# Persistent Memory in Repeated Child-Robot Conversations

**Iolanda Leite, André Pereira, Jill Fain Lehman**
Disney Research Pittsburgh, USA
{iolanda.leite, andre.pereira, jill.lehman}@disneyresearch.com

**ABSTRACT**
Persistent memory is a critical mechanism in long-term human-robot interaction. In this work, we investigate how a robot can use information from prior conversations with the same child to foster a sense of relationship over time. To address this question, we conducted a repeated interaction study with three experimental conditions: a baseline control condition, in which the robot retains no information between conversations and relies on a typical elicitation-response paradigm; a persistence condition, in which children experience the same topic flow but with some robot turns that refer back to prior shared events; and a pro-active persistence condition, in which the robot attempts to offer its own feelings and opinions pro-actively and congruently with what it knows about the child. Our results indicate age differences with respect to the measures of interest. During conversations with the robot, older children who were assigned to the persistence conditions exhibited more positive affect, while younger children showed more positive affect in the control condition. Moreover, in a set of comparative judgments among robots they had played with, children in the augmented persistence condition considered PIPER to be the most intelligent and their favorite more often than children in the other conditions, overall, but the effect was more evident in the older children.

**CCS Concepts**
•**Human-centered computing** → **User studies; Empirical studies in HCI;** *Natural language interfaces;*

**Author Keywords**
Human-robot interaction; memory; repeated interaction

**INTRODUCTION**
In human-human conversation we use a variety of available strategies in service of myriad goals: to inform, to effect action, to modify beliefs, to evoke emotion, to build connection. Who says what, and how and when, contributes to how we feel about the other person and how we think the other feels about us [8]. In relationships that involve repeated conversation over time, persistent memory for information previously shared has its own power. Minimally, persistence allows a conversant to know or signal that s/he is remembered; more importantly, persistence allows her/him to build and act from a model of common values and experiences. As human-robot dialog faces the challenges of long-term interaction, understanding how to use prior conversation to foster a sense of relationship is key because whether robots remember what we've said, as well as how and when they expose that memory, will contribute to how we feel about them.

Figure 1: Child interacting with PIPER.

The study presented here is an initial effort to understand how two basic conversational strategies that require persistent memory effect young children's experiences with a robot over time. All children have four conversations with the same robot, PIPER, interspersed with other activities out of the robot's view. In a baseline control condition (C), PIPER relies on a typical elicitation-response paradigm to work through a pre-determined set of topics about each activity (e.g., "What did you do?" <child names activity> "That's cool. Did you like it?") and retains no information between conversations. In the persistence condition (P), children experience the same topic flow but with some PIPER turns that refer back to prior shared events (e.g., "Hey, done with <previously mentioned activity>, I see. Did you like it?"). In the pro-active persistence condition (P+), PIPER goes further and attempts to offer its own feelings and opinions congruently with what it knows about the child, depending on the child's self-report at the end of the activity (e.g., "Hey, done with <previously mentioned activity>, I see. I think that game is really <hard/fun>).



*IDC '17,* June 27-30, 2017, Stanford, CA, USA



ISBN 978-1-4503-4921-5/17/06…$15.00
DOI: http://dx.doi.org/10.1145/3078072.3079728

We found that during conversations with the robot, older children who were assigned to the persistent-memory conditions exhibited more positive affect, while younger children showed more positive affect in conversations with simple elicitation and response. Self-report questions collected after the first and last conversations showed that children in all conditions liked the robot, but that the likeability gains were higher for children in the persistent-memory conditions overall, and caused primarily by larger gains for the older children. Finally, in a judgment task comparing the three robots that children interacted with across activities, children in the persistence conditions considered PIPER to be the most intelligent and their favorite more often than children in the control conditions. Despite this general trend, young children in the control condition were more likely to favor the other robots. More broadly, we find that a multiple measures approach, examining phenomena at different levels of granularity, may be an important technique in understanding repeated interactions with children.

## BACKGROUND

### Information Elicitation

Information elicitation can serve multiple goals in dialogue, one of them being the establishment of a common ground between people [7]. Kiesler [13] argues that communication between robots and users will benefit if robots have a model of common ground with the user. Bickmore and Cassell [3] investigated the effects of conversational strategies such as small talk to establish common ground and improve trust in human-agent interaction. The use of social dialogue by the agent positively impacted users' perception of trust, especially among extrovert adults. More recently, Matsuyama et al. [19] proposed an architecture that allows an agent to build rapport with users by eliciting personal information. Using strategies such as self-disclosure or reference to a previous shared experience, SARA (Socially-Aware Robot Assistant) elicits user preferences and goals to make personalized recommendations about which conference sessions the user should attend. We note that these results were obtained with adults, and it remains unclear whether these findings would hold for children.

One of the practical benefits of information elicitation in human-machine interaction is personalization [14]. Clabaugh and Matarić [6] investigated the impact of elicitation for personalizing interactions between students and a robotic tutor. They define interactive personalization as "the process by which an intelligent agent adapts to the needs and preferences of an individual user through eliciting information directly from that user about his or her state." In a preliminary experiment to explore the elicitation frequency of learning sensitive information by college students, a social robot either elicited learning information by asking direct questions to the students 25% of the time, 50% of the time or never elicited learning sensitive information (and asked other questions instead). Contrary to their initial hypothesis, the results showed that high elicitation levels contributed to participants' positive perception of the learning session and the robot. Elicitation appears natural in a learning environment where question and answering is common, but frequent elicitation might have a different impact in other types of human-robot conversation.

