Autonomous Disengagement Classification and Repair in Multiparty Child-Robot Interaction

Iolanda Leite[1], Marissa McCoy[2], Monika Lohani[2], Nicole Salomons[1],
Kara McElvaine[2], Charlene Stokes[2], Susan Rivers[2] and Brian Scassellati[1]

***Abstract*— As research on robotic tutors increases, it becomes more relevant to understand whether and how robots will be able to keep students engaged over time. In this paper, we propose an algorithm to monitor engagement in small groups of children and trigger disengagement repair interventions when necessary. We implemented this algorithm in a scenario where two robot actors play out interactive narratives around emotional words and conducted a field study where 72 children interacted with the robots three times in one of the following conditions: control (no disengagement repair), targeted (interventions addressing the child with the highest disengagement level) and general (interventions addressing the whole group). Surprisingly, children in the control condition had higher narrative recall than in the two experimental conditions, but no significant differences were found in the emotional interpretation of the narratives. When comparing the two different types of disengagement repair strategies, participants who received targeted interventions had higher story recall and emotional understanding, and their valence after disengagement repair interventions increased over time.**

I. Introduction

Student engagement is a key element for academic success [1]. Maintaining engagement is particularly important in learning situations in which constant practice is necessary for students to achieve proficiency. In the past years, virtual and robotic embodied tutors have become popular tools to promote learning due to their ability to effectively engage learners. Robotic tutors have shown to increase learning gains [2] and elicit a higher sense of presence [3] when compared to virtually embodied versions in similar conditions. For these reasons, research with robotic tutors has recently been increasing.

Social robots enable the design of novel educational settings because they can, for example, interact with a group of students more naturally than typical computer-based applications [4], [5]. The benefits of small group learning have long been recognized [6], [7], but these inherently more social settings also bring additional challenges. While teachers are able to recognize disengagement in a student – or a group of students – and intervene if necessary, it is still unclear whether robots are capable of doing so.

There have been many efforts toward the automatic classification of engagement in children and adolescents [8], [9], [10], [11], under the assumption that if an artificial tutor is able to track student engagement and respond accordingly, learning gains will increase. While most of the efforts so far are focused on modeling student engagement, less attention has been given to incorporating such models into real-time systems and validating their efficacy [12].

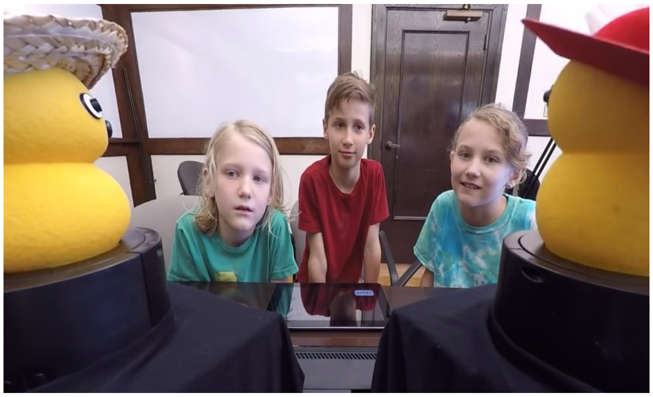

Fig. 1: Children engaged with two social robots acting out an educational story around feeling words.

In this paper, we address the questions of *when* and *how* should an embodied tutor intervene when it automatically detects disengagement in a small group of children. For addressing the question of *when* to intervene, we propose a decision-making mechanism where the disengagement level of each child in the group is modeled as a homeostatic variable. Triggering a disengagement repair intervention depends not only on the value of these homeostatic variables, but also on a set of parameters that can be adjusted to make a robot more or less proactive in interrupting the natural course of the interaction. For the question of *how* to intervene, we investigate two different types of disengagement repair strategies, one in which the robots address one particular child in the group, and another in which the robots address the whole group.

II. Related Work

The use of interventions to repair student disengagement behaviors has been a very prominent research topic in the field of Intelligent Tutoring Systems [13]. Arroyo *et al.* [8], for example, studied the impact of non-invasive interventions, such as presenting performance charts, to prompt students to reflect on their progress.

Robots have the potential to leverage on social cues to keep students engaged in the interaction because of their embodiment. Despite the growing body of research on engagement/disengagement classification in Human-Robot Interaction [9], [11], [14], less attention has been given to

[1] Department of Computer Science, Yale University, New Haven, CT iolanda.leite@gmail.com

[2] Department of Psychology, Yale University, New Haven, CT, USA

when and how robots should respond once they are able to detect user disengagement in real-time. One of the few examples that closes this loop is the work by Szafir and Mutlu [15], who conducted an experiment in which a human-like robot was able to detect drops in user attention through EEG, and then used immediacy cues – raising its volume and using arm gestures – to attempt re-engagement. Participants who received immediacy cues had better story recall than the control group, and females reported higher motivation and rapport. While this is one of the most similar works to the one we report in this paper, Szafir and Mutlu's study was performed with adults interacting individually with the robot, and we are targeting groups of children.

