SOCIAL ROBOTICS

Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders

Shomik Jain*, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, Maja J. Matarić*

Socially assistive robotics (SAR) has great potential to provide accessible, affordable, and personalized therapeutic interventions for children with autism spectrum disorders (ASD). However, human-robot interaction (HRI) methods are still limited in their ability to autonomously recognize and respond to behavioral cues, especially in atypical users and everyday settings. This work applies supervised machine-learning algorithms to model user engagement in the context of long-term, in-home SAR interventions for children with ASD. Specifically, we present two types of engagement models for each user: (i) generalized models trained on data from different users and (ii) individualized models trained on an early subset of the user's data. The models achieved about 90% accuracy (AUROC) for post hoc binary classification of engagement, despite the high variance in data observed across users, sessions, and engagement states. Moreover, temporal patterns in model predictions could be used to reliably initiate reengagement actions at appropriate times. These results validate the feasibility and challenges of recognition and response to user disengagement in long-term, real-world HRI settings. The contributions of this work also inform the design of engaging and personalized HRI, especially for the ASD community.

INTRODUCTION

Socially assistive robotics (SAR) is a promising new subfield of human-robot interaction (HRI), with a focus on developing intelligent robots that provide assistance through social interaction (*1*, *2*). As overviewed in this journal, researchers have been exploring SAR as a means of providing accessible, affordable, and personalized interventions to complement human care (*3*). However, HRI methods are still limited in their ability to autonomously perceive, interpret, and naturally respond to behavioral cues from atypical users in everyday contexts. This hinders the ability of SAR interventions to be tailored toward the specific needs of each user (*4*, *5*).

These HRI challenges are not only amplified in the context of SAR for individuals with autism spectrum disorders (ASD), but ASD is also the context where SAR is especially promising. ASD is a developmental disability characterized by difficulties in social communication and interaction. About 1 in 160 children worldwide are diagnosed with ASD (*6*), with a higher rate of 1 in 59 children in the United States (*7*). Therapists offer individualized services for helping children with ASD to develop social skills through games or storytelling, but such services are not universally accessible or affordable (*8*). To this end, researchers have been actively exploring SAR for children with ASD (*9*). Several short-term studies have already shown SAR to support learning in ASD users (*10*). Moreover, in this journal, Scassellati *et al.* (*11*) reported on a long-term, in-home SAR intervention that helped children with ASD improve social skills such as perspective taking and joint attention with adults.

SAR systems must engage users to be effective. Robot perception of user engagement is a key HRI capability that makes it possible for the robot to achieve the specified goals of the interaction and, in the SAR context, the intervention (*12*). Past work has used rule-based methods to approximate the engagement of children with ASD. For instance, Kim *et al.* (*13*) indirectly assessed engagement by estimating emotional states from audio data. Esteban *et al.* (*12*) gauged engagement by using the frequency of measured variables, such as how many times the child looked at the robot. More recently in this journal, Rudovic *et al.* (*4*) used supervised machine learning (ML) to model engagement in a single-session laboratory study. The study focused on developing post hoc personalized models using deep neural networks and achieved an average agreement of 60% with human annotations of engagement on a continuous scale from −1 to +1.

This article addresses the feasibility and challenges of applying supervised ML methods to model user engagement in long-term, real-world HRI with a focus on SAR interventions for children with ASD. The contributions of this work differ from past work in two key aspects.

First, the methods and results are based on data from month-long, in-home SAR interventions with seven child participants with ASD. Whereas single-session and short-term studies of SAR for ASD are numerous (*14*), the work by Scassellati *et al.* (*11*) in this journal is the only other long-term, in-home SAR for ASD study conducted to date. Long-term, in-home studies and data collections are important for many reasons: They more realistically represent real-world learning environments, they provide more opportunities for the user to learn and interact with the robot, and they produce more relevant training datasets (*15*). Furthermore, long-term, in-home settings present new modeling challenges, given the significantly larger quantity and variance in user data.

Second, this work emphasizes engagement models that are practical for online use in HRI. With a supervised ML approach, models require labeled data for training, which are often expensive or unfeasible to obtain. Previous works reported on models trained and tested on randomly sampled subsets of each participant's data (*4*). However, that approach is impractical for online use if labeled training data for a given user are obtained chronologically after the testing data. In contrast, this work presents, for each user: (i) generalized models trained on data from different users and (ii) individualized models trained on an early subset of the user's data. As detailed in

Interaction Lab, University of Southern California, Los Angeles, CA 90089, USA.
*Corresponding author. Email: shomikja@usc.edu (S.J.); mataric@usc.edu (M.J.M.)

Materials and Methods, models were developed for different numbers of training users in generalized models and varying sizes of early subsets in individualized models. An early subset of the data is defined as the first *X*% of a user's data sorted chronologically. Furthermore, this work also analyzes the temporal structure of model predictions to examine the possibility of initiating reengagement actions (RA) at appropriate times.

The presented engagement models were trained on data from month-long, in-home SAR interventions. During interventions, child participants with ASD played space-themed math games on a touchscreen tablet, while a Stewart platform robot named Kiwi provided verbal and expressive feedback (*16*). The robot's feedback and instruction challenge levels were personalized to each user's unique learning patterns with reinforcement learning over the month-long intervention. All participants showed improvements in reasoning skills and long-term retention of intervention content (*17*, *18*). Over the month-long intervention, we collected an average of 3 hours of multimodal data across multiple sessions for each participant, including video, audio, and performance on the games. As Fig. 1 shows, a USB (universal serial bus) camera mounted on top of the game tablet recorded a front view of the user. Visual and audio features were extracted from the camera data, and performance features were derived from the answers to game questions recorded on the tablet. As detailed in Materials and Methods, the open-source data processing tools used to extract these features are appropriate for online use in HRI contexts (*19*–*21*).

This work frames engagement modeling as a binary classification problem, similar to most previous relevant works (*4*). Participants were annotated as engaged or disengaged in each camera frame using standard definitions of engagement as a combination of behavioral, affective, and cognitive constructs (*9*). A participant was considered to be engaged when paying full attention to the interaction, immediately responding to the robot's prompts, or seeking further guidance from others in the room. The binary labels simplify the representation of participants' behavior, which may vary in degree of engagement and disengagement (*22*). However, temporal patterns in binary labels can provide additional context (*17*); therefore, this work also analyzed the length of time that a participant was continuously engaged and disengaged. Trained annotators labeled engagement, and an interrater reliability of $k = 0.84$ (Fleiss' kappa) was achieved. The Materials and Methods provides additional details about the data and the annotation process.