### Memory and Persistence

A number of authors have highlighted the critical role of memory in long-term human-robot interaction [5, 16, 1]. While several computational models of memory for virtual characters and robots have been proposed in the literature [11, 27, 17, 24, 23, 20], less attention has been given to the empirical effects of memory in repeated human-robot interactions. One exception is the work by Matsumoto et al. [18] who developed a computational model of spatial memory to enable a robot to collect shared experiences with a user and establish common ground. The model was trained with data collected from people window-shopping and then used by a robot to predict the locations that a user might recall in order to provide appropriate directions in a shopping mall. The authors report the results of a study validating the usefulness of this model. More importantly, this work stresses the importance of implicit shared past experiences in human-robot interaction.

Most relevant, Hastie et al. [10] report a study with an artificial tutor that either referred to a previous interaction (*with-memory* condition) or did not reference a previous interaction (*no-memory* condition) while providing assistance to students. The authors found that the references to past events by the tutor increased students' success in the learning task. Despite the learning gains, students in the *with-memory* condition reported liking the tutor less and judged its instructions harder to follow. These are interesting results when considered in combination with those obtained by Clabaugh and Matarić [6]. It appears that the process of elicitation itself is positive, but the way a robot makes use of elicited information in the future is a more delicate matter. The study described in the next section begins to explore this issue with young children.

## METHOD

To explore the effect of referencing prior interactions on young children's experiences with robots, we must contrast dialogs that are reasonably equivalent with respect to features other than the experimental manipulation. Individual differences in life experience and interests are likely to produce free form conversations that lack the necessary degree of commonality. Similarly, developmental differences across the age range we work with (four to ten years old) make it difficult to get comparable task-based dialogs. To keep the topic constant we designed moments of contrast into four dialogs that were interspersed with, and about, a common set of activities in an hour-long session of play. Each non-PIPER activity is a research project with its own goals and hypotheses that was instrumented to provide information necessary to distinguish the PIPER conditions in the dialog that followed it.

### Experimental Design

Although introduced briefly above, we explain the three experimental conditions more fully here before continuing with a description of the participants and procedure:

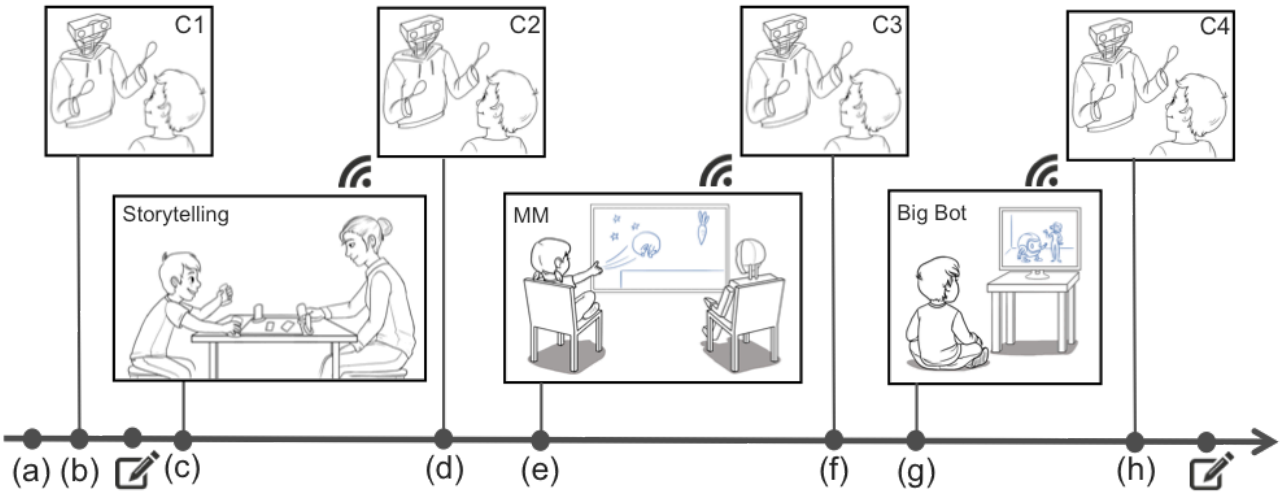



Figure 2: Procedure of the study as described in Section 3.3. Each child has four conversations with PIPER (top row), interspersed with and about the same set of non-PIPER activities. At the end of each activity, children make choices and fill out activity-specific scales that are communicated over the local network to PIPER's dialog manager for use in the P+ condition. Immediately after the first and last conversations, children are asked direct questions about their experience with PIPER.

- *Control (C):* PIPER has no persistent memory, and its language never implies prior interaction. Even if a piece of information is disclosed in a previous conversation, PIPER would need to elicit it again to bring it into the common ground or to predicate behavior on it. It should be noted, however, that all children wore a "name tag" with a character on it and were addressed by that character's name in all conditions. Although it is possible that children would interpret the reuse of their character name as an indicator of familiarity, this personalization did not require persistent memory because the tag was visible. The alternative would have been to use no names, violating expectations in the persistence conditions (P and P+), or to use names only in those conditions, the effect of which might have overshadowed any other contribution of memory [12].
- *Persistent Memory (P):* PIPER is able to remember prior conversations. As a result, it can refer back to previous content without needing to elicit it again, and can predicate its behavior on the basis of previously elicited information. However, in this condition PIPER offers its own opinions only reactively, in congruence with but only after eliciting an opinion from the child. With respect to opinion, then, the P condition is similar to C; both would rely on a dialog path that asked,"Did you like it?" before offering the opinion,"So/Neither did I."
- *Pro-active Persistent Memory (P+)*: PIPER has all of the abilities and behaviors available in P but also offers its own ideas and opinions before the child's are in the common ground. To try to ensure that the comments are congruent with the child's state, PIPER has access to information from the activities that are the topic of conversation but occur out of view. In general that external information was the child's self-report about how much s/he liked the activity. So, for example, if the rating was high/low, pro-active PIPER would follow an eliciting question immediately with a statement containing the same valence: "Hey, what did you think of <activity>? I like/don't like that one."). In short, P+ PIPER acts as if it has a deeper model of the child and can take the risk of exposing its own ideas and opinions first to promote rapport.