More recently, Brown and Howard [16] used interaction features such as speed and validity of submitted answers to monitor the engagement levels of high-school students (13 to 18 years) while completing math exercises in the presence of a robot. Upon detecting disengagement, the robot would employ verbal, nonverbal or a combination of both types of behaviors. The authors found that the verbal-only condition lead participants to complete the exercises faster, but no significant learning gains were found between conditions.

The related work presented so far assumes that the system is interacting with one user at a time. Research on maintaining engagement in multiparty settings with adults has focused on the problem of engagement intention, i.e., whether users are expressing desire to start interacting with the agent, often assuming that participants can join and leave the interaction dynamically [17], [18], [19].

III. System Design

We designed a real-time system that allow robots to (1) monitor the disengagement level of each child within a small group, and (2) decide when to employ repair strategies in order to increase the engagement level of that particular child and, as a consequence, the other children in the group. We define engagement as "the process by which individuals in an interaction start, maintain and end their perceived connection to one another" [20].

A. Real-time Disengagement Detection

We address real-time disengagement detection as a binary classification problem. Let $X_i(t) = [x_1, x_2, ..., x_n]$ be a vector of visual, auditory and contextual features extracted at a time-frame t from the interaction context and behavior of a child i in a group of size N. For every timestep t and each child i, we assume the existence of a data-driven model that receives as input $X_i(t)$ and outputs $y_i(t) \in \{0, 1\}$, a binary value that represents whether children are disengaged from ($y_i(t) = 1$) or engaged in ($y_i(t) = 0$) the interaction with the robot.

B. When to Intervene?

We assume that our disengagement classifier can output whether each child in the group is disengaged or engaged at a small unit of analysis (e.g., 500 msecs) that enables close to real-time decision making by the robot. This means that the same disengagement episode could be "active" for several time intervals. For this reason, a simple if-then rule for triggering disengagement repair actions would result in repetitive and inappropriate social behavior by the robot.

To compute the decision of whether and when a robot should employ a disengagement repair strategy given a history of outputs from the real-time classifier, our approach is to model each child's disengagement level as a homeostatic variable that, when above a certain threshold and within certain domain-specific restrictions, can trigger a repair strategy. We define domain-specific restrictions as the environment conditions that might require fine-tuning depending on the number of children in the group and the application domain. For example, in a math learning task we might limit the total number of disengagement repair interventions, or specify a minimum time interval between two interventions.

The pseudo-code for triggering disengagement repair strategies is presented in Algorithm 1. At every timestep t, the robot starts by receiving the new binary values $y_i(t)$ from the classifier and using these values to update the disengagement level variable $d_i(t)$ of each child i. Then, if the number of interventions did not exceed the maximum number of interventions Θ and the time since the last intervention is greater than the predefined time λ, the robot checks whether the disengagement level of any children in the group is above a threshold value γ. If so, the action of displaying a disengagement repair strategy is triggered. We assume that repair strategies take precedence over the natural course of interaction. This means that if the robot is executing another action, that action is momentarily interrupted and re-started after the repair strategy is over.

Our disengagement intervention algorithm is independent of the number of children in the group and from the context of the interaction. As such, the parameters γ, λ and Θ need to be adjusted depending on user group size and on how "pro-active" the robot is supposed to be regarding interventions. Different robots and application domains may require different types of disengagement repair strategies, so we leave the design of particular strategies out of the system design.

IV. Implementation

The disengagement repair algorithm described above was implemented in a scenario where two socially assistive robots interact with small groups in a storytelling task designed to promote emotional literacy in elementary school children [4]. Children can control the actions of the characters at specific moments of the interaction by selecting different story options presented on a tablet interface (text with an accompanying illustration). The robots act out the story options selected by the children and, while doing so, they use the proposed disengagement intervention algorithm to decide whether or not to interrupt the normal course of the interaction in the attempt to repair disengagement.

We used Robot Operating System (ROS) to handle data stream flows and the communication between the different modules, which we describe in more detail below. Although

Algorithm 1 Pseudo-code of the algorithm for a robot to decide whether and when to select a disengagement action.

```
$NumInterventions = 0, LastInterventionTime = 0$
loop
  for $i = 1 \rightarrow N$ do
    Update $y_i(t)$
    $d_i(t) = d_i(t-1) + \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{otherwise} \end{cases}$
  end for
  if $NumInterventions < \Theta \wedge (t - LastInterventionTime) > \lambda$ then
    if $\max(d_i(t)) > \gamma$ then
      ExecuteInterventionStrategy(i), where $i = \underset{i}{\operatorname{argmax}}(d_i(t))$
      $LastInterventionTime = t$
      $NumInterventions = NumInterventions + 1$
    end if
  end if
end loop
where
$NumInterventions \rightarrow$ total number of interventions since $t = 0$
$LastInterventionTime \rightarrow$ time $t$ of the last intervention (to any user in the group)
$N \rightarrow$ total number of users in the group
$d_i(t) \rightarrow$ cumulative disengagement level of user $i$ at time step $t$
$\gamma \rightarrow$ threshold value
$\lambda \rightarrow$ minimum time between two interventions
$\Theta \rightarrow$ maximum number of interventions allowed
```

the developed architecture can support any number of robotic characters and different group sizes, for simplicity let us assume that we have two robotic characters and a fixed group size of three children.

