? How can models such as these be incorporated into next-generation entertainment and learning environments? These questions suggest important directions for future work.