#### Conversational Participants

Eighty-one children were recruited through postings in physical and online community bulletin boards. Because PIPER dialogs were about the non-dialog activities and condition effects were expected to accumulate over time, 14 children were excluded from the analysis due to failure to complete at least one activity or dialog for personal reasons, or hardware or software failure. In our final sample, $N = 67$ (35 female and 32 male), and children's ages ranged from 4 to 10 years ($M = 7.00$, $SD = 1.77$). Participants were assigned to one of the three study conditions in order to balance with respect to age and gender. The C group consisted of 13 females and 10 males ($M = 7.11$ years, $SD = 1.79$), the P group contained 11 females and 11 males ($M = 7.10$ years, $SD = 1.85$), and the P+ group included 11 females and 11 males ($M = 6.79$ years, $SD = 1.74$). The study was approved by an Institutional Review Board, and children were compensated for their participation.

The other dialog participant was PIPER (see left side of Figure 1), an upper body torso teleoperated via an identical master through a hybrid air-water configuration. The robot has four degrees of freedom (DOF) in each arm and a 2-DOF neck. Stereo cameras mounted on the robot's neck stream real-time video to the operator's head-mounted display, which in turn maps the head orientation of the oper-

ator to the neck servos of the robot. A menu with options that allow the human operator to perform the language understanding capabilities for PIPER is rendered on top of the video stream. The human operator selects the menu option closest to what the child said using joysticks mounted in each arm of the robot. Through this setup, the operator is visually immersed in the robot's physical space and can "puppeteer" the robot's movements behind the curtain in real-time, while PIPER's responses and the storage and reuse of elicited information are all controlled by a dialog system implemented in Unity3D [26].

### Procedure

Because the content of PIPER conversations was grounded in non-dialog activities and information from those activities distinguishes the P and P+ conditions, we explain the experimental procedure with a walk-through of the session the way the child experienced it, as shown in Figure 2.

**(a) Consent, name tags, practice scales:** After parental consent, children were asked to choose a name tag with a familiar character's picture and told that they would be addressed by that character's name during activities. Next, they participated in a short practice session to learn about *Smileyometer* rating scales [22], answering three simple questions designed to elicit responses across the scale's range ("How much do you like ice cream?" "How much do you like broccoli?" and "How do you feel when you stub your toe?"). The experimenter then invited the child to meet PIPER for the first time on their way to a storytelling activity. The robot was strategically located in a corridor to make it appear natural for children to stop by as they went to and from the non-dialog activities.

**(b) Conversation 1 (Baseline):** Except for the farewell ("bye" versus "see you later"), this conversation is identical across experimental conditions. PIPER introduces itself to the child and asks which activity the child is going to do first. When the child replies (or the experimenter intervenes, if the child does not remember), PIPER gives directions to a new room where the activity has been relocated. Once they are out of sight of PIPER and before they reach the storytelling room, children are asked two *Smileyometer* questions - "How much did you like talking to PIPER?" and "How friendly is PIPER?" - in order to get a baseline reaction.

**(c) Activity 1 (Storytelling):** This guided storytelling activity explores different styles of prompting. Children start by selecting the main character of their superhero story from a fixed set of visuals, then new objects are introduced pictorially at fixed time intervals with either a general or story-specific prompt (e.g.,"Ok, let's include a kitten." versus "What if this kitten drinks that potion?"). At the end, children answer *Smileyometer* questions about how much they like the activity and how brave their character was in the story they created. The name of the chosen superhero character and the self-reported value for the first question are communicated to PIPER and determine the dialog flow if the child is in the P+ condition. A full description of the research aims and results for this activity can be found in [25]. As children leave the Storytelling room they are told that the next activity will involve a new robot, named Sammy, and a video game called *Mole Madness*.

**(d) Conversation 2 (Storytelling)**

When the child approaches PIPER after Storytelling, the robot uses a different greeting for control and persistence conditions ("Hi there, <character name>" versus "Hey <character name>, you're back!"). It then introduces the topic of storytelling. In the C condition this must be done by elicitation ("What activity did you just do?"). In the P condition, however, PIPER remembers that the child was going to Storytelling and simply asks whether the child had fun during the activity. For children in the P+ group, PIPER predicates its opening using the information provided by the child's *Smileyometer* choice at the end of the activity. If children selected one of the two higher ends of the scale, PIPER says "You look like you had fun in Storytelling!" and otherwise, "Hmm, doesn't look like Storytelling was much fun."

Another critical moment in this conversation happens a few turns later when PIPER asks about the contents of the child's story, in particular the superhero character selected by the child. In the C and P conditions, this is done through direct elicitation ("So, tell me your story. Who was in it?"), but in the P+ condition PIPER proactively offers its own value judgment of the child's character by injecting the experimenter-provided information during the elicitation ("So, tell me your story. Who was in it? Wait, wait, was <child's choice> in it? I hope <child's choice> was in it!!").

Differences between conditions also occur at parting. PIPER ends C conversations with a simple "Have fun," but P and P+ conversations end with an explicit reference to seeing the child again after the next activity.

**(e) Activity 2 (*Mole Madness*):** This cooperative, speech-based videogame is played with a very different kind of robot, introduced as Sammy. The goal of the game is to move an animated mole through its environment using a small set of keywords. Children play multiple levels with different versions of Sammy's behavior, and complete *Smileyometer* scales about how much they liked the game and how good a player Sammy is. These values are transmitted to PIPER to be used in the P+ condition.

**(f) Conversation 3 (*Mole Madness*):** As in Conversation 2, greetings immediately establish whether PIPER remembers where the child was going (P and P+) or not (C). In the case of P+, PIPER goes further by adding either a positive comment about the game, or commenting on how hard the game is, depending on the child's *Smileyometer* data.