A. Behavioral Feature Extraction

A Microsoft Kinect V2 sensor was used to extract the visual and auditory features necessary to classify the disengagement of each child, namely *Voice Activity*, *Smiles*, *Lean Forward*, *Lean Back*, *Look Up*, *Look Down* and *Look at Robots*. Before streaming the features to the Disengagement Classifier, this module eliminates potential false positive bodies tracked by the sensor by considering only the closest three bodies, and uses information from the face and body coordinates to ensure that the features of each child are being consistently tracked in the same relative position (e.g., left, middle, right). This module streams data to the Disengagement Classifier, running as a ROS node in a different machine, at approximately 30 Frames Per Second (FPS).

B. Disengagement Classifier

To classify disengagement, we used an SVM-based model trained offline with data from a gender-balanced dataset of 21 children interacting with the robots in small groups of three. The training data was collected in a preliminary data collection study in which the robots' perception of the environment was limited to the tablet inputs [14]. The SVM model (type C-SVC, parameter $C = 1$ and $\gamma = 0.5$) with a Radial Basis Function (RBF) kernel was trained using Scikit-learn library [21]. By using the set of visual and auditory features described above, we are able to classify whether each child in the group is disengaged or not with an average of 60% accuracy [14].

The disengagement classifier outputs a value for each child in the group based on the behavioral features of that child[1]. Because the model was built using 500msec time intervals, the features were averaged and converted to a binary value in that time window. The binary value of a feature at time t reflects what happened in the majority of the interval. For example, *Voice Activity* is set to 1 when the child is speaking in the majority of audio frames received from the Behavioral Feature Extraction module for that interval.

C. Action Selection

The action selection module controls the flow of the interaction in a constant loop with two main tasks. First, it continuously monitors and updates disengagement levels using Algorithm 1, and triggers disengagement repair interventions whenever necessary. After pilot tests with a few groups of children interacting with the system, we set up the algorithm parameters as follows: $NumInterventions = 2$ for each story option (i.e., 6 for the whole interaction), $\gamma = 10$, $\lambda = 30$ sec.

The second main task of the this module is to manage the storytelling by interpreting the story scripts, sending the appropriate story lines to the robot controllers and communicating with the tablet interface to display the right story options and perceived user input. When the child selects a new story option on the tablet, that story option is played out by the robots. When the robots finish playing out a scene, the following story options are presented on the tablet as text with an accompanying illustration.

D. Robot Controller

The robot controller is responsible for conveying the animations and speech behaviors in the different robotic characters that play out the stories. We used two MyKeepons (see Figure 1), Leo and Berry, with programmable servos controlled by an Arduino board. Each robot has four degrees of freedom: it can pan and roll to the sides, tilt forward and backward, and bop up and down. To complement the pre-recorded utterances, we developed several non-verbal behaviors such as idling, talking and bouncing.

E. Disengagement Repair Strategies

With the help of elementary school teachers from the school in which we conducted the user study reported below, we designed two sets of disengagement repair strategies: *general* and *targeted*. In general interventions, the robots address the whole group and make generic comments that imply responsibility of all participants, while in targeted interventions they directly address the child with the highest level of disengagement. Additionally, in general interventions both robots look at each one of the children in the group, and then utter a verbal comment without targeting any particular child (e.g., looking at each other or looking at the group

[1] We tested models combining the features of the target child and the other children in the group, and the model performance did not increase.

TABLE I: Examples of verbal comments employed by the robots in general (addressing the group) and targeted (addressing one child) disengagement repair strategies.

General	**Targeted**
Berry: I can't hear you Leo!	*Berry*: Hey, We're trying to work here!
Leo: Let's go over that bit again.	*Leo*: Hello?
Berry: Do they seem distracted to you? *Leo*: Yeah, a little bit.	*Berry*: Can you please pay attention?
Leo: Why aren't they listening to us?	*Leo*: Why aren't you listening to us?

TABLE II: Summary of the study design with regard to the presence of disengagement repair interventions – Yes (Y) or No (N) – and difficulty level of the story content – Easy (E), medium (M) or Hard (H).

Condition	**Session 1**		**Session 2**		**Session 3**	
	Dis.?	Level	Dis.?	Level	Dis.?	Level
Control	N	E	N	M/H	N	M/H
Targeted	N	E	Y	M/H	Y	M/H
General	N	E	Y	M/H	Y	M/H

but without making eye contact with any of the children). In targeted interventions, both robots orient themselves to the most disengaged participant before making the verbal comment.

