Considering Alternate Futures to Classify Of-Task Behavior as Emotion Self-Regulation: A Supervised Learning Approach

JENNIFER L. SABOURIN, JONATHAN P. ROWE, BRADFORD W. MOTT, AND JAMES C. LESTER

Over the past decade, there has been growing interest in real-time assessment of student engagement and motivation during interactions with educational software. Detecting symptoms of disengagement, such as off-task behavior, has shown considerable promise for understanding students' motivational characteristics during learning. In this paper, we investigate the affective role of off-task behavior by analyzing data from student interactions with CRYSTAL ISLAND, a narrative-centered learning environment for middle school microbiology. We observe that off-task behavior is associated with reduced student learning, but preliminary analyses of students' affective transitions suggest that off-task behavior may also serve a productive role for some students coping with negative affective states such as frustration. Empirical findings imply that some students may use off-task behavior as a strategy for self-regulating negative emotional states during learning.

Based on these observations, we introduce a supervised machine learning procedure for detecting whether students' off-task behaviors are cases of emotion self-regulation. The method proceeds in three stages. During the first stage, a dynamic Bayesian network (DBN) is trained to model the valence of students' emotion self-reports using collected data from interactions with the learning environment. In the second stage, a novel simulation process uses the DBN to generate *alternate futures* by modeling students' affective trajectories as if they had engaged in fewer off-task behaviors than they did during their actual learning interactions. The alternate futures are compared to students' actual traces to produce labels denoting whether students' off-task behaviors are cases of emotion self-regulation. In the final stage, the generated emotion self-regulation labels are predicted using off-the-shelf classifiers and features that can be computed in run-time settings. Results suggest that this approach shows promise for identifying cases of off-task behavior that are emotion self-regulation. Analyses of the first two phases suggest that trained DBN models are capable of accurately modeling relationships between students' off-task behaviors and self-reported emotional valence in CRYSTAL ISLAND. Additionally, the proposed simulation process produces emotion self-regulation labels with high levels of reliability. Preliminary analyses indicate that support vector machines, bagged trees, and random forests show promise for predicting the generated emotion self-regulation labels, but room for improvement remains. The findings underscore the methodological potential of considering alternate futures when modeling students' emotion self-regulation processes in narrative-centered learning environments.

1. INTRODUCTION

One-on-one, face-to-face human tutoring has long been considered the gold standard for effective instruction because of its significant pedagogical benefits compared to traditional classroom lectures [Bloom, 1984]. The benefits of expert tutoring have been hypothesized to derive from increased levels of feedback and scaffolding, including support for knowledge self-repair and increased rates of problem completion [VanLehn, 2011]. While it is infeasible for every student to have his or her own personal human tutor, the intelligent tutoring systems community has sought to bridge this gap by endowing educational software with the pedagogical capabilities of expert human tutors

[VanLehn, 2006; VanLehn, 2011]. These efforts have led to significant improvements in educational software's effectiveness, and in some cases intelligent tutoring systems have produced learning gains comparable to human tutors [Bloom, 1984]. However, as with any educational tool, student learning is dependent on how effectively the software is used. Even advanced educational software will not help students who fail to use it properly because they are disengaged.

Students demonstrate a broad range of learning behaviors when interacting with intelligent tutoring systems [Aleven, McLaren, Roll, & Koedinger, 2006; Beal, Mitra, & Cohen, 2007; Bunt & Conati, 2003; Graesser, Person, & Magliano, 1995]. In particular, there has been growing interest in how student motivation affects learning and problem solving. This line of research has raised important questions about how and why students disengage from educational software, as well as the cognitive impacts of disengagement [Baker, Corbett, Koedinger, & Wagner, 2004; Beal, Qu, & Lee, 2006; Muldner, Burleson, Van de Sand, & VanLehn, 2010]. Disengagement can take a variety of forms, including hint abuse [Aroyo et al., 2007; Beal et al., 2006], off-task conversation [Baker, 2007] and gaming the system [Baker et al., 2004a]. In general, students who abuse or disengage from an intelligent tutoring system learn less effectively than students who do not disengage [Baker et al., 2004; Cocea, Hershkovitz, & Baker, 2009; Gong, Beck, Heffernan, & Forbes-Summers, 2010]. Consequently, a growing body of research has investigated techniques for automatically detecting and preventing harmful learning behaviors such as gaming the system [Baker, Corbett, Roll, & Koedinger, 2008; Beal et al., 2006; Beal et al., 2007; Cetintas, Luo, Xin, Hord, & Zhang, 2009].

Recent work investigating off-task behavior and student emotion has begun to raise questions about whether off-task behavior is universally unproductive for learning [Baker et al., 2011]. On the one hand, empirical findings suggest that off-task behavior is associated with boredom, which has been shown to be harmful for learning [Baker, D'Mello, Rodrigo, & Graesser, 2010]. On the other hand, recent findings have suggested that going off-task may alleviate negative affect, which could in turn benefit learning [Baker et al., 2011; Sabourin, Rowe, Mott, & Lester, 2011]. A plausible explanation is that some students use off-task behavior as a coping strategy for negative learning emotions. Along these lines, there is evidence that some students may be more effective at regulating their emotional states than other students [Meyer & Turner, 2006].

Furthermore, some types of off-task behavior have been observed to be more harmful than others [Baker et al., 2004a].

These observations highlight the importance of distinguishing between off-task behavior that is associated with positive affective and learning outcomes and off-task behavior that is unproductive. By employing models that distinguish between unproductive off-task behavior and productive off-task behavior, intelligent tutoring systems can utilize pedagogical strategies that appropriately respond to student disengagement and hypothetically yield both improved affective and learning outcomes.

We investigate the affective role of off-task behavior by analyzing data from student interactions with CRYSTAL ISLAND, a narrative-centered learning environment for middle school microbiology. We find that off-task behavior has an overall detrimental effect on student learning, but observations of students' affective transitions suggest that off-task behavior may serve a productive role for some students coping with negative affective states such as frustration. In some cases, we find preliminary evidence that students use off-task behavior as a strategy for self-regulating their emotional states during learning.

Based on these findings, we present a supervised machine learning procedure for detecting whether students' off-task behaviors are cases of emotion self-regulation. The method proceeds in three stages. During the first stage, a dynamic Bayesian network (DBN) is trained to predict the valence of students' emotion self-reports during interactions with the learning environment. In the second stage, a novel simulation process uses the trained DBN to generate hypothetical affective trajectories from modified student interaction data. The generated affective trajectories are *alternate futures* that might have occurred had the student performed fewer off-task behaviors after each self-report. The alternate futures are modeled under an assumption that each student would perform fewer off-task behaviors under the influence of an "optimal" tutor. The hypothetical affective trajectories are compared to students' actual affective trajectories in order to determine whether students' off-task behaviors impacted their emotional states. These comparisons are used to generate binary labels for each interval of off-task behavior, indicating whether the off-task behavior was emotionally beneficial (i.e., self regulatory) or unproductive. In the final stage, off-the-shelf classification techniques are employed to predict the emotion self-regulation labels. This supervised learning analysis exclusively uses features that would be available in run-time settings, unlike the DBN which does not adhere to run-time constraints.

Findings indicate that this novel procedure shows promise for identifying cases of off-task behavior involving emotion self-regulation. An empirical analysis of the procedure's first phase demonstrates that DBNs can accurately and reliably predict the valence of students' emotion self-reports using data from the CRYSTAL ISLAND learning environment. Furthermore, we observe that the alternate future simulation process produces reliable labels denoting whether students' off-task behaviors are cases of emotion self-regulation. A preliminary investigation of the procedure's third stage suggests that support vector machines, bagged trees, and random forests have potential for improving the precision of emotion self-regulation detectors with modest tradeoffs in recall, but room for improvement remains. The results provide initial support for the proposed method of identifying off-task behaviors as cases of emotion self-regulation. The findings also highlight the potential of continued investigation into educational data mining methods that employ alternate futures.