Fig. 1. Long-term, real-world SAR intervention setup. In this month-long, in-home study, child participants with ASD played math games on a touchscreen tablet, while a socially assistive robot used multimodal data to provide personalized feedback and instruction (*37*).

This article focuses on post hoc models of user engagement based on data from month-long, in-home SAR interventions. The presented approaches are suitable for online perception of engagement and are intended to inform the design of more engaging and personalized HRI. The contributions of this work especially aim to improve SAR's effectiveness in supporting learning by children with ASD.

RESULTS

This work presents two types of supervised ML models of user engagement in long-term, real-world HRI intended for online implementation: (i) generalized models trained on data from different users and (ii) individualized models trained on data from early subsets of the users' interventions. On average, these models achieved area under the receiver operating characteristic (AUROC) values of about 90%. AUROC is a commonly used ML metric for binary classification problems; specifically, it measures the probability that the models would rank a randomly chosen engaged instance higher than a randomly chosen disengaged instance (*23*). To evaluate these two approaches, we also implemented models trained on random samples of all user data. Random sampling yielded significantly higher recall for disengagement compared with generalized and individualized models. This is likely because the month-long, in-home setting led to a large variance in both engagement states and recorded data. Variance in data manifested not only across participants but also within each participant, highlighting an important characteristic of real-world HRI in the ASD context. Despite the lower recall and higher variance for disengagement, temporal patterns in model predictions can be used to reliably initiate RA at appropriate times.

Observed user engagement

Over the course of the month-long, in-home intervention, participants were engaged an average of 65% of the time during the child-robot interactions. However, engagement varied considerably across participants and for each participant, as shown in Fig. 2. Average engagement for participants ranged from 48 to 84%, with an SD of 14%. Analyzing each participant's engagement chronologically over 10% increments also showed an SD of 15%. Moreover, all participants had a significant ($P < 0.01$) decrease in engagement over the month-long intervention, as determined by a regression *t* test and shown by the plotted trend line. For example, participant 2 was engaged 82% of the time in the first 10% and only 19% of the time in the last 10% of the month-long intervention.

This substantial variance in user engagement over the course of a long-term, real-world study indicates the need for online recognition of and response to disengagement. This study observed higher engagement for all participants shortly after the robot had spoken. Specifically, participants were engaged about 70% of the time when the robot had spoken in the previous minute but less than 50% of

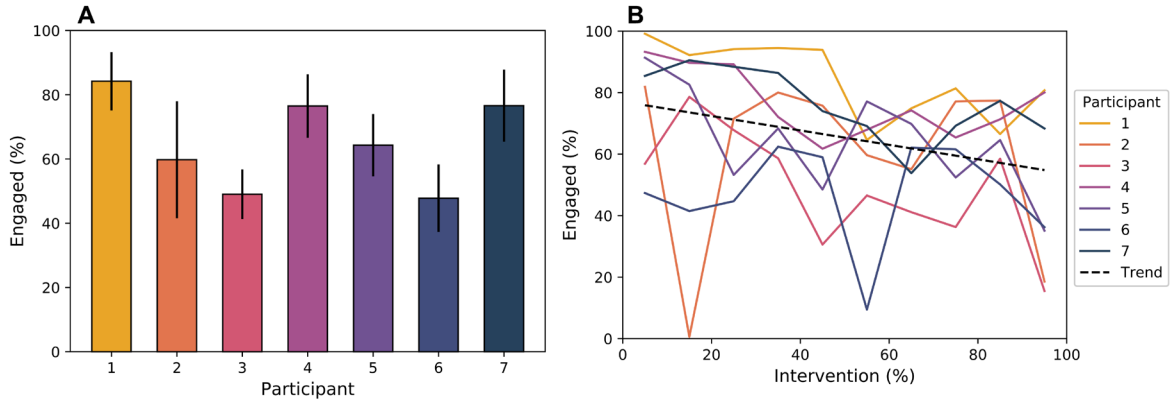

Fig. 2. Engagement by participant. This study observed a significant variance in engagement across participants (error bars, SD) (**A**) and for each participant (**B**). A decreasing trend ($P < 0.01$) in engagement was also observed over the month-long intervention (B), indicating the need for online engagement recognition and response.

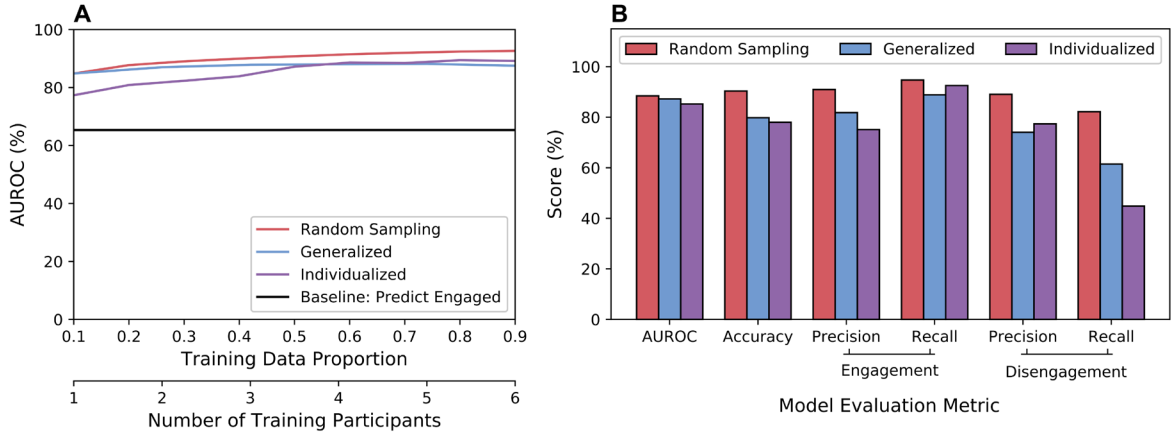

Fig. 3. Model results. Generalized models trained on different users and individualized models trained on early subsets of the intervention achieved comparable AUROC values with those of models trained on random samples of all users' data (**A**) but had much lower recall for disengagement (**B**).

the time when the robot had not spoken for over a minute. This validates the use of appropriately timed robot speech as a tool for eliciting and maintaining user engagement.