These two different categories of disengagement repair will comprise the two experimental conditions of the study reported in this paper. As such, to the same group of children, the robots will employ strategies from only one of these groups. Table I contains examples of the verbal comments used by the robots in the two disengagement repair categories. Once the action selection module triggers an intervention, the selection of specific behaviors within each category is performed randomly, with the two robots having the same probability of uttering the verbal comments to account for potential spacial bias of the children who are closest to one of the robots.

V. Experimental Evaluation

We conducted a repeated interaction study in which groups of children interacted with the robot system described above. This study had two main goals. First, we wanted to have a proof-of-concept of our autonomous disengagement repair algorithm working in real-time in a classroom environment. Our second goal was to investigate the impact of type of repair strategies (i.e., targeted versus general) in small groups.

A. Participants

A total of 72 children were recruited from an elementary school in the United States where RULER [22], a social and emotional learning program that inspired the learning content that the robots deliver to children through the interactive stories, has been implemented. The participants in the study were first through third grade students (35 female, 37 male) between the ages of 6 and 9 (M = 7.5, SD = .83).

B. Study Design

The study was a between subjects design with participants randomized into small groups of 3 and assigned to one of the following conditions: **targeted** (disengagement repair strategies directed specifically toward the disengaged child, N = 27), **general** (disengagement repair directed at the whole group, N = 27), and a **control** condition (no disengagement repair interventions, N = 18). The conditions were counter-balanced for age and gender, and groups were exclusively composed of participants from the same grade level. Group formations remained the same through the entire study.

Participants interacted with the robots a total of three sessions, approximately once per week. Repeated sessions were used to avoid potential effects of interacting with a robot for the first time. For the same reason, the first session was treated as a baseline: even in the disengagement repair conditions, the robots did not employ any disengagement repair interventions. Additionally, all participants in session 1 were assigned to the interactive story with the easiest learning content. In the following two sessions, the session content was counter-balanced in each condition to account for effects of difficulty level over time (see Table II).

C. Session Contents

The goal of our scenario is to create an interactive safe and judgment free space for children to develop their emotional intelligence skills by trying out various situations in emotionally-charged situations that they might find in the real world. Each pedagogical session featured an interactive narrative around a different feeling word that is part of the RULER curriculum: Inclusion (easy), Cooperation (medium), and Frustration (hard). The scripts were reused from a previous experiment that investigated the effects of children interacting alone or in small groups in this scenario, but in which the robots were not capable of detecting disengagement in real-time and respond accordingly [4]. The difficulty level is reflected by the number of core ideas and story details, as well as by the number of additional characters mentioned in the story by the two main robot characters. Each session consists of an introductory scene followed by three different story options for children to explore. The options impact the story line, and consequently, the feelings of the story characters in different ways.

D. Procedure

Consent forms were distributed in the classrooms of teachers that had agreed to participate in the study. A project aide retrieved groups of participants from their classrooms and brought them to the experiment room located in the school library. After receiving verbal consent, the experimenter introduced the participants to the robots, Leo and Berry, and explained how the session would proceed. The robots started by acting out an introduction scene, and then participants were instructed to select what would happen among three different options presented on the tablet with text and accompanying illustrations. Participants were instructed to explore all the story options. The experimenter remained in the room, but out of sight of the participants for the duration of the robot interaction. The experimenter was responsible for recording the interaction using 3 HD cameras (one for each participant), making sure the Kinect sensor was tracking all participants, and controlling the beginning and end of the interaction, since the robots' behavior was fully autonomous.

The experimental sessions lasted roughly 30 minutes. During about half of this time, children interacted with the robots in a small group, and in the remaining 15 minutes each participant was interviewed individually by an experimenter in a separate room by additional experimenters. During the interview, the experimenter used small cards, with the same illustration as the tablet during the storytelling interaction, to represent the various scene choices for the story session. Interviews were conducted following a standardized protocol comprised of the same series of questions (one open-ended question, followed by two closed-ended questions) for each of the four scenes (i.e., Introduction, Option 1, Option 2, Option 3). In the order that each participant selected the story options, the experimenter asked: a) what happened in the story when that option was selected, b) what color of the mood meter (RULER tool) they thought the main story character was in and c) what word would they use to describe how the character was feeling. During each interview session, participants answered a total of 12 questions (4 scenes per session), and a total of 36 questions over the course of the study. If a child did not respond within 10 seconds of the question being asked, the experimenter would inquire, "Would you like me to repeat the question or would you like to move on". All interviews were audio-recorded and transcribed verbatim for coding.

1) Reliability Coding: Two independent coders analyzed the interview transcriptions from the three sessions, counting the number of core characters (e.g., Leo, Berry), core narrative ideas (e.g., Leo does not know anyone, everyone is staring at Leo's clothes), correct and incorrect feeling words, event details and extra-event details. The two coders overlapped in 25% of the data for reliability analysis. Reliability between coders was measured using the Intra-class Correlation Coefficient test for absolute agreement using a two-way random model. All the coded variables for each interview session had high reliabilities. The lowest agreement was found in the number of correct details ($ICC(2,1) = 0.597, p < .001$), and maximum agreement was reached for the number of mentioned core characters ($ICC(2,1) = 0.965, p < .001$). Given acceptable agreement between the two coders in the overlapping interviews, data from one coder was randomly chosen for analysis.