The paper is organized as follows. Section 2 describes related work investigating off-task behavior in intelligent tutoring systems. Section 3 discusses background on narrative-centered learning environments, as well as issues related to off-task behavior in these environments. Section 4 describes CRYSTAL ISLAND, a narrative-centered learning environment for middle school microbiology. Sections 5 and 6 describe a study that was conducted with middle school students using CRYSTAL ISLAND, as well as initial findings investigating off-task behavior and student affect in the environment. Section 7 describes the modeling procedure that was used to generate labels denoting whether off-task behaviors were unproductive or cases of emotion self-regulation. Section 8 reports empirical findings about using the proposed method to identify cases of emotion self-regulation. Section 9 discusses the findings, as well as limitations of the work. Section 10 concludes the paper with a description of future directions.

2. OFF-TASK BEHAVIOR IN INTELLIGENT TUTORING SYSTEMS

Initial work examining off-task behavior in intelligent tutoring systems has distinguished between several ways that students can go off-task, including off-task solitary behavior, off-task conversation, inactivity, and gaming the system [Baker et al., 2004a]. *Gaming the system* occurs when a student exploits the features of educational software (e.g., hint requests) in order to make progress without learning the requisite concepts. Empirical analyses have indicated that off-task behavior generally leads to lower aggregate learning

outcomes. However, a series of studies found that only gaming the system correlated negatively with learning outcomes among the off-task behaviors listed above [Baker et al., 2004a; Cocea et al., 2009]. Additionally, analyses have suggested that *type of off-task behavior* is more useful for predicting learning than *total time spent off task*. Cocea et al. [2009] compared gaming the system behaviors against other off-task behaviors. They found that the two categories had different impacts on learning, including learning specific concepts as well as overall learning on a subject. Their findings were reproduced by Gong et al. [2010]. Further work has indicated that the harmful impacts of gaming the system may not be universal [Baker et al., 2008a]. There is evidence that some students are able to game the system without negatively impacting their learning gains. However, it remains unclear what factors distinguish these students from others.

These observations have coincided with several efforts to devise models for detecting and responding to off-task behaviors in intelligent tutoring systems. For example, work by Beal et al. [2007] examined hidden Markov models (HMMs) for classifying student engagement levels. They distinguish between four categories of students: sustained-high engagement, sustained-medium engagement, increasing engagement, and sustained-low/decreasing engagement. These clusters imply that students experience different types of engagement patterns during their learning interactions. Additional analyses suggested that the sustained-low/decreasing engagement cluster was associated with low-achievement students, corroborating other findings regarding off-task behavior.

Machine-learning techniques have been investigated for detecting and intervening in gaming behavior [Baker, Corbett, & Koedinger, 2004; Baker et al., 2006; Baker et al., 2008a]. Detecting gaming the system involves discriminating between learning behaviors that are exploitive and learning behaviors that are standard use. Baker et al. [2008a] used a latent response model to represent relationships between gaming behavior, student actions, and learning outcomes. The model accurately distinguished between non-gaming, harmful-gaming and non-harmful gaming students at a level significantly better than chance. The model was then used to inform an adaptive gaming intervention system.

Non-adaptive approaches for reducing gaming behavior (e.g., adding a fixed delay during which students cannot ask for help [Murray & VanLehn, 2005]) have generally proved to be ineffective [Baker et al., 2006]. One proposed explanation is that fixed approaches harm non-gaming students and lead to the development of advanced gaming

behaviors in other students. Rather than modify the design of an entire tutoring system, individual cases of harmful gaming behavior can be addressed directly by employing accurate and reliable detection models. Scooter the Tutor detects harmful gaming the system behaviors and intervenes with supplementary learning activities and messages of displeasure [Baker et al., 2006]. Scooter the Tutor was compared to a baseline system without adaptive gaming interventions, and a marginal reduction in the frequency of gaming behaviors was observed among students who interacted with Scooter the Tutor. Despite this reduction in gaming, there was no significant impact on student learning gains.

Recent work has examined emotional components of off-task behavior. Baker et al. [2010; 2011] and Rodrigo et al., [2007] have investigated affect and off-task behavior in several computer-based learning environments. Observers noted whether students were in one of seven emotional states: *boredom*, *confusion*, *frustration*, *engaged concentration*, *surprise*, *delight* and *neutral*. The study found that students were most likely to be *bored*, *confused*, or *frustrated* while performing gaming the system behaviors. Additionally, *boredom* and *confusion* were significant predictors of students' future gaming behavior. This same line of research has begun to investigate whether off-task behavior is an effective strategy for regulating specific negative affective states in intelligent tutoring systems. Baker et al. [2011] found evidence that off-task behavior may reduce subsequent levels of *boredom*. These findings, combined with evidence that some types of off-task behavior are harmful for learning and some types are not harmful [2004a; 2008a], suggest that off-task behavior may not be as universally detrimental to learning as was previously believed.

Recent work has examined emotional components of off-task behavior. Baker et al. [2010; 2011] and Rodrigo et al. [2007] investigated affect and off-task behavior in several computer-based learning environments. Observers noted whether students were in one of seven emotional states: *boredom*, *confusion*, *frustration*, *engaged concentration*, *surprise*, *delight* and *neutral*. The study found that students were most likely to be *bored*, *confused*, or *frustrated* while performing gaming the system behaviors. Additionally, *boredom* and *confusion* were significant predictors of students' future gaming behavior. This same line of research has begun to investigate whether off-task behavior is an effective strategy for regulating specific negative affective states in intelligent tutoring systems. Baker et al. [2011] found evidence that off-task behavior may reduce subsequent

Fig 1. CRYSTAL ISLAND narrative-centered learning environment.

levels of *boredom*. These findings, combined with evidence that some types of off-task behavior are harmful for learning and some types are not harmful [Baker, 2004a; Baker et al., 2008a], suggest that off-task behavior may not be as universally detrimental to learning as was previously believed.

3 NARRATIVE-CENTERED LEARNING

Off-task behavior has been the subject of growing attention in traditional intelligent tutoring systems, but less is known about off-task behavior in narrative-centered learning environments. Narrative-centered learning environments are a class of game-based learning environments that embed subject matter and problem solving within interactive story scenarios [Aylett, Louchart, Dias, Paiva, & Vala, 2005; Marsella, Johnson, & Labore, 2000; Rowe, Shores, Mott, & Lester, 2011]. This approach aims to foster student engagement and capitalize on individuals' inherent abilities to interpret, recall, and reason about narrative structures [Bruner, 1990; Mandler & Johnson, 1988; O'Neill & Riedl, 2011]. Narrative-centered learning environments have been observed to yield positive learning, problem solving, and engagement outcomes [Hickey, Ingram-Goble, & Jameson, 2009; Ketelhut, 2007; Rowe et al., 2011].

While narrative-centered learning environments offer several attractive qualities, they must be carefully designed to avoid narrative elements that detract from learning. Narrative-centered learning environments often provide rich interactive environments, realistic physics, and engaging characters, but these same features can introduce *seductive details* [Rowe, McQuiggan, Robison, & Lester, 2009; Harp & Mayer, 1998].

For example, students may allocate extraneous attention to characters and objects in the world, limiting cognitive resources available for learning. Alternatively, students may spend excessive time manipulating objects to explore physics simulations that underlie gameplay. In the latter case, object manipulation may not contribute to the problem-solving task, and thus the behaviors can be considered off task. In this manner, off-task behavior in narrative-centered learning environments differs from off-task behavior in traditional intelligent tutoring systems. Unlike off-task conversation and inactivity, off-task behavior in narrative-centered learning environments appears superficially similar to engaged behavior; the student is interacting with the virtual environment, but not focusing on the primary learning task. For these reasons, off-task behavior in narrative-centered learning environments is challenging for instructors to detect. This observation underscores the importance of software-based solutions for detection and intervention.