Generalized and individualized model results

This work presents generalized and individualized models of user engagement using data from long-term, in-home SAR interventions. As detailed in Materials and Methods, generalized models were developed by training on data from a given subset of users and then testing on different users. Individualized models were developed by sorting a user's data chronologically and using an early subset for training and a later subset for testing the model. We designed these two approaches to be feasible for online use in HRI; the labeled data required for supervised ML models are practical to obtain in both cases. To evaluate these approaches, we also implemented models trained on random samples of all users' data, despite random sampling not being feasible for online use.

As shown in Fig. 3A, generalized and individualized models achieved about 90% AUROC. For generalized models, the number of training users had little effect on AUROC; models trained on six users resulted in only a 3% improvement in AUROC over models trained on one user. On the other hand, individualized models performed better with additional data; models trained on the first 10% of a user's data only achieved 77% AUROC, whereas models trained on the first 50% of a user's data achieved 87% AUROC. Overall, both generalized and individualized models achieved AUROC values comparable with those of models trained on random samples, which obtained 90% AUROC by training on as little as 30% of data across all users.

However, generalized and individualized models differed from models trained on random samples when considering other ML evaluation metrics such as precision and recall. For a given class (engagement or disengagement), precision measures the proportion of predictions of the class that are correct, and recall measures the proportion of actual instances of the class that are predicted correctly. As Fig. 3B shows, there is an especially large difference between models in recall for disengagement. On average, training on random samples resulted in 82% recall for disengagement, whereas training on different users and early subsets resulted in only 61 and 45% recall, respectively. This indicates that generalized and individualized models would produce a high number of false negatives for detecting disengagement if implemented online in HRI. Tables S4 to S6 contain detailed model results for all approaches and evaluation metrics.

Variance in data across users, sessions, and engagement states

This work's long-term, real-world setting resulted in significantly different means and variances of data across participants, sessions,

and engagement states, as shown in Fig. 4. The figure compresses recorded data with high face-detection confidence to two dimensions using principal components analysis (PCA), a commonly used unsupervised dimensionality reduction technique. Plotting compressed data reveals limited overlap between two participants (Fig. 4A) and two sessions from the same participant (Fig. 4B). In addition, Fig. 4C shows a higher variance in data when participants are disengaged, which may explain the low recall values for disengagement reported in the previous section. Figures S8 and S9 show similar visualizations for all participants and all sessions for the same participant in Fig. 4B.

Statistical analysis confirmed that both the means and variances of features differed significantly ($P < 0.01$) across participants, sessions, and engagement states. We used a one-way analysis of variance (ANOVA) to test differences in means, and an F test for differences in variance. Tests were performed on the principal components of all data.

Detecting disengagement sequences using temporal patterns

This study demonstrates the importance and feasibility of detecting longer sequences of disengagement using the temporal structure of model predictions. Engagement sequences (ES) are periods in the interaction when the user was continuously engaged, whereas disengagement sequences (DS) are periods in the interaction when the user was continuously disengaged. Fig. 5 shows that the duration of ES had an interquartile range of 5.0 to 27.0 s, whereas the duration of DS had an interquartile range of 2.5 to 9.5 s. This work defines long DS as having a duration greater than the upper quartile (9.5 s) and short DS as having a duration less than the lower quartile (2.5 s). Long DS accounted for 75% of the total time that users were disengaged, whereas short DS accounted for only 5% of the total time that users were disengaged. This suggests that reengagement strategies should focus on long DS despite the presence of many shorter sequences.

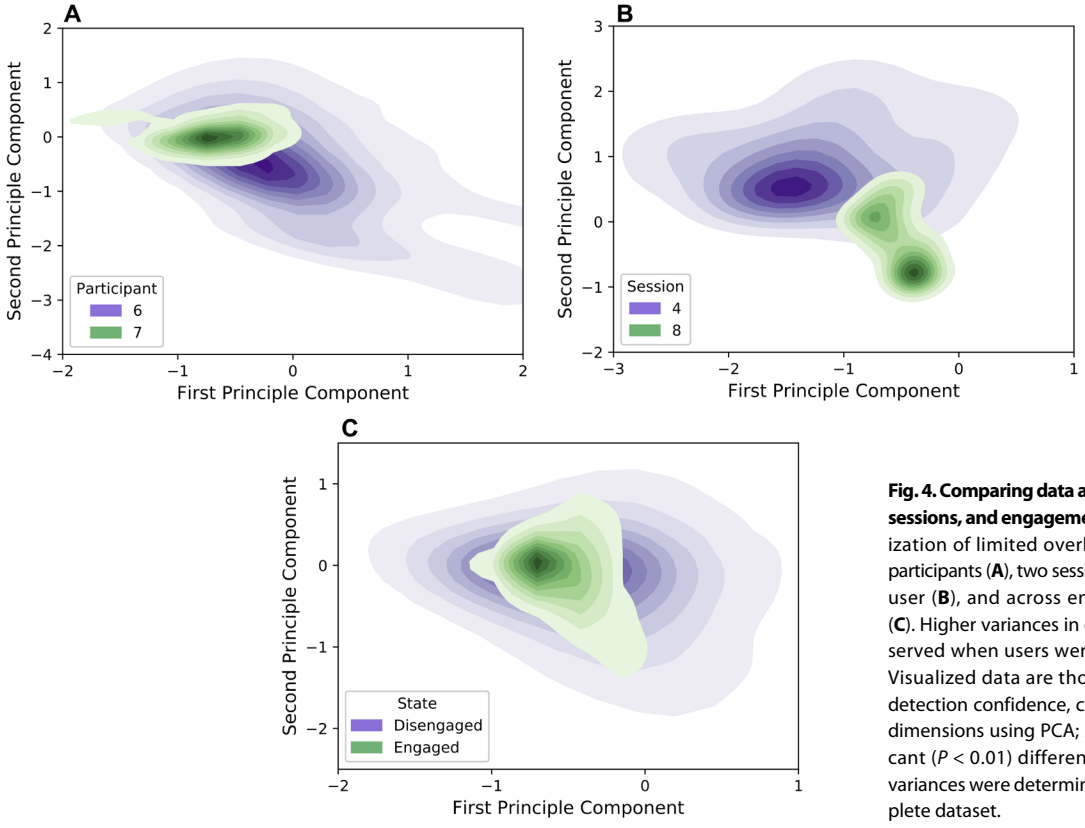

Fig. 4. Comparing data across participants, sessions, and engagement states. A visualization of limited overlap in data of two participants (**A**), two sessions from the same user (**B**), and across engagement states (**C**). Higher variances in data were also observed when users were disengaged (C). Visualized data are those with high face detection confidence, compressed to two dimensions using PCA; statistically significant ($P < 0.01$) differences in means and variances were determined using the complete dataset.