E. Measures

Our exploratory analysis included recall and understanding metrics extracted from the post-interaction individual interviews, and behavioral metrics extracted during the interaction with the robots.

1) Recall and Understanding: The post-interaction measures focused on participants ability to logically reconstruct a comprehensive narrative, and interpret the emotional state of the characters for each of the narrative choices. The interview questions were coded and then combined into the following dependent variables:

Narrative Structure Score (NSS) – the ability of participants to "logically recount the fundamental plot elements of the story" [23], [24]. For each session *S* and participant *i*, responses were coded for the presence of core characters and ideas. NSS was computed using the following formula:

$$NSS_{S,I} = \frac{Mentioned(CoreCharacters + CoreIdeas)}{All(CoreCharacters + CoreIdeas)}$$

A perfect NSS of 1.0 indicates that the child was able to recall all the core characters and main ideas in all four open-ended questions of that interview. The average number of characters in each story was three (Leo, Berry, and one additional character for the medium and hard narratives), while the number of core ideas varied depending on the story difficulty level.

Emotional Understanding Score (EUS) – participants' ability to correctly recognize and label characters' emotional states. Responses to the interview questions "what color of the Mood Meter do you think <robot character> was in?" and "what word would you use to describe how <robot character> was feeling?" were used to calculate this metric[2]. These responses were coded into the $ColorScore$, with participants receiving +1 if the correct Mood Meter color was provided, and -1 if an incorrect color was given. The second direct question was used to create a $FeelingWordScore$, which measures how proficient children are at enumerating affective words to particular Mood Meter quadrants (e.g., enthusiastic, peaceful). In the $FeelingWordScore$, participants received +1 or -1 for the first provided feeling word, and +0.5 or -0.5 points for each additional appropriate or inappropriate word, respectively. EUS was then calculated by summing these two scores:

$$EUS_{S,I} = ColorScore + FeelingWordScore$$

Higher EUS indicates that more Mood Meter colors and corresponding feeling words were accurately identified. This score is not bounded because participants were free to use any number of words in their responses.

2) Behavioral Metrics: The participant videos collected during the robot interaction were synchronized with automatically generated logs that contain the times in which the robots employed disengagement repair interventions. We conducted a post-hoc automatic behavioral analysis of the video recordings using a commercially available facial expression analysis software[3] that outputs levels of *valence*

[2] The Mood Meter is a quadrant with the x-axis representing low to high pleasantness and the y-axis indicating low to high energy. Participants were asked to categorize the emotional state of the robot characters based on the following colors: Yellow (pleasant, high energy), Green (pleasant, low energy), Blue (unpleasant, low energy), or Red (unpleasant, high energy).

[3] http://www.noldus.com/human-behavior-research/products/facereader

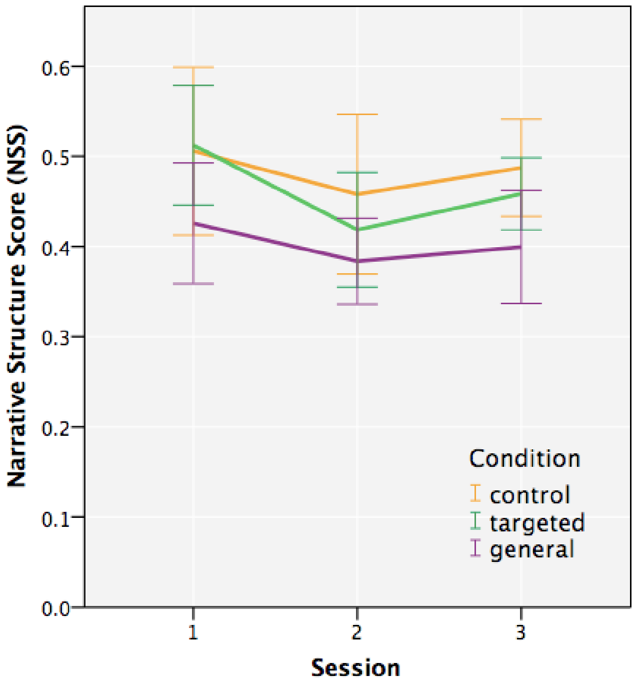

Fig. 2: Average Narrative Structure Scores (NSS) for participants in each condition on every interaction session.