A second contrasting moment in this conversation occurs when PIPER says that the mole needs a vacation and asks the child to suggest a replacement character for the game. After the child responds, PIPER offers an alternative. In the C condition the alternative is always Iron Man, a superhero that is not one of the Storytelling characters. In the P condition, PIPER suggests Trash Can Guy, a character that is always the villain introduced in Storytelling and thus a character that can

be known simply by remembering that the child did the earlier activity. In the P+ condition, PIPER suggests the specific character the child chose, thus reintroducing the shared preference established in Conversation 2 back into the common ground.[1]

This dialog ends when PIPER offers to tell its own story, but the experimenter mentions that it's time for them to go to BigBot's Big Adventure. As in prior conversations, PIPER acknowledges departure with either a simple goodbye (C), or a reference to the next time they will meet (P, P+).

**(g) Activity 3 (Bigbot's Big Adventure):** This problem-solving task has children talk to screen-based characters in an interactive story. Zoe and Smallbot are in Zoe's tree house when their friend, Bigbot, runs out of battery power. Children help Zoe and Smallbot get Bigbot back to the lab by choosing the right tool to solve problems that come up along the way. A full description of the research aims and results for this activity can be found in [4]. At the end of the activity, children complete *Smileyometer* scales about whether they liked the game and whether they thought it was easy to solve the problems, results of which are communicated to PIPER.

**(h) Conversation 4 (Bigbot's Big Adventure):**

The first turns continue to differentiate conditions. In C, a greeting without a familiarity marker is followed by eliciting the name of the last activity and then the child's success in completing it. In P, the greeting refers directly to the information disclosed at the end of Conversation 3, then PIPER elicits success. In P+, PIPER also greets familiarly and knowledgeably ("Hi, again <character name>. I want to hear about Bigbot.") but extends the turn with a pro-active judgment: "I hope you saved him!" if the child found the task easy, and "I gotta tell you I do not like that game," if the child did not.

Later PIPER acknowledges that this is the last dialog and either elicits the child's favorite activity for the session (C and P), or self-interrupts the elicitation (P+) and guesses the child's preference based on the full history of the child's self-reports ("I bet it was storytelling!"). Note that although some information about the child's opinion for each activity was explicitly elicited and remembered in the P condition, PIPER does not integrate and front that knowledge in this final dialog with those children.

**(i) Post-conversation Measures:** After the fourth conversation and out of sight of PIPER, children were given the same *likeability* and *friendliness* scales as in (a). They were also given four stickers and asked to give each sticker to one of the robots they had talked to during the session (PIPER, Sammy, or Smallbot). Children were told that they could offer any number of stickers to any of the bots, and were verbally prompted with the intended meaning of each award. In particular, the stickers showed a pair of glasses ("Who is smarter? Give the glasses to the bot you think is smartest."), a birthday cake ("Who is more fun? Give the birthday cake to the bot you would invite to your birthday party."), an apple ("Who would be a better teacher? Give the apple to the bot you would chose to be your teacher."), and a gold medal ("Who did you like the most? Give the gold medal to your favorite bot").

[1]In the story retell in Conversation 2, children are asked who their character was, so in theory it would have been possible to store and reuse that information in Conversation 3 in the P condition as well. However, not all children will remember or name their character during the retell. We would have had to add an extra turn to P to make sure the name was introduced into the common ground for PIPER to reuse it, creating an imbalance in either dialog length or topic flow across conditions. We chose, instead, to use Trash Can Guy in P, making the back reference without the need for an extra turn.

### Behavioral Measures

Conversations with PIPER were audio and video-taped. Thus, in addition to the self-report measures collected at points (a) and (h), we were able to compute behavioral measures for the child's expression of positive or negative affect in the turns where PIPER's dialog was different across conditions[2]. For empirical purposes, we consider that such *critical turns* begin when PIPER finishes an utterance specific to one or two experimental conditions, and end at the beginning of the robot's next utterance.

Two coders with experience in behavioral analysis marked segments where children were clearly expressing positive or negative affect. We excluded negative affect from subsequent analysis because such annotations occurred in less than 5% of the critical turns, regardless of coder. With respect to positive affect, reliability between coders was moderate ($\alpha = 0.69$), with most of the differences occurring in judgments about children in the control condition in Conversation 2. For clarity of exposition, and because their overall patterns are otherwise consistent, we present one annotator's results. By combining her annotations with the start and end times for PIPER's dialog, we computed the presence of positive affect in each critical turn based on whether there was an annotation that overlapped at least 25% of that turn duration. We extracted a total of 616 critical turns for all participants in Conversations 2 to 4. The average number of critical turns per conversation was about three and corresponded to approximately 30% of the total conversation in each condition.

## RESULTS

Our experience with young children suggests the importance of multiple measures in drawing conclusions about the effect of any experimental condition on emotion or behavior [15]. To that end, likeability, friendliness and positive affect measures look at change over time with PIPER, while the sticker data looks at cumulative perception of PIPER versus other robots in the session. The different measures also look for change at different levels of granularity, with likeability and friendliness measuring large-scale change in attitude from the first to the last conversation, and the affect data offering a window into conversation-by-conversation change. Our experience also suggests that differences due to age are likely to be present in language interactions; as a result, statistics that

[2]In Conversation 1, only the farewell differentiates between the persistence and control conditions. Because many children walked away as soon as the experimenter said it was time to leave, missing PIPER's actual parting words, we include critical turns only for Conversations 2, 3, and 4 in the analysis.