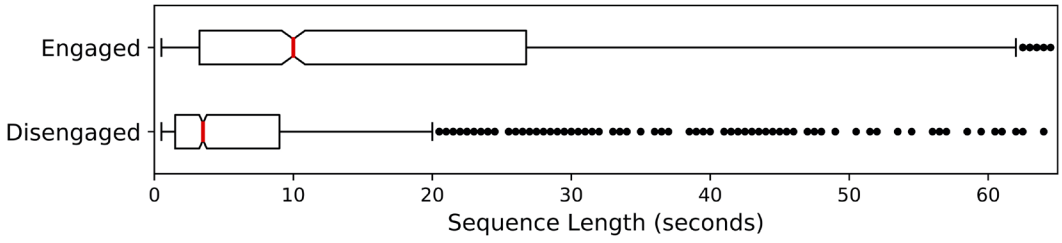

Fig. 5. Sequences of engagement and disengagement. Median duration of ES (11.0 s) was significantly ($P < 0.01$) higher than the median duration of DS (4.0 s). Despite the prevalence of shorter sequences, DS longer than the upper quartile (9.5 s) accounted for 75% of the total time users were disengaged.

The results and insights from the data suggest the following strategy for determining when to initiate RA: (i) average a model's predicted probability of engagement over a given window and then (ii) initiate RA if the engagement probability is less than a given threshold. This approach should maximize long DS with RA and minimize the percentage of ES with RA. Other considerations include the percentage of short DS with RA, the median duration of DS with RA, and the median elapsed time in DS before RA. The window length and threshold will affect these evaluation metrics, so the choices for these parameters should depend on the intervention design and implemented RA.

Figure 6 presents a post hoc analysis of the proposed reengagement strategy. For example, suppose this study initiated RA if the predicted engagement probability was less than 0.35 on average for a 3-s window. This approach would have led to RA in 73% of long DS. The median duration of DS with RA would have been 25 s, and RA would have occurred 2.5 s into these sequences. However, RA would also have occurred in 5% of short DS and 15% of ES.

Varying the window lengths and thresholds highlights the trade-off between maximizing RA in DS and minimizing RA in ES. The window length was negatively correlated with the percentage of long DS ($r_s = -0.74$) and ES ($r_s = -0.88$) with RA for a fixed threshold, as shown in Fig. 6A. On the other hand, the threshold was positively correlated with the percentage of long DS ($r_s = +1.00$) and ES ($r_s = +1.00$) with RA for a fixed window length, as shown in Fig. 6B. The reported results are based on generalized models trained on six users. Tables S7 and S8 contain results for both generalized and individualized models with additional window length and threshold combinations.

Different modalities and model types

Over the month-long, in-home SAR interventions, we collected a rich multimodal dataset from which we derived visual, audio, and game performance features to model engagement. To assess each modality's importance, we created separate models using each feature group. As Fig. 7A shows, all modalities together outperformed each individual modality. However, models created using only visual features outperformed those created using audio or game performance features by about 20% AUROC. These results support related work in this journal that also found visual features as the most significant but multiple modalities as complimentary (*4*).

Moreover, analyzing individual features revealed that the results of this work could largely be replicated using only seven key features. Feature analysis was performed using Pearson's correlation coefficient (r), and key features were determined using a threshold of $|r| > 0.20$. The key features are the elapsed time in a session, the number of people in the environment, the direction of the user's eye gaze, the distance from the camera to the user, the elapsed time since the robot last talked, the count of incorrect responses to game questions, and the confidence value with which the user's face is being detected in the camera frame. Models using only these seven key features achieved AUROC values within 5% of the results described above.

In addition, this work explored several supervised ML model types but found tree-based models to be the most successful. The following conventional model types were considered: naïve Bayes, *K*-nearest neighbors, support vector machines, neural networks, logistic regression, random decision tree forests, and gradient boosted decision trees. Of these, gradient boosted decision trees had the highest AUROC, as shown in Fig. 7B, and are the basis for model evaluation metrics reported in previous sections. We also explored sequential models—such as hidden Markov models, conditional random fields, and recurrent neural networks—but found these to be less effective than conventional static models.