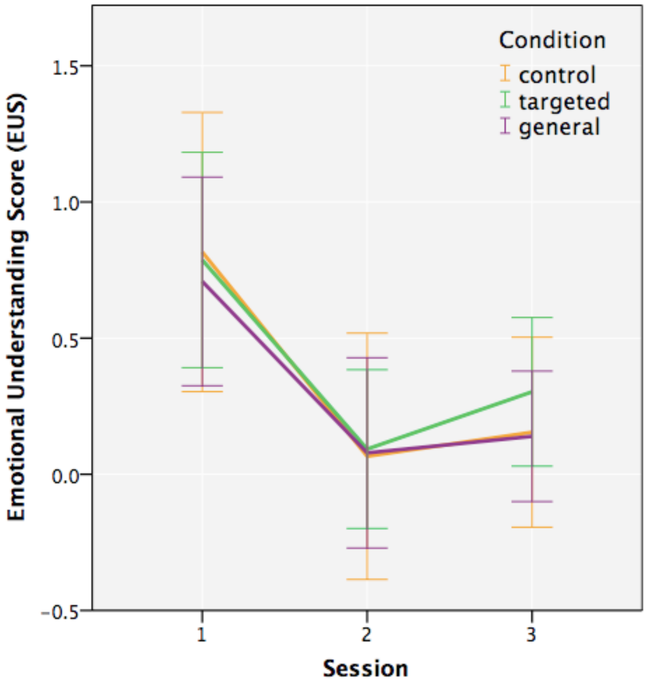

Fig. 3: Average Emotional Understanding Scores (EUS) for participants in each condition on every interaction session.

and *arousal* of tracked faces at 30 FPS. Valence values can range from -1 (most negative) to 1 (most positive), and arousal ranges from 0 (low energy) to 1 (high energy). Because we did not expect the robot's disengagement repair interventions to affect the average valence or arousal of the whole interaction, we focused the analysis of these variables in the 20 seconds following a robot disengagement repair strategy, which encompasses the average duration of children's responses to these interventions. For each participant belonging to one of the two experimental conditions, we averaged his/her valence and arousal for the 20-second windows right after a disengagement repair was employed by one of the robots.

VI. Results

We examined the effect of our manipulation on the recall, understanding and behavioral metrics using analysis of variance (ANOVA). No significant gender differences were found in the results reported below.

A. Recall and Understanding

1) Narrative Structure: A one-way ANOVA was conducted to investigate the impact of study condition and Narrative Structure Score (NSS). There was a statistically significant difference between the two study conditions, $F(2, 207) = 5.418, p < .01, \eta^2 = .05$. A Bonferroni post-hoc test revealed that NSS was significantly higher in the control condition ($M = .48, SD = .15$) compared to the general condition ($M = .40, SD = .15, p < .01$), and that NSS was significantly higher in the targeted than in the general condition ($M = .46, SD = .15, p < .05$). There were no statistically significant differences between the control and targeted groups.

When considering the session number as a within-subjects factor (and keeping study condition as a between-subjects factor), there was a statistically significant main effect of time in a repeated measures ANOVA, $F(2, 64) = 6.25, p < .05, \eta^2 = .16$, and no significant interaction effect between time and condition, $F(4, 130) = .57, p = .69, \eta^2 = .02$. Post-hoc tests applying Bonferroni correction revealed that NSS was significantly higher in session 1 ($M = .48, SD = .17$) when compared to session 2 ($M = .42, SD = .15, p < .05$), but no significant differences were found between session 3 ($M = .45, SD = .12$) and the first two sessions. It is relevant to note that session 1 was our baseline session (easier story content and similar treatment for all participants), and the two following sessions were counter-balanced for story content difficulty. Figure 2 shows NSS over time for participants in each study condition.

2) Emotional Understanding: To investigate the impact of study condition on Emotional Understanding Score (EUS), we conducted a one-way ANOVA. There was no statistically significant difference between study condition in EUS, $F(2, 207) = .21, p = .811, \eta^2 = .00$. When including time as a within-subjects factor, a repeated measures ANOVA revealed again a significant main effect on time, $F(2, 64) = 11.43, p < .01, \eta^2 = .26$, but no interaction effect between time and study condition, $F(4, 130) = .12, p = .97, \eta^2 = .00$. A Bonferroni post-hoc test revealed that EUS was significantly lower in session 2 ($M = .10, SD = .83, p < .01$) and in session 3 ($M = .21, SD = .64, p < .01$), when compared to session 1 ($M = .76, SD = .98$), our easier baseline session.

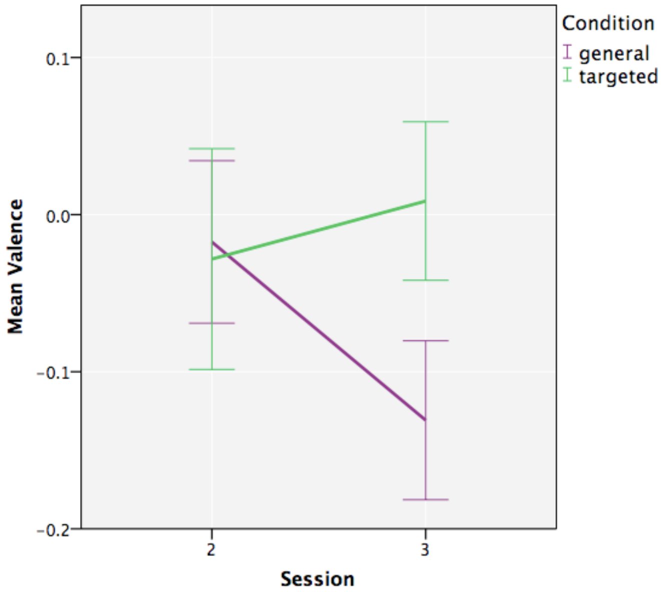

Fig. 4: Average Valence for participants in the two experimental conditions in the moments after the robots employed disengagement repair interventions.