4 CRYSTAL ISLAND

For the past several years, the authors and their colleagues have been designing, implementing, and conducting empirical studies with CRYSTAL ISLAND [Rowe et al., 2011; Rowe et al., 2009; Sabourin et al., 2011a; Sabourin, Mott, & Lester, 2011]. CRYSTAL ISLAND (see Figure 1) is a narrative-centered learning environment built on Valve Software's Source™ engine, the 3D game platform for Half-Life 2. CRYSTAL ISLAND features a science mystery set on a recently discovered volcanic island. The curriculum underlying CRYSTAL ISLAND's mystery is derived from the North Carolina state standard course of study for eighth-grade microbiology. CRYSTAL ISLAND's premise is that a mysterious illness is afflicting a research team stationed on a remote island. The student plays the role of a visitor who recently arrived on the island in order to see her sick father. However, the student gets drawn into a mission to save the entire research team from the spreading outbreak. The student explores the research camp from a first-person viewpoint and manipulates virtual objects, converses with characters, and uses lab equipment and other resources to solve the mystery. As the student investigates the mystery, she completes an in-game diagnosis worksheet in order to record findings, hypotheses, and a final diagnosis. This worksheet is designed to scaffold the student's problem-solving process, as well as provide a space for the student to offload any findings gathered about the illness. The mystery is solved when the student submits a complete, correct diagnosis and treatment plan to the camp nurse.

To illustrate the behavior of CRYSTAL ISLAND, consider the following situation. Suppose a student has been interacting with the virtual characters in the story world and learning about infectious diseases. In the course of having members of the research team become ill, she has learned that a pathogen is an agent that causes disease in its host and can be transmitted from one organism to another. As the student concludes her introduction to infectious diseases, she uncovers a clue while speaking with a sick patient that suggests the illness may be coming from food items the sick scientists recently ate. Some of the island's characters are able to help identify food items and symptoms that are relevant to the scenario, while others are able to provide helpful microbiology information. The student discovers through a series of tests that a container of unpasteurized milk in the dining hall is contaminated with bacteria. By combining this information with her knowledge about the characters' symptoms, the student deduces that the team is suffering from an *E. coli* outbreak. The student reports her findings back to the camp nurse, and they discuss a plan for treatment.

5 CORPUS COLLECTION

As part of an investigation of narrative-centered learning, a study was conducted with 450 eighth grade students from two North Carolina middle schools. Students interacted with the CRYSTAL ISLAND narrative-centered learning environment. After removing instances of incomplete data, the final corpus included data from 400 students. Of these, there were 194 male and 206 female participants. The average age of the students was 13.5 years (SD = 0.62). At the time of the study, the students had not yet completed the microbiology curriculum in their classes.

5.1 METHOD

A week prior to the interaction, students completed a series of pre-study questionnaires including a test of prior knowledge, as well as several measures of personal attributes. *Personality* was measured using the Big Five Personality Questionnaire, which represents personality along five dimensions: openness, conscientiousness, extraversion, agreeableness and neuroticism [McCrae & Costa, 1993]. *Goal orientation*, which refers to the extent that a student values mastery of material and successful performance outcomes when engaged in learning activities, was also measured [Elliot & McGregor, 2001]. Students' *emotion regulation* strategies were measured with the Cognitive

Emotion Regulation Questionnaire [Gernefski & Kraati, 2006] which measures the extent to which each of nine common strategies are used by an individual. Students also completed a researcher-generated curriculum test to assess their domain content knowledge prior to interacting with CRYSTAL ISLAND.

During the study, students interacted with CRYSTAL ISLAND for 55 minutes or until they completed the mystery. During their interaction they received an in-game prompt asking them to report on their emotional state at regular seven-minute intervals (Figure 2). This prompt was described to students as an "experimental social network" that was being pilot tested on CRYSTAL ISLAND. Students selected from one of seven emotional states: *anxious*, *bored*, *confused*, *curious*, *excited*, *focused*, and *frustrated*. They were also asked to type a short "status update." There was no actual cross-student communication enabled by this interface.

Immediately after completing their interaction with CRYSTAL ISLAND, students were given a post-interaction curriculum test with questions identical to the pre-test. They also completed several questionnaires related to their interest and understanding of the CRYSTAL ISLAND mystery. These additional measures were not used in the current investigation.

Fig 2. Emotion self-report device.

5.2 IDENTIFYING OFF-TASK BEHAVIOR

CRYSTAL ISLAND provides an open environment, believable characters, and a mystery narrative. Some of the elements that populate the narrative world are primarily aesthetic, and they are not essential for completing the science problem-solving task. For this work, off-task behavior is defined as any activity in the environment that does not contribute toward completing the science mystery. These behaviors are distinguished by identifying non-essential locations in the virtual world and unproductive interactions with objects.

There are several locations that are tangential to the learning task, including a number of outdoor areas and a waterfall. Students are considered to be off-task if they spend an excessive amount of time (e.g., four minutes without entering a building) in any of these locations. We also consider the player character's z-coordinate, or vertical position, in detecting off-task behavior. During the study researchers observed a common off-task behavior involving climbing trees and stacked crates in order to reach the roofs of buildings. Students are considered off-task if their z-coordinates exceed the maximum height that can be reached without climbing in the virtual environment. Non-critical objects in the virtual environment include cacti, crates, medicine bottles, buckets, and trash bins. No virtual characters refer to these objects at any point in the mystery. Any interaction with these objects is considered off-task. Additionally, if a student brings a task-related object (e.g., a contaminated egg) to an unassociated location (e.g., the living quarters) this is considered to be off-task as well.

Each case and total duration of off-task behavior was tagged in the student trace data. Any sequence of off-task behaviors that occurred less than 30 seconds apart was considered to be a continuous duration of off-task behavior. This decision sought to account for possible "on-task" actions occurring during segments of off-task behavior. For example, a student might rest a cactus on a sick patient (off-task), return to the dining hall to pick up a sandwich (on-task), return to the infirmary (on-task), and rest the sandwich on top of the patient as well (off-task). After aggregating sequences of off-task behavior, periods of off-task behavior totaling less than two seconds were discarded. This decision sought to avoid penalizing accidental interactions with off-task objects. No behavior occurring in the first five minutes of interaction was considered off-task in order to allow students ample time to explore the environment.

6 EXPLORATORY ANALYSIS OF OFF-TASK BEHAVIOR AND EMOTION

In order to assess student learning, paired t-tests were conducted. The tests compared student's pretest (M = 6.6, SD = 2.3) and posttest (M = 8.6, SD = 3.4) scores, and they indicated that students' learning gains from using CRYSTAL ISLAND were statistically significant, $t(399) = 12.5$, $p < 0.0001$. Based on the definition of off-task behavior described previously, the proportion of time that each student spent off-task was calculated. On average, students spent 5.1% (SD = 5.16) of their time engaged in off-task behavior, with a range of 0% to 63.2%.

An initial examination of off-task behavior was conducted using students from only the first school (N = 260). Results resembled findings reported from other investigations of off-task behavior in alternate intelligent tutoring systems [Baker et al., 2010]. Off-task behavior was found to negatively correlate with students' normalized learning gains, $r(258) = -0.18$, $p = 0.004$. There was no evidence that low prior-knowledge students engaged in more off-task behavior, as the correlation between time off-task and pre-test score was not statistically significant, $r(258) = -0.08$, $p = 0.21$. This result contrasted with a previous investigation of off-task behavior using an earlier version of the CRYSTAL ISLAND learning environment [Rowe et al., 2009].

The results also highlighted evidence that off-task behavior may have a significant affective component. In particular, total time off-task was negatively correlated with *curiosity* $r(258) = -0.12$, $p = 0.04$ and *frustration,* $r(258) = -0.13$, $p = 0.04$. This result was surprising given prior work that demonstrated *frustration* as a trigger for off-task behavior [Baker et al., 2010]. The finding prompted an examination of whether off-task behavior helps alleviate *frustration* in the CRYSTAL ISLAND environment.