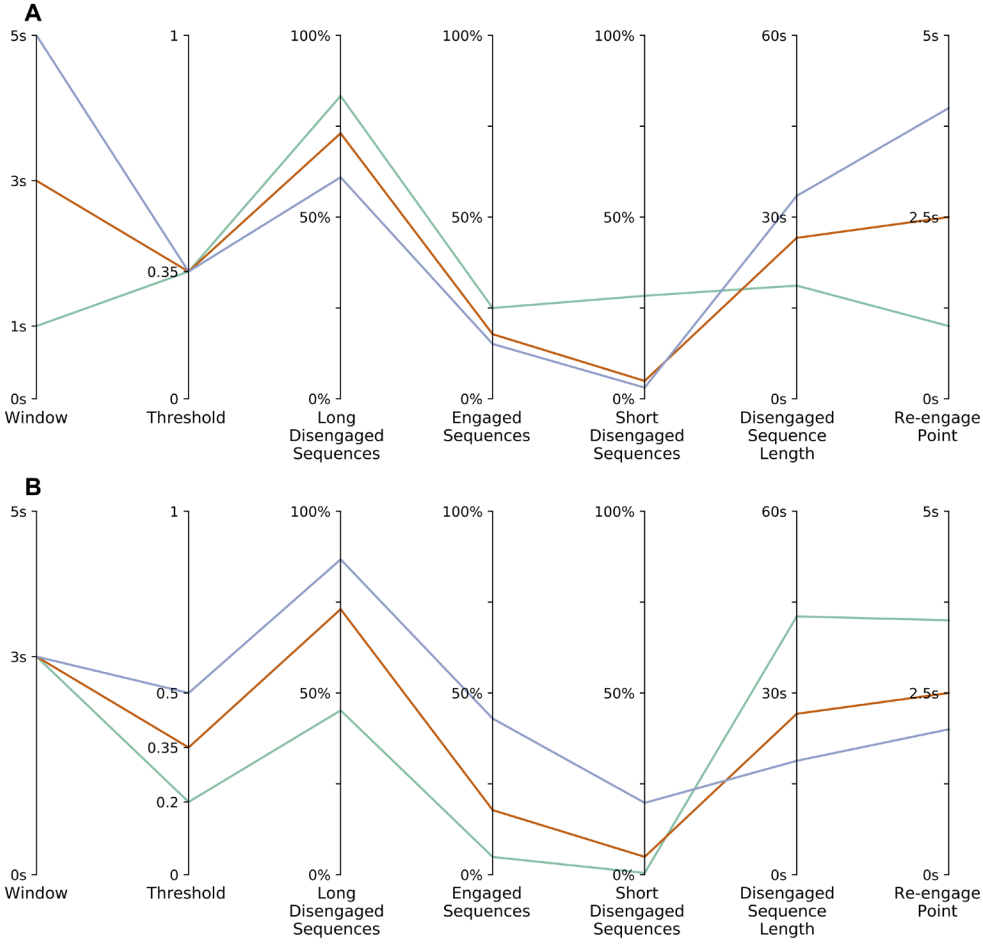

Fig. 6. Reengagement strategy results. Post hoc analysis of strategy to reengage users if predicted engagement probability was less than a threshold on average over a window. As shown by the orange line in (**A**) and (**B**), a threshold of 0.35 and window of 3 s would have reengaged users in not only 73% of long DS but also 15% of ES. Varying window lengths (A) and thresholds (B) illustrates the trade-off between maximizing reengagement in disengaged sequences and minimizing reengagement in engaged sequences. Results based on generalized models trained on six participants.

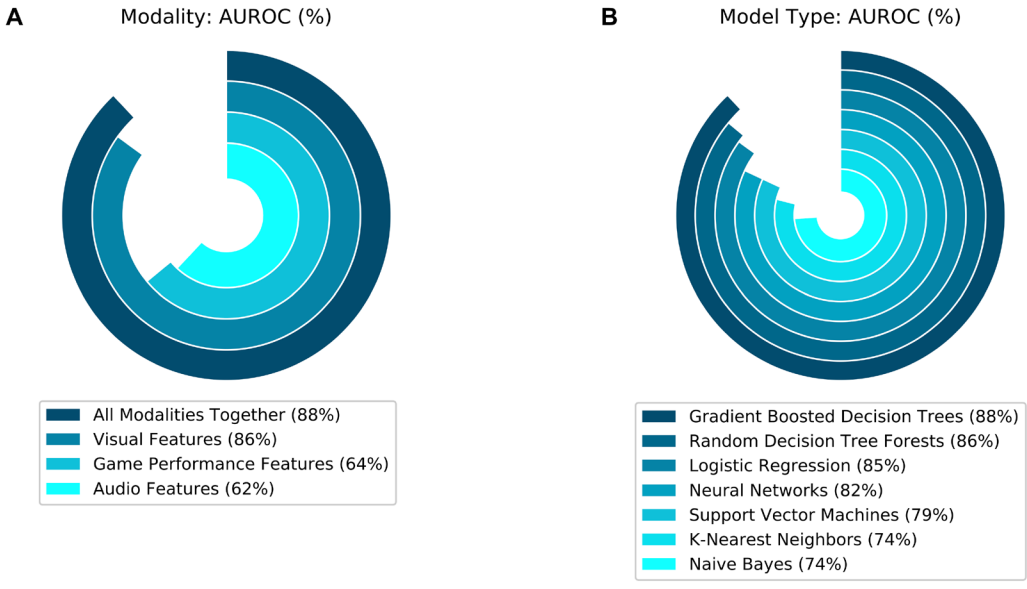

Fig. 7. Results across different modalities and model types. All modalities together outperformed each modality separately, but visual features were the most significant (**A**). Tree-based models were the most successful among conventional supervised ML model types (**B**).

A few alternative modeling approaches were considered as well. An ensemble of generalized and individualized models did not lead to better results than those approaches applied separately. Other explored approaches included (i) rule-based models with the key features, (ii) deep neural networks with reweighting techniques (*24*), and (iii) synthetic oversampling of the disengaged class (*25*). None of these approaches outperformed gradient boosted decision trees.

DISCUSSION

Robot perception of user engagement is a crucial HRI capability previously unexplored in the context of long-term, in-home SAR interventions for children with ASD. This study also differs from previous work by developing supervised ML models intended for real-world deployments. The discussion below focuses on how this work highlights the feasibility and challenges of online recognition and response to disengagement in long-term, in-home SAR contexts. These contributions aim to inform the design of more engaging and personalized HRI and improve SAR's effectiveness in augmenting learning by children with ASD.

Feasibility of online closed-loop implementation

This study presents supervised ML models that are feasible for use in online robot perception of and closed-loop response to disengagement. We developed two types of models for each user: (i) generalized models trained on data from different users and (ii) individualized models trained on data from early subsets of users' interventions. The visual, audio, and performance features along with the labeled training data used in these models can be obtained in online deployments, as discussed further in Materials and Methods. Generalized and individualized models achieved about 90% AUROC in this work's post hoc analysis. Individualized models performed better with additional data, likely because participants had higher engagement in early subsets used for training. Overall, the similar performance of these approaches indicates the possibility of having one model for multiple users. Generalized and individualized models also attained AUROC values comparable to those of models trained on random samples of all participants' data, which is an ideal but impractical method.

A shortcoming of generalized and individualized models is revealed through a 50% false-negative rate for detecting disengagement. This effect is likely due to the higher variance in features when participants were disengaged compared with when they were engaged. Despite the low recall values, a SAR system that accurately recognizes some instances of disengagement can still considerably enhance the interaction by attempting to reengage the user at appropriate times. Analyzing DS, or periods in the interaction when the user is continuously disengaged, shows that 75% of the total time that participants were disengaged occurred during DS that were 10 s or longer. Some examples of participant behavior during these long DS included playing with toys, interacting with siblings, and even abruptly leaving the intervention setting. Shorter DS typically involved brief shifts in participant focus to other aspects of the environment. Moreover, most long DS required caregivers to reengage participants, whereas participants reengaged on their own in short DS. This suggests that reengagement strategies should focus on counteracting longer DS.

This work's post hoc analysis shows that generalized and individualized models could be used to reliably initiate RA during long DS. The presented reengagement strategy would have initiated RA if the average predicted engagement probability over a window of time fell under a set threshold. Using a 0.35 threshold and a 3-s window would have resulted in RA for about 75% of long DS, with the first RA occurring 2.5 s into these sequences on average; however, RA would also have been erroneously initiated in 15% of ES. An exploration of various window lengths and thresholds reveals the trade-off between maximizing RA in DS and minimizing RA in ES. This balance is important for maintaining RA effectiveness, and choices for these parameters should depend on the implemented RA and overall intervention design.

