A BAYESIAN THEORY OF MIND APPROACH TO NONVERBAL COMMUNICATION FOR HUMAN-ROBOT INTERACTIONS

A COMPUTATIONAL FORMULATION OF INTENTIONAL INFERENCE AND BELIEF MANIPULATION

JIN JOO LEE

S.M., Massachusetts Institute of Technology (2011)
M.S., Georgia Institute of Technology (2009)
B.S., Georgia Institute of Technology (2008)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

AUTHOR

Jin Joo Lee
Program in Media Arts and Sciences
June 2017

CERTIFIED BY

Dr. Cynthia Breazeal
Associate Professor of Media Arts and Sciences
Thesis Supervisor

ACCEPTED BY

Dr. Pattie Maes
Academic Head
Program in Media Arts and Sciences
Much of human social communication is channeled through our facial expressions, body language, gaze directions, and many other nonverbal behaviors. A robot’s ability to express and recognize the emotional states of people through these nonverbal channels is at the core of artificial social intelligence. The purpose of this thesis is to define a computational framework to nonverbal communication for human-robot interactions. We address both sides to nonverbal communication, the decoding and encoding of social-emotional states through nonverbal behaviors, and also demonstrate their shared underlying representation.

We use our computational framework to model engagement/attention in storytelling interactions. Storytelling is an interaction form that is mutually regulated between storytellers and listeners where a key dynamic is the back-and-forth process of speaker cues and listener responses. Listeners convey attentiveness through nonverbal backchannels, while storytellers use nonverbal cues to elicit this feedback.

We demonstrate that storytellers employ plans, albeit short, to influence and infer the attentive state of listeners using these speaker cues. We computationally model the intentional inference of storytellers as a planning problem of getting listeners to pay attention. When accounting for this intentional context of storytellers, our attention estimator outperforms current state-of-the-art approaches to emotion recognition.

By formulating emotion recognition as a planning problem, we apply a recent artificial intelligence method of inverting planning models to perform belief inference. We computationally model emotion expression as a combined process of estimating a person’s beliefs through inference inversion and then producing nonverbal expressions to affect those beliefs. We demonstrate that a robotic agent operating under our belief manipulation paradigm more effectively communicates an attentive state compared to current state-of-the-art approaches that cannot dynamically capture how the robot’s expressions are interpreted by the human partner.
A BAYESIAN THEORY OF MIND APPROACH TO NONVERBAL COMMUNICATION FOR HUMAN-ROBOT INTERACTIONS

A COMPUTATIONAL FORMULATION OF INTENTIONAL INFERENCE AND BELIEF MANIPULATION

JIN JOO LEE

THESIS READER

Dr. David DeSteno
Professor of Psychology
Northeastern University
A BAYESIAN THEORY OF MIND APPROACH TO NONVERBAL COMMUNICATION FOR HUMAN-ROBOT INTERACTIONS

A COMPUTATIONAL FORMULATION OF INTENTIONAL INFEREN CE AND BELIEF MANIPULATION

JIN JOO LEE

THESIS READER

Dr. Fei Sha
Assistant Professor of Computer Science
University of Southern California
ACKNOWLEDGMENTS

If I have seen further than others, it is by standing upon the shoulder of giants.

—Isaac Newton

A dissertation is not solely of the author’s ideas and theories. It is a reflection of their academic community: who did they learn from? who inspired them? My academic upbringing was led by a world-class roboticist and in an environment full of dreamers, helpers, and wizards.

I was very fortunate to have apprenticed under Dr. Cynthia Breazeal. She continually challenges herself and never stops learning. She instills that same level of dedication to being a life-long learner in her students. If she could, she would give her students the world to leverage for our education, but instead she offers us endless resources from cutting-edge technology to finding the right person in her vast network to her own graduate school experiences to pull from. In this fertile playground where dreams can grow, she pushed me to be bold and to relentlessly go after that impossible research idea. She grants you autonomy to pursue your own research because she innately trusts you. Giving someone that much faith and confidence transforms them. She helped me find my voice. Thank you for your endless guidance and encouragement for the last eight year.

Its not every day that you can find someone that shares in the same exact passions you have. I had a whole group. This thesis is a celebration of Matt Berlin’s and Jesse Gray’s research work. Not only did they leave a lasting legacy of cognitive robotics in the Personal Robots Group, they still continually mentor the graduate students, helping them achieve their high bar of excellence. They taught me how to navigate the academic game with a collaborative spirit, a kindness and understanding towards others, while respecting your own work and giving it a deserving spotlight on stage. You set my research tone and taste from the get-go. Thank you so much for helping me call both the robots and Somerville my new home.

Sigurður Örn Adalgeirsson is the funniest and most talented person I have ever met, but what makes him stand out is his kindness and generosity in helping others. Oh, and his superhuman ability to read people like a book. He sees your strengths and weaknesses to mold you out of your shortcomings and shows you how to wield your strengths like a deadly harpoon. He taught me everything I know about hardware to POMDPs. Thank you for being my role model and mentor from Day 0.

Palash Nandy is my life guru. Me: Hey Palash, why do people have friends? Palash: Life stories Me: What is the meaning of life? Palash: Life experiences. We are both, in our own way, fascinated about people, and robots are our medium to better understand humanity. As my officemate, I offered him academic advice and paper critiques, while he taught me
to always celebrate my victories, no matter how small, and how to be a loyal friend. Thank you for forcing me to have fun.

William Bradley Knox gave me the courage to continue pursuing research in machine learning. In helping me publish my very first computational journal paper, he gave me the belief and confidence that I can contribute in what I thought was too difficult a field to play in. You also instilled a lasting dedication to academic rigor as well as to the art of good communication. Till this day, whenever I write, I always hear you say “You need to be more precise with your language Jin Joo.”

Nick DePalma commiserated with me as we both pushed hard toward the finish line. I saw you persevere against your personal and professional obstacles and saw you come out victorious every single time. Thank you for showing me how to never give up.

Samuel Spaulding was always receptive to hearing me complain while I would stress eat his entire box of Triscuit. He would patiently listen and then give the most thoughtful advice, an optimistic outlook, or simply say “You can do this Jin Joo.” His positive energy is contagious, and his faith in people unwavering. Thank you for all of your support.

Hae Won Park is a second sister to me. While I was graduating, she protected me in making sure I was not distracted by any lab-related emergencies. She brought me snacks whenever I forgot to eat. She would hug me and tell me so many times that she is proud of me and impressed with my work. She gives ample time and care to her peers and mentees even though she is overloaded with work. As her officemate, I learned her secret. Occasionally, she mutters under her breath “grit grit grit grit.” Sooyeon Jeong, in the most frustrating of situations, will always make things work while simultaneously finding time to help you. Whenever I needed a favor from borrowing her son to run experiments on to covering for a demo, she was always the most reliable person to lean on. Jacqueline Westlund showed me how to be laser-focused in your research goals with efficient productivity. How she manages to successfully balance between her work and life is still a mystery to me. The females in my lab are all Wonder Woman, and I could not have asked for a better set of role models.

Finally, thank you Gelsomini Mirko for helping me develop the prosodic cue detector. I miss your energy and enthusiasm; it was such a joy having you be a part of our robot family.

The depth of my learning came from my robot family but the breath came from the wonderful collaborations I have had over the years. Dr. David DeSteno is a reliable professor that is always on your side, rooting for you to succeed and generously encouraging you with positive feedback. He taught me how to be more rigorous in my data analysis while also clever when designing my user studies. He gave me a deep appreciation for the fundamentals to human social psychology as well as the importance of understanding human behavior through observation. Thank you for giving me this research foundation my very first year as a graduate student and continuing to support me throughout my academic career.

Dr. Fei Sha had no idea who I was until I showed up in front of his office at USC and asked him to teach me about discriminative HMMs. I still remember your example of predicting night-time traffic as a means to explain gradient decent and optimization to me. He is extremely devoted to his students. He video-conferenced into my general’s oral exam
from the waiting room of the emergency hospital. Thank you so much for your dedication and tutelage for the past 5 years.

Through the SAR Expedition, Dr. Brian Scassellati introduced me to so many awesome people. Most notably, Dr. Cliff Nass pointed out to me an underlying flaw in how the HRI community approaches nonverbal behavior: “It’s not grounded in human mental states.” Our conversations lead me down the path that resulted in this body of work. Ross Mead is my brother from another robotics mother, and we love all things Disney. He told me that researchers either “study their own deficiencies or in studying then form one.” I hope we both one day become more normal with our nonverbal behaviors. Malte Jung, Jolie Wormwood, and Leah Dickens, thank you for teaching me the ins-and-outs of human-subjects experiments from creating good logs to the tricks in getting participants to show up.

My thesis work has been shaped by many constructive conversations with many different researchers. Thank you Dr. Sonia Chernova, Dr. Goren Gordon, Stefanos Nikolaidis, Dr. Marc Toussaint, Dr. Jonathan Gratch, Dr. Matt Grimes, Dr. Anca Dragan, Dr. Sidd Srinivasan, and Dr. Andrea Thomaz for patiently listening to my half-baked ideas and sharing your thoughts.

I learned how to build a robot from collaborating with a team of experts, and together we built the awesome Tega platform: Fardad Faridi, Sigurður Örn Adalgeirsson, Stacy Dyer, Luke Plummer, Kris dos Santos, Meng Xi, collaborators from Cooper Perkins Inc. Harald Quintus-Bosz, Robert Hartmann, and Mike Hess, and collaborators from IFRobots Inc. It was a privilege to lead such a wonderful and talented group of people.

My continued education into hardware lead to an incredible side journey to Shenzhen, China. With my colleagues, Artem Dementyev, David Ramsay, Jeremy Rubin, and Nazmus Saquib, we learned from the best: Andrew ‘Bunnie’ Huang, Sean Cross, and Chris ‘Akiba’ Wang. Thank you for teaching me not only about manufacturing processes but also what the free life of a hacker really means.

The crux of this thesis was the data collection of children storytelling. Dr. Paul Harris provided crucial insights into the expected behaviors of children, enabling a flawless study design. Thank you Matthew Sears for assisting as a study proctor and Yiyu Li, Katharine Brush, Emily Holding, and Raina Hall for diligently annotating the videos. A special thank you to Erin, Carla, and Katie for welcoming us into your school as well as the adorable children participants. Thank you so much for sharing your incredibly silly and creative stories.

It takes a village to support someone through a PhD. The staff at the MIT Media Lab loves their students and genuinely cares about their academic development. Polly Guggenheim is the source of endless hugs and encouraging words. She celebrates with you in your academic achievements, fiercely protects you from possible trouble, and knows exactly what to say at the most despairing of times. She is the crazy aunt that you wish you had that always losses her glasses but somehow always finds money to support your projects. She is the heart that keeps our robot family of misfits functioning within tolerable parameters.

Linda Peterson and Keira Horowitz are true student-advocates. They have the most difficult job of wrangling a group of unresponsive procrastinators to do the most important
thing they came here to do: graduate. Thank you so much for keeping me in check and for always having my best interest in mind.

Jessica Barton, Monica Orta, and Amanda Stoll thank you so much for looking out for my social-emotional well-being while at the lab. Kevin Davis and Cornelle King thank you for always being responsive to the resource needs of our group.

My friends and family gave me the love and support I needed to thrive these past eight years. Sunny Jolly is my dear and cherished friend. He shares in your joy; fumes with you in your anger. In knowing me since middle school, he reminds me of my roots, never letting me stray from my core beliefs and values, while also pointing out how much I have grown at every major turn. Thank you for always helping me whenever I needed you.

Thank you Tito Alfaro, Jie Qi, Bianca Datta, Edwina Portocarrero, David Nunez, Jennifer Jacobs, and David Cranor for being my Media Lab cohorts. Our time at the lab has been a roller coaster of emotions, but we have (or will) made it!

Thank you Grove St. for your overwhelming amount of support. You made chicken soup when I was sick, surprised me in celebrating my birthday, brought some stability in my life through our nightly tv shows, and kept my mind off of research when I needed it the most. I am always so happy to come home. DDP gave me a creative outlet to share with my closest of friends. We are a rare and unique group with a diverse set of talents and perspectives. Thank you for creating an open space to share with you my most intimate of feelings. Thank you to all of my friends for sharing with me your life stories and letting me be a part of your life experience. All my friends mean so much to me.

My family has always encouraged me to pursue my own interests. I am so proud of my sister for receiving her medical doctorate degree and my mom who recently beat me in getting her doctorate in theology. Between the three of us we cover the human study of the mind, body, and soul. Dad, your hard work is what inspired me to tirelessly go after my dreams. I will never forget the sacrifice you made in leaving everything you knew behind to rebuild in a foreign world in hopes that my sister and I would have better opportunities. The strength and fortitude I had to finish this degree was a reflection of what you have shown me throughout my life.

On my first day of graduate school, I met Jesse Gray and said, “JG? wait, i’ve read your C6 code.” Even before joining the group, Jesse taught me everything I know about robotics programming. He deeply cares and cherishes those around him, and his empathy toward others is boundless and even extends to the inanimate. How he brings robots to life is grounded in his ability to see the uniqueness of character in everything. His natural insatiable curiosity is contagious. He never lets me forget to appreciate every opportunity I have been given. For all these qualities and more, I feel in love with my life partner and robot soulmate. Thank you for helping me go through the growing pains of graduate school. I could not have done this thesis without your love, optimism, and your repository of cute cat memes.

This research was supported by the Dolores Zohrab Liebmann Fellowship and William Asbjornsen Albert Memorial Fellowship as well as the National Science Foundation (NSF) under Grant CCF-113898.
For mind readers, mind manipulators, and magicians.
Figure 1  **Storytelling Rounds and Turns.** Each participant had multiple rounds of storytelling with different partners. In a dyad session, a pair of students took turns (T1, T2) narrating their story to their partner. In sum, the data collection consists of 3 rounds (with a supplementary 4th round for redo opportunities) totaling 29 dyad sessions, which equates to 58 individual storytelling episodes.

Figure 2  **Classroom poster** on how to be a good listener.

Figure 3  **Story Space Setup.** Setup included three different camera angles, a high-quality microphone, listener & storyteller chairs, and a story book with compounding story elements per page. The bottom-right photo shows how we labeled each chair to further emphasize to a child his/her role as either the storyteller or listener.

Figure 4  **Video-Recordings of Storytelling Interactions.** Three time-synchronized cameras captured the frontal-view of each participant along with a bird’s-eye view.

Figure 5  **Smileyometer.** A pictorial rating scale developed for children [Read et al., 2002]. Please see the proctor’s script Table 20 in Appendix B in how to instruct children to answer with this scale.

Figure 6  **Estimating the Emission-Time of a Prosodic Cue.** Based on the moment of the backchannel response, we extract the last speaking turn of the storyteller and estimate that its terminating edge is the ending-time of the prosodic cue. This is calculated for all prosodic-based cues except for the pause cue, which is roughly estimated to be halfway between the backchannel time and the terminating edge.

Figure 7  **Timeline Visualization of Annotated Behaviors.** *(Left)* Listener’s nonverbal behaviors are continuous labels through the marked start and stop times. *(Right)* Storyteller’s cueing behaviors are event-based labels based on the prosodic cue’s emission-time and gaze onset-time.
Figure 8  **Video-based Human-subjects Experiment.**
From TRUE interactions between a storyteller and listener, we manipulate the absence and falseness of the storyteller context. For the FALSE condition, we replace the original storyteller with the audio and video of a different storyteller. The ABSENT condition completely removes the storyteller context (both audio and video).

Figure 9  **Video Alignment of Listener Across Condition.** We demonstrate how for the listener (in the white t-shirt) we retain his exact behavior across the three conditions: TRUE (top) FALSE (middle) ABSENT (bottom).

Figure 10  **Example of Scoring Accuracy and Latency.** At frame 440, a listener is annotated by experts as transitioning from an attentive to inattentive state. As such, a participant that reports the transition occurring at 250 frames is marked as being incorrect. A participant that made a prediction at 600 frames is accurate with a latency of 160 frames.

Figure 11  **Extracting Data Tuples from a Demonstration.**
The continuous timelines abstractly represent various behaviors of a listener. The red vertical lines are the detected speaker cues of a storyteller that are then merged into multimodal cueing events, shown as green vertical lines. Given a cueing event, we summarize the listener’s overall valance of response between [0.5 - 3.0] seconds after this event. The listener’s attentive state is sampled immediately at the end of this response window, shown as magenta vertical line.

Figure 12  **Predicting Attentive State of Listeners Based on Cue-Strength and Response-Valence.**
The x-axis represents overall listener’s response as either very positive (+4) to very negative (-3). The y-axis represents the likelihood of attention, or conversely as inattention. (Right) Shows the 95% confidence bounds of strong vs weak cue-contexts. For the same listener response (e.g. x=-2), there is a difference in interpretation if we observed it after a weak vs a strong cue. Strong cues buys us greater certainty that the listener is not paying attention (likelihood of 70%-100% vs 50%-70%).
Co-occurring Nonverbal Behaviors of Listeners. Values are the Pearson’s correlation coefficients. Highly correlated behaviors (in yellow) are prolonged brow-raises and away gazes as well as smiles and partner gazes. Uncorrelated behaviors (in dark blue) are partner gazes and away gazes since they are mutually exclusive behaviors.

Graphical Representation of our POMDP Model. POMDPs consist of two processes: 1 the state estimator maintains a belief about the listener’s attentive state. 2 the policy recommends cueing actions, based on the current belief, that will maximize expected long-term reward.

Diagram of the Apprenticeship Learning Algorithm. The apprenticeship learning algorithm first sampling a candidate POMDP from a prior distribution of possible world models. Then solves for its policy using SARSOP from the APPL toolkit [Kurniawati et al., 2008]. It evaluates how likely the POMDP model and its policy can explain the set of demonstrations (i.e., Equation 8). Based on this objective function, the optimization algorithm searches for an improved POMDP model. The search terminates when either the optimization iteration exceeds the maximum number of evaluations or if the improvement in the objective function is less than a relative tolerance tol.

ROC Curve of Our State Estimators. In ROC space, the upper left corner (0,1) represents perfect classification. In general, a curve closer to this point is the better predictor.

Belief Manipulation Model of a Listening Agent. The model consists of two processes: 1 the belief estimator tracks storyteller’s beliefs about the robot’s attentive state through plan inversion 2 the myopic policy selects listening behaviors to affect those beliefs toward a desired inference. Our greedy policy selects a next response that minimizes the distance between the resulting belief and the target belief.
Figure 18  **Scenario 1: Filtering Inference.** The robot starts with no idea (uniform distribution for \( b_0 \)) what the storyteller thinks about the robot. After the storyteller emits a gaze cue \( a_1 \), the belief manipulation policy suggests for the robot to do a *high-positive* response \( o_1 \). This leads to a next belief distribution \( b_1 \) with a mass slightly closer to its belief goal. The robot again responds with a *high-positive* behavior to a gaze cue. And once more, but now at \( b_3 \) the agent is very close to its belief goal. At the next gaze cue, the robot responds with a *neutral* behavior to remain near the target belief \( b_4 \).

Figure 19  **Scenario 1: Smoothing Inference.** Using smoothing or retrospective inference, the listening agent can reason that the storyteller, given this sequence of gaze cues, began the interaction with a moderately confident initial belief that the robot is attentive, \( \mathbb{E}(b_1) = 0.79 \). This is congruent with their policy in Figure 30 in Appendix C since gaze cues are more likely to occur when the storyteller believes that the listener is paying attention.

Figure 20  **Scenario 2: Filtering Inference.** The robot again starts with a uniform distribution for \( b_0 \). After the storyteller emits a gaze cue \( a_1 \), the belief manipulation policy suggests for the robot to do a *high-negative* response \( o_1 \). This leads to a next belief distribution \( b_1 \) with a mass that is moving in the opposite direction of our goal even though the robot did a negative response. Remember, the next belief is a function of *both* cues and responses. As such, the gaze cue, which is indicative of the storyteller believing the robot to be attentive, contributes to this resulting belief distribution. The robot then responds to the next two gaze cues with *low-negative* behaviors. At \( b_3 \), the agent is very close to its belief goal. At this point, if the robot were to do another *low-negative* or *high-negative* response to the next gaze cue, then the resulting belief distribution would actually cause the robot to overshoot its target belief. To continue being perceived as *moderately* not paying attention, the policy instead suggests for the robot to do a *low-positive* response.
Scenario 2: Smoothing Inference. Like the last example, the listening agent can use retrospective inference to determine that the storyteller began with a high confidence initial belief that the robot is attentive, $\mathcal{E}(b_1) = 0.91$. This would explain why a storyteller would continue with the gaze cues even though the listening agent is responding with inattentive behaviors.

Comparing Robot Listeners. Participants told stories to two different robots in back-to-back sessions. Both robots are capable of contingently responding to the storyteller’s cues, but they are different in their listening strategies.

Tega Robot Platform. Initial concept to hardware design to a working prototype.

Autonomous Robot Listener System Diagram. This perception-to-behavior generation pipeline is realtime and detects prosodic- and gaze-based cues of a human storyteller, decides the agent’s response given its policy, and controls the exhibited set of nonverbal behaviors. The message passing between modules uses ROS, an open-source Robot Operating System [Quigley et al., 2009].

Red and Blue. Participants told stories to both robots one at a time. For each participant, the robots were randomly assigned a listening policy and turn order.

Story Space Setup for HRI Study. Setup includes three different camera angles for recording (C1-C3), a high-quality microphone for prosodic cue detection, a tobii eye tracker for gaze cue detection, and a GoPro camera for live-streaming to parents (C4).

Parent Perception of the Listening Robots. The BTOM robot was rated higher in both active listening skill as well as in human-likeness. It was also perceived as slightly more intelligent but not significantly so.

Eight Storybooks. Each scene introduces new characters or events. The illustrated storybooks are a modification of a prior tablet-based storytelling application [Kory, 2014].
Expected Policy of Storytellers. The x-axis is the storyteller’s beliefs, where 0.0 is a high-confident belief that a listener is not paying attention, 1.0 is a high-confident belief that the listener is paying attention, and the middle region represents varying degrees of certainty. The y-axis is the probability of selecting a cueing action. We expect that the value of a strong cue is to influence inattentive listeners. Storytellers will most likely select this cue when they believe the listener is not paying attention. Also, we expect that the value of a gaze cue is to “check in” on listeners but not try to change their state. Storytellers will most likely use this cue when they believe the listener is attentive. And finally, we expect that the value of a weak cue is to help disambiguate listeners’ state in moments of high uncertainty (i.e., the middle region).

POMDP Policy $\Pi^5 : b \rightarrow a$. The likelihood of selecting a strong cue is higher when storytellers have a belief of an inattentive listener, which is congruent with our expectation that the value of a strong cue is to influence inattentive listeners. The likelihood of selecting a weak cue is mostly uniform across all beliefs. Gaze cues are more likely when storytellers are confident that the listener is attentive. This is congruent with our expectation that the value of a gaze cue is to not change state but instead to gather information. Note: The final resulting policy is different from our initial policy since it depends on the learned POMDP world dynamics.