In order to investigate relationships between off-task behaviors and affect transitions we utilized a measure of transition likelihood, L (Equation 1), which calculates the likelihood of a transition between two states relative to chance [D'Mello, Taylor, & Graesser, 2007]. The L statistic has a maximum value of 1, and its minimum value is $-\infty$. An L-value above zero indicates that a particular transition is more likely to occur than chance. A negative L-value indicates that a state transition is less likely than chance. This statistic is based on Cohen's kappa, and it is frequently used to measure changes in student emotions that occur over time [Baker et al., 2010; D'Mello et al., 2007].

Equation 1. L-value calculation

$$L(Current \rightarrow Next) = \frac{\Pr(Next|Current) - \Pr(Next)}{1 - \Pr(Next)}$$

To examine whether off-task behavior alleviates *frustration* or other negative learning emotions, we defined student states as follows: the current state is comprised of the student's emotion self-report at time t_n and whether or not the student went off-task between time t_n and t_{n+1}. The next state is comprised of the student's emotion self-report at time t_{n+1}. The likelihood of transitioning between these two states was calculated using the L statistic described above.

The analysis revealed that students who reported *frustration* at time t_n and subsequently went off-task were most likely to report feeling *focused* seven minutes later at t_{n+1} (Figure 3a). These observations are consistent with a hypothesis that off-task behavior helps to alleviate *frustration*. The finding lends support to the premise that some students use off-task behavior as a way to productively cope with negative affect. A possible explanation for the finding is that students employ emotion self-regulation strategies by taking breaks from challenging tasks, exploring the virtual environment, and returning "refreshed" at later times to re-engage in problem-solving activities. Students who did not go off-task after reporting *frustration* did not appear to reap this same benefit. *Frustrated* students who stayed on-task were most likely to report *boredom* at the next self-report (Figure 3a). An emotion transition from *frustration* to *boredom* may indicate that a student has disengaged from problem solving altogether.

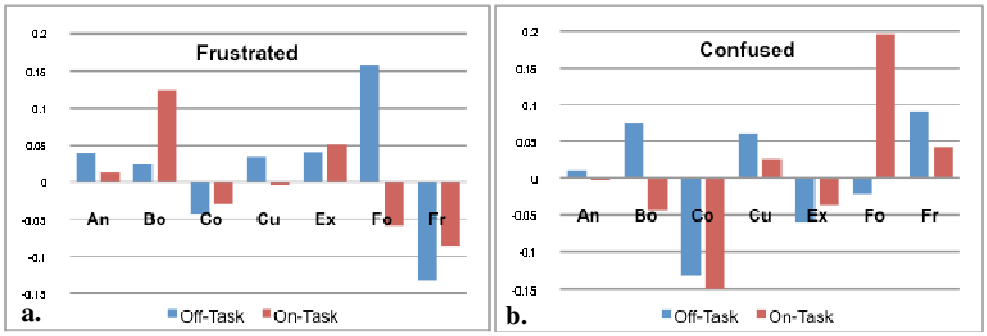

Fig. 3. Graph of transition likelihoods from (a) *frustration* and (b) *confusion* based on off-task behavior.

The hypothesis that students use off-task behavior as a productive strategy for regulating negative affect was not supported when examining affect transitions from the state of *confusion* (Figure 3b). Confused students who remained on-task were most likely to report *focus* at the next self-report, but confused students who went off-task were most likely to report *boredom*. An affect transition from *confusion* to *boredom* may signify a student reaching an impasse and giving up. These observations also suggest that students who persevered through *confusion* achieved positive affective benefits for doing so. A notable distinction between the *frustration* transitions and the *confusion* transitions is that *frustration* is generally considered harmful for learning, but *confusion* is considered productive for learning despite its negative valence [Baker et al., 2010].

These findings indicate that off-task behavior is not universally effective for self-regulating negative affect, but the findings also imply that some students may experience emotional benefits from off-task behavior under particular circumstances (e.g., when experiencing *frustration)*. A promising direction for future investigation is exploring whether students' individual differences impact the relationships between off-task behavior and emotion transitions. In a follow up investigation that compared students with above-median learning gains and below-median learning gains, the high-learning students tended to transition from *frustration* to *focus* after going off-task. On the other hand, students in the low-learning group were no more likely than chance to transition from *frustration* to *focus* whether they went off-task or remained on-task. While these findings were not statistically significant, the observations raise questions about whether students' individual differences—in prior knowledge, problem-solving ability, personality, self-regulation ability—may affect whether an intelligent tutor should recommend that a student "take a break" from problem solving by going off-task in CRYSTAL ISLAND.

7 MODELING OFF-TASK BEHAVIOR AS EMOTION SELF-REGULATION

After exploring the relationships between off-task behavior and student affect transitions, we devised a procedure for automatically predicting student off-task behaviors that alleviate negative learning emotions. The method proceeds in three stages. During the first stage, the conditional probabilities of a theoretically grounded dynamic Bayesian network are learned to predict the valence of students' emotion self reports using data about off-task behaviors. During the second stage, the dynamic Bayesian network is used

to generate hypothetical affective trajectories, or *alternate futures*, which play a central role in generating labels to annotate off-task behaviors. The labels identify whether each interval of off-task behavior is self-regulatory (i.e., increases emotional valence) or unproductive (i.e., decreases emotional valence or has no impact). During the third stage, supervised machine learning techniques are used to induce predictive models of emotion self-regulation using only features available in run-time settings. The resulting models can be used by an intelligent tutoring system to guide pedagogical decisions about intervening in students' off-task behaviors.

7.1 BAYESIAN MODELING OF OFF-TASK BEHAVIOR AND AFFECT

Bayesian networks have been used to model several aspects of intelligent tutoring systems, including models of student learning [Baker, Corbett, & Aleven, 2008; Corbett & Anderson, 1994], affect [Sabourin et al., 2011b; Conati & Maclaren, 2009], and hinting [Gertner, Conati, & VanLehn, 1998]. Bayesian networks provide a concise graphical representation for reasoning under uncertainty as they explicitly encode relationships between random variables in terms of conditional probability distributions. Bayesian networks are comprised of two primary components: a network structure that encodes the variables and conditional independence relationships among them, and conditional probability distributions that concisely encode the joint probability distribution over all of the model's variables. Network structures and conditional probability distributions can be machine learned using a variety of algorithms [Alpaydin, 2004], or they can be authored by hand.

This work follows a hybrid approach that combines a hand-crafted network structure with conditional probability values that have been machine learned using the Expectation-Maximization (EM) algorithm. A model is crafted to predict students' emotion self-report values by considering their personal attributes, their narrative problem-solving progress, and their off-task behaviors. Once a Bayesian network has been obtained that accurately models how these factors are associated with affective experience, it can be used to examine whether students are using off-task behavior to reduce negative affect.

7.2 CONSIDERING ALTERNATE FUTURES

Distinguishing between off-task behaviors that are unproductive and off-task behaviors that are self-regulatory poses notable challenges. In particular, a transition between two emotional states cannot be automatically attributed to the off-task behaviors that occur between emotion self-reports; a range of factors impact student emotions. However, if comparisons can be drawn between two sequences that differ *only* in their patterns of off-task behavior, changes in emotion self-reports can be attributed to the off-task behavior. In order to perform such a comparison, we use the dynamic Bayesian network from the procedure's first stage to simulate students' affective trajectories as if they had performed fewer off-task behaviors than in reality. This approach involves simulating *alternate futures* after each emotion self-report in order to determine whether off-task behavior directly impacted the student's affect transition. In effect, the dynamic Bayesian network is used to generate virtual emotion data to conduct a simulated experiment.