The presented models are also readily interpretable, an important characteristic for facilitating implementation. Interpretability of ML is especially important in the ASD context, where therapists and caregivers need an understanding of the system's behavior to trust and adopt it (*4*). The described models achieved interpretability in two ways: (i) through a simplified feature set and (ii) through the selected model types. First, this work replicated the described model accuracies to within 5% using seven key features, as described in Results. This shows the problem of modeling engagement to be more tractable and provides insights for the design of future HRI studies in similar contexts. Second, we found tree-based ML models to be the most effective. Such methods are comparatively easier to train and interpret compared with more complex ML models (*26*).

Nevertheless, it is unknown whether the effectiveness of tree-based models in this work would generalize to other contexts. The interpretability of this work further demonstrates the feasibility of applying supervised ML to model user engagement online in closed-loop HRI.

Challenges of the real-world in-home context

A long-term, real-world HRI setting raises many modeling challenges due to significant noise and variance in data. The unconstrained home environment in particular presented several unforeseen problems. First, the camera was attached to the top of the game tablet, but its position was frequently disturbed by both caregivers and child participants. For example, some caregivers temporarily turned the camera toward the ground or toward the robot when child participants were taking a break instead of using the system's power switch as instructed. As a result, the camera position varied throughout the study, adding noise to the extracted visual features. The audio data in this study also contained a high level of background noise, including sounds from televisions, pets, kitchen appliances, and lawn mowers. All participating families chose to place the SAR system in their living rooms, so external parties—such as siblings, friends, and neighbors—regularly interrupted sessions. The camera sometimes failed to capture all individuals in the environment because the system was not designed for multiparty interactions. The variance in data was also higher in this study, given that participants were children with ASD, who display atypical and highly diverse behaviors (*4*).

Substantial variance in data leaves supervised ML models vulnerable to overfitting. To mitigate this risk, we used standard ML practices—such as bagging, boosting, and early stopping—as discussed below in Materials and Methods. In spite of the challenges of a real-world setting, generalized and individualized models achieved AUROC values around 90% and could reliably initiate RA in long sequences of disengagement. However, these models only had 50% recall for disengagement. Further improving the false-negative rate is a key area of future work.

Limitations and future work

As discussed in the previous section, a key modeling challenge in real-world HRI is the substantially increased variance in data, especially when users are disengaged. The solution to this problem in traditional ML is to obtain a large sample of labeled training data. This is not always feasible in HRI and is especially complex for atypical user populations. Moreover, this challenge is especially acute in the ASD context, where high variances in behaviors manifest not only between individuals but also within each individual.

Active learning (AL) is a promising approach to this challenge because it seeks to automatically select the most informative instances that need labeling (*14*). Preliminary work has shown AL to successfully train models of user engagement with a small amount of data (*27*). However, AL is yet to be validated in long-term, real-world settings, as discussed previously in this journal (*28*). A future approach could first use supervised ML to train base models on available labeled data from other users or a user's beginning sessions, as done in this work. Then, AL could be applied to decide when to request a label for unseen data. A therapist or caregiver could provide the labels offline after sessions, allowing the model to iteratively improve in a long-term setting.

Ultimately, the most important direction for future work is to deploy ML frameworks online in real-world HRI and SAR. Such deployments are critical for understanding how well models recognize disengagement under realistic constraints of noise, uncertainty, and variance in data. When implemented online, these models could inform the activation of robot RA; specifically, these could entail verbal and nonverbal robot responses such as socially supportive gestures and phrases (*29*). Overall, online recognition of and response to disengagement will enable the design of more engaging, personalized, and effective HRI, especially in SAR for the ASD community.

MATERIALS AND METHODS

Multimodal dataset

The presented engagement models are based on data from month-long, in-home SAR interventions with children with ASD. During child-robot interactions, participants played a set of space-themed math games on a touchscreen tablet, while a 6-degree-of-freedom Stewart-platform robot named Kiwi provided verbal and expressive feedback, as shown in Fig. 1. The study was approved by the Institutional Review Board of the University of Southern California (UP-16-00755), and we obtained informed consent from the children's caregivers. The seven child participants in this work had a clinical diagnosis of ASD from mild to moderate ranges as described in the *Diagnostic and Statistical Manual of Mental Disorders* (*30*). Table S1 reports the ages and genders of the participants: Ages were between 3 years, 11 months and 7 years, 2 months; three were female and four were male. An earlier article provides further details about the SAR system and intervention design, with a focus on how the robot's feedback and instruction challenge levels were personalized to each user's unique learning patterns using online reinforcement learning (*18*).

Over the course of the month-long, in-home study, we collected a large multimodal dataset from which we derived visual, audio, and game performance features to model engagement. Because of numerous technological challenges commonly encountered in noisy real-world studies, this work only considered about 21 hours of interaction from seven participants who had sufficient multimodal data. Participant 4 had the maximum interaction time analyzed (3 hours and 48 min), and participant 6 had the minimum interaction time analyzed (1 hour and 52 min). Data collected in individual sessions were truncated to only use the content between the first and the last game because session data often included unstructured interactions before and after the games. Each participant was given a tutorial session as an introduction to the SAR system; data from that session were not included in the analysis.

A USB camera mounted at the top of the game tablet recorded a front view of the participants. Visual and audio features were extracted from this camera data using OpenFace (*19*), OpenPose (*20*), and Praat (*21*), open-source data-processing tools feasible for online use in HRI. Visual features derived from OpenFace included (i) the face detection confidence value, (ii) eye gaze direction, (iii) head position, and (iv) facial expression features. OpenPose was only used to estimate the number of people in the environment because the camera's field of view centered on the user's face. Audio features derived from Praat included pitch, frequency, intensity, and harmonicity. Game performance features were also derived from system recordings and included the challenge level of the game being played; the count of incorrect responses to game questions; and the elapsed time in a session, in a game, and since the robot last talked. Note S1 lists all visual, audio, and game performance features used for modeling engagement.