Transition Function. The gaze cue does not have much effect on changing the attentive state of listeners since the likelihood of transitioning out of states are very small. The weak cue is more likely to transition inattentive listeners to pay attention, while strong cues have the highest chance of transitioning them.
Figure 32  

**Observation Function.** The x-axis is the overall valence of listener response where $-2$ is a *high-negative* response and $+2$ is a *high-positive* response. The y-axis is the probably of observation.  

*(top)* When a listener is attentive and the storyteller does a gaze cue, there is a small but non-zero probability of observing negative responses, but it is more likely to observe neutral or positive responses. This almost has a uniform distribution from 0 to $+2$, which further indicates how a gaze cue is an information-seeking action. For a weak cue, the overall distribution shifts towards the positive end. For a strong cue, an attentive listener will most likely response in a very positive way.  

*(bottom)* When the state of the listener is inattentive, we will most likely observe very negative behaviors. This is collapsed across all cues because of the simplifying assumption we made in Sidebar 2.  

---
### List of Tables

| Table 2 | **Round Differences.** Each round varied in the amount of instruction, type of partner, and number of story scenes. For example, the second round had a light amount of instruction, students with high storytelling ability were paired with students with low ability (based on teacher’s suggestions), and two story scenes were used. Round 3 paired participants with a partner of opposite gender, if they had not previously. | 45 |
| Table 3 | **List of All Annotated Behaviors.** The selected set of nonverbal behaviors were either found in prior works (see Table 15 and Table 16 in Appendix A) or commonly observed in the storytelling interactions. Each annotation category has a set of mutually exclusive labels and was coded for storytellers (S) and/or listeners (L) or jointly evaluated (joint). | 48 |
| Table 4 | **Effect of Storyteller Context on Inference Accuracy and Latency.** Accuracy is reported as the percentage of participants correctly inferring the transitions of listeners’ attention. Latency is reported as the median time-to-respond when a correct prediction is made. | 57 |
| Table 5 | **Descriptive Statistics and Logistic Regression Models to Estimate Attention from Listener Behaviors.** Total is the collective frequency counts found in the dataset. The Mean Frequency is the average number of occurrences in a storytelling episode (i.e., Total/58). The Mean Duration is the average duration of an emitted behavior in seconds. %Pop refers to the proportion of the population (of the 18 participants) that demonstrated a single instance of the behavior across the repeated interactions. The logistic regression models predicts the listener’s attention based on the normalized duration and/or frequency rate of the observed behavior. Note: the number of observations N for the chi-squared tests (i.e., $\chi^2$(DF, N) are different for each behavioral category since each analysis include periods only from storytelling episodes where at least one instance of the behavior type was observed. | 60 |
Table 6  Descriptive Statistics and the Logistic Regression Model for Individual Speaker Cues. The logistic regression model predicts the likelihood of a response from attentive listeners based on the emitted speaker cues. Total is the collective frequency counts found in the dataset. Rate is the likelihood of response. 

Table 7  Descriptive Statistics and the Binomial Test for Co-occurring Speaker Cues. This table lists the most frequently observed cue combinations. A cue combination is specified through the presence of the cue’s symbolic letter. G:Gaze C:Pitch P:Pause E:Energy F:Filled Pause W:Wordy (for long utterances). A dot represents the absence of that cue. N is the total occurrences of the cue combination found in the dataset.

Table 8  Logistic Regression Models Predicting Attentive State Based on Cues and Responses. The first logistic regression model considers only the main effects of cue-strength and response-valence to predict state. The second model adds an interaction term which represents the relationship between cues and responses.

Table 9  Listener Response Summary. Summary of non-verbal behaviors, as prolonged expressions or occurrences, that are indicative of a child either attentive or inattentive to a peer’s storytelling.

Table 10  Speaker Cue Summary. Summary of multimodal co-occurring cues that children storytellers are observed to use that can elicit a contingent response from listeners at different rates of success.

Table 12  Differences Between State Estimators. All the three models have model parameters to learn from the dataset. The HMM has its initial state, transition function, and emission function. The MHMM has the same set except it has an extra action-emission function. The POMDP also has the same set, but its transition and emission functions are conditioned on both state and action. Note: the POMDP’s policy is not used for state inference, but it is used to shape how the parameters are learned. Both the HMM and MHMM use expectation-maximization to learn their model parameters, while the POMDP uses a generic non-linear optimizer. The number of free parameters grow with model complexity.
Table 13 **Overall Performance of State Estimators.** The prior classifier always predicts the most common class. A general trend can be observed across measures where the POMDP model obtains the highest inference performance, followed by the MHMM model, and then the HMM model.

Table 15 Speaker cues that were either validated or coded for in prior works.

Table 16 Listener responses that were either validated or coded for in prior works.

Table 17 Proctor script for the Instructions stage of the data collection.

Table 18 Proctor script and list of questions used to elicit more details and further a child’s story.

Table 19 Proctor script for the Story Sharing stage of the data collection.

Table 20 Proctor script for the Recall and Questionnaire stages of the data collection.

Table 21 The multilevel logistic regression model to predict whether a correct or incorrect inference would be made based on storyteller-context treatment. The model includes a random intercept and slope that both vary independently at the participant-level to reflect the bias and sensitivity, respectively, of a participant’s inference ability.

Table 22 The multilevel gamma regression model to estimate latency of response based on storyteller-context treatment. Similar to the previous multilevel model (Table 21), within-subject dependencies are accounted for through participant-level variability in both the intercept and slope. But unlike the previous model, the predictor is represented as categorical values through two variables that dummy code the three conditions (with the TRUE condition as the reference category).
Table 23: The ground truth is the non-thresholded filtering inferences generated by our POMDP model. The correlation is the Pearson’s correlation coefficient along with its significance value. The root mean squared error (RMSE) represents the prediction error of the DBN; lower is better. The different models vary in their resolution of belief-states. Their initial distribution over belief-states also vary as either a uniform distribution or as a biased normal distribution with an expected value set to our POMDP’s $b_0$. A random model selects a random belief-state.

**Note:** The 590 is the total number of data instances/tuples (i.e., 60 demonstrations with an average 10 in length ~ 600 data instances).

Table 24: Items for perceived human-likeness from the Godspeed questionnaire [Bartneck et al., 2009].

Table 25: Items for perceived intelligence from the Godspeed questionnaire [Bartneck et al., 2009].

Table 26: A modified version of the Listening Skill subscale from the Active Listening Skill questionnaire [Mishima et al., 1999].
DEFINITIONS

$\Pi^S : b \rightarrow a$  Storyteller’s Policy
$\Pi^L : b \rightarrow o$  Listener’s Policy

AI  Artificial Intelligence
AUC  Area under the receiver operating characteristic
BTOM  Bayesian Theory of Mind
DBN  Dynamic Bayesian network
EM  Expectation-maximization
GLM  Generalized linear models
HMM  Hidden Markov model
HRI  Human-robot interaction
LOOCV  Leave-one-out cross-validation
MAP  Maximum a posteriori
MHMM  Multivariate hidden Markov model
MLE  Maximum likelihood estimate
POMDP  Partially observable Markov decision process
ROC  Receiver operating characteristic
CONTENTS

1 INTRODUCTION 33
  1.1 Outline 33
  1.2 Terminology 34
2 BACKGROUND 37
  2.1 Overview 37
  2.2 Speaker Cues & Listener Responses 38
  2.3 Attention as a Social-Emotional State 38
  2.4 Emotion Recognition 39
     2.4.1 Intra-personal Context 39
     2.4.2 Inter-personal Context 39
  2.5 Emotion Expression 40
     2.5.1 Intra-personal Context 40
     2.5.2 Inter-personal Context 40
     2.5.3 Signaling Paradigm 41
  2.6 Dialog Understanding 41
  2.7 Beliefs and Goals Inference by Plan inversion 42
3 HUMAN BEHAVIOR ANALYSIS 43
  3.1 Data Collection of Peer-to-Peer Storytelling 44
     3.1.1 Participants 44
     3.1.2 Storytelling Task 44
     3.1.3 Study Procedure 44
     3.1.4 Data Annotation 48
  3.2 Effects of Storyteller Context on Inference Performance 52
     3.2.1 Method 52
     3.2.2 Analysis of Human Inference Accuracy 56
     3.2.3 Analysis of Human Inference Latency 57
     3.2.4 Discussion 58
  3.3 Effects of Speaker Cues 59
     3.3.1 Identifying Listener Responses 59
     3.3.2 Identifying Speaker Cues 59
     3.3.3 on Listener Response Regulation 61
     3.3.4 on Listener Response Interpretation 62
     3.3.5 on Listener Attentive State 65
     3.3.6 on Listener Perception 65
     3.3.7 Discussion 66
  3.4 General Discussion 69
     3.4.1 Implications to HRI Design 69
     3.4.2 Implications to Affective Computing 70
     3.4.3 Initial Evidence to Computational Approach 70
4 INTENTIONAL INFERENCE 71
  4.1 Overview 71
  4.2 Informal Sketch 71
## CONTENTS

4.3 Computational Representation as a POMDP  
4.3.1 POMDP World  
4.3.2 Belief Updating  
4.3.3 POMDP Policy  
4.3.4 Summary  
4.4 Apprenticeship Learning for POMDPs  
4.4.1 Likelihood Estimate  
4.4.2 Maximum A Posteriori  
4.5 Evaluation: Comparing State Estimators  
4.5.1 Alternative State Estimators  
4.5.2 Learning Model Parameters  
4.5.3 Training and Testing Framework  
4.5.4 Performance Measures  
4.5.5 Inference Results  
4.6 Discussion  

### BELIEF MANIPULATION

5.1 Overview  
5.2 Computational Representation as a DBN  
5.2.1 Belief Estimator  
5.2.2 Myopic Policy  
5.2.3 Demonstrative Examples  
5.3 Evaluation: Comparing Robot Listeners  
5.3.1 Technology  
5.3.2 Method  
5.3.3 Results  
5.4 Discussion  

### CONCLUSION

6.1 Contributions  
6.2 Future Work  
6.2.1 Immediate Extensions  
6.2.2 Broad Extensions  
6.2.3 Applications  
6.2.4 Long-term Vision  

### BIBLIOGRAPHY

### Appendix

A SUPPLEMENTARY MATERIALS TO BACKGROUND  
B SUPPLEMENTARY MATERIALS TO HUMAN BEHAVIOR ANALYSIS  
C SUPPLEMENTARY MATERIALS TO INTENTIONAL INFERENCE  
D SUPPLEMENTARY MATERIALS TO BELIEF MANIPULATION
A BAYESIAN THEORY OF MIND APPROACH TO NONVERBAL COMMUNICATION FOR HUMAN-ROBOT INTERACTIONS
INTRODUCTION

The field of robotics and artificial intelligence (AI) has primarily been interested in how machines perform human-like tasks. Computers can outperform humans at games like Jeopardy and Go. Robots can pick-and-place objects in an assembly line. Robotics and AI have traditionally been about manipulating objects, navigating environments, or answering knowledge-based queries. But agent-like devices, like Apple’s Siri, Amazon’s Alexa, and Alphabet’s Google Home, operate on a social medium of dialog and are becoming increasingly commonplace as assistive technologies in people’s homes, schools, and at work. Due to our strong and natural tendency to anthropomorphize robots, human-robot interactions are designed to closely mirror human-human communication forms. A major factor in their success as assistive technologies will heavily depend on their ability to effectively communicate with us as social agents. The next robotics challenge is moving away from artificial intelligence and toward social intelligence.

My research work is in developing socially intelligent robots capable of engaging people in social-emotional communication. Much of our social communication is channeled through our facial expressions, body language, gaze directions, and many other nonverbal behaviors. Our very human ability to express and recognize the social-emotional states of others through this nonverbal medium is at the core of social intelligence.

The purpose of this thesis is to define a computational framework to nonverbal communication for human-robot interactions. We address both sides to nonverbal communication, i.e., the decoding and encoding of social-emotional states through nonverbal behaviors, and also demonstrate their shared underlying representation. We make three primary contributions.

First, emotion recognition is constructed as an intentional inference process of agents. We define Intentional Inference as a goal-directed process of forming a social-emotional inference about others while embedded in an interaction. Through a combination of human behavior analyses, human-subjects experiments, and a comparison of computational representations, we demonstrate that the accurate interpretation of emotional states depend on accounting for not only the social context of interaction partners but also adopting the intentional stance to their behaviors.

Second, emotion expression is constructed as a process of influencing others in their inference formation. We define Belief Manipulation as a process of estimating others’ beliefs through inference inversion and producing nonverbal behaviors to affect those beliefs. Through a human-subjects experiment, we demonstrate that a robotic agent operating under this paradigm more effectively communicates a social-emotional state compared to current state-of-the-art approaches in human-agent interactions.
Lastly, we demonstrate a unified computational approach to nonverbal communication where emotion recognition and emotion expression share an underlying representation.

1.1 OUTLINE

We ground our computational framework in storytelling interactions of young children. Storytelling is an interaction form that is mutually regulated between storytellers and listeners where a key dynamic is the back-and-forth process of speaker cues and listener responses. Listeners convey attentiveness, engagement, and understanding through listener responses, also called backchannel feedback, while storytellers use speaker cues, also called backchannel-inviting cues, to elicit these responses.

We pose that storytellers employ plans (albeit short) to influence and infer the attentive state of listeners using these speaker cues. We begin by investigating the direct effects of speaker cues on listeners and how it can also influence perceptions about listeners (Chapter 3: HUMAN BEHAVIOR ANALYSIS). Based on our evidence in the intentionality of storytellers’ behaviors, we computationally model attention recognition as a partially observable Markov decision process (a POMDP problem) of estimating the attentive state of listeners from the perspective of storytellers. We demonstrate the gains in inference accuracy when accounting for this intentional context compared to alternative state estimators that only consider the intra-personal or inter-personal context (Chapter 4: INTENTIONAL INFERENCE).

By formulating emotion recognition as a planning problem, we apply a recent AI method of reasoning about people’s beliefs using Bayesian Theory of Mind (BTOM). We computationally model attention expression as a combined process of estimating storytellers’ beliefs through inference inversion, using a dynamic Bayesian network (DBN), and then producing nonverbal expressions based on a myopic policy to affect those beliefs. This enables a listening robot to communicate an attentive state by first tracking storytellers’ beliefs about the agent based on their cueing behavior. Then by producing appropriate listener responses, the agent can manipulate those beliefs toward a desired perception of an attentive listener. We demonstrate that our BTOM-based listening robot can better communicate an attentive state over a robot that has no internal representation of how its expressions are being interpreted by their human partners (Chapter 5: BELIEF MANIPULATION).

1.2 TERMINOLOGY

Our work studies the interaction dynamics between two people, dyads, engaged in a social interaction. We differentiate between the two persons by referring to the individual of interest as the person in which an inference is being made. The remaining person is then refer to as either the interlocutor, partner, or other. In our storytelling domain, an inference is made regarding the attentive state of listeners and their partners are storytellers.
We address both sides to nonverbal communication, and depending on
the specificity of context, the terminology to address the two processes are
different. The most general terms are mental state inference and expression.
When specific to social-emotional mental states, the terms are emotion recognition
and emotion expression.

We model both processes from an ego-centric view, and each section will
clarify the role an agent is taking. For our storytelling context, this is further
specified as the robot storyteller or robot listener. Furthermore and most im-
portantly, we will consistently use the words cues and responses to differentiate
the source of the emitted nonverbal behavior as either from storytellers
or listeners, respectively.
In this chapter, we review the background necessary to understand the contents of this document as well as frame our work within the research fields of human social psychology, human-agent interaction, affective computing, and artificial intelligence.

We begin by reviewing speaker cues and listener responses that have been studied among adult populations and highlight the limited findings surrounding young children in peer-to-peer interactions (Section 2.2: Speaker Cues & Listener Responses). We are specifically interested in nonverbal behaviors related to the social-emotional state of attention in face-to-face interactions (Section 2.3: Attention as a Social-Emotional State).

Although modern theories of human nonverbal communication emphasize the contextual nature of emotion understanding, current state-of-the-art approaches to emotion recognition model only the intra-personal context of individuals without reference to any contextual elements like the social context of the partner (Section 2.4.1: Intra-personal Context). Recently, a growing number of emotion recognition models incorporate the nonverbal behaviors of both interactants (Section 2.4.2: Inter-personal Context).

Models of emotion expression in human-agent interactions can be similarly categorized into approaches that consider either the intra-personal or inter-personal context. Majority of emotion expression research focus on communicating across modalities (Section 2.5.1: Intra-personal Context), and models that do consider the inter-personal context only focus on the timing aspects in contingently responding to partners (Section 2.5.2: Inter-personal Context). We highlight a limitation to these prior approaches in that they follow a signaling paradigm to emotion expression that cannot dynamically capture how an agent’s behavior is being interpreted by the human partner.

We view nonverbal communication as having the same dynamics as conversation and dialog (Section 2.6: Dialog Understanding). It similarly requires a sharing of representations and an understanding of how to convey a social-emotional message. Our approach to emotion expression follows recent work in AI that reason about people’s beliefs and goals using the principle of rational action (Section 2.7: Beliefs and Goals Inference by Plan inversion). Prior works are able to make this inference by operating in known worlds that are defined either through physical constraints or hand-coded rules. We differentiate our work by highlighting the need to learn the social world from human demonstrations.
**2.2 Speaker Cues & Listener Responses**

**Overview**  A well-known structure in nonverbal communication is the call-response contingencies between speaker cues and listener responses during face-to-face interactions. Speakers elicit feedback from listeners through subtle nonverbal cues [Ward and Tsukahara, 2000]. These speaker cues are signaled nonverbally through changes in intonation, gaze directions, speech pauses, etc. Listeners then respond to these cues either linguistically (e.g., “I see”), para-linguistically (e.g., “mm-hmm”), and nonverbally (e.g., head nods). In conversations, the role of listener responses have been comprehensively characterized in carrying different functions such as signaling understanding, support and empathy, and agreement as well as facilitating conversation flow [Maynard, 1997; Duncan and Fiske, 1977; Dittman, 1972]. However, in our paper, we specifically focus on the role of these backchannels as communicating continued attention, interest, and engagement of listeners [Schegloff, 1982; Kendon, 1967].

**Adults vs Children**  Although extensive research exist studying adult populations, cue-response behaviors have been seldom studied amongst young children, especially in the context of peer-to-peer interaction. When having conversations with adults, children are observed to exhibit different backchanneling behaviors based on their age. More specifically, 11-year-olds were found to provide significantly more listener responses than 7- or 9-year-olds, with a three-fold increase between 7-year-olds and 11-year-olds [Hess and Johnston, 1988]. In a separate study with 2- to 5-year-olds, older preschool children were found to use more head nods and spent more time smiling and gazing at adult speakers. This suggests that older children better understand a listener’s role in providing collaborative feedback [Miller et al., 1985].

Both children and adult listeners have been found to respond more frequently to joint cues over single cues. Co-occurring speaker cues, like simultaneously making eye-contact while changing speech intonation, were found to quadratically increase the likelihood of eliciting a backchannel response [Hess and Johnston, 1988; Gravano and Hirschberg, 2009]. For an organized collection of prior research into speaker cues and listener responses of adults and children, see Table 15 and Table 16 in Appendix A. We extend this work by pioneering the identification of attention-related listener responses as well as speaker cues that children use amongst peers.

**2.3 Attention as a Social-Emotional State**

In our work, we focus on the social-emotional state of listeners in storytelling interactions. Specifically we model their level of engagement, which we interchangeably use with the word attention. However, this should not be confused with joint attention, which is the research problem of inferring what people are attending to in a physical environment [Scassellati, 1999].

**joint attention**
Social behaviors like gaze direction, pointing gestures, and posture are used to establish joint attention between a human partner and a robot. Although these nonverbal behaviors seem similar to the ones we model, they serve more as a mechanism to attend to objects and events in the world rather than to communicate a social-emotional state.

2.4 Emotion Recognition

2.4.1 Intra-personal Context

Human-robot interaction (HRI) researchers depend on emotion recognition technologies to better understand people’s emotional experience. In affective computing, emotions are typically discretize into a basic set of anger, surprise, happiness, disgust, sadness, and fear, while states like boredom, confusion, frustration, engagement, and curiosity are considered to be non-basic [D’mello and Kory, 2015]. Emotion recognition systems infer these discrete set of emotional states by detecting behaviors such as prototypical facial expressions through facial muscle action units [Sariyanidi et al., 2015]. Based on a recent survey, facial expressions are still the primary modality used for affect detection but models also incorporate gaze behaviors, body movements, voice features, spoken language, and bio-signals like electrodermal activity [D’mello and Kory, 2015]. Of the 90 systems reported, 93% of approaches focus on these within-person features.

This representation follows a classical theory in human nonverbal communication of nonverbal leakage, where emotional states are direct influencers of exhibited nonverbal behaviors. Traditional perspectives, like those of Paul Ekman, focus on the nonverbal expressions of single individuals without reference to any contextual elements like setting, cultural orientation, or other people [Ekman, 1984]. In contrast, modern theories emphasize the contextual nature of emotion understanding where greater accuracy comes from interpreting expressions with reference to the social context [Barrett et al., 2011; Hassin et al., 2013].

Prior emotion recognition models represent this perspective in capturing only the within-person features, i.e., only considering the intra-personal context. This approach is more appropriate for applications in understanding consumer reactions to advertisements and other passive media content [Affectiva Inc., 2017]. But social robots are not passive systems in situated human-robot interactions but are active participants where their own generated behaviors are contributing as context and can even influence the expressiveness of their human partners [Celiktutan and Gunes, 2015].

2.4.2 Inter-personal Context

To better capture the inter-personal context of emotion understanding, a growing amount of work has started to model the behaviors of both interactants to recognize social-emotional states like trust [Lee et al., 2013], rap-
port [Yu et al., 2013], and bonding [Jaques et al., 2016]. But, although the behaviors of both interactants are now being considered as co-observations, they are typically represented as either a pair of independent events or as dyadic features like the number of conversational turns. As such, their approaches do not take advantage of the call-response structure of social interactions but instead model their behaviors as flat observations for non-temporal models like support vector machines or feedforward neural networks. Although this interpersonal dynamic is a foundation when modeling other domains such as turn-taking [Thórisson, 2002] or conversational structure [Otsuka et al., 2007], emotion recognition models, for dyadic interactions, currently do not represent this causal structure between people’s behaviors.

2.5 Emotion Expression

Models of emotion expression in human-agent interactions can be similarly categorized into approaches that consider either the intra-personal or inter-personal context.