Consider the following example: a student is midway through interacting with CRYSTAL ISLAND. She reports *confusion* at one self-report (SR_i) and at the next opportunity (SR_{i+1}) she reports feeling *focused*. In the interval between SR_i and SR_{i+1} she engaged in off-task behavior. With just this information, one cannot determine whether the off-task behavior caused the positive transition into a *focused* state. However, suppose that the student's trace data is altered by removing all instances of off-task behavior that occurred between time SR_i and SR_{i+1}.The altered data is next provided to the dynamic Bayesian network from the procedure's first stage. The DBN provides a prediction about what the student's affective state might have been at the next time step (PO_{i+1}) if she had not gone off-task (Figure 4). If the model predicts a positive state such as *focused* or *curious*, one could conclude that the off-task behavior did not cause the positive change

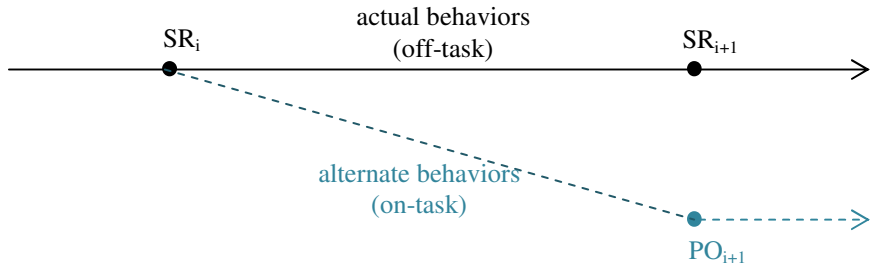

Fig. 4. Comparison of affective outcomes from an actual sequence of behaviors and simulated alternate behaviors.

in emotion. On the other hand, if the model predicts that the student would feel *frustrated* if she had not gone off-task, then one has obtained evidence that the off-task behavior brought about the positive affect transition. This latter observation identifies a possible case of emotion self-regulation, and a case where intervention by an intelligent tutor may be undesirable. This comparison can be used to label the interval's off-task behavior as self-regulatory or unproductive depending on the outcome.

7.3 REAL-TIME PREDICTION

While consideration of alternate futures provides evidence about self-regulation of negative emotion, it is not a technique that can be used at run-time. The technique cannot be directly used at run-time because it involves comparisons of students' emotional states after they have already gone off-task. An intelligent tutoring system requires models that can determine whether an off-task behavior will likely be a case of emotion self-regulation before the off-task behavior or self report occur. This type of predictive model provides the ability to decide whether to intervene or remain idle at the onset of off-task behavior.

In order to obtain such a model, we employ supervised machine learning techniques to predict labels generated during the procedure's second stage. The learned model is restricted to using predictor features that can be calculated during run-time (Figure 5). This approach is similar to work on contextual guess and slip, which uses information from future practice opportunities to generate class labels for training runtime models

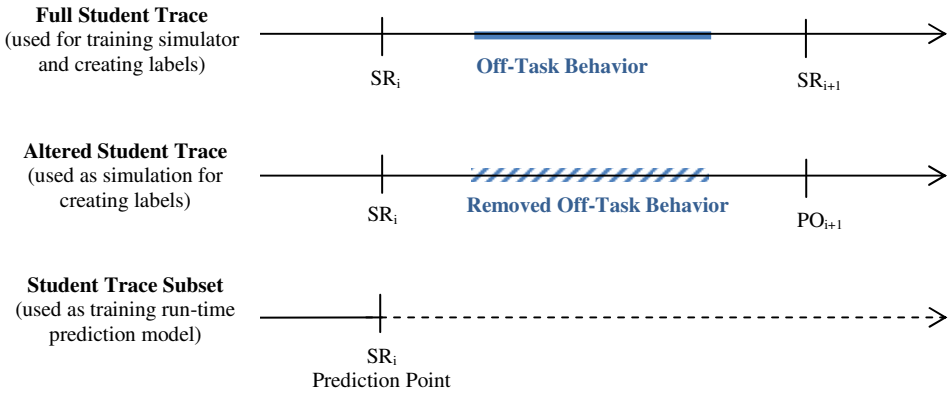

Fig 5. Modifications and uses of student trace data

[Baker et al., 2008b]. A notable distinction with the current work is that contextual guess and slip models employ a Bayesian analysis procedure to generate data labels, whereas the current approach uses machine learned dynamic Bayesian networks and alternate future simulations to generate the data labels.

8 FINDINGS

The corpus described in Section 5 was used to examine the proposed procedure for modeling off-task behavior as emotion self-regulation. This section reports findings from empirical investigations of each of the three stages: (1) developing a dynamic Bayesian network to accurately predict students' affective states, (2) generating possible alternate futures to distinguish cases of unproductive and self-regulatory off-task behavior, and (3) training models that predict whether off-task behaviors are self-regulatory or unproductive at run-time. We report significant findings for the first two stages, as well as preliminary results for the third stage. These findings underscore the potential of using alternate futures to investigate emotion self-regulation in intelligent tutoring systems.

8.1 PREDICTING STUDENT AFFECT

The CRYSTAL ISLAND corpus includes a total of 2,886 emotion self-reports from 400 students. Student reports spanned the full range of available emotion choices. *Focused* (23.1%) was the most frequently reported emotion. Following were reports of *curiosity* (19.1%), *confusion* (16.2%), *frustration* (15.7%), *excitement* (13.1%), *boredom* (8.4%) and *anxiety* (4.5%). Overall, emotions with positive valence (*focused*, *curious*, and *excited*) accounted for 55.3% of emotion self-reports.

The corpus was randomly split into two equal-sized data sets, DS_1 and DS_2. Equal numbers of students from each school were included in each data set. The first data set, DS_1, was designated for training models and generating labels during the procedure. The second data set, DS_2, was designated for validation and assessing the reliability of the generated self-regulation labels.

During the procedure's first stage, a dynamic Bayesian network structure was handcrafted that included variables related to three main sources: student personal traits, in-game progress, and off-task behaviors.

- **Personal Attributes.** These fixed attributes were obtained from students' scores on questionnaires completed prior to using CRYSTAL ISLAND. The features included four

measures of goal orientation: mastery avoidance, mastery approach, performance avoidance, and performance approach. Goal orientation describes how students approach learning tasks, and it has been shown to significantly impact student emotions during learning [Dweck & Leggett, 1988; Elliot & Pekrun, 2007]. Three personality features expected to have close relations to affective dispositions were also included: openness, agreeableness, and conscientiousness. In particular these traits relate to how individuals approach novel situations (e.g., learning tasks) and react to feedback (e.g., a student is told they have given an incorrect answer). The questionnaires were scored according to their individual scoring instructions. The resulting ordinal values were used to create an even ternary split of students on each attribute (Low, Medium, High). The resulting discretized labels were used to train the model.

- **In-Game Progress.** These attributes were calculated from students' log data, and they summarized student actions in CRYSTAL ISLAND up until the time of particular emotion self-reports. The attributes characterized important actions taken, such as *TestsRun, BooksViewed,* and *GoalsCompleted.* They also included features quantifying student progress on problem-solving milestones, such as *SuccessfulTest* and *WorksheetChecks.* Most of these attributes were split into equal ternary groups (High, Medium, and Low) resulting in three discrete values for each variable. The *TestsRun* and *WorksheetChecks* attributes had four possible values: High, Medium, Low and Zero. The High, Medium, and Low values represented an even ternary split of students who performed the actions at least once, and the Zero value represented students who did not perform the action at all. The *SuccessfulTest* feature was binary, indicating whether a student performed a laboratory test that turned out positive.
- **Off-Task Behaviors.** Two attributes related to students' off-task behavior were calculated. The first feature was a binary attribute measuring whether or not the student went off-task since the previous self-report. The second attribute measured the proportion of time the student spent off-task since the most recent self-report. This attribute is discretized into five buckets, each representing different proportions of off-task behavior.