In this work, a participant was annotated to be engaged or disengaged using standard definitions of engagement as a combination

of behavioral, affective, and cognitive constructs (*9*). Specifically, a participant was considered to be engaged when paying full attention to the interaction, immediately responding to the robot's prompts, or seeking further guidance from others in the room. The first author of this work annotated whether a participant was engaged or disengaged for the seven participants. To verify the absence of bias, two additional annotators independently annotated 20% of the data for each participant; interrater reliability was measured using Fleiss' Kappa, and a reliability of $k = 0.84$ was achieved between the primary and verifying annotators. Table S2 contains the specific criteria followed by all annotators.

Modeling approaches

This work applied and evaluated conventional supervised ML methods to model user engagement in month-long, in-home SAR interventions for children with ASD. First, we applied a few preprocessing techniques to the multimodal dataset described in the previous section to address missing data and possible errors in the fusion of modalities. Although we obtained data for each camera frame at a standard rate of 30 frames per second, this work considered the median value of features and annotations in overlapping 1-s intervals (i.e., 0 to 1, 0.5 to 1.5, 1 to 2 s, etc.). The following features were added for each interval: (i) the variance of continuous-valued features in the interval and (ii) an indicator for whether discrete-valued features changed in the interval. This also addressed the problem of low OpenFace confidence for detecting the user's face; low confidence occurred in 38% of camera frames overall but in only three continuous frames on average. Furthermore, all features were standardized to have zero mean and unit variance because raw values were measured on different scales. The means and variances of each feature needed for standardization were obtained with respect to the training set to maintain the feasibility of online implementation.

To model user engagement, this work used two supervised ML approaches that are practical for online implementation in closed-loop HRI: (i) generalized models trained on data from different users and (ii) individualized models trained on data from early subsets of the users' interventions. Generalized models were implemented by training on data from a given subset of M participants. The models were then tested on the remaining N users not in the training subset. We considered all possible combinations of participants; because there were seven participants, values of M and N ranged from 1 to 6. Individualized models were developed by sorting a user's data chronologically and using an early subset for training and later subset for testing the model. In particular, we applied this evaluation to training sets from the first 10 to the first 90% of a user's data, in increments of 10%. We used this approach to standardize the training sets across differences in participant interaction times; future implementations could use beginning sessions as training data instead. The generalized and individualized approaches are both feasible for online use in HRI deployments; the labeled training data required for supervised ML models can be obtained in both cases. To evaluate these approaches, we also implemented models trained and tested on distinct random samples of all users' data despite being this approach impractical for online use. The proportions of training data evaluated in the random sampling approach also ranged from 10 to 90%, in increments of 10%.

Using the generalized, individualized, and random sampling approaches, this work implemented several supervised ML model types. All considered model types are reported in Results; gradient boosted decision trees were the most successful. Specifically, we implemented gradient boosted decision trees with early stopping and bagging (*31*). Boosting algorithms train weak learners sequentially, with each learner trying to correct its predecessor. Early stopping partitions a small percentage of the training data for validation and ends training when performance on the validation set stops improving. Bagging fits several base classifiers on random subsets of the training data and then aggregates the individual predictions to form a final prediction. These techniques were adopted to mitigate the increased risk of overfitting in high-variance datasets, as was the case in this long-term, in-home study.

We implemented the ML models in Python using the following libraries: Scikit-learn version 0.21.3 (*32*), XGBoost version 0.90 (*31*), Hmmlearn version 0.2.1 (*33*), CRFsuite version 0.12 (*34*), TensorFlow version 1.15.0 (*35*), and Keras version 2.2.4 (*36*). All models were implemented with default hyperparameters, as specified in table S3. We used default hyperparameters because the variance in data caused commonly used strategies such as cross validation and grid search to overfit to the training data. All reported model results are from Scikit-learn implementations, except for gradient boosted decision trees, which we implemented using XGboost for improved computational performance. Neural networks were also implemented in TensorFlow and Keras and had similar performance to the reported results from Scikit-learn. Sequential models were explored using Hmmlearn, CRFsuite, and Keras.

SUPPLEMENTARY MATERIALS

robotics.sciencemag.org/cgi/content/full/5/39/eaaz3791/DC1
Note S1. List of multimodal features.
Fig. S1. Comparing data across users.
Fig. S2. Comparing data across sessions.
Table S1. Participant demographic information.
Table S2. Engagement annotation criteria.
Table S3. Model hyperparameters.
Table S4. Generalized model results.
Table S5. Individualized model results.
Table S6. Random sampling model results.
Table S7. Reengagement strategy evaluation using generalized models.
Table S8. Reengagement strategy evaluation using individualized models.
Data S1. Dataset for engagement models.
Data S2. Descriptions of columns in data S1.
Reference (*38*)

REFERENCES AND NOTES

1. D. Feil-Seifer, M. J. Matarić, Defining socially assistive robotics, in *Proceedings of the 2005 IEEE 9th International Conference on Rehabilitation Robotics* (IEEE, 2005), pp. 465–468.
2. M. J. Matarić, B. Scassellati, *Springer Handbook of Robotics* (Springer, 2016), pp. 1973–1994.
3. M. J. Matarić, Socially assistive robotics: Human augmentation versus automation. *Sci. Robot.* **2**, eaam5410 (2017).
4. O. Rudovic, J. Lee, M. Dai, B. Schuller, R. W. Picard, Personalized machine learning for robot perception of affect and engagement in autism therapy. *Sci. Robot.* **3**, eaao6760 (2018).
5. C. -M. Huang, B. Mutlu, Learning-based modeling of multimodal behaviors for humanlike robots, in *HRI '14 Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction* (ACM, 2014), pp. 57–64.
6. Autism spectrum disorders (Tech. rep., World Health Organization, 2018).
7. Data and statistics on autism spectrum disorder (Tech. rep., Centers for Disease Control and Prevention, 2019).
8. M. B. Ospina, J. K. Seida, B. Clark, M. Karkhaneh, L. Hartling, L. Tjosvold, B. Vandermeer, V. Smith, Behavioural and developmental interventions for autism spectrum disorder: A clinical systematic review. *PLOS ONE* **3**, e3755 (2008).
9. B. Scassellati, H. Admoni, M. Matarić, Robots for use in autism research. *Annu. Rev. Biomed. Eng.* **14**, 275–294 (2012).
10. J. J. Diehl, L. M. Schmitt, M. Villano, C. R. Crowell, *Res. Autism Spectr. Disord.* **6**, 249–262 (2012).