2.5.1 Intra-personal Context

In the field of human-robot interaction, a majority of emotion-related work focus on robots expressing emotions across modalities like using non-anthropomorphic colored LEDs to communicate basic emotions [Johnson et al., 2013] or producing multimodal behaviors, like gestures that accompany facial expressions, to better convey a robot’s emotional state [Costa et al., 2013]. Furthermore, a recent survey on the computational approaches in HRI reports a lack of “generative computational systems that make use of these channels for robot emotion expression” [Thomaz et al., 2016].

In the virtual agents community, researchers similarly study how different multimodal sets of nonverbal behaviors can better convey attentiveness of a listening agent [Oertel et al., 2016]. But the community has a stronger emphasis on computational approaches that consider the inter-personal dynamics.

2.5.2 Inter-personal Context

In contingently responding to their human partners, interactive agents are found to engender feelings of rapport and support more fluid conversations [Gratch et al., 2007; Al Moubayed et al., 2009]. Prior approaches treat the behaviors of interaction partners, namely their speaker cues, as a timing mechanism to determine an upcoming backchannel opportunity for a listening agent [Morency et al., 2010]. However, beyond its function for stimulus-response contingency, we demonstrate that speaker cues themselves can modify how the agent’s response will be interpreted.
Mimicking agents, in a similar way, use their partner’s behaviors as triggers for response. But they add a reciprocity to their behaviors by matching the quality of that response [Bailenson and Yee, 2005; Schroder et al., 2012].

2.5.3 Signaling Paradigm

To date, work on emotion expression in human-robot interaction has followed the dominant paradigm of emotion as a signaling mechanism and emotion expressions are seen as signals that reveal processes and states that would otherwise be hidden.

— Jung [2017].

All these prior approaches to emotion expression, whether implicitly designed through user studies or explicitly modeled through probabilistic or rule-based methods, follow a signaling paradigm. These agents are capable of broadcasting a social-emotional state using its nonverbal behaviors, but it cannot dynamically model how its behaviors are being interpreted by the human partner over the course of the interaction. The limitations of this approach are best exemplified through the following set of questions:

1. How does an agent know when it has successfully expressed an emotion? If unsuccessful, how can it act to repair this unwanted perception?
2. How can an agent go beyond the maximal expression of a social-emotional state but instead pursue a gradation of perception?
3. How does a listening robot know when to stop nodding? Or how much to nod?

A model capable of answering these questions will have to track how the robot is perceived by the human partner throughout the interaction as well as how its behaviors directly influence those perceptions.

2.6 Dialog Understanding

In the field of natural language understanding, dialogues compared to monologues are regarded as less difficult to understand since conversations enable an interactive alignment of linguistic representations [Garrod and Pickering, 2004]. Conversational agents model how to reason about their partner’s knowledge and also plan their utterances to influence their understanding about the contents of the conversation. These pro-active speaking agents can tailor their utterances to make sure their message is properly conveyed as well as decide when to ask for user feedback to clarify the current dialog state [Buschmeier and Kopp, 2014].

We view nonverbal communication as having the same dynamics as verbal conversation; it similarly requires a sharing of representations and interactive alignment when convey a social-emotional message.
A recent movement in artificial intelligence research is applying a concept called inverse planning or plan inversion to reason about people’s goals and beliefs [Baker and Tenenbaum, 2014]. Humans show a strong natural ten-
dency, as early as one-years-old, to interpret the behaviors of others as in-
tentional goal-directed actions. The principle of rational action proposes that goals are achieved by the most efficient means within the constraints of the environment [Csibra, 2003]. In observing a sequence of actions and adopting the intentional stance [Dennet, 1981], we can predict the intentions, beliefs, and goals of others through this expectation that they are behaving as a rational agent.

Research in human–robot interactions have used this mentalistic reason-
ing to better predict human behaviors as well as communicate a robot’s own beliefs, goals, and intentions [Thomaz et al., 2016; Mutlu et al., 2016]. A robot with a planning model of how to efficiently grasp a cup can invert this model to generate grasping trajectories that are easily interpretable by a human observer [Dragan et al., 2013]. In re-using its own internal cognitive architecture for planning, a robot can simulate from a person’s visual perspective and select appropriate actions to influence their beliefs in order to win a competitive game [Gray and Breazeal, 2014]. Lastly, a robot with a predictive forward model of teammates can infer their goals and beliefs in an engine-
assembly task to better assist them and improve the overall efficiency of the human–robot team [Adalgeirsson, 2014].

Prior works are able to reason about people by operating in known worlds that are defined either through physical constraints or hand-coded rules. For example, the shortest path defines a navigation task of searching for a partic-
cular food-truck on campus [Baker and Tenenbaum, 2014]. Grasping tasks are constrained through efficient trajectories [Dragan et al., 2013]. Geometrically defined line-of-sight and occlusions define perspective-tasking [Gray and Breazeal, 2014]. A specific order of operations define an engine-building task [Adalgeirsson, 2014]. And finally, hand-coded rules are used to capture classroom dynamics between bullies, students, and teachers [Pynadath et al., 2013]. We differentiate our work from these prior works by highlighting the need to learn the social world from human demonstrations.
In this chapter, we deeply investigate how the behaviors of storytellers are related to the process of making inferences about listeners. More specifically, we examine how speaker cues influence perceptions about listeners as either a first-person or third-person observer. Furthermore, we examine how the cues more directly effect the listener during the interaction itself. We make the following set of hypotheses:

Hypothesis Set

1. Storyteller context enables for more accurate inferences about listeners.
2. Speaker cues can regulate the responses of listeners.
3. Speaker cues can modify the interpretation of listener behavior.
4. Speaker cues can influence the attentive state of listeners.
5. Speaker cues predict storytellers’ later perceptions about listeners.

To support our human-subjects experiments and human behavior analyses, we gathered a corpus of peer-to-peer storytelling interactions of young children (Section 3.1). Through a video-based human-subjects experiment, we demonstrate that a third-person observer is most accurate in evaluating the attentive state of these young listeners when able to observe them in the context of the storyteller (Hypothesis 1 see Section 3.2). But what is it about the storyteller’s behaviors that add to this inference about listeners?

We first identify the listener responses (Section 3.3.1) and speaker cues (Section 3.3.2) exhibited by our young population. We find that co-occurring speaker cues can elicit responses from listeners at different rates of success, suggesting a means for storytellers to regulate a social interaction (Hypothesis 2 see Section 3.3.3). We also find that the strength of a speaker cue can modify the interpretation of listener backchannels (Hypothesis 3 see Section 3.3.4). We also observe that different cues have higher chances of toggling inattentive listeners to immediately pay attention (Hypothesis 4 see Section 3.3.5). Lastly, we find that the cueing behaviors of storytellers predict their later evaluation about how much their partner paid attention to their story (Hypothesis 5 see Section 3.3.6).
3.1 DATA COLLECTION OF PEER-TO-PEER STORYTELLING

3.1.1 Participants

Children of typical development were recruited from a Boston public elementary school whose curriculum already included an emphasis on storytelling. A total of 18 students from a single kindergarten (K2) classroom participated in the study. The average age was 5.22 years-old (SD = 0.44) and 61% were male. Overall, 10 participants identified as White, 3 as Black or African American, 2 as Hispanic or Latino, 1 as Asian, 1 as Mixed, and 1 not specified.

3.1.2 Storytelling Task

Over a span of five weeks, each participant completed at least three rounds of storytelling with different partners and storybooks. The storybooks are a series of colored pictures with illustrated characters and scenes that children can use to craft their own narratives (see Figure 3 for an example storybook and Table 28 in Appendix B for the complete set). In the first round, students were paired with friends and told a story using only one scene from the storybook. Subsequent rounds varied in the level of instructions, partner type, and in the number of additional story scenes (see Table 2 for details on round variations).

In a dyad session, the pair of students took turns narrating a story to their partner with each turn generating a storytelling episode (see Figure 1). In total, our data collection consists of 58 storytelling episodes, where the average length of a child’s story was 1 minute and 17 seconds.

3.1.3 Study Procedure

A dyad session averaged 15-minutes in duration and consisted of four stages: Instruction, Story Construction, Story Sharing, and Recall & Questionnaire.

Figure 1: Storytelling Rounds and Turns. Each participant had multiple rounds of storytelling with different partners. In a dyad session, a pair of students took turns (T1, T2) narrating their story to their partner. In sum, the data collection consists of 3 rounds (with a supplementary 4th round for redo opportunities) totaling 29 dyad sessions, which equates to 58 individual storytelling episodes.
Table 2: Round Differences. Each round varied in the amount of instruction, type of partner, and number of story scenes. For example, the second round had a light amount of instruction, students with high storytelling ability were paired with students with low ability (based on teacher’s suggestions), and two story scenes were used. Round 3 paired participants with a partner of opposite gender, if they had not previously.

**Instruction** During the classroom’s free-choice time, pairs of students were brought to the school’s 1-on-1 sessions office. Students were informed that they are practicing telling stories as well as practicing good listening skills. The proctor reviewed what constitutes being a good listener based on an existing classroom concept of “body still,” “voices quiet,” “eyes watching,” and “ears listening” (see classroom poster Figure 2). The proctor then described, using an example storybook, how to make up a narrative using the illustrated characters and setting. In addition, the students watched a pre-recorded video of an adult telling an example story. Each student then randomly drew a token from a hat that indicated their storybook selection for the current dyad session (see Table 17 in Appendix B for proctor script).

**Story Construction** Randomly paired with one of the participants, proctors went into separate locations (also randomly selected) to help participants generate a story about their book. In these 1-on-1 sessions, proctors asked questions to elicit more detail and further their story line. Proctors were provided a list of questions as possible prompters, and depending on a child’s storytelling ability, either none, few, or all the questions were used in a session (see Table 18 in Appendix B for list of questions). The goal of this stage was to help equalize differences in the population’s storytelling ability. We wanted to avoid an individual’s difficulty with the task itself hampering
Figure 3: **Story Space Setup.** Setup included three different camera angles, a high-quality microphone, listener & storyteller chairs, and a story book with compounding story elements per page. The bottom-right photo shows how we labeled each chair to further emphasize to a child his/her role as either the storyteller or listener.

the social dynamics of storytelling. By providing varying degrees of guidance, each participant entered into the next stage ready with a practiced story in-hand.

**Story Sharing**  In the *story space* (as shown in Figure 3), each participant of the dyad took turns telling their story. In view of both participants, a coin was flipped to randomly determine who went first. At a small table, the first storyteller sat on the *storyteller chair*, and the other participant sat on the *listener chair*. The storybook was placed on a tabletop book stand, easily visible by both participants. The proctor then described their particular roles:

**Storyteller role:** *Your important job as the Storyteller is to be like a teacher. You want to make sure your classmate is paying attention and understands your story.*

**Listener role:** *Your important job as the Listener is to use your good listening skills and to pay attention because later i’m going to ask you some questions about his/her story.*

The proctor then left the room to help promote natural peer-interactions by removing adult supervision. After the first storytelling episode was
complete, the proctor re-entered to instruct the participants to switch seats and briefly reiterate the exchanged roles. The goal of emphasizing the two roles was to set a common understanding across participants as this peer-interaction setting may have been unfamiliar to them (see Table 19 in Appendix B for proctor script).

**Recall & Questionnaire** At the end of the dyad session, proctors individually asked the participants to recall their partner’s story to the best of their ability. Furthermore, we asked them to rate how well their partner was as a storyteller and as a listener. Using a smiley-face based 5-point Likert scale (see Figure 5), we asked the following questions:

1. When you were the storyteller, how was your partner at paying attention to your story?
2. When you were the storyteller, how was your partner at understanding your story?
3. When you were the listener, how was your partner at making sure you were paying attention to the story?
4. When you were the listener, how was your partner at making sure you understood the story?

**Figure 5: Smileyometer.** A pictorial rating scale developed for children [Read et al., 2002]. Please see the proctor’s script Table 20 in Appendix B in how to instruct children to answer with this scale.
### Table 3: List of All Annotated Behaviors

The selected set of nonverbal behaviors were either found in prior works (see Table 15 and Table 16 in Appendix A) or commonly observed in the storytelling interactions. Each annotation category has a set of mutually exclusive labels and was coded for storytellers (S) and/or listeners (L) or jointly evaluated (joint).

<table>
<thead>
<tr>
<th>Category</th>
<th>Labels (default behavior)</th>
<th>S</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaze</td>
<td>book, partner, away</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Posture</td>
<td>upright, toward, away, other</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Nod</td>
<td>none, nod</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Eyebrows</td>
<td>neutral, raise, furrow</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Mouth</td>
<td>neutral, smile, frown, other</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Voicing</td>
<td>silence, storyteller’s voice, listener’s voice, both</td>
<td>joint</td>
<td></td>
</tr>
<tr>
<td>Task</td>
<td>on-task, off-task</td>
<td>joint</td>
<td></td>
</tr>
<tr>
<td>Attentive State</td>
<td>listening, not listening, speaking-turn</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Prosodic Cue</td>
<td>pitch, energy, pause, filled pause, long utterance, other</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

### 3.1.4 Data Annotation

From the video recordings of the three time-synchronized cameras shown in Figure 4, coders used a video-annotation software called ELAN [Brugman and Russel, 2004] to mark the start and stop times for all the behaviors listed in Table 3 except for the prosodic cues. For each storytelling episode, the behaviors of both listener and storyteller were annotated by multiple coders; we achieved moderate levels of agreement (Fleiss’ $\kappa = 0.53$). The coders were instructed to annotate the video of only one participant and only one behavior category at a time. Coders annotated the perceived attentive state of listeners as well as their gaze direction, posture shifts, nods, eyebrow movements, smiles & frowns, and short utterances. Storytellers were annotated for their gaze- and prosodic-based speaker cues (see Figure 7 for a timeline visualization of the annotated nonverbal behaviors).

For the attentive state annotation, a listening label meant that the participant was paying attention to the storyteller’s story. It is important to note that our state labels include when a listener takes a speaking-turn as a mutually exclusive event. This enables us to filter out behaviors related to conversational structure like turn-yielding cues, which have been previously demonstrated to be different from backchannel-inviting cues [Gravano and Hirschberg, 2009]. Based on the Task annotation, we further exclude moments from our analyses when both of the children participants are completely off-task from the storytelling activity.

We developed a custom program to help coders easily annotate when and what type of prosodic cue is detected in the storyteller’s speech. The pro-
Figure 6: Estimating the Emission-Time of a Prosodic Cue. Based on the moment of the backchannel response, we extract the last speaking turn of the storyteller and estimate that its terminating edge is the ending-time of the prosodic cue. This is calculated for all prosodic-based cues except for the pause cue, which is roughly estimated to be halfway between the backchannel time and the terminating edge.

gram played back the audio recording of a storytelling episode, and coders were asked to simulate in realtime being a listener and mark the moments when they want to backchannel by simply tapping on the space bar. After this simulation, coders then reviewed the audio snippets surrounding these moments to reflect on what prompted their backchannel and categorize their reasoning from the following prosodic cues:

- pitch (intonation in voice, change in tone)
- energy (volume of voice, softness/loudness)
- pause (pause in speech, long silence)
- filled pause (e.g., ‘um’, ‘uh’, ‘so’, ‘and’)
- long utterance (a long contiguous speech segment)
- other

This stimulus-based coding is a method for annotators to identify when they wanted to backchannel (backchannel moment) as well as categorize the reasoning behind what caused their wanting to backchannel (speaker cue event). The null-space that is not marked has an implied default label of ‘none.’

Three coders underwent this simulation, and we followed the Parasocial Consensus approach [Huang et al., 2010] to build consensus of when backchannel opportunities occurred. More specifically, each of our three coders’ registered backchannel times were added as a vote on a consensus timeline with a duration of one second around the central moment. An area in the timeline with more than two total votes was counted as a valid backchannel moment.
From these backchannel moments along with the voicing annotations listed on Table 3, we estimated the emission time of prosodic cues (see Figure 6 for more detail).

To capture the complete cue context embodied by storytellers, we combined their prosodic cues with their physical gaze cues to make sets of multimodal cues. Based on the prosodic cue emission-time and gaze onset-time, events were considered to be co-occurring and merged if they are within an empirically found 1.3 seconds of each other. When looking at the period between back-to-back gaze cues exhibited by storytellers, we found a minimum separation time of 1.5 seconds. This established an upper bound of a merge window when trying to collect co-occurring cues. Beyond this window we start encroaching on cues that could be a part of the next cueing instance. The cue times within this window are averaged to reflect the collective moment of emission for this multimodal cue.
Figure 7: Timeline Visualization of Annotated Behaviors. (Left) Listener’s non-verbal behaviors are continuous labels through the marked start and stop times. (Right) Storyteller’s cueing behaviors are event-based labels based on the prosodic cue’s emission-time and gaze onset-time.
3.2 Effects of Storyteller Context on Inference Performance

In this section, we demonstrate that the accurate inference of the attentive state of listeners depends on accounting for the social context of storytellers. Through a video-based experiment, we manipulate the presence, absence, or falseness of speakers from original interactions between storytellers and listeners. Although the listener’s behaviors remain exactly the same, we expect third-party observers will have different perceptions about the listener’s attentive state across these conditions. We hypothesize the following:

**Main Hypothesis**: Inferences about a listener’s attentive state is most accurate when observing both the storyteller’s and listener’s behaviors of a social interaction and least accurate when missing the social context of the storyteller (i.e., storyteller context).

We quantify inference performance as a function of speed and accuracy and demonstrate that both measures improve when observing the true storyteller context to the listener’s behaviors.

3.2.1 Method

3.2.1.1 Participants

Participants were recruited online through Amazon Mechanical Turk. Turk Workers were from the United States to ensure cultural relevance. To limit the participation pool to high quality workers, their qualification requirements met the following:

- # of approved HITs (Human Intelligence Tasks) greater than 5000
- Approval rating from former requesters greater than 98%

From the 542 Turk workers that submitted to the HIT task, 36 individuals were rejected for not fully completing all parts of the task or for not properly following the task’s instructions. The average age of the remaining 506 participants is 38 years old (SD = 11). Nearly half (56%) are parents and gender balance is close to half (53% female). Below we detail two exclusion principles applied in removing further participants from our analysis.

3.2.1.2 Study Procedure

The online survey-based experiment took an average 19 minutes (SD = 12) to complete the following three parts: Affect Recognition Assessment, Training Exercise, and Inference Task.
AFFECT RECOGNITION ASSESSMENT  The Diagnostic Analysis of Non-verbal Behavior (DANVA2) is an assessment to measure an individual’s non-verbal affect recognition ability [Nowicki and Duke, 1994]. The evaluation consists of viewing a series of facial expressions as well as listening to paralinguistic vocal expressions of children to identify the expressed emotion: happiness, sadness, anger, or fear. Individuals are scored based on the number of items incorrectly identified from the 24 different pictures of children’s faces and 24 different recordings of children’s voices. Participants took this assessment through a web-based flash program that would present the stimuli and record their multiple choice response.

Overall, participants scored a mean error of 2.9 (SD = 2.0) in recognizing children’s facial expressions and 4.8 (SD = 2.6) in recognizing children’s vocal expressions. To ensure our population consists of individuals of average affect recognition ability, 23 participants that scored an error greater than two standard deviations from the population’s mean, on either the face or voice subtests, were excluded from our analysis below.

TRAINING EXERCISE  To familiarize participants to the procedure of the primary inference task, they first experienced a similar task but on a simple example video. The example video contained animated illustrations on how to sound out basic vocabulary words like “cat” and “mat.” As a training exercise, participants were asked to carefully watch this video and immediately pause it when they heard the word “bat.” They were then instructed to report the number in the upper-left hand corner of the video, which represented the video frame corresponding to the paused scene.

Overall, participants were on average 41 frames, or 1.4 seconds, away from the exact moment of the target event (SD = 147 frames or 4.9 seconds). Participants that were within two standard deviations from the population’s mean response frame passed this training exercise. As a measure of task adherence to filter out low-quality Turk workers, the 22 participants that failed to meet this criteria were excluded from our analysis below.

INFERENCE TASK  Participants were asked to watch a series of short videos (each around 30 seconds in duration) of different listeners. Participants were told that in all the videos listeners are at first paying attention to the narrator, but we would like to know when/if they stop being attentive. Following the same procedure introduced in the training exercise, participants reported the paused frame number, which represents the moment they perceived a listener transitioning from attentiveness to inattentiveness. They also had the option of reporting that they believed that the listener was paying attention the entire time.

† There is a bit of irony in using a contextless test to exclude participants from a study that is investigating the importance of context to emotion recognition. However, it is possible to make an inference (of lesser accuracy) in contextless situations. This exclusion is to ensure a population of typically development.
Figure 8: Video-based Human-subjects Experiment. From TRUE interactions between a storyteller and listener, we manipulate the absence and falseness of the storyteller context. For the FALSE condition, we replace the original storyteller with the audio and video of a different storyteller. The ABSENT condition completely removes the storyteller context (both audio and video).

3.2.1.3 Experiment Design

From an original interaction between a storyteller and listener, we manipulate the presence, absence, or falseness of the storyteller context. Although the listener’s behaviors remain the same, we investigate how an observer’s perception about the listener’s attentive state changes across the different contextualizations. As a within-subject study design, a participant viewed a video from each of the three conditions but of three different listeners and in a random order. In using three different listeners, we can generalize our results as being beyond a listener-specific phenomenon. Our three conditions are defined as the following:

1. **TRUE (control):** Participants view the original interaction between a storyteller and listener. With access to both of their behaviors, participants make an inference about the listener’s attentive state.

2. **ABSENT:** Participants can only view the listener and make an inference based solely on the listener’s behaviors.

3. **FALSE:** Participants view an unmatched interaction where the original storyteller is replaced with one from a different storytelling episode.
Figure 9: Video Alignment of Listener Across Condition. We demonstrate how for the listener (in the white t-shirt) we retain his exact behavior across the three conditions: TRUE (top) FALSE (middle) ABSENT (bottom).

From the video recordings collected in Section 3.1, we create a set for the TRUE condition with the audio and video (AV) of the original storyteller, a set for the ABSENT condition with the storyteller’s AV removed, and a set for the FALSE condition with the AV of a different storyteller (see Figure 8).

It is important to note that although the audio recordings capture both of the storyteller’s and listener’s voices, in general only the storyteller is speaking and the listener is quiet. We also took further precautions in selecting video snippets that did not include any moments where the storyteller asks a direct question or the listener interrupts the story. To preserve the illusion that the FALSE condition is from a real interaction, we avoided moments containing any dialog-related coordination.

All the videos were composed and edited to allow an observer to easily see the facial expressions of both the storyteller and listener while also preserving the gaze cues (by arranging images to mimic the original interaction geometry). Demonstrated in Figure 9, we make sure that a listener’s behavior between conditions remained exactly the same.

3.2.1.4 Dependent Measures

Based on the hand-annotated attentive state label from trained experts (see Section 3.1.4), the video snippets contain a single moment where the listener transitions from attentiveness to inattentiveness as illustrated in Figure 10. From a participant’s report on where they believe the transition point to be, we define two measures for inference performance:
Figure 10: Example of Scoring Accuracy and Latency. At frame 440, a listener is annotated by experts as transitioning from an attentive to inattentive state. As such, a participant that reports the transition occurring at 250 frames is marked as being incorrect. A participant that made a prediction at 600 frames is accurate with a latency of 160 frames.