B. Behavioral Measures

To investigate the variation of children's valence and arousal in the moments right after the robots employed a disengagement repair behavior, we conducted within-subjects ANOVAs with session number (2 or 3) as the within-subjects factor and condition (targeted or experimental) as a between-subjects factor.

1) Valence: We found no significant main effect on session number in participants' valence, $F(1, 121) = 1.78, p = .19, \eta^2 = .01$. A significant interaction effect was found between session and study condition, $F(1, 121) = 10.32, p < .01, \eta^2 = .08$ (see Figure 4), with children in the targeted condition experiencing significantly higher valence in session 3 ($M = .02, SD = .25$) than in session 2 ($M = -.03, SD = .25$), and participants in the general condition experiencing lower valence in session 3 ($M = -.15, SD = .25$) than in session 2 ($M = -.02, SD = .22$).

2) Arousal: There was no significant main effect on session and arousal, $F(1, 121) = 2.15, p = .15, \eta^2 = .02$, nor in the interaction between session and condition, $F(1, 121) = .16, p = .69, \eta^2 = .00$. Although not statistically significant, arousal is lower in session 3 ($M = .36, SD = .08$) when compared to session 2 ($M = .38, SD = .09$) collapsed across the two experimental conditions.

VII. Discussion

Our results contrast with previous research with adults who found that higher interaction recall is achieved when robots can monitor participant's engagement and employ social cues when attention drops [15]. However, it is important to frame these results within the context of groups of children. First of all, interruptions are disruptive by nature [25] and children of this age might not have developed the capacity to quickly recover from them, especially in the presence of their peers and without adult supervision. In this line of reasoning, Kennedy and colleagues [26] found that more social behavior is not always reflected in increased learning gains in child-robot interaction. In fact, we noticed that oftentimes the robot's interventions lead to higher disengagement: participants responded to the robots (e.g., by denying that they were disengaged or asking how the robots knew) or started talking to each other about what just happened. These behaviors were more frequent in session 3 than in session 2. Finally, an alternative explanation is that the selected parameters of our algorithm, such as maximum number of disengagement repair interventions and minimum time between interventions, were not optimal for this case. Fine-tuning the algorithm parameters differently could have impacted children's recall and understanding abilities in a different manner.

When comparing the two types of disengagement repair strategies, interventions targeting one particular child lead to higher recall and understanding gains than interventions addressing the whole group. These results can be explained by the diffusion of responsibility theory, which suggests that the mere presence of others decreases the pressure on individuals to respond because of a sense of shared responsibility [27], a phenomenon also known as the bystander effect. Research has shown that young children are also influenced by the bystander effect [28]. For instance, they are more likely to exhibit prosocial behavior when personal responsibility increases [29]. In our study, it could have been the case that the sheer presence of the other group members (i.e., bystanders) reduced the likelihood of successful interventions addressing the entire group.

Additionally, behavioral metrics indicate that targeted interventions might be more successful over time. In the moments right after the robots employed a disengagement repair behavior, participants' valence tended to increase from one session to the other in the targeted condition, but it decreased for children in the general condition. One possible interpretation is that general interventions tend to penalize participants who are engaged, and over time this can have a negative effect. However, future experiments are necessary to investigate whether these trends would persist over longer periods of time.

VIII. Conclusion

The contributions of this work were twofold. First, we presented an algorithm that allows social robots to monitor disengagement in small groups of children and decide when to intervene in the attempt to re-engage children in the interaction. Second, we investigated the impact of targeted (to a particular child) versus general (addressing the group) disengagement repair strategies in a field study where the same groups of children interacted with two autonomous robots once a week over three weeks.

Overall, our results suggest that interrupting the natural course of the interaction can be extremely costly, especially in multiparty child-robot interaction. As children realize that the robots are socially aware of their behavior, they start addressing the robots more often (and more socially). Previous research also suggests that this can be a source of distraction [26]. In other words, robot social behavior leads to more human social behavior and, in learning environments, that comes with a cost. Further research is needed to understand what the optimal trade-offs are with regard to both the number and type of disengagement repair interventions, so that educational robots with augmented perceptive capabilities can make a positive impact on children's learning.

References

[1] J. Finn and K. Zimmer, "Student engagement: What is it? why does it matter?" in *Handbook of Research on Student Engagement*, S. L. Christenson, A. L. Reschly, and C. Wylie, Eds. Springer US, 2012, pp. 97–131. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-2018-7_5

[2] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati, "The physical presence of a robot tutor increases cognitive learning gains," in *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci 2012)*. Cognitive Science Society, 2012.