Table 1: Mean and Standard Deviations (in parentheses) of the perceived likeability and friendliness measures collected after the first and last conversations with PIPER. Results in bold represent the Mean across all conditions.

| | | 4 to 6 years | | 7 to 10 years | | All ages | |
|---|---|---|---|---|---|---|---|
| | | **c1** | **c4** | **c1** | **c4** | **c1** | **c4** |
| Likeability | C | 4.14 (.95) | 4.64 (.74) | 4.54 (.52) | 4.62 (.65) | 4.34 | 4.63 |
| | P | 4.39 (.77) | 4.31 (.99) | 4.25(.75) | 4.58 (.90) | 4.32 | 4.45 |
| | P+ | 4.82 (.60) | 4.64 (.81) | 3.91(.94) | 4.46 (.69) | 4.37 | 4.55 |
| | | **4.45** | **4.53** | **4.23** | **4.55** | **4.34** | **4.54** |
| Friendliness | C | 4.57 (.85) | 4.93 (.27) | 4.85 (.38) | 4.92 (.28) | 4.71 | 4.93 |
| | P | 4.62 (.87) | 4.54 (.66) | 4.67 (.65) | 4.67 (.65) | 4.65 | 4.61 |
| | P+ | 5.00 (.00) | 5.00 (.00) | 4.64 (.67) | 4.73 (.65) | 4.82 | 4.87 |
| | | **4.73** | **4.82** | **4.72** | **4.77** | **4.73** | **4.80** |

collapse across age may show no effect because the trends in different age groups are at odds. At the same time, there are simply not enough children in our study for full condition X ages analyses. As a compromise we explore age-related trends by dividing participants into two groups (4 to 6 and 7 to 10 years) based on the child developmental literature [21] and re-examine the data with respect to age group subsequent to the computation of main effects. Such analyses are particularly important when top-level results are not significant.

### Likeability and Friendliness

The *Smileyometer* data on perceived likeability and friendliness was converted into two variables using a 5-point scale where 1 was the least positive smiley and 5 was the most positive. We conducted a repeated-measures ANOVA to investigate the effects of these two measures, collected after the first and the last conversations with PIPER, across study condition (C, P and P+) and age group.

For perceived likeability, we found a significant main effect for conversation $F(1,68) = 4.48, p < .05, \eta^2 = .07$, such that children reported to like PIPER more after the last conversation ($M = 4.54, SD = .85$) than after the first conversation ($M = 4.34, SD = .80$) regardless of their assigned condition and age. No significant interaction effect was found between conversation and condition, $F(2,68) = .27, p = .77, \eta^2 = .01$, nor in the interaction between conversation and age, $F(1,68) = 1.60, p = .21, \eta^2 = .02$. However, a significant interaction effect was found between conversation, condition and age group, $F(2,68) = 3.39, p < .05, \eta^2 = .10$. In particular, if we examine the top of Table 1, we see that the younger children showed a gain in mean likeability only in the control condition, while the same degree of gain occurred in the older children only in the P+ condition.

Regarding perceived friendliness, we found no significant main effect for conversation $F(1,68) = 1.55, p = .22, \eta^2 = .02$, nor in the interaction between conversation and condition, $F(2,68) = 1.67$, $p = .20, \eta^2 = .05$, the interaction between conversation and age group, $F(1,68) = .10$, $p = .76, \eta^2 = .00$, or in the interaction between the three factors, $F(2,68) = 1.07, p = .35, \eta^2 = 0.03$. As can be seen in the second half of Table 1, initial scores for this measure were so high that there was less room for positive change than in the case of likeability. Of course, there was more room for negative change, but our primary constraint in designing interactions is always that the children have an enjoyable experience, so all versions of PIPER's dialog were phrased in a friendly manner.

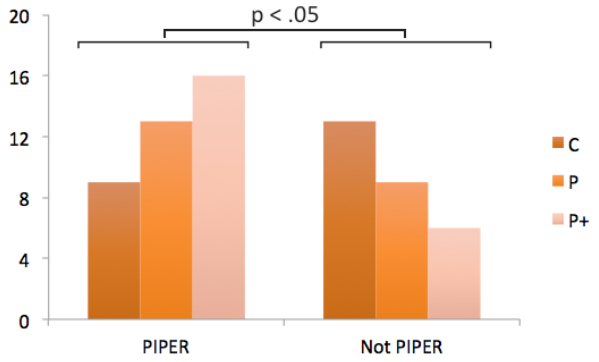



Figure 3: Number of sticker glasses received by PIPER in comparison with the other robots, by study condition.

### Award Stickers

To analyze the sticker attribution data, we created a contingency table with one categorical variable for each sticker and the frequencies of children that offered that sticker to PIPER versus either of the other two robots (not-PIPER). We then conducted Pearson's chi-square tests with these categorical variables to examine the relation between the study condition and the number of stickers that PIPER received compared to the other robots. The relation between the number of glasses stickers (Who is smarter?) that PIPER received across study condition was significant, $\chi^2(2) = 5.24, p < .05$. Children in the persistence conditions were more likely to give the glasses sticker to PIPER while children in the control condition were more likely to give the glasses to a different robot (see Figure 3). The relation was not significant for the cake (Who is more fun?), $\chi^2(2) = .82, p = .33$, apple (Who would be a better teacher?), $\chi^2(2) = .74, p = .35$ or gold medal (Who is your favorite?) stickers, $\chi^2(2) = 3.53, p = .09$.