1. Accuracy: A response frame after the transition point is marked as correct and elsewhere as incorrect, including the option of reporting the listener as attentive for the entire time. Accuracy is a dichotomous variable, where a value of 0 means incorrect and 1 means correct.

2. Latency: Latency is measured as the distance between the response frame and the target frame. This difference represents the participant’s hesitation or delay and is only calculated for correct predictions.

In accordance to our hypothesis, we expect an increasing trend (TRUE > FALSE > ABSENT) where participants achieve their best inference performance in both accuracy and latency with the TRUE condition and their worst inference performance with the ABSENT condition. We anticipate that participants will have a difficult time with the FALSE condition since the disjointed sets of speaker cues to listener responses will either delay or confuse their predictions. But since it still contextualizes the listener’s behaviors to a storyteller, we hypothesize that a FALSE, or unmatched, context is still better than no context (ABSENT).

3.2.2 Analysis of Human Inference Accuracy

We examine the ability of storyteller-context treatment to predict an increasing trend of accuracy rates using generalized linear models (GLM). A multilevel (i.e., mixed-model) logistic regression was performed to determine the effect of storyteller context on the likelihood of participants making a correct inference about listeners’ attentive state while controlling for within-subject dependencies from repeated measures (see Table 2 in Appendix B for model details). Based on our expectation that inference accuracy increases across
3.2 Effects of Storyteller Context on Inference Performance

<table>
<thead>
<tr>
<th>Conditions</th>
<th>TRUE</th>
<th>FALSE</th>
<th>ABSENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>measures</td>
<td>58.4% correct</td>
<td>57.7% correct</td>
<td>51.0% correct</td>
</tr>
<tr>
<td>Latency</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>frames</td>
<td>96</td>
<td>107</td>
<td>99</td>
</tr>
</tbody>
</table>

Table 4: Effect of Storyteller Context on Inference Accuracy and Latency. Accuracy is reported as the percentage of participants correctly inferring the transitions of listeners’ attention. Latency is reported as the median time-to-respond when a correct prediction is made.

treatment groups, the predictor variable is contrast coded as ordered values [-1 0 1] to model a linear trend where having access to the TRUE context yields the highest accuracy rate while having access to no context (ABSENT) yields the lowest.

Based on the Wald Chi-square statistic, the logistic regression model is statistically significant, \( \chi^2(1) = 4.15, p^* = 0.04 \), which indicates a linear relationship exists between our expected order of storyteller-context treatment and their accuracy rates. As shown in Table 4, an ascending trend in inference accuracy is observed with the TRUE condition obtaining the highest percentage of participants that correctly predict the attentive state of listeners and the ABSENT condition obtaining the lowest.

3.2.3 Analysis of Human Inference Latency

Similar to the trend analysis described for accuracy, we examine the ability of storyteller-context treatment to predict an increasing trend of inference latency values. The latency observations are positive whole numbers and have a skewed distribution since the highest density of observations are found closest to the target frame and then drop-off over time. Given the nature of the data, we used a gamma GLM (vs the typical normal distribution assumption) with storyteller context as the primary predictor of the log latency while again controlling for within-subject dependencies. We expected an increasing trend where participants experienced the greatest delays in the ABSENT condition, followed by the FALSE condition, and with the TRUE condition obtaining the lowest latencies, but no significant trend was found \( \chi^2(1) = 0.01, p = 0.94 \).

However, rather than looking for a trend, we instead looked for any differences between the conditions. By treating the predictor as a categorical variable, a statistically significant gamma GLM was found, \( \chi^2(2) = 6.35, p^* = 0.04 \) \( \dagger \) (see Table 22 in Appendix B for model details). There is a significant difference between the TRUE and FALSE conditions, \( \dagger \) There are two degrees-of-freedom since the three conditions are dummy coded as two categorical predictor variables.
\[ t(767) = 2.21, p^* = 0.03 \], with the TRUE condition obtaining lower latencies (\( \mu = 96 \) frames) than the FALSE condition (\( \mu = 107 \) frames). No significant difference was found between the TRUE and ABSENT conditions.

3.2.4 Discussion

By changing the storyteller context in which listener behaviors occur, we can manipulate observers to a delayed or even incorrect appraisal about the listener’s attentive state. Participants were most accurate when observing both the storyteller and listener, and they were least accurate when missing the storyteller context. When presented with a false storyteller context, participants were again less accurate but also slower in being able to form a correct inference. This demonstrates the extent to which we can degrade an observer’s perception about a listener’s true emotional state.

We expected participants to be also slower in making correct predictions when missing out on the storyteller context. But observers achieved similar speeds as when witnessing the true context. If we view inference latency as an operationalization of confidence, we interpret this result as meaning that observers felt similarly confident about their appraisals in these two conditions. With no context to doubt their judgments, observers could have felt confident in their evaluation of listeners, but they were more likely to be incorrect.
3.3  **EFFECTS OF SPEAKER CUES**

Our video-based human-subjects experiment demonstrated that the accurate interpretation of listeners’ attentive state depend on accounting for the social context of storytellers. But what is it about the storyteller’s behaviors that add to this inference process? In this section, we examine how speaker cues directly effect the listener during the interaction itself as well as predict storytellers’ later perceptions about listeners.

### 3.3.1 Identifying Listener Responses

We first identify nonverbal behaviors exhibited by listeners that are related to their attentive state. Or more formally, we use a logistic regression analysis to identify nonverbal behaviors (the explanatory variables) that can predict the perceived attentive state of listeners (the predicted outcome). A logistic regression analysis finds the best linear model to describe the relationship between the outcome and explanatory variables. Based on the fitted coefficients, we can determine how much the explanatory variables can predict the outcome.

For each of the annotated listener behavior listed in Table 3, a logistic regression analysis was performed to predict attention (0/1) based on the behavior’s normalized duration and frequency rate. Normalized duration and frequency rate of behaviors were observed during contiguous/block periods of either attentive or inattentiveness. For nonverbal behaviors that are quickly expressed likes nods (or any behavior with an average duration less than 90 seconds), the frequency rate was the only predictor.

Shown in Table 5, gazes, leans, brow-raises, smiles, nods, and utterances are nonverbal behaviors that significantly predict listeners’ attention. Based on the sign of the coefficients (B) and significance (p) of the explanatory variables, we determine that frequent partner-gazes, frequent forward-leans, frequent brow-raises, prolonged smiles, frequent nods, and frequent utterances are associated with attentive listeners. In contrast, prolonged gazes-away from the partner, frequent away-leans, and prolonged brow-raises are associated with inattentive listeners. Interestingly, brow-raises can hold opposite associations depending on its form of emittance.

### 3.3.2 Identifying Speaker Cues

In Section 3.1.4: Data Annotation, adult-coders detected and annotated the speaker cues from the storytellers, but which ones do children understand and know to respond to during the real interaction? In our next set of analyses, we examine which speaker cues, taken singly or in combination, were observed to elicit a contingent backchannel from the children listeners. We registered a contingent backchannel as being observed if the listener responded with any of the previously found attentive behaviors within
only from storytelling episodes where at least one instance of the behavior type was observed.

For the chi-squared tests, \( \chi^2 \) are different for each behavioral category since each analysis includes periods for the chi-squared test on the normalized duration and/or frequency rate of the observed behaviors. The number of participants that demonstrated a single instance of the behavior across the repeated interactions. The logistic regression models are performed on the entire dataset across the repeated interactions. The logistic regression models provides a single instance of the behavior across the repeated interactions. The Mean Frequency is the average number of occurrences in a storytelling episode (i.e., Total/Behavior). Total is the collective frequency counts found in the dataset. The Mean Duration is the average duration of an individual behavior in seconds, Total refers to the population of the behaviors across the repeated interactions (i.e., Total/Behavior).

Table 5: Descriptive Statistics and Logistic Regression Models to Estimate Attention from Listener Behaviors

<table>
<thead>
<tr>
<th>Behavior</th>
<th>Total</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Median</th>
<th>25%</th>
<th>50%</th>
<th>75%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smiling</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Brow Furrow</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Brow Raise</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Lean Toward</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Lean Away</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Gaze Partner</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Gaze Listener</td>
<td>122</td>
<td>233</td>
<td>72</td>
<td>27.05</td>
<td>10</td>
<td>20</td>
<td>30</td>
<td>40</td>
</tr>
</tbody>
</table>

Logistic Regression Models

Overall frequency, Total Mean, Mean %
Table 6: Descriptive Statistics and the Logistic Regression Model for Individual Speaker Cues. The logistic regression model predicts the likelihood of a response from attentive listeners based on the emitted speaker cues. Total is the collective frequency counts found in the dataset. Rate is the likelihood of response.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>1.89</td>
<td>0.65</td>
<td>0.08</td>
<td>0.09</td>
<td>1.33</td>
<td>1.05</td>
</tr>
<tr>
<td>t-stat</td>
<td>5.35</td>
<td>2.16</td>
<td>0.22</td>
<td>0.31</td>
<td>2.13</td>
<td>2.25</td>
</tr>
<tr>
<td>p-value</td>
<td>$p^*=8.7 \times 10^{-8}$</td>
<td>$p^*=0.03$</td>
<td>$p=0.82$</td>
<td>$p=0.76$</td>
<td>$p^*=0.03$</td>
<td>$p^*=0.02$</td>
</tr>
<tr>
<td>N</td>
<td>174</td>
<td>147</td>
<td>52</td>
<td>122</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>rate</td>
<td>0.76</td>
<td>0.59</td>
<td>0.58</td>
<td>0.51</td>
<td>0.59</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Logistic Regression Model

As expected, some of the speaker cues from our prior analysis—energy and pause—do not offer significant predictive ability when examined in isolation. However, young children have been previously observed to respond more often in stronger cue contexts where two or more cues are co-occurring [Hess and Johnston, 1988].

Using the set of multimodal cues (extracted in Section 3.1.4), we examine the ability of cue combinations to predict that an attentive listener would...
backchannel. The likelihood of observing a combination of cues is much smaller than individual cues, resulting in small sample sizes for each unique combination. Rather than performing a logistic regression analysis, we use the binomial exact test to determine whether the response rate of a cue combination is greater than an expected rate of 0.5. As shown in Table 7, the one-sided binomial test indicates that the co-occurring cues Pitch-Energy, Gaze-Pause, Gaze-Pitch, Gaze-Pitch-Pause, and Gaze-Pitch-Energy have response rates significantly higher than the expected rate. Interestingly, as the number of co-occurring cues increases (1 → 2 → 3), the likelihood of receiving a response also increases (0.68 → 0.82 → 0.93‡). Stronger the cue context, the more likely a listener will respond.

### 3.3.4 on Listener Response Interpretation

In this section, we examine how speaker cues can modify the interpretation of listener responses. More specifically, how do cues and responses in their joint meaning influence perceptions about listeners’ attention. As such, we first examine the ability of cues and responses as independent explanatory variables to predict perceived state (Section 3.3.4.2: Only Main Effects), and then we compare what happens when we add an interaction term that represents the relationship between cues and responses (Section 3.3.4.3: With Interaction Effects).

To fully model how the unique combinations of cue-response pairs affect listener perception, we need much more data. As such, we use informative heuristics to create high-level meta-behaviors that define a smaller range of possible behavior combinations (Section 3.3.4.1: Data Tuples).

---

Averaged response rates of only the significantly predictive cue combinations
3.3.4.1 *Data Tuples*

Based on our analysis in Section 3.3.3, we categorize multimodal cues based on their number of co-occurring cues as either weak, moderate, or strong. For example, a Gaze-Pitch-Energy multimodal cue is a strong cue-context, Gaze-Pause is a moderate cue-context, and any single cue is a weak cue-context.

Based on our analysis in Section 3.3.1, we categorize listener response combinations based on their overall valence. Measured as a sum of individual valences, a forward-lean (+), prolonged smile (+), and an away-gaze (-) observed from a listener within [0.5 - 3.0] seconds after an emitted cue is represented with a total value of +1, which is an overall weak positive response. By accounting for both attentive and inattentive behaviors, we roughly measure the magnitude and direction of listeners’ overall expressions during this time window.

To capture how cue-response pairs influence perceived state, we extract whether coders annotated listeners as being attentive or inattentive at the end of this time window (see Figure 11 for a helpful visualization). This allows us to analyze how the perception about listeners change after witnessing their response to the cue-context. In summary, for the following analyses, we use data tuples of <Cue, Response, State>:

- **Cue:** number of co-occurring speaker cues as either weak, moderate, or strong cue-contexts [1 to 3]
- **Response:** measure of listener’s response to a cue as an overall valence rating [-3 to +4]
- **State:** perception of listener’s attentive state sampled immediately after the response window [0 or 1]
Table 8: Logistic Regression Models Predicting Attentive State Based on Cues and Responses. The first logistic regression model considers only the main effects of cue-strength and response-valence to predict state. The second model adds an interaction term which represents the relationship between cues and responses.

3.3.4.2 Only Main Effects

We examine the ability of cue-strength and response-valence as independent explanatory variables to predict listeners’ attentive state. The overall logistic regression model was statistically significant, \( \chi^2(2) = 71.4, p^* = 3.2e^{-16} \), where response-valence is the predictor that best explains state (see Table 8). One unit increase in the response-valence makes the listener 2.91 times more likely to be perceived as paying attention. This result is not surprising since listeners’ behaviors are, of course, good predictors of their attentive state. But this analysis also serves as a means to validate our method of measuring response as an overall valence rating.

3.3.4.3 With Interaction Effects

In adding an interaction term to our previous logistic regression model, we find that the overall model is again statistically significant, \( \chi^2(3) = 78.4, p^* = 6.74e^{-17} \), but can explain more of the variance \( R^2 = 19.5\% \) compared to \( R^2 = 17.8\% \) of the previous model. As shown in Table 8, the interaction term is significant (\( p^* = 0.02 \)), which indicates that the predictive power of the listener’s response is modified by the cue context. As shown in Figure 12, the strong-cue curve approaches areas of higher likelihood (i.e., the limits of the y-axis) more quickly than the other curves, especially in comparison to the weak-cue curve. This means, that for the same quality of listener response, stronger cues facilitate higher levels of certainty as to whether listeners are attentive or inattentive.
Figure 12: Predicting Attentive State of Listeners Based on Cue-Strength and Response-Valence. The x-axis represents overall listener’s response as either very positive (+4) to very negative (-3). The y-axis represents the likelihood of attention, or inversely as inattention. (Right) Shows the 95% confidence bounds of strong vs weak cue-contexts. For the same listener response (e.g., x=-2), there is a difference in interpretation if we observed it after a weak vs a strong cue. Strong cues buys us greater certainty that the listener is not paying attention (likelihood of 70%-100% vs 50%-70%).

3.3.5 on Listener Attentive State

We examine whether speaker cues can immediately influence inattentive listeners to pay attention. We performed a logistic regression analysis to see if cue-strength could predict the next state of inattentive listeners. This analysis only included data from moments when the listener is perceived to be inattentive at the moment when the storyteller emitted a cue. Although no significant effect was found \[ \chi^2(1) = 1.89, p = 0.33 \], we observed that weak cues overall have a 19% chance of toggling inattentive listeners to immediately pay attention, moderate cues have a 23% chance, and strong cues have the highest chance at 33%.

3.3.6 on Listener Perception

We examine whether storytellers’ speaker cues are predictive of their later evaluation regarding their partner’s level of attention to their story. As discussed in Section 3.1.5: Recall & Questionnaire, we asked participants at the end of the storytelling sessions to rate, using a 5-point Likert scale, how well their partner payed attention to and understood their story. We performed a multi-level linear regression analysis to see whether the rates in which storytellers emit weak, moderate, or strong cues predict their later ratings about listeners. We controlled for within-subject dependencies since storytellers can individually varying in their cueing rates. We found that speaker cues are predictive of storytellers’ later judgement about listeners’ level of attention to their story \[ \chi^2(3) = 8.85, p^* = 0.04 \] but does not predict listeners’ level of understanding \[ \chi^2(3) = 2.31, p = 0.52 \].
3.3.7 Discussion

listening behaviors of young children  We identified nonverbal behaviors of young listeners that are indicative of their attentive state toward their partners’ storytelling (see summary in Table 9). We determined the relevant form in which these nuanced behaviors are emitted as either prolonged expressions or as occurrences. Of the behaviors identified, the most unexpected result is the opposing meaning behind frequent vs prolonged brow-raises. However, we found that prolonged brow-raises most often co-occur when listeners are also looking away from storytellers (see Figure 13 for a correlation map); their joint emission can serve as a strong signal of an inattentive listener.

<table>
<thead>
<tr>
<th>Attentive Behaviors</th>
<th>Inattentive Behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequent partner gazes</td>
<td>prolonged away gazes</td>
</tr>
<tr>
<td>frequent forward leans</td>
<td>frequent away leans</td>
</tr>
<tr>
<td>frequent brow raises</td>
<td>prolonged brow raises</td>
</tr>
<tr>
<td>prolonged smiles</td>
<td>frequent nods</td>
</tr>
<tr>
<td>frequent utterances</td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Listener Response Summary. Summary of nonverbal behaviors, as prolonged expressions or occurrences, that are indicative of a child either attentive or inattentive to a peer’s storytelling.

regulating the response of listeners  By examining prosodic- and gaze-based cues, we identified multimodal speaker cues, taken singly or in combinations, that can elicit a response from listeners (see summary in Table 10). Some prosodic cues like pauses in speech or changes in energy are too subtle for young children to perceive, but their cueing context can be strengthened in adding co-occurring behaviors like a gaze cue. We confirm prior work in demonstrating that children respond more often in stronger cue-contexts [Hess and Johnston, 1988]. However, we differentiate our work by observing cues that young listeners not only respond to but also use when telling stories to peers.

<table>
<thead>
<tr>
<th>Single Cue</th>
<th>rate</th>
<th>Dual Cues</th>
<th>rate</th>
<th>Tri Cues</th>
<th>rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch</td>
<td>0.59</td>
<td>Pitch-Energy</td>
<td>0.66</td>
<td>Gaze-Pitch-Energy</td>
<td>0.93</td>
</tr>
<tr>
<td>Filled Pause</td>
<td>0.59</td>
<td>Gaze-Pause</td>
<td>0.89</td>
<td>Gaze-Pitch-Pause</td>
<td>0.93</td>
</tr>
<tr>
<td>Long Utterance</td>
<td>0.76</td>
<td>Gaze-Pitch</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gaze</td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Speaker Cue Summary. Summary of multimodal co-occurring cues that children storytellers are observed to use that can elicit a contingent response from listeners at different rates of success.
Figure 13: Co-occurring Nonverbal Behaviors of Listeners. Values are the Pearson’s correlation coefficients. Highly correlated behaviors (in yellow) are prolonged brow-raises and away gazes as well as smiles and partner gazes. Uncorrelated behaviors (in dark blue) are partner gazes and away gazes since they are mutually exclusive behaviors.

**Magnifying Certainty about Listeners’ Attentive State**

We demonstrated how the cue-context modifies the interpretation of backchannels. More specifically, we observed that the strength of a cue magnifies the predictive power of listener’s response by adding greater certainty when estimating attention. For the same valence of listener response, there is a difference in the perception about the listener’s state if it is observed after a weak cue versus a strong cue. Strong cues, in essence, buys us greater certainty that the listener is attentive or inattentive.

**Influencing the Attentive State of Listeners**

We did not find statistically significant support of the immediate effects of speaker cues to change the state of inattentive listeners. However, we observed an overall trend of stronger cues having higher chances to transition them. Our analysis only captured the one-shot abilities of these cues. We expect that a temporal model can better predict the compounding effect of a sequence of cues on the attention of listeners.

**Predicting Storytellers’ Perception about Listeners**

Speaker cues are predictive of storytellers’ later judgement about the level of attention their partner had to their story. Depending on how to view the cueing behavior of storytellers, we can interpret this result in two different ways. If we view these cues as nonverbal signals that reveal underlying mental states, then they can reveal what storytellers think about listeners. However, if we view these cues as *intentional* behaviors of storytellers, then
the employment of cues by storytellers are not random acts but are related to how they form inferences about their partner. This interpretation supports our original claim that storytellers’ employ a policy to infer the attentive state of listeners using these speaker cues, while the former interpretation supports the capacity to reason about storytellers’ beliefs regarding their partner.
To conclude this chapter, we discuss the implications of our results when designing robots listeners and storytellers for child-robot interactions as well as when designing computational models for emotion recognition. Furthermore, we discuss the evidence in support of our computational design decisions for intentional inference and belief manipulation. For convenience, we list our hypothesis set here:

**Hypothesis Set**

1. **Storyteller context enables for more accurate inferences about listeners.** (Upheld)
2. **Speaker cues can regulate the responses of listeners.** (Upheld)
3. **Speaker cues can modify the interpretation of listener behavior.** (Upheld)
4. **Speaker cues can influence the attentive state of listeners.** (Partial)
5. **Speaker cues predict storytellers’ later perceptions about listeners.** (Upheld)

### 3.4.1 Implications to HRI Design

**Attention-related nonverbal behaviors** When estimating a child’s level of engagement or attention to a task, **Robot storytellers** will benefit from detecting our identified attention-related nonverbal behaviors as predictive features. **Robot listeners**, especially if being modeled as a peer companion, can effectively express and communicate an attentive (or even an inattentive) state using these signals. Pointedly, in the case of communicating with head nods, children have been previously observed to more likely use this behavior at older ages when interacting with adults (67% with 2-year-olds, 82% with 3-year-olds, and 100% with 4-year-olds [Miller et al., 1985]). However, in our peer-to-peer context, only 39% of the 5- & 6-year-olds emitted nods (see Table 5). This implies not only a developmental factor when designing appropriate robot behaviors for children but also a consideration of how their behaviors can impress a perception of being a peer.

**A greater role of speaker cues** Our findings demonstrate a greater role speaker cues can have in human-robot interactions beyond its current use as a simple timing mechanism. Since co-occurring cues have a higher likelihood of eliciting a response from listeners, **robot storytellers** can manipulate their production of nonverbal speaker cues to deliberately gain more information about a child’s attentive state. In moments of high uncertainty about the listener, a social robot can plan to emit an appropriate cue-context to elicit a response that can reduce state uncertainty. Through speaker cues, social robots can regulate a social interaction to better understand their partner’s affective state.

Furthermore, **robot listeners** that behave stochastically by following these response rates can more likely be perceived as a more believable and human-
like listener over an agent that responds to any cue, agnostic to its signaling strength. However, experimental manipulation is required to validate this expectation.