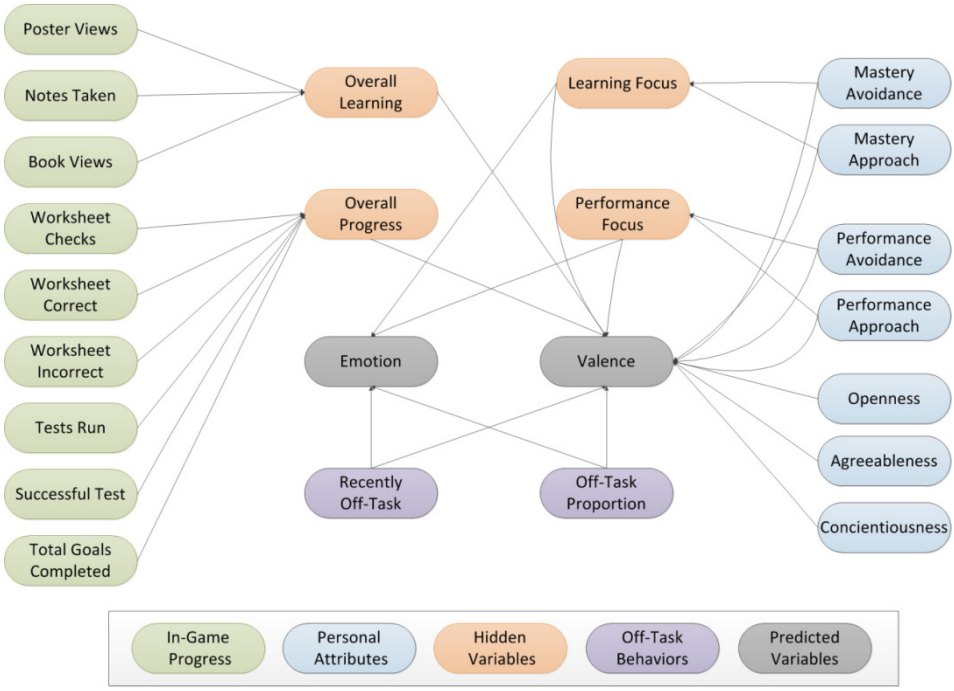

Fig. 6. Structure of one time slice from dynamic Bayesian network.

In addition to observable states, each time slice of the dynamic Bayesian network included four hidden states. Two of these states represented summaries of student progress in CRYSTAL ISLAND in terms of learning progress and goal completion. The other two hidden states characterized students' achievement goals during learning. The network's structure was based on prior work examining Bayesian prediction of student affect [Sabourin et al., 2011b]. A single time slice from the network is shown in Figure 6. Temporal relationships were also added between temporally adjacent emotion and valence nodes. A high-level illustration of the dynamic Bayesian network depicts the temporal connections in Figure 7.

The dynamic Bayesian network was created using the GeNIe modeling environment developed by the Decision Systems Laboratory of the University of Pittsburgh (http://dsl.sis.pitt.edu). After hand-crafting the structure of the dynamic Bayesian network, the parameters were learned using the EM algorithm provided by GeNIe. The model was evaluated using 10-fold cross-validation. In this technique, a model's parameters are trained using data from 90% of the student corpus. The predictive accuracy of the model is then evaluated on the remaining 10% of the corpus. This

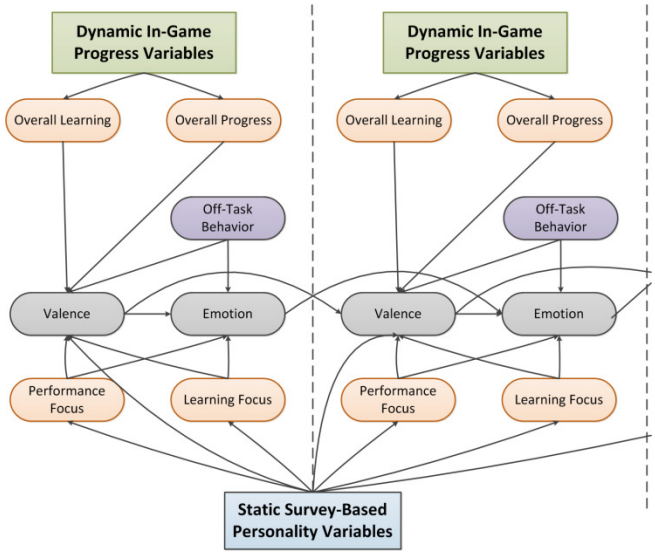

Fig. 7. Structure of dynamic Bayesian network.

approach is designed to provide an accurate measure of how well a trained model will extend to future, unseen populations. The evaluation of the DBN showed that it was able to predict emotion labels with 34.7% accuracy and emotional valence with 80.6% accuracy. This performance offers statistically significant ($p < 0.001$) improvement over the most-frequent baseline measures, which for DS_1 are 22.8% for emotion and 54.4% for valence.

While the accuracy of emotion label prediction is significantly better than baseline, it is likely too low for practical use in guiding interventions. However, the 80.6% accuracy rate for valence prediction could be used in a tutorial setting. As a result, we decided to focus on emotion valence for the remainder of the investigation.

8.2 CONSIDERING ALTERNATE FUTURES

In order to distinguish between cases where students should be permitted to go off-task and cases where students should be discouraged from going off-task, we compared students' actual affective outcomes (as indicated by self-report SR_i) with predicted affective outcomes (PO_i). Predicted outcomes were based on alternate futures, which were hypothetical scenarios where students performed fewer off-task behaviors than they did in reality. The comparisons indicated whether off-task behaviors may have been

responsible for changes in student emotions. The comparison results were also proxies for modeling the consequences of intervening in unproductive off-task behavior.

When creating alternate futures for a student, only off-task behaviors occurring between the affective state being predicted (SR_{i+1}) and the previous self-report (SR_i) are modified. The procedure assumes that an "optimal" tutorial intervention delivered during this interval would reduce or eliminate off-task behavior; the data is altered accordingly. Since reductions in off-task behavior can fall into several possible levels, we consider multiple possible futures for each student.

The two features that represent off-task behavior are discrete. One binary feature measures whether or not the student went off task since the last self report. The other feature has five possible values representing the discretized proportion of time spent off task. When reducing or eliminating the amount of off-task behavior in a student's data, we consider every possible reduction in off-task behavior level. For example, if a student spends a large amount of time off-task and has a value in the fourth bucket for proportion off-task, we simulate scenarios where her off-task behavior is reduced to the third, second, and first buckets. The binary feature is correspondingly adjusted to account for possible elimination of off-task behavior.

The modified data is input to the dynamic Bayesian network to generate a predicted affective valence for the student's next emotion self report. The predicted emotional valence is then compared with the actual emotional valence. If the predicted valence is lower than the actual valence, the transition (and associated off-task behavior) reflects evidence of self-regulation. This represents a case where intervening in off-task behavior may actually do more harm than good. If the predicted valence is higher or equal to the actual valence, the off-task behavior is considered unproductive from an affective standpoint.

In total there were 498 cases of off-task behavior in DS_1. With the possibility of different levels of reduction in off-task behavior, 656 alternate futures were generated. Using the procedure described above, 19.7% of the alternate futures were labeled as self-regulatory. In order to test the reliability of the labels assigned to the off-task behavior intervals, a second set of comparison labels was generated using the data in DS_2. A second dynamic Bayesian network was trained on DS_2 using the same approach described above. The alternate futures from DS_1 were then input to this second model to generate new labels signifying self-regulation.

If the second model's labels, which were generated using a separate corpus than the first model, were highly similar to those produced by the first model, then we consider the emotion self-regulation labels to be reliable. This validation process seeks to avoid biased labels that are closely related to the particular corpus and model used to generate them. When the two sets of labels were compared, the models agreed 88.2% of the time, with a kappa value of .64. This represents a reasonable level of reliability, suggesting that the labels appear to reflect a real phenomenon: off-task behavior as emotion self-regulation.