[3] J. Kennedy, P. Baxter, and T. Belpaeme, "Comparing robot embodiments in a guided discovery learning interaction with children," *International Journal of Social Robotics*, vol. 7, no. 2, pp. 293–308, 2014.

[4] I. Leite, M. McCoy, M. Lohani, D. Ullman, N. Salomons, C. Stokes, S. Rivers, and B. Scassellati, "Emotional storytelling in the classroom: Individual versus group interaction between children and robots," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 75–82. [Online]. Available: http://doi.acm.org/10.1145/2696454.2696481

[5] T. Ribeiro, A. Pereira, A. Deshmukh, R. Aylett, and A. Paiva, "I'm the mayor: A robot tutor in enercities-2," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, ser. AAMAS '14. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1675–1676. [Online]. Available: http://dl.acm.org/citation.cfm?id=2615731.2616120

[6] G. W. Hill, "Group versus individual performance: Are n+1 heads better than one?" *Psychological Bulletin*, vol. 91, no. 3, p. 517, 1982.

[7] P. Dillenbourg, "What do you mean by collaborative learning?" *Collaborative-learning: Cognitive and Computational Approaches.*, pp. 1–19, 1999.

[8] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, and B. P. Woolf, "Repairing disengagement with non-invasive interventions," in *Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*. Amsterdam, The Netherlands, The Netherlands: IOS Press, 2007, pp. 195–202.

[9] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 2011, pp. 305–311.

[10] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester, "The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 42–49. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663264

[11] L. J. Corrigan, C. Basedow, D. Kuster, A. Kappas, C. Peters, and G. Castellano, "Perception matters! engagement in task orientated social robotics," in *Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on*. IEEE, 2015, pp. 375–380.

[12] R. Baker and L. Rossi, "Assessing the disengaged behaviors of learners," *Design Recommendations for Intelligent Tutoring Systems*, vol. 1, p. 153, 2013.

[13] M. Cocea and S. Weibelzahl, "Disengagement detection in online learning: validation studies and perspectives," *Learning Technologies, IEEE Transactions on*, vol. 4, no. 2, pp. 114–124, 2011.

[14] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 99–105. [Online]. Available: http://doi.acm.org/10.1145/2696454.2696466

[15] D. Szafir and B. Mutlu, "Pay attention!: Designing adaptive agents that monitor and improve user engagement," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 11–20. [Online]. Available: http://doi.acm.org/10.1145/2207676.2207679

[16] L. Brown and A. M. Howard, "Engaging children in math education using a socially interactive humanoid robot," in *Humanoid Robots (Humanoids), 2013 13th IEEE-RAS International Conference on*. IEEE, 2013, pp. 183–188.

[17] M. E. Foster, A. Gaschler, and M. Giuliani, "How can i help you': comparing engagement classification strategies for a robot bartender," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 255–262.

[18] Q. Xu, L. Li, and G. Wang, "Designing engagement-aware agents for multiparty conversations," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2233–2242.

[19] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... um... hesitations," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14. New York, NY, USA: ACM, 2014, pp. 2–9. [Online]. Available: http://doi.acm.org/10.1145/2663204.2663241

[20] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 12, pp. 140 – 164, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0004370205000512

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] S. E. Rivers, M. A. Brackett, M. R. Reyes, N. A. Elbertson, and P. Salovey, "Improving the social and emotional climate of classrooms: a clustered randomized controlled trial testing the RULER approach," *Prev. Sci.*, vol. 14, no. 1, pp. 77–87, Feb. 2013.

[23] K. A. McCartney and K. Nelson, "Children's use of scripts in story recall," *Discourse Process.*, 1981.

[24] F. McGuigan and K. Salmon, "The influence of talking on showing and telling: adultchild talk and children's verbal and nonverbal event recall," *Appl. Cogn. Psychol.*, 2006.

[25] J. P. Borst, N. A. Taatgen, and H. van Rijn, "What makes interruptions disruptive?: A process-model account of the effects of the problem state bottleneck on task interruption and resumption," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '15. New York, NY, USA: ACM, 2015, pp. 2971–2980. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702156

[26] J. Kennedy, P. Baxter, and T. Belpaeme, "The robot who tried too hard: Social behaviour of a robot tutor can negatively affect child learning," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 67–74. [Online]. Available: http://doi.acm.org/10.1145/2696454.2696457

[27] J. M. Darley and B. Latane, "Bystander intervention in emergencies: diffusion of responsibility." *Journal of personality and social psychology*, vol. 8, no. 4p1, p. 377, 1968.

[28] M. Plötner, H. Over, M. Carpenter, and M. Tomasello, "Young children show the bystander effect in helping situations," *Psychological science*, p. 0956797615569579, 2015.

[29] G. Maruyama, S. C. Fraser, and N. Miller, "Personal responsibility and altruism in children." *Journal of Personality and Social Psychology*, vol. 42, no. 4, p. 658, 1982.