3.4.2 Implications to Affective Computing

An important contribution in this chapter is emphasizing the contextual nature of emotion understanding. Both the storyteller’s and listener’s behaviors are necessary to form accurate inferences about the listener’s attention. We demonstrate this in a two-part process. First, through a human-subjects experiment that demonstrated how accurate inferences about listeners’ attentive state depend on including the social context of storytellers. Second, through a finer analysis, we found that storytellers’ cueing behaviors modify the interpretation of listeners’ response by magnifying our certainty about their attentive state.

To gain an accurate understanding about the social-emotional states of others, social robots will need to model how their own behaviors affect their human partners. We argue for a move away from the pervasive contextless approach to emotion recognition, especially for human-robot interactions. Socially situated robots are not passive observers but contribute to the interaction context and can pursue a pro-active form of inference. A robot’s awareness of the social effects of its actions has a critical role in affective computing.

3.4.3 Initial Evidence to Computational Approach

If we interpret some of our findings within the language of a planning agent, we observe that:

1. Speaker cues can elicit responses at different rates of success; the action space of a storytelling agent.
2. Certain cues magnifying the certainty about listeners’ attentive state; information-seeking actions reduce an agent’s uncertainty.
3. Speaker cues can influence the attentive state of listeners; state-changing actions enable an agent to reach its social goal.

This initial evidence informed our selection of using a partially observable Markov decision process to model the intentional inference of storytellers. Furthermore, we found that speaker cues are predictive of storytellers’ later judgement about listeners’ attentive state. This demonstrates that it is possible for our belief manipulation model to reason about storytellers’ beliefs about listeners from their employment of speaker cues.
INTENTIONAL INFERENCE

CHRISTOPHER: Cryptography is the science of codes.
YOUNG ALAN TURING: Like secret messages?
CHRISTOPHER: Not secret. That’s the brilliant part.
Messages that anyone can see, but no one knows what they mean, unless you have the key.
YOUNG ALAN TURING: How is that different from talking?
When people talk to each other they never say what they mean. They say something else. And you’re supposed to just know what they mean. Only, I never do. So how is that different?

— from the Imitation Game movie

4.1 OVERVIEW

We computationally model the intentional inference process of storytellers as a partially observable Markov decision process (POMDP). Although POMDP models are typically used to provide policies to agents operating in known physical worlds, we are solving for storytellers’ policy in how to behave in a social-emotional domain. As such, we learn from their demonstrations to determine how to best make an inference about listeners. We then compare our intentional inference model to estimate the attentive state of listeners to current state-of-the-art approaches to emotion recognition.

4.2 INFORMAL SKETCH

POMDPs provide a mathematical framework for modeling goal-oriented actions in situations where the underlying state of the world cannot be directly observed [Kaelbling et al., 1998]. Given a complete and correct model of the world dynamics and goal structure, a POMDP policy yields an optimal way to behave within the environment based on the agent’s belief about the world state.

Prior approaches have typically used POMDPs to provide policies to agents optimizing the trade-off between exploring an environment to help reduce state uncertainty (e.g., sensing an unknown rock) and exploiting the information gained to maximize rewards related to goal-achievement (e.g., collecting a desired rock). But rather than interacting with an environment, a storyteller (our agent) is interacting with people, and we model the agent as having a social goal of wanting listeners to pay attention.

By observing how listeners respond to the agent’s cueing actions, the agent can estimate their attentive state. The agent’s actions can change the state of the listener (state-changing actions) or seek a response that reduces its uncertainty about the listener’s current state (information-seeking actions). Assuming social costs exist in behaving in an irrational manner, the POMDP policy yields the best way, based on these social constraints, to produce speaker cues for a storytelling agent that wants the listener to pay attention.
We detail the construction of the parameter space and dynamics used to encode our storytelling domain as a \textit{POMDP world} (Section 4.3.1: \textit{POMDP World}). For ease of explanation, we assume that we know the correct \textit{POMDP world} in this section. We then describe how we model a storyteller’s belief about a listener’s attentive state given this formulation (Section 4.3.2: Belief Updating). Assuming that a storyteller is driven by an interaction goal, we describe how we solve for a policy operating within the behavioral constraints of this social context (Section 4.3.3: \textit{POMDP Policy}).

When we refer to an agent in this chapter, we take the perspective of the storyteller and model their ego-centric view of an interaction. We start with a helpful narrative illustrating a storytelling session from our data collection but told from the storyteller’s point-of-view:

\textit{Stacy the Storyteller wants to tell Lenny the Listener her favorite storybook. Stacy wants Lenny to pay attention to her story (goal). At the start, she is pretty confident that he is following along (initial belief). A few moments later, Stacy looks up from the book to glance at Lenny (gaze cue), and his posture is slumped and leaning away from the book (high-negative response). Stacy is starting to think that Lenny is maybe no longer engaged (belief update). At her next utterance, Stacy energetically says “the lion gave out a loud ROARRRRR” and looks over at Lenny (strong cue). Lenny excitedly leans toward the book and asks “and thennn?” (high-positive response). Stacy is now pretty confident that Lenny is paying attention (goal achievement). Throughout the storytelling, Stacy continues to glance over at Lenny (infinite-horizon policy).}

4.3.1 \textit{POMDP World}

\textbf{Parameter Space} Throughout the interaction, the storyteller is trying to assess whether the listener is paying attention or not. Our \textit{state space} is two discrete states of \textit{Listening} or not \textit{Listening}: \{\textit{L}, \textit{\overline{L}}\}.

We discretize the \textit{action space} as \textit{A}: \{gaze, weak, strong\}, where ‘gaze’ is our information-seeking action and ‘weak’ & ‘strong’ cues are our state-changing actions. Based on the observed cueing rates from Section 3.3: Effects of Speaker Cues, we selected that strong-actions are multimodal cues consisting of combined gaze and prosodic cues, weak-actions are solely prosodic cues, and gaze-actions are just gaze cues.

The storyteller can observe the listener’s response as five different observations of \textit{O}: \{high-negative, low-negative, neutral, low-positive, high-positive\}. This \textit{observation space} follows our prior method from Section 3.3.4.1: Data Tuples of measuring listener response as an overall valence rating except we reduce our original range of \([-3 \text{ to } +4]\) to \([-2 \text{ to } +2]\).\footnote{The remapping consists of capping at an absolute max value of two.}
**Social World Dynamics**  As described in the above scenario, the cueing actions of the storyteller can influence the listener’s attentive state. The probability that a cue causes a change in the listener’s state is represented by the *transition function* \( T: p(s'|s, a) \), which encodes the likelihood that the listener’s next state will be at \( s' \) when the storyteller takes a cueing action \( a \) when the listener is in state \( s \).

The storyteller infers the listener’s attentive state by observing the non-verbal responses throughout the interaction. As demonstrated in Section 3.3.4: on Listener Response Interpretation, the interpretation of listener’s responses also depend on accounting for the storyteller’s cueing-context, which is represented in the *observation function* \( \Theta: p(o|s, a) \). This encodes the distribution of response behaviors \( o \) to a cueing action \( a \) when the listener is in state \( s \). The observation function can capture how conjointly emitted speaker cues (i.e., strong actions) can increase the likelihood of a positive response from listeners.

### 4.3.2 Belief Updating

Throughout the interaction, the storyteller maintains a belief about the listener’s attentive state, which is characterized as the following probability distribution:

\[
b(s) = \begin{cases} 
  p_L & \\
  p_T = 1 - p_L 
\end{cases}
\]

In our scenario, the storyteller’s current belief about the attentive state of the listener is influenced by how Lenny responds to her cueing actions. Upon taking an action and receiving an observation, the storyteller updates her current belief based on two factors: 1) the likelihood of that action causing a state change, 2) the likelihood of receiving that observation in response to that action at the next state:

\[
b' \leftarrow \text{BeliefUpdate}(b, a, o), \quad \text{shorthand operation} \\
b(s') \propto \Theta(a, s', o) \sum_{s \in S} T(s, a, s') b(s)
\]

This forward updating of beliefs is known as *Bayesian filtering* and starts with the initial belief \( b_0 \). In our illustration, the starting belief of the storyteller is her expectation that the listener is attentive from the get-go. This initial belief can represent any biases people have coming into an interaction. With novel partners, this bias can represent a social norm or standard. An initial belief with no expectations would be represented as a uniform distribution over states.
4.3.3 POMDP Policy

The representations defined thus far describe the forward dynamics of our problem domain; how the storyteller’s beliefs about the listener change given the sequence of actions and responses. But our storyteller is driven by a social goal of wanting Lenny to pay attention to her story. How does an agent plan in this social world? Given the world dynamics along with a goal and a set of behavioral constraints, a policy can be calculated by maximizing the expected short-term and long-term gains of cueing actions leading the listener to a desired state.

**Reward Function** Our storyteller wants Lenny to pay attention to her story. Through her set of cueing actions, the storyteller can influence the listener toward a desired attentive state. However, her use of these actions are bounded by the constraints and rules governed by this social context. For example, the storyteller could always use strong cues at every opportunity to ensure listener’s attention. This is conceptually similar to constantly asking someone “Are you still paying attention,” which can negatively impact interaction flow as well as frustrate a diligent listener. In fact, we observe that these strong cues are used sparingly by storytellers.

Continuing in our POMDP formulation, we can represent how taking actions in certain states has associated social costs through the reward function \( R(s, a) \), which assigns positive and negative rewards of taking action \( a \) when the listener is in state \( s \). An example function that captures our intuitions about our problem domain is as follows:

\[
R(s, a) = \begin{cases} 
R(L, \text{gaze}) = -1 & R(\overline{L}, \text{gaze}) = -1 \\
R(L, \text{weak}) = -3 & R(\overline{L}, \text{weak}) = -2 \\
R(L, \text{strg}) = -5 & R(\overline{L}, \text{strg}) = -4
\end{cases}
\]

We expect that strong cues are overall the most costly action and gaze cues the least. Since state-changing actions are more appropriate and beneficial when the listener is not paying attention \( \overline{L} \), they are less costly when taken in that particular state.

**Goal Achievement** Typically, AI planning problems terminate when a particular state is reached, i.e., a goal state. In our POMDP formulation, we instead use an action-based termination where a goal is considered to be reached once a special stopping action is taken.

Once a storytelling agent believes with high enough certainty that the listener is paying attention, then it should take the ‘predict-listening’ action (or predict for short). If the listener is indeed paying attention, then this state-action configuration generates a high positive reward.† A high penalty

† As Kaelbling et al. [1998] points out, “The reward function may seem strange; the agent appears to be rewarded for merely believing that it is in good states. However, because the state estimation is constructed from a correct observation and transition model of the world, the belief state represents the true occupation probabilities for all states...”
is given if the storyteller makes an incorrect prediction. The only value of this terminating action is to collect reward since its observation function and transition function are uniformly distributed across all states, thereby revealing no helpful differentiating information regarding state as well as results in a belief of high uncertainty. Note: our predict action is not an observable behavior that a storyteller does in our problem domain, but it is a necessary mechanism for a POMDP to encode goal achievement.

**INFINITE HORIZON POLICY** The belief-based reward function below defines the immediate value of performing an action given our current belief.

\[
\rho(b, a) = \sum_{s \in S} b(s)R(s, a)
\]  

This is sufficient information for an agent operating on a short-term/greedy policy, which is called a myopic policy. For an agent operating on a longer-term policy, it has to know how to maximize not only the the short-term gains of actions but also the future gains of actions leading to desired beliefs. Formally, the value of taking action \(a\) while holding belief \(b\) is defined as the Q-function:

\[
Q(b, a) = \rho(b, a) + \gamma \sum_{o \in O} \Theta(o|b, a)V(b')
\]

The Q-function expresses the value of belief-action pairs as a sum of the immediate value of actions and its discounted future value when factoring in all possible observations leading to a next belief. The discount factor \(\gamma\) controls how far to look into the future (before the value of rewards becomes negligible in influencing the policy). To obtain the value function, an agent simulates, over different initial beliefs, the successive beliefs it can reach by trying every action and observation sequence, or more formally:

\[
V(b) = \arg \max_{a \in A} Q(b, a)
\]

Algorithms such as Value Iteration and Policy Iteration (i.e., POMDP solvers) are examples of approximation methods to compute this value function [Smallwood and Sondik, 1973; Hauskrecht, 2000; Ross et al., 2009]. Based on this Q-function, an agent executes an optimal policy by selecting actions of maximum value. But for our work, we use a soft-max policy:

\[
p(a|b) \propto \exp(\beta Q(b, a))
\]

This is a policy based on a probability distribution over actions given a belief. This stochastic selection of actions is more appropriate for our target domain
Figure 14: Graphical Representation of our POMDP Model. POMDPs consist of two processes: [1] the state estimator maintains a belief about the listener’s attentive state. [2] the policy recommends cueing actions, based on the current belief, that will maximize expected long-term reward.

inverse temperature

since people do not always behave optimally. The $\beta$ parameter, known as the inverse temperature, defines the degree to which an agent sticks to the policy. As $\beta$ approaches infinity, the agent executes the exact optimal policy, and as it approaches zero, the agent execute a random policy where actions are selected from an uniform distribution.

AN ONGOING POLICY Once an agent believes with high enough certainty that the listener is paying attention, then the best action is to predict. But given our current formulation, this stops policy execution since our desired goal has been reached. However in real interactions, storytellers continue with their cueing behaviors since a listener’s attentive state is dynamic and could change over time. The inference process is continuous.

We transform our policy into an ongoing policy by augmenting the solved value function. To prevent the selection of the predict action, we set its value across all beliefs to negative infinity. This effectively drives the probability of choosing this action down to zero based on Equation 7. In redistributing the probability to the remaining actions, the second best action becomes more likely. As such, if the storyteller believes with high confidence that the listener is paying attention, then instead of doing a predict action, the ongoing policy will suggest the next best action of a gaze cue, which is an appropriate behavior for our problem domain.
4.3.4 Summary

Our POMDP construction of storytellers’ intentional inference process is summarized below (also see Figure 14):

- **State Space** $S$: $\{L, \overline{L}\}$ is the set of listener’s possible attentive states.

- **Action Space** $A$: $\{gaze, weak, strong, predict\}$ is the set of nonverbal cueing actions available to storytellers, including the terminating action ‘predict’ to encode goal achievement.

- **Observation Space** $O$: $\{hi\_neg, lo\_neg, neutral, lo\_pos, hi\_pos\}$ is the set of listener’s nonverbal responses measured as an overall valence rating. Notice that these are observations for the storyteller.

- **Transition Function** $T$: $p(s'|s, a)$ is the state-transition distribution encoding how cueing actions of the storyteller can influence the listener’s attentive state.

- **Observation Function** $\Theta$: $p(o|s, a)$ encodes how listeners respond to a cue when in a particular state.

- **Reward Function** $R$: $(s, a)$ assigns immediate state-dependent rewards of taking actions. This encodes the behavioral constraints and goals of the social context.

- **Initial Belief** $b_0$ captures the storyteller’s bias regarding the listener’s attentive state at the start of the interaction.

- **Discount factor** $\gamma \in [0, 1]$ represents the importance of obtaining future rewards over immediate rewards.

- **Ongoing Policy** $\Pi^S: b \rightarrow a$ defines how storytellers should continuously act based on their goal. Given inverse temperature $\beta$, it expresses a soft-max policy suggesting a stochastic selection of actions (minus the predict action) based on the current belief about the listener’s attentive state.

$\lambda = (S, A, O, T, \Theta, b_0, R, \gamma, \beta)$ is the full set of POMDP parameters.
Thus far, we have made the assumption that the parameters of the POMDP model for our storytelling domain is fully known and correct. But especially for social-emotional domains, agents will need the capacity to learn across different individuals, situations, and cultures.

Agents can learn by either directly experiencing the world or by observing others. A teacher can also guide its experience, scaffold new demonstrations, or provide feedback/critique [Chernova and Thomaz, 2014]. For our work, we follow an apprenticeship learning paradigm where an agent learns from an expert demonstrating the task [Makino and Takeuchi, 2012]. This approach makes a key assumption that the expert has perfect knowledge of the world and therefore the expert’s behaviors are based on an optimal policy based on this world knowledge. Their learning algorithm searches for a world model that can best explain the experts’ actions as well as the observed responses. From a small set of demonstrations, this apprenticeship learning approach can learn correct estimates to an unknown POMDP.

4.4.1 Likelihood Estimate

For our storytelling domain, a demonstration consists of pairs of actions and observations $D$: $(a_1 o_1, a_2 o_2, \cdots, a_t o_t)$. This captures the sequence of storyteller’s cueing behaviors and listener’s nonverbal responses. The sto-

![Diagram of the Apprenticeship Learning Algorithm.](figure15.png)
rytellers in our data collection are the “experts” demonstrating how to act given the social goal of wanting listeners to pay attention.† The apprenticeship learning algorithm evaluates how well our current world hypothesis $\lambda$ along with its induced policy $\Pi_{\lambda}$ can explain our demonstration sequence. We estimate the likelihood of a demonstration given the world and policy $p(D|\lambda, \Pi_{\lambda})$ as function of two factors (see Equation 8):

1. the likelihood that the policy would have suggested to do the demonstrated action at its current belief.

2. the likelihood of having seen the demonstrated response to the demonstrated action according to the POMDP model given its next belief.

$$p(D|\lambda, \Pi_{\lambda}) = \prod_{i=1}^{L} p(a_i|b_i) \times \prod_{i=1}^{L} p(o_i|b_i, a_i)$$

$$= \prod_{i=1}^{L} \Pi_{\lambda}(a_i|b_i) \times \prod_{i=1}^{L} \sum_{s' \in S} \Theta(a_i, s', o_i) \sum_{s \in S} T(s, a_i, s') b_i(s)$$

For an entire dataset, the log-likelihood of each demonstration is summed together to represent how well the candidate model fits (i.e., the overall objective function). It is important to note that when calculating the action likelihood, we use the ongoing policy as described in Section 4.3.3: An Ongoing Policy. In natural interactions, storytellers do not demonstrate the ‘predict’ action. There is no direct and observable signal of their inference decision. As such, the action likelihood is evaluating the appropriateness of the policy’s second best action.

4.4.2 Maximum A Posteriori

To find a model that best explains the set of demonstrations, we search over a prior distribution of possible world models for one that maximizes the overall log-likelihood of the dataset, which is called maximum a posteriori (MAP) estimation. This search is performed using a generic nonlinear-optimization algorithm (COBYLA) that maximizes this objective function over a provided prior distribution (see Figure 15 for an overall diagram of the apprenticeship learning algorithm.).

† See Section 3.3.3: Story Sharing on how we instruct children storytellers from our data collection of this social goal.
We compare our intentional inference approach to emotion recognition against current state-of-art approaches. As described in Section 2.4: Emotion Recognition, current approaches to emotion recognition fall under two schools of thought. Intra-personal models focus solely on the behaviors of the individual of interest. Inter-personal models further capture the co-presence of the individual’s partner but represented their behaviors as flat observations. We match these emotion-modeling perspectives to machine learning algorithms that are commonly used. To setup fair and representationally-similar comparisons against our POMDP model, we stay in the algorithmic family of Markov models. We represent the intra-personal modeling approach as a standard hidden Markov model (HMM) and the inter-personal model approach as a multivariate hidden Markov model (MHMM). We state the following hypothesis:

**Main Hypothesis:** Inferences about listeners’ attentive state is most accurate when modeling the intentional inference process of storytellers and least accurate when not accounting for their added social context.

After describing our alternative state estimators (Section 4.5.1: Alternative State Estimators), we detail how we learn the unknown parameters of our models (Section 4.5.2: Learning Model Parameters). The training and testing procedures used for model assessment evaluates the predictive power of a model to infer the attentive state of listeners (Section 4.5.3: Training and Testing Framework).

### 4.5.1 Alternative State Estimators

We briefly describe our alternative state estimators and their capacity to represent our storytelling domain compared to our POMDP model. For our work, we use the Bayes Net Toolbox to represent, learn, and make inferences for our HMM and MHMM models [Murphy, 2001].

**HMM** *Hidden Markov models* represent a sequence of observations as probabilistically emitted from a temporal progression of hidden states. An HMM model consists of three probability distributions: initial states \( b_0 \): \( p(s_0) \), state transitions \( T \): \( p(s'|s) \), and observation emissions \( \Theta \): \( p(o|s) \). Compared to the state estimator component of POMDPs, HMMs model a similar process except that the likelihood of observations and transitions are not conditioned on an action. This means it cannot represent how the storyteller’s cueing actions can influence listeners’ state nor their observed response behaviors.

**MHMM** *Multivariate hidden Markov models* are an extension of the standard HMM. They include multiple observations as separate random variables per emission. Our MHMM model adds an extra probability distribution of \( p(a|s) \). This distribution captures how the storyteller’s cueing behaviors are
### Table 12: Differences Between State Estimators

All the three models have model parameters to learn from the dataset. The HMM has its initial state, transition function, and emission function. The MHMM has the same set except it has an extra action-emission function. The POMDP also has the same set, but its transition and emission functions are conditioned on both state and action. **Note:** the POMDP’s policy is not used for state inference, but it is used to shape how the parameters are learned. Both the HMM and MHMM use expectation-maximization to learn their model parameters, while the POMDP uses a generic non-linear optimizer. The number of free parameters grow with model complexity.

<table>
<thead>
<tr>
<th>Learning Method</th>
<th>HMM</th>
<th>MHMM</th>
<th>POMDP</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Free Param.</td>
<td>11</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Unknown Param.</td>
<td>(b_0)</td>
<td>(b_0)</td>
<td>(b_0)</td>
</tr>
<tr>
<td></td>
<td>(p(s'</td>
<td>s))</td>
<td>(p(s'</td>
</tr>
<tr>
<td></td>
<td>(p(o</td>
<td>s))</td>
<td>(p(o</td>
</tr>
<tr>
<td></td>
<td>(p(a</td>
<td>s))</td>
<td>(p(a</td>
</tr>
</tbody>
</table>

Related observations to the state of listeners. But it does not model the structure and dynamics of the cueing actions like their state-changing properties.

### 4.5.2 Learning Model Parameters

In general, supervised machine learning consists of using labeled training data to estimate parameter values (*parameter learning* or fitting) as well as define model-specific variables (*model selection*). For our work, the model structures (i.e., the relationship between the variables) are already defined as they represent the different hypotheses (HMM, MHMM, POMDP) we have regarding emotion recognition. As such, we only focus on only learning the parameter values.

#### 4.5.2.1 Training Data

For each storytelling episode from our data collection, we extracted data tuples of \(<a_1 o_1 s_1, a_2 o_2 s_2, \ldots, a_t o_t s_t>\). This sequence captures the temporal progression of storyteller’s cueing actions, listener’s immediate re-
sponses, and the listener’s resulting next states (see Figure 11 as well as Section 3.3.4.1: Data Tuples for more details on the method of data extraction)\(^\ddagger\).

The action-observation pairs are used to train the MHMM and POMDP models while only the observations are used for the HMM model. The associated state labels serve as the ground truth to listeners’ attentive state. They are used to evaluate the inference performance of our state estimators.