Although the smartness metric was the only one to achieve significance across conditions, the raw data suggest some age patterns that might be more robust in a larger sample. Table 2 shows the number of stickers received by PIPER versus the other robots, divided by the three study conditions and age. Note that the older children contribute more to the glasses win for PIPER in the persistence conditions, while the younger children contribute more to the not-PIPER win in the control condition. Moving right in the table, the pattern for cake (fun) seems the same across age groups, but the pattern for apple (teacher) is inverted: older children in all conditions voted for PIPER but younger children in all conditions voted for a different robot. Finally, when we look at which robot receives the medal as favorite, the raw scores for all children (bottom two rows) follow the same pattern as glasses - preference for

Table 2: Contingency table with the number of stickers that each child offered to PIPER and to the other robots after the four conversations, divided by age group and condition. (A few children insisted on splitting a sticker across PIPER and one of the non-PIPER robots. In these cases, a point was awarded to each.)

| | | Glasses (smarter) | | | Cake (fun) | | | Apple (teacher) | | | Medal (favorite) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **C** | **P** | **P+** | **C** | **P** | **P+** | **C** | **P** | **P+** | **C** | **P** | **P+** |
| 4 to 6 | PIPER | 3 | 5 | 7 | 4 | 2 | 2 | 4 | 2 | 3 | 3 | 7 | 3 |
| | Not PIPER | 8 | 6 | 4 | 7 | 9 | 9 | 7 | 9 | 8 | 8 | 4 | 8 |
| 7 to 10 | PIPER | 6 | 8 | 9 | 4 | 3 | 4 | 9 | 8 | 7 | 4 | 5 | 9 |
| | Not PIPER | 5 | 3 | 2 | 8 | 8 | 7 | 3 | 3 | 4 | 7 | 5 | 2 |
| All | PIPER | 9 | 13 | 16 | 8 | 5 | 6 | 13 | 10 | 10 | 7 | 12 | 12 |
| | Not PIPER | 13 | 9 | 6 | 15 | 17 | 16 | 10 | 12 | 12 | 15 | 9 | 10 |

PIPER versus non-PIPER robots in persistence conditions - but the patterns by age are not as pronounced.

### Affect

A one-way ANOVA was conducted to investigate the impact of study condition (C, P or P+) on the presence of positive affect in children during the critical turns in the latter three conversations. There was no statistically significant difference between conditions, collapsed across conversation, $F(2,218) = 2.40, p = .09, \eta^2 = .02$. Despite the lack of main effect, we nevertheless expect repeated interaction with the robot to show a novelty effect, per [9, 16], and planned comparisons with conversation number as a within-subject factor, keeping study condition as a between-subject factor. Of course, this was possible only because the topics in each conversation were the same for all conditions.

Results showed a statistically significant main effect for conversation number in a repeated measures ANOVA, $F(2,69) = 11.08, p < .05, \eta^2 = .20$, but no significant interaction effect between conversation number and condition, $F(4,140) = 1.24, p = .30, \eta^2 = .04$. Post hoc analyses applying Bonferroni correction indicated that the proportion of positive affect was significantly higher in Conversation 2 ($M = .48, SD = .38$) than Conversation 3 ($M = .31, SD = .31$) and Conversation 4 ($M = .24, SD = .30$), $p < .05$, as novelty wore off, but that the overall difference in affect between Conversations 3 and 4 were not significant, $p = .55$.

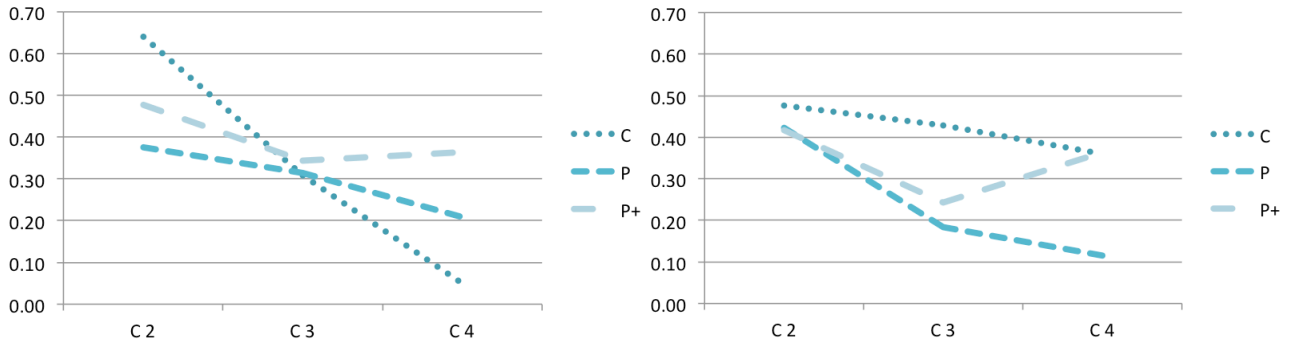
Based on both the condition X conversation effect and *a priori* differences in language sophistication with age, we carried out individual planned comparisons for each conversation, considering age group as a factor. For the critical turns of Conversation 2, we found no significant main effect for condition, $F(2,68) = 1.14, p = .34, \eta^2 = .03$, nor in the interaction between condition and age group, $F(2,68) = .50, p = .61, \eta^2 = .01$. Similarly, the analysis of variance conducted for Conversation 3 yielded no main effect for condition, $F(2,68) = .84, p = .44, \eta^2 = .03$, or in the interaction between condition and age, $F(2,68) = 1.31, p = .28, \eta^2 = .04$.

In Conversation 4, however, there was a significant main effect for condition, $F(2,67) = 3.56, p < .05, \eta^2 = .11$. Post hoc analyses applying Bonferroni correction indicated that the proportion of positive affect was significantly higher ($p < .05$) in children assigned to the P+ condition ($M = .36, SD = .25$) than children in P ($M = .16, SD = .27$), but no significant differences were found between either of the persistence conditions (P and P+) and the C condition ($M = .21, SD = .33$). We also found an interaction effect between condition and age, $F(2,67) = 4.07, p < .05, \eta^2 = .12$. The age-distinct patterns of positive affect over time can be seen clearly in Figures 4a and 4b. The older children show a typical novelty effect in the control condition, slower attenuation in the persistence conditions, and a small reversal of trend for P+ in the final conversation. In contrast, younger children's affect remains fairly steady in the control condition and attenuates quickly in the persistence conditions, except for P+ in the last conversation, where they show an even more pronounced reversal than the older children.