11. B. Scassellati, L. Boccanfuso, C.-M. Huang, M. Mademtzi, M. Qin, N. Salomons, P. Ventola, F. Shic, Improving social skills in children with ASD using a long-term, in-home social robot. *Sci. Robot.* **3**, eaat7544 (2018).
12. P. G. Esteban, P. Baxter, T. Belpaeme, E. Billing, H. Cai, H.-L. Cao, M. Coeckelbergh, C. Costescu, D. David, A. De Beir, Y. Fang, Z. Ju, J. Kennedy, H. Liu, A. Mazel, A. Pandey, K. Richardson, E. Senft, S. Thill, G. Van de Perre, B. Vanderborght, D. Vernon, H. Yu, T. Ziemke, How to build a supervised autonomous system for robot-enhanced therapy for children with autism spectrum disorder. *Paladyn* **8**, 18–38 (2017).
13. J. C. Kim, P. Azzi, M. Jeon, A. M. Howard, C. H. Park, Audio-based emotion estimation for interactive robotic therapy for children with autism spectrum disorder, in *URAI 2017 14th International Conference on Ubiquitous Robots and Ambient Intelligence* (IEEE, 2017) pp. 39–44.
14. C. Clabaugh, M. J. Matarić, Escaping Oz: Autonomy in socially assistive robotics. *Annu. Rev. Control Robot. Auton. Syst.* **2**, 33–61 (2019).
15. K. Dautenhahn, Some brief thoughts on the past and future of human-robot interaction. *ACM Trans. Hum. Robot Interact.* **7**, 4:1–4:3 (2018).
16. E. S. Short, M. J. Matarić, Towards autonomous moderation of an assembly game, presented at Workshop on Groups in Human-Robot Interaction at RO-MAN 2016, New York, NY, 26 to 31 August 2016.
17. C. Clabaugh, S. Jain, B. Thiagarajan, Z. Shi, L. Mathur, K. Mahajan, G. Ragusa, M. J. Matarić, Month-long, in-home socially assistive robot for children with diverse needs, in *Proceedings of the 2018 International Symposium on Experimental Robotics (ISER)* (Springer, 2018), pp. 608–618.
18. C. Clabaugh, K. Mahajan, S. Jain, R. Pakkar, D. Becerra, Z. Shi, E. Deng, R. Lee, G. Ragusa, M. J. Matarić, Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders. *Front. Robot. AI* **6**, 110 (2019).
19. T. Baltrusaitis, A. Zadeh, Y. C. Lim, L.-P. Morency, OpenFace 2.0: Facial behavior analysis toolkit, in *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)* (IEEE, 2018), pp. 59–66.
20. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields (2018); https://arxiv.org/abs/1812.08008.
21. P. Boersma, Praat, a system for doing phonetics by computer. *Glot Int.* **5**, 341–345 (2002).
22. A. Ben-Youssef, C. Clavel, S. Essid, M. Bilac, M. Chamoux, A. Lim, UE-HRI: A new dataset for the study of user engagement in spontaneous human-robot interactions, in *ICMI'17 Proceedings of the 19th ACM International Conference on Multimodal Interaction* (ACM, 2017), pp. 464–472.
23. A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.* **30**, 1145–1159 (1997).
24. M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning (2018); https://arxiv.org/abs/1803.09050.
25. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique (2011); https://arxiv.org/abs/1106.1813.
26. Y. Li, L. Yang, B. Yang, N. Wang, T. Wu, Application of interpretable machine learning models for the intelligent decision. *Neurocomputing* **333**, 273–283 (2019).
27. O. Rudovic, M. Zhang, B. Schuller, R. W. Picard, Multi-modal active learning from human data: A deep reinforcement learning approach (2019); https://arxiv.org/abs/1906.03098.
28. C. Clabaugh, M. J. Matarić, Robots for the people, by the people: Personalizing human-machine interaction. *Sci. Robot.* **3**, eaat7451 (2018).
29. L. Brown, R. Kerwin, A. M. Howard, Applying behavioral strategies for student engagement using a robotic educational agent, in *2013 IEEE International Conference on Systems, Man, and Cybernetics* (IEEE, 2013), pp. 4360–4365.
30. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (2013).
31. T. Chen, C. Guestrin, XGBoost: A Scalable Tree Boosting System, *KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 785–794.
32. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
33. hmmlearn: Hidden markov models in python (2015).
34. N. Okazaki, CRFsuite: A fast implementation of conditional random fields (2007).
35. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems (2015); software available from tensorflow.org.
36. F. Chollet, Keras (2015); https://keras.io.
37. M. O'Brien, K. Tobin, Science nation: Socially assistive robots for children on the autism spectrum (National Science Foundation, 29 October 2018).
38. T. Baltrušaitis, M. Mahmoud, P. Robinson, Cross-dataset learning and person-specific normalisation for automatic action unit detection, in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)* (IEEE, 2015), vol. 6, pp. 1–6.

Acknowledgments: The authors thank K. Mahajan and L. Mathur for their help with data analysis; K. Peng and J. Keller for their assistance with annotations; and T. Groechel, L. Klein, and R. Pakkar for their advice and support. In addition, the authors are very grateful to G. Ragusa for a key role in the original study design, recruitment, assessments, and more. The entire research team thanks the children and families who participated in the study that generated the dataset. **Funding:** This research was supported by the National Science Foundation Expedition in Computing Grant NSF IIS-1139148. **Author contributions:** S.J. led this work's conceptualization and investigation. S.J., B.T., and Z.S. processed the data, implemented the methods, and analyzed the results. C.C. designed the study and managed the deployments that generated the datasets used in this work. M.J.M. was the leading faculty advisor for the overarching study on SAR for ASD, and all conducted research, data collection, and analysis. S.J., B.T., Z.S., and M.J.M. all contributed to the writing of this article. **Competing interests:** M.J.M. is a co-founder of Embodied Inc. but has not been involved with the company since December 2016. C.C. is now a full-time employee of Embodied Inc. but was not involved with the company while the reported work was done. **Data and materials availability:** All data needed to evaluate the conclusions that can be released under USC IRB policies are included in the article or the Supplementary Materials. Please contact S.J. and M.J.M. for questions about other materials.

Submitted 4 September 2019
Accepted 22 January 2020
Published 26 February 2020
10.1126/scirobotics.aaz3791

Citation: S. Jain, B. Thiagarajan, Z. Shi, C. Clabaugh, M. J. Matarić, Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Sci. Robot.* **5**, eaaz3791 (2020).

Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders

Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh and Maja J. Mataric

Sci. Robotics **5**, eaaz3791.
DOI: 10.1126/scirobotics.aaz3791