8.3 PREDICTING SELF-REGULATION LABELS

In order to train models for predicting the labels at run-time, off-the-shelf supervised machine learning techniques were employed. Supervised learning techniques require a single set of "gold standard" labels that classify each observation. In order to generate one set of labels, the predictions from the two models were consolidated. In nearly all cases there was agreement between the two models and possible futures. In cases where the two models or possible futures did not agree, the most frequent label was selected. In the very small minority (<1%) of cases where neither label occurred more frequently, we labeled the off-task behavior interval as unproductive. This choice was a conservative approach motivated by the observation that off-task behavior is generally associated with decreased learning. On this basis, it is likely preferable to err on the side of intervention in cases of uncertainty. After consolidating the labels, 19.5% of off-task behavior cases were labeled as self-regulatory.

As part of a preliminary investigation of the third stage, we trained models for predicting emotion self-regulation by leveraging the same features that were used to train the dynamic Bayesian network in the procedure's first stage. The features were altered to leverage information that would be available in run-time settings by including information only up until the time that a prediction was made. Off-the-shelf classification algorithms for predicting these labels were compared, including random forests, support vector machines, and bagged decision trees. All of the models were trained using the WEKA machine learning toolkit [Hall et al., 2009]. In particular, we sought to obtain a model that could identify as many cases of emotion self-regulation as possible while avoiding misclassifications of instances that may be unproductive for learning. This objective prioritizes improving precision and recall of the self-regulation class while

Table 1. Trained classifiers and performance metrics, ordered by recall on no-evidence class

	Self-Regulation Evidence		Unproductive		
Model	**Precision**	**Recall**	**Precision**	**Recall**	**Accuracy**
Baseline (most frequent)	0.000	0.000	0.805	1.000	80.5%
Support Vector Machine	0.222	0.042	0.806	0.964	78.5%
Random Forest	0.259	0.074	0.809	0.949	77.9%
Bagged Tree	0.286	0.105	0.812	0.936	77.5%

maintaining a high recall of the unproductive class. The accuracy, precision and recall of each trained model are shown in Table 1. The models were compared to a baseline model that predicted the most frequent label: unproductive off-task behavior.

While several models offered improvements in precision for recognizing self-regulation, none of the models achieved greater overall accuracy than the baseline model. Among the trained models, support vector machines achieved the greatest accuracy, and all three techniques observed improvements in precision at detecting self-regulation. While these findings do demonstrate that supervised learning techniques can achieve modest improvements in precision with small reductions in recall, considerable room for improvement remains. Several directions exist for improving the performance of these predictive models to a level that could be used in an intelligent tutoring system. A natural next step includes considering a broader pool of predictor features when training the classifiers. In particular, students' log data is amenable to producing a diverse range of metrics that characterize students' problem solving behaviors in the CRYSTAL ISLAND environment. While the current set of features have proven effective for predicting emotional valence in dynamic Bayesian networks, alternate features may need to be considered to obtain accurate predictive models for the procedure's third stage. The current investigation also did not explore automated techniques for feature selection, which offer potential to narrow down the most promising predictors from a pool of candidates. Ensemble learning techniques are another promising approach, as they combine the predictive strengths of multiple models.

9 DISCUSSION

Results from investigating the first two stages of the procedure for modeling off-task behavior as emotion self-regulation are especially encouraging. Theoretically grounded

dynamic Bayesian networks predicted students' emotional valence in CRYSTAL ISLAND with high levels of accuracy. Furthermore, the proposed method for obtaining emotion self-regulation labels based on *alternate futures* produced reliable results. These findings support the promise of empirical approaches for identifying off-task behaviors that are cases of emotion self-regulation. A preliminary analysis of the procedure's third stage yielded modest results, but several directions for continued investigation remain. The analysis did reveal that improvements in identifying self-regulatory off-task behavior can be obtained with small reductions in recall of unproductive off-task behavior. In addition to explorations of additional classification techniques and predictor features, other evaluation metrics may be considered to identify models with acceptable trade-offs. Potential directions for future work include selecting models based on different evaluation metrics, as well as incorporating the models into run-time systems in order to assess resulting student outcomes in the learning environment. As these models improve, it will become increasingly important to identify what level of predictive accuracy and recall is necessary to satisfy the pedagogical requirements of runtime systems; the current predictive accuracies are too low for operational tutoring systems, as they underperform a majority class baseline. Furthermore, it remains to be seen which predictive performance metrics should be optimized in order to yield affect-sensitive intelligent tutors with the strongest positive impact on student learning.

In this work, we have operated under the assumption that off-task behavior is generally unproductive for learning, and it should be discouraged unless strong evidence exists to the contrary. However, it is possible that tutorial interventions may be harmful in settings like narrative-centered learning environments because they interrupt typical gameplay sequences and student flow. Consideration of alternate off-task behavior intervention strategies may help to guide further development in this area.

A limitation of this work is that it only sought to identify cases of successful regulation of negative emotion. There may have been cases where a student should have gone off task but did not have the emotion self-regulation skills to make this decision. There may be circumstances when students should be encouraged to go off task in order to regulate affect. These cases were not reflected the current investigation and represent an important area for future study.

10 CONCLUSION

This work examined the relationships between student affect and off-task behavior in narrative-centered learning environments. We reported empirical results that suggest off-task behavior is associated with positive affect transitions from the emotional state of *frustration*, but the opposite is true for off-task behavior occurring after self-reported *confusion*. The findings suggested that off-task behavior may be an effective emotion self-regulation strategy under some circumstances. However, an aggregate association between off-task behavior and decreased learning implied that principled approaches for distinguishing between unproductive off-task behavior and self-regulatory off-task behavior are needed.

We proposed a novel supervised machine learning procedure for classifying off-task behaviors that are cases of emotion self-regulation. During the first stage, dynamic Bayesian networks are trained to predict student emotion self-reports from trace data. We found that a theoretically grounded DBN was capable of predicting emotional valence with high accuracy, and it predicted emotional states at levels significantly better than a baseline approach.

The DBN was used to generate *alternate futures* that simulated students' affective trajectories as if they had performed fewer off-task behaviors than in reality. The alternate futures were compared to students' actual affective trajectories in order to generate labels indicating whether off-task behaviors were cases of emotion self-regulation. In a validation process involving a second DBN trained from a distinct data set, the labels were found to be highly reliable. The two DBN models produced labels with high levels of agreement between one another. During the procedure's third stage, off-the-shelf classification models were trained to predict the labels using features that would be available in run-time settings. Support vector machines, bagged trees, and random forests achieved a promising balance between precision in detecting self-regulatory off-task behavior and recall in detecting unproductive off-task behavior, but considerable room for improvement remained. The overall findings underscore the methodological potential of using empirically generated *alternate futures* for analyzing students' emotion self-regulation processes.

In future work, we will investigate additional predictor features for training the DBN and label classification models. Additionally, we will investigate alternate evaluation metrics and classification techniques to more effectively predict generated emotion self-

regulation labels. If significant improvements in predictive accuracy can be obtained, we intend to incorporate the models into the CRYSTAL ISLAND learning environment in order to assess their ability to guide interventions for off-task behavior, shape students' affective states, and yield improved student learning outcomes.

ACKNOWLEDGEMENTS

The authors wish to thank members of the IntelliMedia Group for their assistance, Omer Sturlovich and Pavel Turzo for use of their 3D model libraries, and Valve Software for access to the Source™ engine and SDK. This research was supported by the National Science Foundation under Grants REC-0632450, DRL-0822200, and IIS-0812291. This material is based upon work supported under a National Science Foundation Graduate Research Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Additional support was provided by the Bill and Melinda Gates Foundation, the William and Flora Hewlett Foundation, and EDUCAUSE. Additional thanks to Ryan Baker for helpful discussions about modeling off-task behavior and affect.

REFERENCES

ALEVEN, V., MCLAREN, B.M., ROLL, I., and KOEDINGER, K.R. 2006. Toward meta-cognitive tutoring: A model of help-seeking with a cognitive tutor. *Internaitonal Journal of Artificial Intelligence in Education, 16*, 101-128.

ALPAYDIN, E. 2004. *Introduction to Machine Learning*, MIT Press.