## DISCUSSION

Looking across all of the results presented in the previous section, a consistent picture emerges in which younger and older children have different basic reactions to the conversational strategies associated with persistent memory in PIPER's interactions. Overall children's cumulative experience with the robot led to a significant increase in their likeability ratings, but older children's ratings were more likely to increase in the persistent conditions (particularly P+) while younger children's ratings were more likely to increase if they were in the control condition. Similarly, cumulative experience led children to judge PIPER as more intelligent than the other robots with whom they interacted, but the effect was more pronounced if you were older and in a persistent condition and less pronounced if you were younger and in the control. Looking at positive affect conversation by conversation, all children showed evidence of a novelty effect, but affect dropped more steeply in C for older children, and in P and P+ for younger children, at least until Conversation 4. Indeed, there is clearly something about what children experience in the P+ condition in Conversation 4 that is special, irrespective of age.

What was it about Conversation 4 that mattered and why did it only matter for P+ in that culminating context? Recall that the difference between C and the other conditions is persistent memory, which means that if topic flow across conditions is to be kept as similar as possible, C PIPER will have to elicit information that P and P+ PIPER do not. The essential difference between P and P+ is that P+ has a pro-active conversational strategy, but to create that distinction we provided PIPER with additional information in the P+ condition without it having to be elicited. Thus, while we intended to create a different persona for P+ PIPER, one that sounded more varied in its conversation and appeared to be more vested in building a relationship by risking exposing its opinions first, we inadvertently created a situation in Conversation 4 in which P PIPER was hamstrung. The important moment in Conversation 4, the culmination of the use of persistent memory, occurs when PIPER moves to the topic of the child's fa-




(a) 7 to 10 year olds

(b) 4 to 6 year olds

Figure 4: Proportion of critical turns with positive affective reactions divided by age groups for each conversation with PIPER (excluding baseline Conversation 1) in the control (C), persistence (P) and augmented persistence (P+) conditions.

vorite activity. In C this is done through simple elicitation. In P+ it is done by performing a computation over all the stored information that has been provided from the various activities and making a guess. In P, however, PIPER also uses simple elicitation, like C, despite having all the same information that is available in P+. In P the information was gained over accumulated elicitations rather than by unseen means, but to use it would violate the strategy difference between the conditions. In other words, by Conversation 4 the difference between P and P+ was less about whether each version had the information necessary to perform the computation of which activity to guess was the child's favorite, than that P was not allowed to express the results of that computation.

We conjecture that if we had allowed P PIPER to follow the elicitation of the final value at the start of the conversation (how easy it was to help Bigbot) with a proactive evaluation of the child's experience during the entire session, P PIPER's measures would have mirrored P+ PIPER's, even as it violated the distinction between the two conditions. This is conjecture, however, and not just because we didn't try it, but because the origin of the memories might matter. In P+ the information that went into the guess was self-report via *Smileyometer* scales, but in P the information was self-report in conversation. We observed that not all children in the C and P conditions were consistent with regard to how they answered with the scale and what they said in subsequent elicited response. Absent a conversational turn that grounded the child's opinion in situ, P+ PIPER was effectively making an inference and that inference could have been (and occasionally was) wrong. Had we let P PIPER guess as well, the fact that it was using grounded values – that the child knew s/he had provided those values to PIPER – might have made the personalization more powerful.

Stepping back, the data suggest to us that although older and younger children have different baseline preferences, persistence matters when the content it makes available matters, regardless of age. The preference for the C condition in the younger children may be because simple elicitation and response is a familiar and easily understood format. Or it may be that any conversation that elicits more responses will, absent other factors, be more enjoyable to them. It is also possible that the youngest children did not notice the familiarity markers, or noticed them but did not understand them as intended, or recognized the intent but did not build up a model of mutual familiarity over time. But while small intimations of familiarity might not have had much cumulative impact, the moment of personalization that resonated – the moment when they understood that PIPER knew something important about them – that use of memory had a noticeable effect. The story is different for the older children, but consistent nonetheless. They prefer the kinds of conversation that are more typical for their ages, and the control condition's simplicity and predictability is less enjoyable, albeit no less friendly. The steady accumulation of markers of familiarity and personalization make for a better experience, and the guess in Conversation 4 signals another instance that matters, but it is less impactful because of the cumulative effect.

## CONCLUSION

We studied how different conversational strategies, with and without the ability to refer to prior events, effect children's experience with a robot over repeated interactions. To create a common backdrop for evaluation, each child played the same games and activities and then had conversations about them with the robot. Our method allowed the robot's dialogs to be both automated and predicated on information that distinguished the experimental conditions. We find that the small expressions of familiarity associated with persistent memory were important only to the older children, but that a personalized observation made on the basis of accumulated information had a positive effect across the age range.

The work reported here is a preliminary study in which we were as interested in developing a new methodology for evaluating repeated child-robot conversations as we were in understanding the outcomes. Thus, a number of limitations must be acknowledged. First, to have both repeated interactions and non-trivial common conversation across subjects, we had to create a setup that was complex in both conception and execution. That complexity is why we decided to recruit

children across the age range of interest within the duration of the study rather than try to run a smaller number of ages first and the remainder at some future time. As a consequence, our results suggest several age differences that a larger sample of participants per age might have helped clarify. Although exact replication of this study will be difficult for any group, including our own, studies exploring the same conversational phenomena are needed to verify the interpretation these data suggest.

Second, it seems clear that the more conversations there are, the less likely it is that simple pre- versus post measures will uncover important patterns of change related to cumulative effect. We collected multiple measures at different levels of granularity for this reason, but the research community is still developing techniques and standards in this regard [16, 2]. We acknowledge that the measures used in this study are unlikely to be the only ones relevant to understanding the cumulative effect of conversational variables over repeated interaction between robots and children.