For each of the 58 storytelling episodes, a valid demonstration sequence is defined as a contiguous set of action-observation pairs of a minimum length of 5. It is possible for a storytelling episode to not generate any demonstrations if it has insufficient action-observation pairs or generate multiple demonstrations if in the middle of the episode both participants were momentarily off-task (violating our contiguous rule). In total, we have 60 demonstrations of an average length of 10 (min of 5 and max of 26).

\[\text{Sidenote 1 (Imperfect State Label)}\]

To obtain our state label, a third-party adult observer continuously rate throughout the interaction whether the listener is paying attention or not. Other common methods to obtain state labels are to prompt individuals to act out an emotion or induce specific emotions through experimental methods. A perfect ground truth label would be knowing the listeners’ exact true state, but annotations are typically used as a proxy measure.

4.5.2.2 Optimization Algorithms

For parameter learning, we employ the general technique of searching for a set of parameters that maximizes the likelihood of observing the training data, also known as the maximum likelihood estimate (MLE). For the HMM and MHMM models, we use a common optimization algorithm called expectation-maximization (EM) to iteratively adjusted parameters toward a MLE [Rabiner, 1989]. Our POMDP model employs a generic nonlinear-optimization algorithm to perform its MAP search (see Section 4.4.2: Maximum A Posteriori).

To set up fair training conditions across all models, the optimization algorithms had the same terminating criterions of not exceeding 1000 evaluations and of stopping when an optimization iteration changes its objective function value by less than a relative tolerance of \(1e^{-4}\).

4.5.2.3 Known vs Unknown Parameters

POMDPs can be decomposed into two components [Kaelbling et al., 1998]:

1. the state estimator that updates beliefs about states
2. the solved policy for action selection

\[^\ddagger\] We follow the method as described in Section 3.3.4.1: Data Tuples except that a single vote counted as a valid backchannel moment (i.e., consensus level of 1). For our human data analyses, we wanted high confidence in the detection of a cue through two-coder agreement. However for our training data, we relaxed this constraint as it caused the frequency of cues in an episode to become very sparse.
Since our primary goal is to evaluate the inference performance of our state estimators, we focus on learning only the parameters \((b_0, T, \Theta)\) related to this first component for our POMDP model. As such, we assume that the parameters related to policy formation \((R, \gamma, \beta)\) are known (see Appendix C for details on how the known parameters were selected). The number of parameters to estimate grow with model complexity. As such, our HMM model has the fewest and our POMDP model has the most (as shown in Table 12).

**Side note 2 (Reduction of Unknown Parameters for POMDP)**

Based on Equation 9, our POMDP model should have 31 free parameters to learn since \(N_{\text{states}} = 2, N_{\text{actions}} = 3,\) and \(N_{\text{obsrvs}} = 5\). However, the method of our learning algorithm enables the model designer to define parameter relationships that are highly correlated but unreliably estimated due to insufficient data. In particular, our dataset does not have many examples of how inattentive listeners respond to specific cueing actions. We make the assumption that if the resulting state is \(L\), listener’s response (or lack of one) is the same regardless of the cueing action. This simplification reduces the number of free parameters to learn from 31 to 23.

\[
\begin{align*}
N_{\text{params}} &= N_{\text{states}}(N_{\text{states}} - 1)N_{\text{actions}} + \\
&= N_{\text{states}}(N_{\text{obsrvs}} - 1)N_{\text{actions}} + \\
&= N_{\text{states}} - 1
\end{align*}
\]

from transition function

from observation function

from initial states

### 4.5.3 Training and Testing Framework

In Section 4.5.3.1: Cross-Validation, we detail our leave-one-out cross-validation (LOOCV) method to evaluate the inference performance of our models (model assessment). To setup fair comparisons, we use filtering inferences for all the models (Section 4.5.3.2: Inference Method) and set their prior distributions using similar pre-counts (Section 4.5.3.3: Initial Conditions). See Algorithm 1 for details of our training and testing framework used for each of our state estimators.

#### 4.5.3.1 Cross-Validation

When modeling human behavior, collecting and annotating enough real-world data is a challenge. In cases when the sample size is small, the method of cross-validation (CV) is often used to estimate prediction error by partitioning the dataset into subsets, and in multiple rounds, each subset acts as the testing set while the remaining is used as the training set. One benefit of using cross-validation is that the model can be trained from almost the whole dataset. For our model assessment, we leave one demonstration out and train on the remaining 59. The resulting state estimator then makes an inference on the omitted test demonstration. This process is repeated such that each demonstration is left out once and tested.
Algorithm 1: Training and Testing Framework

\[ M = \# \text{ of simulations} \]
\[ N = \# \text{ of demonstrations} \]
\[ D = \{D_1, D_2 \ldots D_N\}, \text{containing action-observation pairs} \]
\[ S = \{S_1, S_2 \ldots S_N\}, \text{corresponding ground truth labels} \]

Initialize prior distributions
Sample candidate models \( \Theta_1 \cdot \cdot \cdot \Theta_M \) from prior

for \( i = 1, 2, \ldots, M \) do
    initialModel = \( \Theta_i \)
    for \( j = 1, 2, \ldots, N \) do
        Set \( \text{test} = \{ D_j \} \)
        Set \( \text{train} = D \setminus \{ D_j \} \)
        trainedModel = learnParameters\( (\text{train}, \text{initialModel}) \)
        beliefs\( j = \text{filteringInference}(\text{test}, \text{trainedModel}) \)
    end for
    auc\( i = \text{calculateAUCMeasure(beliefs}, S) \)
    accuracy\( i = \text{calculateAccMeasure(beliefs}, S) \)
    f1score\( i = \text{calculateF1Measure(beliefs}, S) \)
end for

\[ \text{AUC}^* = \argmax_i (\text{auc}) \]
\[ \text{ACC}^* = \argmax_i (\text{accuracy}) \]
\[ \text{F1SCORE}^* = \argmax_i (\text{f1score}) \]

4.5.3.2 Inference Method

Different types of queries can be made against a dynamic Bayesian model:

Viterbi

\( \text{Viterbi} \) path: Given the history of observations to the present time, what is the most likely current state?

Filtering

\( \text{Filtering} \) or online inference: Given the history of observations to the present time, what is the belief of the current state?

Smoothing

\( \text{Smoothing} \) or retrospective inference: Given the history of observations to the present time, what is the belief about this past state?

For our evaluation, we use the filtering method to make an inference about the current state. Since our target application is enabling social robots to appropriately act in realtime based on their current estimates about the attentive state of listeners, we evaluate our state estimator’s capacity to form these online inferences.

4.5.3.3 Initial Conditions

All of our optimization algorithms are sensitive to initial conditions and perform a local search given a starting point. To increase the chance of finding a good set of parameters for our state estimators, we also perform a macro-level search by trying out different starting configurations. More specifically, we ran 1000 simulations (i.e., tried out 1000 candidate models) where the initial state, transition, and observation probabilities were randomly drawn from prior distributions.
Priors were set with broad distributions centered at values close to the MLE of parameters. To set similar priors across models, the pre-observation counts (i.e., the initial $\alpha$ vector) used to set the POMDP’s beta and Dirichlet distributions were tallied and averaged into a collapsed form for the HMM and MHMM models. For example, transitions for $\bar{L} \rightarrow L$ for a POMDP is conditioned on actions and are expressed with the following set of priors:

$$p_{\text{gaze}} \sim \text{Beta}(2, 8), \quad p_{\text{weak}} \sim \text{Beta}(4, 6), \quad p_{\text{strg}} \sim \text{Beta}(6, 4)$$

These pre-observation counts are tallied and averaged for the HMM and MHMM models to yield a transition prior of $p_{L \rightarrow L} \sim \text{Beta}(4, 6)$.

### 4.5.4 Performance Measures

We measure inference performance as a function of accuracy, the F1-score, and the area under the receiver operating characteristic (ROC) curve (AUC). Since our class distribution is skewed, with listeners demonstrating attentiveness $0\%$ of the time, we primarily focus on measures that are more robust to class imbalance in our discussions (i.e., F1-score and AUC). We also follow the convention of setting the positive class condition to be that of the rarer class. As such, the performance measures reflect how well a state estimator detects inattentive listeners.

Since the ground truth label is binary, we convert beliefs to state predictions based on a threshold value between $[0, 1]$. This threshold defines the class boundary. Given a performance measure, an optimal threshold is one that maximizes this specific metric.

### 4.5.5 Inference Results

Out of the 1000 simulations of trying various initial conditions (i.e., possible model candidates), the best HMM, MHMM, and POMDP models for a given metric are listed in Table 13. Overall, all the state estimators perform moderately

<table>
<thead>
<tr>
<th>Measure</th>
<th>Prior</th>
<th>HMM</th>
<th>MHMM</th>
<th>POMDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.80</td>
<td>0.85</td>
<td>0.84</td>
<td>0.86</td>
</tr>
<tr>
<td>F1 Score</td>
<td>0.00</td>
<td>0.58</td>
<td>0.59</td>
<td>0.63</td>
</tr>
<tr>
<td>AUC</td>
<td>n.a.</td>
<td>0.79</td>
<td>0.80</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Table 13: **Overall Performance of State Estimators.** The prior classifier always predicts the most common class. A general trend can be observed across measures where the POMDP model obtains the highest inference performance, followed by the MHMM model, and then the HMM model.
well in classifying the attentive state of listeners (see ROC curves in Figure 16).

But as hypothesized, a general trend can be observed across measures where the POMDP model obtains the highest inference performance, followed by the MHMM model, and then the HMM model.

Similar to the trend analysis described in our previous experiment (in Section 3.2.2), we test for a hypothesized trend of linearly increasing inference performances across our expected order (HMM < MHMM < POMDP). Since our demonstrations are not independent samples but consist of multiple demonstrations of particular listeners, we perform a multi-level linear regression analysis that controls for within-listener dependencies from these repeated measures. This analysis tests whether a significant ascending trend exists in the average performance of state estimators across the different listeners.

We test for this trend on the F1-score measure. Based on the Wald Chi-square statistic, a statistically significant performance trend is found \( \chi^2(1) = 4.76, p^* = 0.03 \), where the POMDP \( [\mu = 0.512, SD = 0.272] \) is the most accurate estimator across listeners, the MHMM is next \( [\mu = 0.481, SD = 0.273] \), and then the HMM \( [\mu = 0.478, SD = 0.267] \).
4.6 DISCUSSION

We compared our intentional inference model to estimate the attentive state of listeners against current state-of-the-art computational approaches to emotion recognition. We find that a state estimator capable of accounting for the co-presence of storytellers’ behaviors (i.e., the MHMM model) outperforms one that only models the nonverbal expressions of listeners (i.e., the HMM model). But the most accurate interpretation of listeners’ attentive state is achieved when capturing the intentional context of storytellers (i.e., the POMDP model).

We posed that storytellers employ a policy to infer the attentive state of listeners. As such, we computationally model emotion recognition as an intentional inference process of storytellers using a classic planning algorithm of a partially observable Markov decision process. This is an ego-centric model that takes the perspective of storytellers, who themselves are making inferences about listeners during a real interaction. Although POMDP models are typically used to provide policies for agents operating in known physical worlds, we use an apprenticeship learning paradigm to learn from human demonstrations. In learning the inference policy of storytellers, we were able to better estimate the social-emotional attentive state of listeners.

**Limitation 1: Implicitly Modeling the Passage of Time**

The intentional inference model makes the assumption that the listener’s mental state is stationary until the next action-response observation. The passage of time is implicitly represented. The model depends on cueing events occurring at an informative rate, but this could lead to missed transitions. Although these cues help highlight the possible “rich points” of an interaction, the model would benefit from including other cueing sources from the environment as well as incorporating unsolicited listener behaviors.

**Limitation 2: Non-Gradient Based Optimization Method**

As Domingos [2012] points out about optimization methods: “It is common for new learners to start out using off-the-shelf optimizers, which are later replaced by custom-designed ones.” Although POMDPs were introduced in the last two decades [Kaelbling et al., 1998], its popularity in robotic applications have primarily been in solving policies for known worlds, and limited research exist on how to learn models with partial or mixed observability from human demonstrations [Atrash and Pineau, 2010; Nikolaidis and Shah, 2013].

Although our optimization method was at a disadvantage, we believe that what our approach lacked in terms of guaranteeing to find a local maxima for the POMDP parameters, it compensated by having a more appropriate representation of the domain.
Limitation 3: Known Reward Function  We only learned the parameters related to a POMDP’s state-estimator and did not learn the variables related to policy formation, mainly the reward function. However, learning the reward function, which represents the social costs of performing cueing actions, will be necessary as it can change across individuals, cultures, and other contexts. As a generalization of inverse reinforcement learning [Abbeel and Ng, 2004], the apprenticeship learning algorithm [Makino and Takeuchi, 2012] supports learning the transition, observation, and reward functions. However, to obtain a reasonable model, the learning algorithm will need either strong priors on the unknown parameters or significantly more data.

Limitation 4: Modeling One Emotional State  For our work, we modeled only one social-emotional state of attention, but social interactions can include multiple and simultaneous emotional states. For example, a listener can be attentive yet also bored. The state space of a POMDP model can represent the discrete set of emotional states that can hold true for a storytelling interaction, but a growing state space also means an exponential growth of unknown parameters to learn.

Limitation 5: Learning from Annotated Behaviors  A challenge when modeling human behavior is collecting and annotating enough real-world social interactions. But in identifying the relevant speaker cues and listener responses of our storytelling domain, we can develop the automated detection for these specific behaviors. This can enable a robot to use our current model to bootstrap its learning, then in realtime it can learn a more refined model through interactions with new people.

Lessons Learned: Modeling Emotions of Children vs Adults  Our model needs demonstrations of when listeners are attentive and also inattentive to reason along the belief spectrum. When modeling social-emotional behaviors, an unforeseen benefit in working with children is their level of emotional expressivity. The children were overt and obvious in their expressions of inattention, allowing for easier annotation of their attentive state.

Furthermore, although the storytelling episodes on average lasted only 1 minute and 17 seconds, listeners overall demonstrated inattentiveness 20% of the time. In our prior work modeling inter-personal trust among adults [Lee et al., 2013], social interactions resulted in a majority of participants having higher degrees of trust toward their novel partner while a mere 12% demonstrated having lower degrees of trust. We conjecture that children compared to adults exhibit more honest social-emotional expressions that are not yet subdued from social display rules.
Social regulation of emotion refers to one individual (the regulator) deliberately attempting to change the emotional response of another individual (the target)” [Reeck et al., 2016]. A growing recognition in cognitive science is the push to better understand how people regulate the emotions of others and its connectedness to self-regulation. Prior work has already demonstrated that children can catch mental state such as curiosity or a growth mindset from robots [Gordon et al., 2015a; Park et al., 2017b]. But robots can also participate in actively regulating the emotions of others [Gordon et al., 2015b].

Social robots can use our learned policy of storytellers to not only make an intentional inference about their human partners but also pro-actively get listeners to pay attention. An immediate extension of this work is to demonstrate how a storytelling robot using the POMDP policy is more successful in getting listeners to pay attention compared to a robot that arbitrarily uses speaker cues throughout the storytelling.
BELIEF MANIPULATION

5.1 OVERVIEW

We computationally model emotion expression as a combined process of estimating people’s beliefs through inference inversion and then producing nonverbal expressions to affect those beliefs. In the last chapter, we demonstrated how to model emotion recognition as an intentional inference process of storytellers. Based on our learned policy of storytellers, we can track storytellers’ beliefs about a robot’s attentive state from their employment of speaker cues. An agent can then produce appropriate listener responses to manipulate those beliefs toward a desired perception of being an attentive listener. Our belief manipulation model makes two key assumptions:

1. The listening agent’s partner (i.e. the human storyteller) has a goal of wanting the robot to pay attention.

2. Our formulation of intentional inference as a POMDP is a predictive model of how storytellers behave in our domain.

Given these assumptions, we describe our computational formulation for belief manipulation, which takes the perspective of a robot listener and models its ego-centric view of a storytelling interaction. Through a human-subjects experiment, we demonstrate how a listening robot operating under our belief manipulation paradigm more effectively communicates an attentive state compared to current state-of-the-art approaches.

5.2 COMPUTATIONAL REPRESENTATION AS A DBN

Our belief manipulation model consists of two parts. First, the belief estimator tracks the storyteller’s beliefs about the attentive state of a listening robot through plan inversion (Section 5.2.1: Belief Estimator). Second, the myopic policy selects the best listening response to affect those beliefs toward a desired inference (Section 5.2.2: Myopic Policy).

Our belief estimator follows a Bayesian Theory of Mind (BTOM) framework [Baker and Tenenbaum, 2014] to infer beliefs using a dynamic Bayesian network (DBN). Although this prior work is capable of making inferences about the beliefs and goals of others through plan inversion, we make the simplifying assumption that we know the goal of storytellers and therefore only focus on inferring their beliefs. We extend this prior work by adding on top of their representation how an agent could act on those beliefs to manipulate them toward a desired inference.
Figure 17: Belief Manipulation Model of a Listening Agent. The model consists of two processes: [1] the belief estimator tracks storyteller’s beliefs $b$ about the robot’s attentive state through plan inversion [2] the myopic policy selects listening behaviors $o$ to affect those beliefs toward a desired inference. Our greedy policy selects a next response that minimizes the distance between the resulting belief and the target belief.

5.2.1 Belief Estimator

The goal of the belief estimator is to infer what the storyteller’s think about the robot’s attentive state, which we will simply refer to as the storyteller’s belief. Rather than reasoning about their beliefs in continuous space like our POMDP model, we instead discretize this to a set of belief-states $B$ for our DBN model.

Under the theory of plan inversion, by observing storytellers’ policy unfolding through their speaker cues, we can estimate their beliefs about the robot. The soft-max policy (see Equation 7) describes how storytellers are likely to behave given their current belief. This defines the probability of action emission $p(a|b)$ for our DBN model. Conceptually, this is similar to a HMM where the sequence of observed action emissions represent a temporal progression of hidden belief-states.

But for our domain, storytellers’ beliefs transition not only as a function of their current belief and their emitted action but also depends on the listener’s response. The POMDP’s belief update function (see Equation 2) captures this belief transition, which is used to populate the transition function $p(b'|b, a, o)$ for our DBN model. The POMDP model, from which we derive these DBN components, achieved the highest AUC measure from Section 4.5.5: Inference Results (see Appendix C for more details on its model parameters and resulting policy).
5.2.1.1 **DBN Setup**

The forward dynamics of storytellers’ beliefs about a robot is captured by the DBN graphically shown in Figure 17. The complete set of model parameters to represent this model consist of:

- **B**: is a finite set of discretized storytellers’ beliefs, or belief-states for short. For example, an interval of 0.25 generates a set of belief-states \( \{b^1 = 0.00, b^2 = 0.25, b^3 = 0.50, b^4 = 0.75, b^5 = 1.00\} \) where \( b^1 \) is a highly confident belief that the listener is not paying attention and \( b^5 \) is a highly confident belief that the listener is attentive.

- **A**: is the action space of storytellers, which is the same as our POMDP model minus the predict action \( A: \{\text{gaze, weak, strong}\} \).

- **O**: is the action space of our listening agent, but we stick to our POMDP’s symbolic convention of these being observations to the storyteller. \( O: \{\text{hi_neg, lo_neg, neutral, lo_pos, hi_pos}\} \).

- **\( b_0 \)**: is the initial probability distribution over the storyteller’s beliefs. In essence, this is the agent’s starting belief about the storyteller’s belief. For example, a uniform distribution across belief-states reflects an agent starting an interaction with no idea what the storyteller thinks about the robot.

- **\( \Pi^S: P(a|b) \)** is the action emission probability of our storyteller given their current belief-state, which is our POMDP soft-max policy from Equation 7 but sampled at belief-states \( B \).

- **\( T: P(b'|b, a, o) \)** is the belief-state transition function modeling how we expect storytellers’ belief will change. It is a function of their last belief \( b \), their emitted action \( a \), and how the listening agent responds \( o \) to that action. To determine the probability of transitioning from \( b \) to \( b' \), we computed all possible action and observation combinations from \( b \), counted the frequency of transitions yielding \( b' \) based on our belief update in Equation 2, and then normalize this frequency over all next belief-states.

As a sanity check to ensure that the DBN setup is correct, we compared its inferences about our dataset against ones made by our POMDP model (see Appendix D for more details).
5.2.2 Myopic Policy

A listening agent has a goal of communicating an attentive state to the storyteller. Given the storyteller’s belief (about the agent), how can the agent change this belief to a desired one? A listening agent could make a long-term policy and form a sequence of optimal responses to manipulate the storyteller’s belief. But, our work first demonstrates how the agent could make a short-term policy. Our myopic policy, $\Pi^L : b \rightarrow o$, immediately selects the next response that yields a transition to a belief that is closest to our target belief.

An agent that wants to express a target attentive-state to its partner will have a belief-state target of $b_{\text{goal}} \in B$. A myopic policy assumes that an agent only has one action to achieve its goal. As such, our agent selects the next response $o$ to the storyteller’s cueing action $a$ to one that immediately transitions their current belief-state $b$ to a $b'$ that is closest to $b_{\text{goal}}$. It solves:

$$\arg \min_{o \in O} |E[T(b'|b, a, o)] - b_{\text{goal}}|$$

This function selects a listener response that minimizes the distance between the expected value of storyteller’s next belief and the target belief. In summary, $\lambda = (B, A, O, \Pi^L, T, b_0, b_{\text{goal}})$ is the set of parameters necessary to represent our belief manipulation model.

5.2.3 Demonstrative Examples

To better understand the resulting behaviors and inferences of our belief manipulation model, we simulate sequences of storyteller’s cueing actions to observe how the model tracks the storyteller’s beliefs as well as its recommendations for how the listening agent should respond given its goal. For our graphical demonstrations below, we use a discretization interval of 0.05 for our belief-states, and the listening agent has an initial probability distribution that is uniform over the belief-states.

### Sidenote 3 (Selecting a Belief Space Interval)

The interval value sets the level of granularity in which a model can reason about people’s beliefs. A larger interval will yield a model that reasons about people’s belief like on a 5-point scale ranging from very inattentive, somewhat inattentive, uncertain/neutral, somewhat attentive, and very attentive. Or even a 3-point scale of yes, maybe, and no. Although a smaller interval allows for greater precision for belief manipulation, this is limited by the actions and observations and their ability to cause movement (transitions) along this belief spectrum. The interval value also determines the sampling frequency of the POMDP policy. This sampling rate should be high enough to capture the changes and shifts of the policy from belief-state to belief-state.
**Scenario 1** For the first set of graphs, we examine how a listening agent with a $b_{\text{goal}}$ of 0.93 responds to a sequence of gaze cues. We also look at how it tracks the storyteller’s beliefs about the agent as both filtered and smoothed inferences. This belief goal value represents a target inference of being perceived as an attentive listener with high confidence.