ARROYO, I., FERGUSON, K., JOHNS, J., DRAGON, T., MEHERANIAN, H., FISHER, D., BARTO, A., MAHADEVAN, S., and WOOLF, B.P. 2007. Repairing disengagement with non-invasive interventions. *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* 195-202.

BAKER, R.S. 2007. Modeling and understanding students' off-task behavior in intelligent tutoring systems, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1059-1068.

BAKER, R.S., CORBETT, A.T., and ALEVEN, V. 2008. More accurate student modeling through contextual estimation of slip and guess probabilities in Bayesian knowledge tracing. *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems,* 406-415.

BAKER, R.S., CORBETT, K.R., KOEDINGER, K.R., EVENSON, S., ROLL, I., WAGNER, A., NAIM, M., ... BECK, J. 2006. Adapting to when students game an intelligent tutoring system, *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 392-401.

BAKER, R.S., CORBETT, A.T., KOEDINGER, K.R., and WAGNER, A. 2004. Off-task behavior in the cognitive tutor classroom: When students game the system. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 383-390.

BAKER, R.S., CORBETT, A.T., ROLL, I., and KOEDINGER, K.R. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction, 18*, 287-314.

BAKER, R.S., D'MELLO, S.K., RODRIGO, S.K., and GRAESSER, A.C. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies, 68*, 223-241.

BAKER, R.S., MOORE, G., WAGNER, A., KALKA, J., SALVI, A., KARABINOS, M., ASHE, C., and YARON, D. 2011. The dynamics between student affect and behavior occurring outside of educational software. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction,* 14-24.

BEAL, C.R., MITRA, S., and COHEN, P.R. 2007. Modeling learning patterns of students with a tutoring system using Hidden Markov Models. In *Proceedings of the 13th International Conference of Artificial Intelligence in Education*, 238-245.

BEAL, C.R., QU, L., and LEE, H. 2006. Classifying learner engagement through integration of multiple data sources. *Proceedings of the National Conference on Artificial Intelligence,* 2-8.

BLOOM, B.S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4-16.

BRUNER, J.S., 1990. *Acts of Meaning*, Cambridge, MA: Harvard University Press.

BUNT, A. and CONATI, C. 2003. Probabilistic student modeling to improve exploratory behavior. *User Modeling and User-Adapted Interaction, 13*, 269-309.

CETINTAS, S., LUO, S., XIN, Y., HORD, C., and ZHANG, D. 2009. Learning to identify students off-task behavior in intelligent tutoring systems, *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 701-703.

COCEA, M., HERSHKOVITZ, A., and BAKER, R.S. 2009. The impact of off-task and gaming behaviors on learning: Immediate or aggregate. *Proceedings of the 14th International Conference on Artificial Intelligence in Education,* 507-514.
CONATI, C. and MACLAREN, H. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction, 19*, 267-303.
CORBETT, A.T. AND ANDERSON, J.R. 1994. Knowledge tracing: Modeling the Acquisition of Procedural Knowledge. *User Modeling and User-Adapted Interaction, 4*, 253-278.
D'MELLO, S., TAYLOR, R.S., and GRAESSER, A.C. 2007. Monitoring affective trajectories during complex learning. *Proceedings of the 29th Annual Meeting of the Cognitive Science Society*, 203-208.
DWECK, C.S. and LEGGET, E.L. A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256-273.
ELLIOT, A. and MCGREGOR, H.A. 2001. A 2 x 2 achievement goal framework, *Journal of Personality and Social Psychology, 80*, 501-519.
ELLIOT, A. and PEKRUN, R. 2007. Emotion in the hierarchical model of approach-avoidance achievement motivation. *Emotion in Education,* P. SCHUTZ and R. PEKRUN, eds., London: Elsevier, 57-74.
GERNEFSKI, N. and KRAATI, V. 2006. Cognitive Emotion Regulation Questionaire: Development of a short 18-item version. *Personality and Individual Differences, 41*, 1045-1053.
GERTNER, A., CONATI, C., and VANLEHN, K. 1998. Procedural help in Andes: Generating hints using a Bayesian network student model. *Proceedings of the 15th National Conference on Artificial Intelligence*.
GONG, Y., BECK, J., HEFFERNAN, N., and FORBES-SUMMERS, E. 2010. The fine-grained impact of gaming on learning. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, 194-203.
GRAESSER, A.C., PERSON, N.K., MAGLIANO, J.P. Collaborative dialogue patterns in naturalistic one-to-one tutoring, *Applied Cognitive Psychology, 9*, 495-522.
HARP, S. and MAYER, R.E. 1998. How seductive details do their damage: A theory of cognitive interest in science learning. *Journal of Educational Psychology, 90*, 414-434.
HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., and WITTEN, I. The WEKA data minig software: An update. *SIGKDD Explorations, 11*, 2009.
HICKEY, D.T., INGRAM-GOBLE, A.A., and JAMESON, E.M. 2009. Designing assessments and assessing designs in virtual educational environments. *Journal of Science Education and Technology, 18*, 187-208.
KETELHUT, D.J. The impact of student self-efficacy on scientific inquiry skills: An exploratory investigation in River City, a multi-user virtual environment. *Journal of Science Education and Technology, 6*, 99-111.
MANDLER, J. and JOHNSON, N. 1988. Remeberance of things parsed: Story structure and recall. *Journal of Cognitive Psychology, 9*, 111-151.
MARSELLA, S.C., JOHNSON, W.L., and LABORE, C.M. 2000. Interactive pedagogical drama. *Proceedings of the 5th International Conference on Intelligent Virtual Agents*, 305-315.
MCCRAE, R. and COSTA, P. 1993. *Personality in Adulthood: A Five-Factor Theory Perspective,* New York: Guilford Press.

MEYER, D. and TURNER, J. 2006. Reconceptualizing emotion and motivation to learn in classroom contexts, *Educational Psychology Review, 18*, 377-390.
MULDNER, K., BURLESON, B., VAN DE SAND, B., and VANLEHN, K. 2010. An analysis of gaming behaviors in an intelligent tutoring system. *Proceedings of the 10th International Conference on Intelligent Tutoring Systems*, 184-193.
MURRAY, R.C. and VANLEHN, K. Effects of dissuading unnecessary help requests while providing proactive help, *Proceedings of the 12th International Conference on Artificial Intelligence in Education,* 887-889.
O'NEILL, B. and RIEDL, M.O. 2011. Toward a computational framework of suspense and dramatic arc. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, 256-263.
RODRIGO, M., BAKER, R.S., LAGUD, M., LIM, S., MACAPANPAN, A., PASCUA, S., SANTILLANO, J., ...VIEHLAND, N. Affect and usage choices in simulation problem-solving environments, *Proceedings of the 13th International Conference on Artificial Intelligence in Education,* 145-152.
ROWE, J.P., MCQUIGGAN, S.W., ROBISON, J.L., and LESTER, J.C. 2009. Off-task behavior in narrative-centered learning environments. *Proceedings of the 14th International Conference on Artificial Intelligence and Education,* 99-106.
ROWE, J.P., SHORES, L.R., MOTT, B.W., and LESTER, J.C. 2011. Integrating learning, problem solving, and engagement in narrative-centered learning environments, *International Journal of Artificial Intelligence in Education, 21*, 115-133.
SABOURIN, J.L., MOTT, B.W., and LESTER, J.C. 2011. Modeling learner affect with theoretically grounded dynamic Bayesian networks. *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction*, 286-295.
SABOURIN, J.L., ROWE, J.P., MOTT, B.W., and LESTER, J.C. 2011. When off-task in on-task: The affective role of off-task behavior in narrative-centered learning environments. *Proceedings of the 15th International Conference on Artificial Intelligence and Education*, 534-536.
VANLEHN, K. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197-221.
VANLEHN, K. 2006. The behaviour of tutoring systems, *International Journal of Artificial Intelligence in Education, 16*, 227-265.