Finally, although long-term is by definition repeated, repeated is not necessarily long-term. We cannot assume that the results reported here will apply to long-term child-robot interactions, and future work should address that issue. Of course, long-term interaction with children must entail a theory of how persistent an item of elicited information should be, i.e., how long it should be held to be true given developmental change. Absent such a theory, we see focusing on repeated interaction over a period of time in which what the robot remembers is still likely to be what the child believes, as a useful beginning.

## REFERENCES


1. P. Baxter and T. Belpaeme. Pervasive memory: the future of long-term social HRI lies in the past. In *Third International Symposium on New Frontiers in Human-Robot Interaction at AISB*, 2014.
2. P. Baxter, J. Kennedy, E. Senft, S. Lemaignan, and T. Belpaeme. From characterising three years of HRI to methodology and reporting recommendations. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 391–398, March 2016.
3. T. Bickmore and J. Cassell. Relational agents: A model and implementation of building user trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 396–403, New York, NY, USA, 2001. ACM.
4. E. J. Carter, J. Hyde, and J. K. Hodgins. Investigating the use of interactive features for children's television programming. In *Proceedings of the The 16th International Conference on Interaction Design and Children*, IDC '17, New York, NY, USA, 2017. ACM.
5. G. Castellano, R. Aylett, K. Dautenhahn, A. Paiva, P. W. McOwan, and S. Ho. Long-term affect sensitive and socially interactive companions. In *Proc. of the 4th International Workshop on Human-Computer Conversation*, 2008.
6. C. Clabaugh and M. J. Matarić. In *The 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016.
7. H. H. Clark. *Using language*. Cambridge university press, Cambridge, 1996.
8. S. R. Fussell and R. M. Krauss. Coordination of knowledge in communication: effects of speakers' assumptions about what others know. *Journal of personality and Social Psychology*, 62(3):378, 1992.
9. R. Gockley, A. Bruce, J. Forlizzi, M. Michalowski, A. Mundell, S. Rosenthal, B. Sellner, R. Simmons, K. Snipes, A. C. Schultz, and J. Wang. Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1338–1343, Aug 2005.
10. H. Hastie, M. Y. Lim, S. Janarthanam, A. Deshmukh, R. Aylett, M. E. Foster, and L. Hall. I remember you!: Interaction with memory for an empathic virtual robotic tutor. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '16, pages 931–939, Richland, SC, 2016. International Foundation for Autonomous Agents and Multiagent Systems.
11. W. C. Ho, K. Dautenhahn, M. Y. Lim, P. A. Vargas, R. Aylett, and S. Enz. An initial memory model for virtual and robot companions supporting migration and long-term interaction. In *The 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 277–284, Sept 2009.
12. T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro. Interactive robots as social partners and peer tutors for children: A field trial. *Human-computer interaction*, 19(1):61–84, 2004.
13. S. Kiesler. Fostering common ground in human-robot interaction. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communicationn (RO-MAN)*, pages 729–734. IEEE, 2005.
14. G. Kreutler and D. Jannach. Personalized needs elicitation in web-based configuration systems. In *Mass Customization: Challenges and Solutions*, pages 27–42. Springer, 2006.
15. I. Leite and J. F. Lehman. The robot who knew too much: Toward understanding the privacy/personalization trade-off in child-robot conversation. In *Proceedings of the The 15th International Conference on Interaction Design and Children*, IDC '16, pages 379–387, New York, NY, USA, 2016. ACM.
16. I. Leite, C. Martinho, and A. Paiva. Social robots for long-term interaction: A survey. *International Journal of Social Robotics*, 5(2):291–308, 2013.
17. M. Y. Lim. *Memory Models for Intelligent Social Companions*, pages 241–262. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

18. T. Matsumoto, S. Satake, T. Kanda, M. Imai, and N. Hagita. Do you remember that shop?: Computational model of spatial memory for shopping companion robots. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, pages 447–454, New York, NY, USA, 2012. ACM.

19. Y. Matsuyama, A. Bhardwaj, R. Zhao, O. J. Romero, S. A. Akoju, and J. Cassell. Socially-aware animated intelligent personal assistant agent. In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 224, 2016.

20. M. Oliveira, G. H. Lim, L. S. Lopes, S. H. Kasaei, A. M. Tomé, and A. Chauhan. A perceptual memory system for grounding semantic representations in intelligent service robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2216–2223, Sept 2014.

21. J. Piaget. *The moral judgement of the child*. Simon and Schuster, 1997.

22. J. C. Read and S. MacFarlane. Using the fun toolkit and other survey methods to gather opinions in child computer interaction. In *Proceedings of the 2006 Conference on Interaction Design and Children*, IDC '06, pages 81–88, New York, NY, USA, 2006. ACM.

23. S. Rosenthal, S. Skaff, M. Veloso, D. Bohus, and E. Horvitz. Execution memory for grounding and coordination. In *Proceedings of the 8th ACM/IEEE International Conference on Human-robot Interaction*, HRI '13, pages 213–214, Piscataway, NJ, USA, 2013. IEEE Press.

24. D. Stachowicz and G. J. M. Kruijff. Episodic-like memory for cognitive robots. *IEEE Transactions on Autonomous Mental Development*, 4(1):1–16, March 2012.

25. M. Sun, I. Leite, J. Lehman, and B. Li. Collaborative storytelling between robot and child: A feasibility study. In *Proceedings of the The 16th International Conference on Interaction Design and Children*, IDC '17, New York, NY, USA, 2017. ACM.

26. J. Whitney, T. Chen, J. Mars, and J. Hodgins. A hybrid hydrostatic transmission and human-safe haptic telepresence robot. In *2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden (to appear)*. IEEE, 2016.

27. R. Wood, P. Baxter, and T. Belpaeme. A review of long-term memory in natural and synthetic systems. *Adaptive Behavior*, 20(2):81–103, 2011.