Figure 18: **Scenario 1: Filtering Inference.** The robot starts with no idea (uniform distribution for $b_0$) what the storyteller thinks about the robot. After the storyteller emits a gaze cue $a_1$, the belief manipulation policy suggests for the robot to do a *high-positive* response $o_1$. This leads to a next belief distribution $b_1$ with a mass slightly closer to its belief goal. The robot again responses with a *high-positive* behavior to a gaze cue. And once more, but now at $b_3$ the agent is very close to its belief goal. At the next gaze cue, the robot responses with a *neutral* behavior to remain near the target belief $b_4$.

Figure 19: **Scenario 1: Smoothing Inference.** Using smoothing or retrospective inference, the listening agent can reason that the storyteller, given this sequence of gaze cues, began the interaction with a moderately confident initial belief that the robot is attentive, $E(b_1) = 0.79$. This is congruent with their policy in Figure 30 in Appendix C since gaze cues are more likely to occur when the storyteller believes that the listener is paying attention.
**Scenario 2** For the next set of graphs, the listening agent has a $b_{\text{goal}}$ of 0.20, which is a target inference of being perceived as an *inattentive* listener with moderate confidence. The storyteller again does a sequence of gaze cues.

**Figure 20: Scenario 2: Filtering Inference.** The robot again starts with a uniform distribution for $b_0$. After the storyteller emits a gaze cue $a_1$, the belief manipulation policy suggests for the robot to do a *high-negative* response $o_1$. This leads to a next belief distribution $b_1$ with a mass that is moving in the opposite direction of our goal even though the robot did a negative response. Remember, the next belief is a function of *both* cues and responses. As such, the gaze cue, which is indicative of the storyteller believing the robot to be attentive, contributes to this resulting belief distribution. The robot then responds to the next two gaze cues with *low-negative* behaviors. At $b_3$, the agent is very close to its belief goal. At this point, if the robot were to do another *low-negative* or *high-negative* response to the next gaze cue, then the resulting belief distribution would actually cause the robot to overshoot its target belief. To continue being perceived as *moderately* not paying attention, the policy instead suggests for the robot to do a *low-positive* response.

**Figure 21: Scenario 2: Smoothing Inference.** Like the last example, the listening agent can use retrospective inference to determine that the storyteller began with a high confidence initial belief that the robot is attentive, $\mathbb{E}(b_1) = 0.91$. This would explain why a storyteller would continue with the gaze cues even though the listening agent is responding with inattentive behaviors.
5.3 Evaluation: Comparing Robot Listeners

Through a human-subjects experiment, we evaluate whether a robotic agent operating under our belief manipulation paradigm (or BTOM robot for short) can more effectively communicate an attentive state compared to current state-of-the-art approaches. As discussed in Section 2.5.3: Signaling Paradigm, current approaches to emotion expression for human-agent interactions predominately follow a signaling paradigm. A limitation of a signaling robot is its inability to know when it has successfully conveyed attentiveness to the human partner. While a BTOM robot can modulate its listening behaviors because it can track what the storyteller currently thinks about the robot. As such, we have the following hypotheses:

Hypothesis Set

1. A listening robot capable of belief manipulation more effectively communicates attentiveness over a signaling robot (Primary Effect).
2. A listening robot capable of belief manipulation is perceived as more human-like and intelligent over a signaling robot (Secondary Effect).
5.3.1 Technology

Major technologies to support this evaluation are the development of the Tega robot platform and the autonomous system for the perception-to-behavior pipeline.

5.3.1.1 Tega: Robot Platform

As both project manager and system integrator, the author lead a team of designers, animators, and engineers to design and develop a new social robot platform. Through its small size and furry exterior, Tega (as shown in Figure 23) is a research platform designed to support long-term interactions with young children. The robot leverages smart-phone technology to not only graphically display facial expressions but also for computation, which includes behavioral control, sensor processing, and motor control to drive its five degrees of freedom: head up/down, body-tilt left/right, body-lean forward/back, body-extend up/down, and body-rotate left/right. For increased perceptual awareness, we augmented the phone’s ability with an external camera that can capture high-definition images with a wider field-of-view.

To withstand long-term continual use, the efficient battery-powered system can last up to six hours before requiring charge. We designed for more robust and reliable actuator movements, including the ability to expand and contract rapidly through a lead-screw design between the torso and the head. This enables the robot to generate consistent and expressive behaviors over longer periods of time.

5.3.1.2 Autonomous System

Our autonomous system (see Figure 24) supports the perception-to-behavior generation pipeline for the Tega robot. In realtime, the system detects prosodic- and gaze-based cues of a human storyteller, decides the agent’s response given its policy, and controls the exhibited set of nonverbal behaviors.
Figure 24: **Autonomous Robot Listener System Diagram.** This perception-to-behavior generation pipeline is realtime and detects prosodic- and gaze-based cues of a human storyteller, decides the agent’s response given its policy, and controls the exhibited set of nonverbal behaviors. The message passing between modules uses ROS, an open-source Robot Operating System [Quigley et al., 2009].

**PERCEPTION LAYER** The prosodic cue detection module uses openSMILE [Eyben et al., 2013], an open-source audio feature-extraction software, to capture realtime speech features from a high-quality microphone (the MXL AC404 USB conference microphone). Four rule-based models were developed to detect the following prosodic speaker cues: pitch, energy, long pauses, and long utterances†. The parameters for these models were determined by training and testing against the children’s voices from our storytelling dataset (see our prior work for more details [Park et al., 2017a]).

The gaze cue detection module uses custom software to capture gaze targets using the Tobii eye tracker [Tobii AB Inc., 2017]. The module reports when a person starts and stops looking at a calibrated region of space that is either labeled as the storybook or the robot.

**COGNITION LAYER** The Action Space Mapper module combines the onset of a gaze-cue event along with any of the prosodic-cue events using an empirically determined merge window of 500 milliseconds. These combined cues are then translated into our action space, where strong-actions are multimodal cues consisting of both gaze and prosodic cues, weak-actions are solely prosodic cues, and gaze-actions are solely gaze cues. The module has some activation energy in that we limit the rate in which it can report/trigger a cueing action (at an empirically determined rate of 2 seconds).

The policy module decides how a listening agent should respond to the storyteller’s cueing actions given its policy. The set of listener responses are the high-level valence ratings defined in Section 5.2.1.1: DBN Setup.

**ACTION LAYER** The Tega robot is capable of exhibiting all the listening behaviors listed in Table 9 as single **behavioral units** and as **combinatorial sets**. The behavioral units **lean-forward**, **lean-away**, **gaze-partner**, **gaze-away**, **behavioral units**

† We were not able to detect for filled pauses such as *um* or *uh* as this is still an ongoing area of research [Audhkhasi et al., 2009]
gaze-book, smile, brow-raise† can be held as poses while nod and utterance are one-time triggers. Modifiers to a behavioral unit include setting the pose duration, delaying the onset of the behavior‡, and specifying parameters like the type of utterance (e.g., ‘oh’ or ‘uh-huh’) or a particular lean-forward animation.

A combinatorial set is made up of multiple behavioral units. For example, a combinatorial set of {lean-forward, gaze-partner} with no modifiers will have the robot exhibit a randomly selected lean-forward animation that is held at a default duration and blended with a looking-at subroutine that orientates the full robot (eyes, head, and body) to look-at a 3-dimensional world coordinate specified by the gaze target.

Each of the high-level listener responses are mapped to multiple combinatorial sets, and the Response to Behavior Mapper module randomly selects which one to command the Tega robot.

5.3.2 Method

5.3.2.1 Participants

Children and parents from the greater Cambridge/Boston community were recruited to take part in the experiment. In total, we had 16 child-parent pairs, and the average age of the children is 5.63 years-old (SD = 0.72) and 63% were male. Two participants requested that their parents join in the storytelling interaction with the robot. Since our interaction design supports

† Although the Tega robot does not have eyebrows, we emulate a brow raise by enlargening its graphical eye pupils.
‡ This is necessary to ensure that combinatorial sets exhibit behaviors that all reach their individual behavioral apexes at the same time.
only dyadic interactions with young children and cannot attend to an adult over-hearer, we excluded them from our analysis.

5.3.2.2 Study Procedure

A study session averaged 30 minutes in duration and consisted of four stages: Instruction, Story Construction, Story Sharing with the first robot, and Story Sharing with the second robot. The procedure for this experiment closely mirrored the study procedures used for our data collection in Section 3.1.3.

Instruction Children were informed that they will tell a story to two different robots in back-to-back sessions. To impress the notion as being peers, Blue and Red (as shown in Figure 25) were introduced as being close to the participant’s age “but in robot years.” The robots were described as currently unable to use human words and instead use robots noises like “beep or boop” or “uh-huh.”

Story Construction The study proctor then assisted the participant in generating a story about his/her selected storybook. This mirrored the procedures used for our data collection.

Story Sharing In “Story Land” (as shown in Figure 26), the participant told his/her story to the first robot and later again to the second robot.
Participants were instructed that their important role as a storyteller is to “make sure that Red/Blue is paying attention because you want Red/Blue to understand your story.” Importantly, this emphasis helped ensure that storytellers are likely to have a goal of wanting the robot to pay attention, which is an underlying assumption of the belief manipulation model.

The proctor then left the room to help promote natural peer-interactions. After the first robot session, participants were instructed to wait outside while the proctor prepared the second robot.

5.3.2.3 Experimental Manipulations

As a within-subjects study design, the participants told the same story to two different robots. Both of the robots know *when* to respond but are different in *how* they respond. More specifically, both of the robots contingently respond to storytellers’ cues. They are only different in their employed listening strategy, which is defined by the policy module as shown in Figure 24.

The policy of the signaling robot is to always respond with a *high-positive* listening behavior. The policy of the BTOM robot follows our belief manipulation model and can suggest *high-positive* behaviors along with more subdued listening behaviors (i.e., *low-positive* and *neutral*).

To reduce potential biases related to participants’ preference in the appearance of the robot, Red and Blue were randomly assigned a listening policy for each participant. Furthermore, the order in which participants were exposed to the two different policies was randomly selected and not revealed to the study proctor.

5.3.2.4 Dependent Measures

We measured parents’ perceptions about the two listening robots. Parents were instructed to watch a live video-recording of their child interacting with the robots. Immediately after a storytelling session with either Red or Blue, they filled out a questionnaire regarding this robot’s capabilities. Our dependent measures consists of eight items to assess active listening skill, five for human-likeness, and five for intelligence.

**Perception of Active Listening Skills** To support our primary hypothesis, we measured parents’ perception about a robot’s active listening skill through a questionnaire developed by Mishima et al. [1999]. We modified their *Listening Skill* sub-scale to reflect appropriate language regarding a robot listening to their child’s story. Parents rated statements such as “The robot tended to listen to my child’s story seriously” and “When my child was hesitating, the robot give him/her a chance” on a 5-point Likert scale ranging between strongly disagree and strongly agree. The full list of questions are in Table 26 in Appendix D.

**Perception of Human-Likeness and Intelligence** To support our hypotheses regarding secondary effects, we also measured perceptions
5.3 Evaluation: Comparing Robot Listeners

5.3.1 Results

A one-tailed paired-samples t-test was conducted to evaluate the hypothesized differences between the BTOM and signaling robots (see Figure 27 for graphical results).

In support of our primary hypothesis, parents rated the BTOM robot higher in active listening skill compared to the signaling robot; \( \mu = 3.76, SD = 0.53 \) vs \( \mu = 3.24, SD = 0.69 \) with \( t(14) = 2.62, p^* = 0.01 \). Furthermore, the BTOM robot is perceived as more humanlike than the signaling robot; \( \mu = 4.11, SD = 0.69 \) vs \( \mu = 3.76, SD = 0.83 \) with \( t(14) = 2.09, p^* = 0.03 \). Although rated as slightly higher in intelligence, a significant difference was not found between the BTOM robot and the signaling robot; \( \mu = 3.80, SD = 0.88 \) vs \( \mu = 3.59, SD = 0.84 \) with \( t(14) = 0.93, p = 0.18 \).

Figure 27: Parent Perception of the Listening Robots. The BTOM robot was rated higher in both active listening skill as well as in human-likeness. It was also perceived as slightly more intelligent but not significantly so.

of human-likeness and intelligence through the Godspeed questionnaire developed by Bartneck et al. [2009]. The questionnaire captures the impressions of robots using a semantic differential scale, which indicates a person’s position on a 5-point scale between two bipolar words. Items for perceived intelligence consist of rating between bipolar words like incompetent vs competent and ignorant vs knowledgeable. Items for perceived anthropomorphism include machinelike vs humanlike and artificial vs lifelike. The full list of questions are in Table 24 and Table 25 in Appendix D.
5.4 DISCUSSION

A robotic agent capable of belief manipulation can effectively communicate a social-emotional state of attention to human partners. In reasoning about the intentional cueing actions of storytellers, robots can interpret and predict their beliefs about its attentive state. Through a mentalistic representation of how storytellers interpret listener responses, a listening agent can produce the appropriate behaviors for perceivers in order to manipulate their beliefs toward a desired inference.

A signaling agent cannot represent how its behaviors are being interpreted by the storyteller over-the-course of an interaction. Continually broadcasting a social-emotional state can be a somewhat effective but naïve means of communication. Active listening skills involve not only attentive body language but also giving storytellers space to tell their story [Robertston, 2005]. A robot in its over-aggressiveness to communicate a message of attention more frequently collides with the child’s message of his/her story. While a robot capable of understanding message acceptance supports a more fluid storytelling interaction.

LIMITATION: DITHERING & RESOLVING TIES Our current representation can lead to a situation where the agent’s policy cannot reach the goal belief-state and instead dithers back-and-forth between two surrounding belief-states that are reachable. This is a classic controls problem of oscillating around a target from overdamping and underdamping. If the agent had a larger action space (i.e., more listener responses), it can have finer control to manipulate along the belief spectrum.

Furthermore, rather than using a myopic policy, a listening agent could instead solve a long-term policy to manipulate storytellers’ beliefs. But an optimal sequence of responses implies that social cost can be associated to how listeners express attentiveness. People certainly vary in their behaviors as listeners; some are very overt while others subtle. Different situations and cultures call upon a certain set of social display rules that can govern the level of expressivity. A listening agent can encode these situational and personality-based constraints through a reward function. With this reward function and the DBN world, a longer-term policy can be solved by maximizing the expected short-term and long-term gains of listener responses leading the storyteller to a desired inference.

Also, in associating costs to listener responses, we can better determine how to resolve ties when multiple listener responses can transition to belief-states that are equally distant from the goal. Rather than using our current implementation of randomly selecting a listener response in this situation, we can instead choose the least costly action.

FUTURE WORK: SMOOTHING INFERENCES TO DETECT SOCIAL BIAS As illustrated in Section 5.2.3: Demonstrative Examples, the DBN model through retrospective inference can infer storytellers’ initial beliefs and biases about lis-
teners. Through experimental manipulation [Paepcke and Takayama, 2010], we can pre-condition participants to expect a robot to be either a good or bad listener. Then we can evaluate whether the model can accurately predict their expectations. Awareness of these preconceived notions about social robots can enable a robot to repair this user expectation during one-shot interactions or to track improvements in long-term interactions.
CONCLUSION

6.1 CONTRIBUTIONS

We claim three primary contributions of this thesis work. First, we introduce the intentional context to emotion recognition. Through a combination of human behavior analyses, human-subjects experiments, and a comparison of computational representations, we demonstrate that the accurate interpretation of emotional states depend on accounting for not only the social context of interaction partners but also adopting the intentional stance to their behaviors. We computationally model emotion recognition as an intentional inference process of agents using a classic planning algorithm of POMDPs. Although these models are typically used to provide plans to agents operating in known physical worlds, we demonstrate how to learn from human demonstrations to solve for inference policies of our social world.

Second, in formulating emotion recognition as a planning problem, we are able to apply recent probabilistic AI methods of plan inversion to reason about what people think about their partner’s social-emotional state. We computationally model emotion expression as a joint process of estimating people’s beliefs through inference inversion and planning nonverbal expressions to affect those beliefs. Through a human-subjects experiment, we demonstrate that a listening robot using our belief manipulation model can effectively communicate a social-emotional state of attention.

Lastly, we demonstrate a unified computational approach to nonverbal communication where the shared representation for both attention recognition and attention expression operate as a function of the storyteller’s beliefs, which drive both the policy of storytellers and listeners.

6.2 FUTURE WORK

6.2.1 Immediate Extensions

FUTURE WORK 1: INCREASING THE EXPRESSIVITY OF MODELS  The expressivity of the POMDP model was limited by the amount of demonstration data. In reducing our action and observation space to high-level meta-behaviors, we were able to better learn the model parameters. But we expect that the primary reason we observed narrow margins of improvement between our state estimators is because the models can only extract so much accuracy from these high-level behaviors. If we went beyond these meta-behaviors, the POMDP model could have a richer understanding in how specific cue combinations affect the listener’s attentive state and how it can also elicit specific listener responses (like a gaze-cue begets a gaze-response). The
belief manipulation model can then recommend specific sets of nonverbal behaviors to exhibit for a listening agent.

**Future Work 2: Evaluating Interaction Repair Capabilities**

A reasonable question to wonder is whether it is worth developing a complex computational approach to emotion expression. Could we have achieved the same results through a set of hand-crafted rules? The primary benefit in using machine learning algorithms is being able to learn from human demonstrations, which enable a framework to scale to different and larger domains that can outpace the rule-crafting from experts. Our evaluation focused on demonstrating our approach in a typical normal interaction. But a comprehensive system is capable of recovering from failures.

Failures in our listening domain could consist of listeners momentarily drifting off from the interaction, leading the storyteller to think they are not paying attention. Our belief manipulation model would be able to handle this “wrench” thrown into the system and immediately respond with a high-positive behavior to compensate for its unintended last behavior. However a rule-based approach can only respond to the next cueing action as dictated by its model. A rule-based method would have to cover all possible action-response sequences to be capable of repairing this interaction. But our model goes to the heart of what we are trying to repair: the storyteller’s belief.

**Future Work 3: Modeling Verbal and Nonverbal Communication**

A cornerstone of this thesis is demonstrating that social-emotional communication has the same dynamics as conversation and dialog. It similarly requires a sharing of representations and interactive alignment to understand how to convey a social-emotional message. An immediate extension of this work is to go beyond nonverbal cues and responses and integrate speech. Currently, agents capable of producing verbal and nonverbal behaviors primarily view nonverbal behaviors like deixis or beat gestures as coordinating with dialog [Cassell et al., 2001; Bevacqua et al., 2010]. However, nonverbal behaviors are not simply accompaniments to speech, they themselves are meaningful to the content of the message. In modeling both verbal and nonverbal channels, agents can better understand the communicative intent of their human partners and also effectively communicate back to them.

**Future Work 4: Modeling Other Social-Emotional States**

Our computational framework can capture other social-emotional states beyond attention. Criteria when applying this framework to other domains are:

1. an interlocutor has a social goal or desired outcome
2. there are associated social costs to their behavior
3. their behavior can elicit information or cause change to their partner’s social-emotional state.
As an example of modeling liking, let’s say Stacy wants Lenny to like her (social goal). Stacy does not want to be too obvious in professing her crush to Lenny; this might scare him off (social cost). Stacy can try different social behaviors like giving him a playful jab on the shoulder to see if he reciprocates (information-seeking actions) or be really helpful and friendly to him in hopes he grows to like her back (state-changing actions). Using a POMDP model, we can model Stacy’s policy in how she infers and influences others to like her. In inverting this model, we can recognize whenever she is trying to get someone to like her as well as predict whether she believes that she is succeeding or not given her behavior. Lenny can then modify his behavior to communicate either his interest or disinterest and manipulate her beliefs. Playing hard-to-get is in essence applying a social cost to Lenny’s behavior when creating a long-term plan to communicate his interest.

A good starting point in identifying social-emotional states that can be easily represented by our framework is investigating known call-response behaviors of various social interactions. For example, synchrony and mimicking are well-established forms of social reciprocity that are related to rapport and bonding [Bailenson and Yee, 2005]. Since mimicking is a behavioral response that communicates rapport, do mimicking-eliciting cues exist? If so, we can model this dynamic between mimickers and mimickees and how it is related to establishing and estimating rapport.

Furthermore, in marital literature, “bids” are known as acts to signal an attempt to interact with a person. Bids always imply a response, and the quality of that response is indicative of the affective quality of the interaction, especially of the emotional health of a marriage [Jung et al., 2012]. We can model this bidding behavior as a goal-direct process of trying to establish and infer the emotional connectedness of a long-term relationship.

**Future Work:** Cross-training & Online Personalization

Our computational model currently captures the general policy of storytellers and how listeners in general react. This is the foundation upon which personalized models can build upon.

To develop personalized models of different listeners, a storytelling robot can use our generalized model at first and then gradually learn a listener’s particular user-model (i.e., their unique observation and transition functions). In essence, our model serves as a prior to online learning methods like the MEDUSA algorithm that can incrementally improve the model through new interactions [Atrash and Pineau, 2010].

To learn the polices of different storytellers, a listening robot would instead learn their unique reward functions. A robot listener could intentionally weave in-and-out of being attentive or inattentive to observe the storyteller’s use of speaker cues. Updating their personalize reward function would be straightforward since the state-action pairs are fully observable.

A robot can learn how a particular child behaves as both the listener and storyteller if they take turns playing each of the roles. This cross-training of iteratively switch roles has already been demonstrated to be success-
ful in aligning the mental models of dyads in a cooperative place-and-drill task [Nikolaidis and Shah, 2013]. Robots capable of role playing or pretend play with children could be an effective method to converge on social-emotional mental models.

6.2.2 Broad Extensions

**Future Work 6: Planning and Inference in Unknown Worlds**  
The method of plan inversion in AI has had success in a wide range of domains such as navigation and manipulation. However, the application of this method has been limited to domains where a planning model is already known and defined through physical constraints or hand-coded rules. Our work demonstrates how to first learn the planning model from demonstrations to then invert for inference. This opens the opportunity to model endless situations that enable robots to reason about the everyday intentional behaviors of people.

**Future Work 7: Learning through Other Modes of Interaction**  
Our current learning paradigm is based on observing two people demonstrating the storyteller and listener roles. However this learning can be achieved in many different ways.

*(Learning from Self Experience)* Rather than learning from others, an agent can instead learn by trial and error. Much like how a navigating robot can try out all possible movement actions to reach a goal destination. A socially awkward robot can create different social plans to see how they play out in reaching a social goal. For our domain, this would consist of trying out different sequences of cues to evaluate whether the storytelling robot can get listeners to eventually pay attention. But to prevent over-taxing the interaction partner, the robot will most likely need to bootstrap its learning through other methods first.

*(Learning from Tutelage)* By learning from a teacher’s critique or shaping [Chernova and Thomaz, 2014], a storytelling robot can practice using speaker cues to get a teacher to pay attention and receive positive/negative feedback on its selected behaviors. Reinforcement learning has had success with human teachers scaffolding robots to learn cooking and navigational tasks [Thomaz and Breazeal, 2008; Chernova and Veloso, 2008; Knox et al., 2013]. But how can a teacher socially guide robots to learn not a physical task but a social-emotional skill? Can robots use social referencing [Thomaz et al., 2005] to bootstrap its learning by using the emotional reactions of teachers as a means to appraise its own social behaviors? Can positive & negative feedback come from nonverbal behaviors like smiles & frowns? Applying existing reinforcement learning techniques to the domain of social-emotional learning can be a promising area for future research.
(Learning from Triac Interactions) If we continue in our approach to learn from observing two people, then the social robot will need to model how to participate in triac interactions. In this social configuration, the robot would take the role of either a second listener or an overhearer. This will allow the robot to more naturally engage in social interactions where the parent is the storyteller and the child is the listener and vice versa. Through triac interaction, a social robot can learn by observing the remaining dyad.

6.2.3 Applications

APPLICATION SPACE 1: STORYTELLING COMPANION FOR YOUNG CHILDREN Artificial listening and storytelling agents in the past have had success in engaging young children in reading and storytelling. One of these is a program called Reading Education Assistance Dogs (R.E.A.D) where children practice reading aloud to a trained therapy dog [Jalongo, 2012]. Children generally view dogs as non-threatening and non-judgmental creatures. In creating a more comfortable and less stressful reading situation, children feel more at ease reading aloud to dogs instead of peers and teachers. The program showed promising results as two schools in a longitudinal study had students that improved their reading scores significantly.

Also, Teddy Ruxpin was a popular storytelling teddy bear in the 1980’s. By inserting an audiotape in the back, Teddy Ruxpin would move his mouth and his eyes while playing back a recorded story. Although Teddy Ruxpin was a popular educational toy, its capabilities stopped at simple playback of behaviors when reciting stories to passive listeners.

Social robots can significantly add to the interactive quality of these storytelling sessions as both an active listener and pro-active storyteller.

APPLICATION SPACE 2: LISTENING ROBOTS FOR COGNITIVE THINKING In many different situations, speaking aloud to oneself helps think through ideas, practice conversations, prepare for presentations, and engage in brainstorming. For example, in programming, “rubber duck debugging” is the method of explaining a programming problem to an inanimate object like a rubber duck which helps in discovering a solution [Hunt and Thomas, 1999]. By activating the social centers of the mind, listening robots can assist in cognitive thinking.

6.2.4 Long-term Vision

My personal goal is to develop socially intelligent robots that can help people better understand social interactions. Social interactions are messy, full of joy, rewarding, and even painful. It takes a lifetime of learning to know how to relate with others, make friends, and establish long-term relationships.

Good communication is a foundation of relationships, and a great communicator understands that people have diverse mental models. They can tailor and personalize their message in a way that allows the individual to better
understand what they are trying to convey. When meeting new and different people, we try to fit our repository of mental models for the closest fit when communicating with this novel person while also spawning new models when meeting a very unique individual. These representations help us understand people, reason about their behaviors, and predict their responses. However, people are limited in their capacity to retain a diverse set and rely on stereotypical fits that can result in poor understanding, expectation biases, and ultimately social conflicts. But robots and computers can scale, enabling them to become helpful moderators in identifying when mental models are clashing or when common ground has not been established. Although there is growing concern that artificial intelligence can acquire social biases like race and gender, I believe robots can help us better realize our diversity.

We have just scratched the surface in pushing computational models toward more human-levels of social-emotional understanding and expression. By bring in more relevant contextual information, we are getting closer, but these are only a handful of variables toward fully understanding the mental models of people. In modeling and simulating the social physics of the mind, robots will hopefully one day be able to understand our very human beliefs, intentions, and desires.
-Jin Joo Lee


Starkey Duncan and Donald Fiske. *Face-to-Face Interaction: Research, Methods, and Theory*. Routledge, 1977. (Cited on page 38.)


APPENDIX
### Speaker Cues

<table>
<thead>
<tr>
<th>Nonverbal Cue</th>
<th>Speaker (Age)</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech Pauses</td>
<td>adult</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok et al., 2013]</td>
</tr>
<tr>
<td>Clause endings</td>
<td>adult</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td>Utterance end</td>
<td>adult</td>
<td>[Ward and Tsukahara, 2000]</td>
</tr>
<tr>
<td>Long Utterance end</td>
<td>adult</td>
<td>[Gravano and Hirschberg, 2009]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td>Low-pitch region</td>
<td>adult</td>
<td>[Ward and Tsukahara, 2000]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Truong et al., 2011]</td>
</tr>
<tr>
<td>High-pitch region</td>
<td>adult</td>
<td>[Truong et al., 2011]</td>
</tr>
<tr>
<td>Higher pitch</td>
<td>adult</td>
<td>[Gravano and Hirschberg, 2009]</td>
</tr>
<tr>
<td>Pitch / Pitch slope</td>
<td>adult</td>
<td>[de Kok et al., 2013]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td>Rising intonation</td>
<td>adult</td>
<td>[Ward and Tsukahara, 2000]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Gravano and Hirschberg, 2009]</td>
</tr>
<tr>
<td>Higher speech intensity</td>
<td>adult</td>
<td>[Gravano and Hirschberg, 2009]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td>Dropping speech intensity</td>
<td>adult</td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td>Energy / Energy slope</td>
<td>adult</td>
<td>[de Kok et al., 2013]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td>Lower noise-to-harmonics</td>
<td>adult</td>
<td>[Gravano and Hirschberg, 2009]</td>
</tr>
<tr>
<td>Eye gaze</td>
<td>adult</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Hjalmarsson and Oertel, 2012]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok et al., 2013]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Truong et al., 2011]</td>
</tr>
</tbody>
</table>

Table 15: Speaker cues that were either validated or coded for in prior works.
<table>
<thead>
<tr>
<th>Nonverbal Response</th>
<th>Listener (Age)</th>
<th>Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>'um-hmm', 'yes', 'okay'</td>
<td>(2-5) &amp; (7-11) &amp; adults</td>
<td>[Miller et al., 1985]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2011]</td>
</tr>
<tr>
<td>laughter</td>
<td>(7-11)</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td>smiles</td>
<td>(2-5) &amp; (7-11) &amp; adults</td>
<td>[Miller et al., 1985]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2011]</td>
</tr>
<tr>
<td>frowns</td>
<td>(7-11)</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td>eyebrow raised / movement</td>
<td>(7-11) &amp; adults</td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2011]</td>
</tr>
<tr>
<td>eye-contact / gaze</td>
<td>(2-5) &amp; (7-11)</td>
<td>[Miller et al., 1985]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Morency et al., 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2011]</td>
</tr>
<tr>
<td>head nods / movement</td>
<td>(2-5) &amp; (7-11) &amp; adults</td>
<td>[Miller et al., 1985]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[Hess and Johnston, 1988]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2010]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[de Kok and Heylen, 2011]</td>
</tr>
</tbody>
</table>

Table 16: Listener responses that were either validated or coded for in prior works.
### Stage 1: INSTRUCTIONS

**PROCTOR 1** Hi, I’m (Proctor 1) and this is (Proctor 2).

**PROCTOR 1** Today we are here to practice telling stories and practice good listening skills.

**PROCTOR 1** Does that sound like fun? AWESOME *high five.* If you guys do a good job, at the end, I’ll give you a robot sticker.

**PROCTOR 1** Remember in class, we learned that being a good listener means . . . Do it with me! . . . body still, voices quiet, eyes watching, and ears listening.

_Accompany with corresponding listening hand-gestures_

**PROCTOR 1** Okay great! These are all the different types of stories you can tell.

_Show story scenes like a deck of cards_

**PROCTOR 1** Let’s pretend that I had to tell a story about this picture.

_Show example story scene_

**PROCTOR 1** This one let’s you tell a story about a mom, dad, and baby alien on the red planet. With these characters, I can make up a story about this picture. And here is an example of me doing that!

_Put headphones on students and play video example_

**PROCTOR 1** You both will take turns telling your own story about a different picture. There is no right or wrong story here. I just want you to have fun and tell each other whatever creative, fun, story you want to tell!

**PROCTOR 2** Let’s see what kind of story you will tell. Reach in this hat and pick only one thing inside. No peeking!

_Students pick tokens from hat_

**PROCTOR 2** ⟨Student 1⟩, you’re going to tell a story about . . .

⟨Student 2⟩, you’re going to tell a story about . . .

_Describe storybook’s characters and setting_

**PROCTOR 1** We are going to help you make up your stories. ⟨Student 1⟩, go with ⟨Proctor 2⟩, and ⟨Student 2⟩ you are with me.

---

**Table 17:** Proctor script for the Instructions stage of the data collection.
Stage 2: STORY CONSTRUCTION

PROCTOR  I’m going to help you create your own story. I’ll start you off!
          Once upon a time . . .

*Prompters below used to assist with story creation*

[Q1] What do you want the character’s name to be?
[Q2] Who do you want the main character to be?
[Q3] Tell me a bit about this character.
[Q4] Where does your story happen? Where is this?
[Q5] What is happening in this picture?
[Q6] What else do you see in this picture?
[Q7] Then what happened?
[Q8] What did [a character] do then?
[Q9] How does [a character] feel when that happened?
[Q10] Why do you think [a character] did that?

PROCTOR  Then we finish with “The End.”

PROCTOR  You told me a great story. Can you tell your classmate that story?

Table 18: Proctor script and list of questions used to elicit more details and further a child’s story.

Stage 3: STORY SHARING

PROCTOR  Let’s see who will tell their story first. (Student 1) are heads and (Student 2) are tails.

*Flip coin*

PROCTOR  [Storyteller’s name] you are going to tell your story first so go sit on the Storyteller Chair.

PROCTOR  [Listener’s name] you are the listener so sit on the Listener Chair.

PROCTOR  Next turn you will switch seats!

*Place storybook on book stand*

PROCTOR  Leave this here and don’t play with it so you both can see.

PROCTOR  Your important job as the Listener is to use your good listening skills and pay attention because later I’m going to ask you some questions about his/her story.

PROCTOR  Your important job as the Storyteller is to be like a teacher. You want to make sure your classmate is paying attention and understands your story.

PROCTOR  Does anyone have any question?

*Answer questions*

PROCTOR  Great! [Storyteller’s name] start telling your story whenever you are ready. I’ll be right outside.

Table 19: Proctor script for the Story Sharing stage of the data collection.
Stage 4: RECALL & QUESTIONNAIRE

**PROCTOR**
Okay, I want you to tell me the story your classmate just told you. Its okay if you don’t remember the whole thing. Just try your best!

*Turn on audio recorder*

**PROCTOR**
Great! Okay, now I have some questions about the storytelling we just did. I want you to try to be honest with your answers. Your answers are a secret. I will be the only one that knows. I promise I won’t tell anyone. And no one is going to get into trouble.

**PROCTOR**
Remember when it was your turn to sit on the storyteller chair and tell your story?

**PROCTOR**
Q1: How was ⟨ Student ⟩ at paying attention to your story?

**PROCTOR**
Was he/she Bad?

*Point to corresponding smiley face*

**PROCTOR**
Was he/she Little Bad?

*Point to corresponding smiley face*

**PROCTOR**
Was he/she Good?

*Point to corresponding smiley face*

**PROCTOR**
Was he/she Really Good?

*Point to corresponding smiley face*

**PROCTOR**
Or Great?

*Point to corresponding smiley face*

**PROCTOR**
Record which smiley face was pointed to

*Repeat for the following questions*

**PROCTOR**
Q2: How was ⟨ Student ⟩ at understanding your story?

**PROCTOR**
Remember when it was your turn to sit on the listener chair and listen to ⟨ Student’s ⟩ story?

**PROCTOR**
Q3: How was ⟨ Student ⟩ at making sure you were paying attention to his/her story?

**PROCTOR**
Q4: How was ⟨ Student ⟩ at making sure you understood his/her story?

Table 20: Proctor script for the Recall and Questionnaire stages of the data collection.
Figure 28: Eight Storybooks. Each scene introduces new characters or events. The illustrated storybooks are a modification of a prior tablet-based storytelling application [Kory, 2014].
**Trend Analyses of Human Inference**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Binomial</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link</strong></td>
<td>Logit</td>
</tr>
</tbody>
</table>

**Model**

\[ y_{im} = \beta_{0m} + \beta_{1m}X_{im} + \epsilon_{im} \quad i = 1, 2, \ldots, N = 1383 \quad m = 1, 2, \ldots, M = 461 \]

\[ \beta_{0m} = \beta_{00} + b_{0m} \]

\[ \beta_{1m} = \beta_{10} + b_{1m} \]

<table>
<thead>
<tr>
<th>Variables</th>
<th>Param. Est.</th>
<th>Std. Err.</th>
<th>95% conf. interval</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.23</td>
<td>0.06</td>
<td>0.13</td>
<td>0.34</td>
<td>4.25</td>
</tr>
<tr>
<td>Speaker Context</td>
<td>0.15</td>
<td>0.08</td>
<td>0.01</td>
<td>0.30</td>
<td>2.04</td>
</tr>
</tbody>
</table>

Table 21: The multilevel logistic regression model to predict whether a correct or incorrect inference would be made based on storyteller-context treatment. The model includes a random intercept and slope that both vary independently at the participant-level to reflect the bias and sensitivity, respectively, of a participant’s inference ability.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link</strong></td>
<td>Log</td>
</tr>
</tbody>
</table>

**Model**

\[ y_{im} = \beta_{0m} + \beta_{1m}X_{1im} + \beta_{2m}X_{2im} + \epsilon_{im} \quad i = 1, 2, \ldots, N = 770 \quad m = 1, 2, \ldots, M = 452 \]

\[ \beta_{0m} = \beta_{00} + b_{0m} \]

\[ \beta_{1m} = \beta_{10} + b_{1m} \]

\[ \beta_{2m} = \beta_{20} + b_{2m} \]

<table>
<thead>
<tr>
<th>Variables</th>
<th>Param. Est.</th>
<th>Std. Err.</th>
<th>95% conf. interval</th>
<th>t-stat</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.43</td>
<td>0.04</td>
<td>4.35</td>
<td>4.51</td>
<td>108.50</td>
</tr>
<tr>
<td>ABSENT (x1)</td>
<td>0.01</td>
<td>0.06</td>
<td>-0.11</td>
<td>0.13</td>
<td>0.23</td>
</tr>
<tr>
<td>FALSE (x2)</td>
<td>0.13</td>
<td>0.06</td>
<td>0.01</td>
<td>0.24</td>
<td>2.21</td>
</tr>
</tbody>
</table>

Table 22: The multilevel gamma regression model to estimate latency of response based on storyteller-context treatment. Similar to the previous multilevel model (Table 21), within-subject dependencies are accounted for through participant-level variability in both the intercept and slope. But unlike the previous model, the predictor is represented as categorical values through two variables that dummy code the three conditions (with the TRUE condition as the reference category).
SUPPLEMENTARY MATERIALS TO INTENTIONAL INFERENCE

REWARD FUNCTION SELECTION

In our work, we assume that the parameters related to policy formation are known. Given a POMDP world based on the MLE of variables from our dataset, we hand-selected the $(R, \gamma, \beta)$ parameters by finding a set of values that formed a POMDP policy that best reflected our expectations of how storytellers act. Our selected set, with the reward function in Equation 11, discount factor of 0.1, and a $\beta$ of 0.8, yielded the policy shown in Figure 29.

$$R(s, a) = \begin{cases} R(L, \text{gaze}) = -0.5 & R(L, \text{gaze}) = -3.0 \\ R(L, \text{weak}) = -1.5 & R(L, \text{weak}) = -1.5 \\ R(L, \text{strg}) = -4.0 & R(L, \text{strg}) = -3.0 \\ R(L, \text{pred}) = +35 & R(L, \text{pred}) = -100 \end{cases}$$

(11)

Figure 29: Expected Policy of Storytellers. The x-axis is the storyteller’s beliefs, where 0.0 is a high-confident belief that a listener is not paying attention, 1.0 is a high-confident belief that the listener is paying attention, and the middle region represents varying degrees of certainty. The y-axis is the probability of selecting a cueing action. We expect that the value of a strong cue is to influence inattentive listeners. Storytellers will most likely select this cue when they believe the listener is not paying attention. Also, we expect that the value of a gaze cue is to “check in” on listeners but not try to change their state. Storytellers will most likely use this cue when they believe the listener is attentive. And finally, we expect that the value of a weak cue is to help disambiguate listeners’ state in moments of high uncertainty (i.e., the middle region).
The POMDP model, from which we derived our belief manipulation model components, achieved the highest AUC measure from Section 4.5.5: Inference Results. The figures below illustrate its model parameters and resulting policy when trained using the full dataset.

Figure 30: POMDP Policy $\Pi^S: b \rightarrow a$. The likelihood of selecting a strong cue is higher when storytellers have a belief of an inattentive listener, which is congruent with our expectation that the value of a strong cue is to influence inattentive listeners. The likelihood of selecting a weak cue is mostly uniform across all beliefs. Gaze cues are more likely when storytellers are confident that the listener is attentive. This is congruent with our expectation that the value of a gaze cue is to not change state but instead to gather information. Note: The final resulting policy is different from our initial policy since it depends on the learned POMDP world dynamics.
(a) State transitions from a gaze cue.

(b) State transitions from a weak cue.

(c) State transitions from a strong cue.

Figure 31: **Transition Function.** The gaze cue does not have much effect on changing the attentive state of listeners since the likelihood of transitioning out of states are very small. The weak cue is more likely to transition inattentive listeners to pay attention, while strong cues have the highest chance of transitioning them.
(a) attentive listener (L).

(b) inattentive listener (Ł).

Figure 32: **Observation Function.** The x-axis is the overall valence of listener response where $-2$ is a high-negative response and $+2$ is a high-positive response. The y-axis is the probability of observation. *(top)* When a listener is attentive and the storyteller does a gaze cue, there is a small but non-zero probability of observing negative responses, but it is more likely to observe neutral or positive responses. This almost has a uniform distribution from 0 to $+2$, which further indicates how a gaze cue is an information-seeking action. For a weak cue, the overall distribution shifts towards the positive end. For a strong cue, an attentive listener will most likely response in a very positive way. *(bottom)* When the state of the listener is inattentive, we will most likely observe very negative behaviors. This is collapsed across all cues because of the simplifying assumption we made in Sidenote 2.
SUPPLEMENTARY MATERIALS TO BELIEF MANIPULATION

SANITY CHECK FOR DBN SETUP

As a validation method to check that the DBN setup is outputting informative estimates (sanity check), we use our inference results from our POMDP model to compare against the DBN’s inferences. The inferences of the POMDP, in essence, represent storytellers’ beliefs about listeners of our dataset. As shown in Table 23, overall the DBN with an initial belief-state that matches our POMDP model (i.e., biased model) obtains the lowest prediction error. The uniform model is at a close second. Both of these DBN models outperform a random model, which achieves our sanity check. Another observation is that when increasing the resolution of belief-states, we level-off on the prediction gains. Our model’s ability to express finer differences between two belief-states is limited by our small action and observation space.

<table>
<thead>
<tr>
<th>Model</th>
<th>Interval</th>
<th>Correlation</th>
<th>RSME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.25</td>
<td>r(590) = 0.01, p = 0.79</td>
<td>0.53</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.25</td>
<td>r(590) = 0.56, p* = 2.41e^-49</td>
<td>0.26</td>
</tr>
<tr>
<td>Biased</td>
<td>0.25</td>
<td>r(590) = 0.60, p* = 2.52 e^-59</td>
<td>0.24</td>
</tr>
<tr>
<td>Random</td>
<td>0.1</td>
<td>r(590) = 0.01, p = 0.81</td>
<td>0.49</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.1</td>
<td>r(590) = 0.57, p* = 2.53e^-52</td>
<td>0.25</td>
</tr>
<tr>
<td>Biased</td>
<td>0.1</td>
<td>r(590) = 0.61, p* = 1.00e^-61</td>
<td>0.23</td>
</tr>
<tr>
<td>Random</td>
<td>0.05</td>
<td>r(590) = 0.03, p = 0.53</td>
<td>0.48</td>
</tr>
<tr>
<td>Uniform</td>
<td>0.05</td>
<td>r(590) = 0.57, p* = 5.01e^-52</td>
<td>0.25</td>
</tr>
<tr>
<td>Biased</td>
<td>0.05</td>
<td>r(590) = 0.61, p* = 2.35e^-62</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 23: The ground truth is the non-thresholded filtering inferences generated by our POMDP model. The correlation is the Pearson’s correlation coefficient along with its significance value. The root mean squared error (RMSE) represents the prediction error of the DBN; lower is better. The different models vary in their resolution of belief-states. Their initial distribution over belief-states also vary as either a uniform distribution or as a biased normal distribution with an expected value set to our POMDP’s b0. A random model selects a random belief-state. Note: The 590 is the total number of data instances/tuples (i.e., 60 demonstrations with an average 10 in length ~ 600 data instances).
## Robot Comparison Study Materials

### Humanlikeness Questionnaire

Please rate your impression of the robot on these scales:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Natural</th>
<th>Natural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fake</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Natural</td>
</tr>
<tr>
<td>Machinelike</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Humanlike</td>
</tr>
<tr>
<td>Unconscious</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Conscious</td>
</tr>
<tr>
<td>Artificial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Lifelike</td>
</tr>
<tr>
<td>Moving rigidly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Moving Elegantly</td>
</tr>
</tbody>
</table>

Table 24: Items for perceived human-likeness from the *Godspeed* questionnaire [Bartneck et al., 2009].

### Intelligence Questionnaire

Please rate your impression of the robot on these scales:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Competent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Incompetent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Competent</td>
</tr>
<tr>
<td>Ignorant</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Knowledgeable</td>
</tr>
<tr>
<td>Irresponsible</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Responsible</td>
</tr>
<tr>
<td>Unintelligent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Intelligent</td>
</tr>
<tr>
<td>Foolish</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sensible</td>
</tr>
</tbody>
</table>

Table 25: Items for perceived intelligence from the *Godspeed* questionnaire [Bartneck et al., 2009].
Active Listening Skill Questionnaire

Please answer to the degree in which you disagree/agree to the following statements:

<table>
<thead>
<tr>
<th>Questions</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) The robot payed attention to my child’s unexpressed feelings.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>2) The robot tended to listen to my child’s story seriously.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>3) In the robot’s mind, the robot was summarizing what my child was saying.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>4) The robot listened to my child by being in his/her shoes.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>5) When my child was hesitating, the robot give him/her a chance.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>6) The robot listened to my child calmly, while he/she was speaking.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>7) The robot payed attention to the changes of my child’s feelings.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
<tr>
<td>8) The robot listened to my child absent-mindedly.</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
<td>□</td>
</tr>
</tbody>
</table>

Table 26: A modified version of the Listening Skill sub-scale from the Active Listening Skill questionnaire [Mishima et al., 1999].
Fin