HUMAN-ROBOT INTERACTION

Personalized machine learning for robot perception of affect and engagement in autism therapy

Ognjen Rudovic[1]*, Jaeryoung Lee[2], Miles Dai[1], Björn Schuller[3,4], Rosalind W. Picard[1]

Robots have the potential to facilitate future therapies for children on the autism spectrum. However, existing robots are limited in their ability to automatically perceive and respond to human affect, which is necessary for establishing and maintaining engaging interactions. Their inference challenge is made even harder by the fact that many individuals with autism have atypical and unusually diverse styles of expressing their affective-cognitive states. To tackle the heterogeneity in children with autism, we used the latest advances in deep learning to formulate a personalized machine learning (ML) framework for automatic perception of the children's affective states and engagement during robot-assisted autism therapy. Instead of using the traditional one-size-fits-all ML approach, we personalized our framework to each child using their contextual information (demographics and behavioral assessment scores) and individual characteristics. We evaluated this framework on a multimodal (audio, video, and autonomic physiology) data set of 35 children (ages 3 to 13) with autism, from two cultures (Asia and Europe), and achieved an average agreement (intraclass correlation) of ~60% with human experts in the estimation of affect and engagement, also outperforming nonpersonalized ML solutions. These results demonstrate the feasibility of robot perception of affect and engagement in children with autism and have implications for the design of future autism therapies.

INTRODUCTION

The past decade has produced extensive research on human-centered robot technologies aimed at achieving more natural human-robot interactions (*1*). However, existing robots, and software that enables robots to recognize human affect in real time (*2*), are still limited and in need of more social-emotional intelligence if they are to interact naturally and socially with people (*3*). Health care is one of the areas, in particular, that can substantially benefit from the use of socially assistive robots, because they have the potential to facilitate and improve many aspects of clinical interventions (*4*). The most recent advances in machine learning (ML) (*5*) and, in particular, deep learning (*6*) have paved the way for such technology.

Various terms for this area have emerged, including socially assistive robotics (*7*), robot-enhanced therapy (*8*), and robot-augmented therapy (*9*). The main role of social robots that we examine is to engage children in interactive learning activities that supplement and augment those delivered by therapists. We focus on children with autism spectrum condition (ASC) (*10*), which affect 1 in 64 in the United States, with a ratio of 4:1 males:females (*11*). Children with ASC have persistent challenges in social communication and interactions, as well as restricted and repetitive patterns of behavior, interests, and/or activities (*10*). Many therapists encourage children to engage in learning through play (*12*), traditionally with toys as open-ended interaction partners. Recently, social robots have been used to this end because many children with ASC find them enjoyable and engaging, perhaps due to their human-like yet predictable and nonthreatening nature (*13*).

A typical robot-assisted autism therapy for teaching emotion expressions to children with ASC proceeds as follows: A therapist uses images of facial and body expressions of basic emotions (e.g., sadness, happiness, and fear), as shown by typically developing children. Then, the robot shows expressions of these emotions to the child, and the therapist asks the child to recognize the emotion. This is followed by the mirroring stage, in which the child is encouraged to imitate the robot's expressions. If successful, the therapist proceeds to the next level, telling a story and asking the child to imagine what the robot would feel in a particular situation. These steps are adopted from the theory of mind concept (*14*), designed to teach perspective-taking ("social imagination")—a challenge for many children with ASC. Other therapy designs include applied behavioral analysis and pivotal response treatment (*15*); however, using humanoid and other robotic solutions as part of therapy is still in the experimental stage (*16*). The progress, in part, has been impeded due to the inability of current robots to autonomously perceive, interpret, and naturally respond to human behavioral cues. Today, this has been accomplished in the so-called Wizard of Oz (WoZ) scenario (*17*), in which a therapist or a person "behind the curtain" controls the robot via a set of preprogrammed behaviors and prompts, such as the robot waving at or saying something to the child. This makes the interaction less natural and potentially more distracting for the child and therapist. Thus, there is a need for (semi)autonomous and data-driven robots that can learn and recognize the child's behavioral cues and respond smoothly (*18*), a challenge discussed in several pioneering works on robot-assisted therapy (*17*, *19*).

Automated analysis of children's behavioral cues relies on ML from sensory inputs (e.g., cameras and microphones) capturing different modalities (face, body, and voice) of child's behaviors (*20*). In the standard supervised ML approach, these inputs are first transformed into numerical representations (called "input features") and paired with target outputs (e.g., engagement labels). These data pairs are used to train ML models [e.g., a support vector machine (SVM) (*5*)], which learn a mathematical function that maps the input features onto target outputs. The learned model is then applied to new input features, extracted from held-out recordings of the child's behaviors, to estimate target outputs (engagement levels). For instance, Sanghvi *et al.* (*21*) used a data set of postural expressions of children playing chess with a robot; they extracted the upper body silhouette and trained a set of weak classifiers for engagement estimation. Kim *et al.* (*22*) used SVM to estimate emotional states of children with ASC from their audio data and assess their social

[1]MIT Media Lab, Cambridge, MA 02139, USA. [2]Department of Robotic Science and Technology, Chubu University, Kasugai, Aichi 487-8501, Japan. [3]Department of Computing, Imperial College London, London SW7 2AZ, UK. [4]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany.
*Corresponding author. Email: orudovic@mit.edu

engagement while playing with a robot. Other works adopted a similar approach to estimate engagement based on the children's facial expressions (*23*), body movements (*24*), autonomic physiology (*25*), and vocalizations (*26*). More recently, Esteban *et al.* (*8*) used the NAO robot and gaze direction, facial expressions, body posture, and tone of voice to classify stereotypical behaviors and "social" engagement (based on eye contact and verbal utterances) of children with ASC. Essentially, the main focus of these studies was on robot appearance and interaction strategy, while children's behavioral cues were used as a proxy of their affect and engagement (*7*). Our work takes a different approach by using supervised ML to estimate levels of affective states and engagement coded by human experts.

In this work, we constructed a personalized deep learning framework, called the personalized perception of affect network (PPA-net), that can adapt robot perception of children's affective states and engagement to different cultures and individuals. This is motivated by our previous work (*27*), which found large cultural and individual differences in affect and engagement of children with ASC during robot-assisted therapy. Our data came from 35 children diagnosed with autism: 17 from Japan (C1) and 18 from Serbia (C2). It includes synchronized (i) video recordings of facial expressions, head movements, body movements, pose, and gestures; (ii) audio recordings; and (iii) autonomic physiology—heart rate (HR), electrodermal activity (EDA), and body temperature (T)—measured on the nondominant wrist of the child. We used this multimodal data set of children with ASC (MDCA) (*27*) to train and evaluate our framework. These data were coded for valence, arousal, and engagement on a continuous scale from −1 to +1 by five human experts who reviewed the audiovisual recordings of the therapy sessions. The quality of the coding was measured using intraclass correlation (ICC), type (3,1), which ranges from 0 to 1, and is commonly used in behavioral sciences to assess coders' agreement (*28*). The average agreement scores (and their SD) computed from the pair-wise ICC of the coders were as follows: valence (0.53 ± 0.17), arousal (0.52 ± 0.14), and engagement (0.61 ± 0.14). These individual data annotations were then aligned among the coders to produce the gold standard labels (a single annotation for each target dimension: valence, arousal, and engagement) that we used for the robot learning. Further details about the data and coding process are provided in notes S2 and S3.

The workflow of the envisioned ML robot–assisted autism therapy consists of three key steps: sensing, perception, and interaction (Fig. 1).

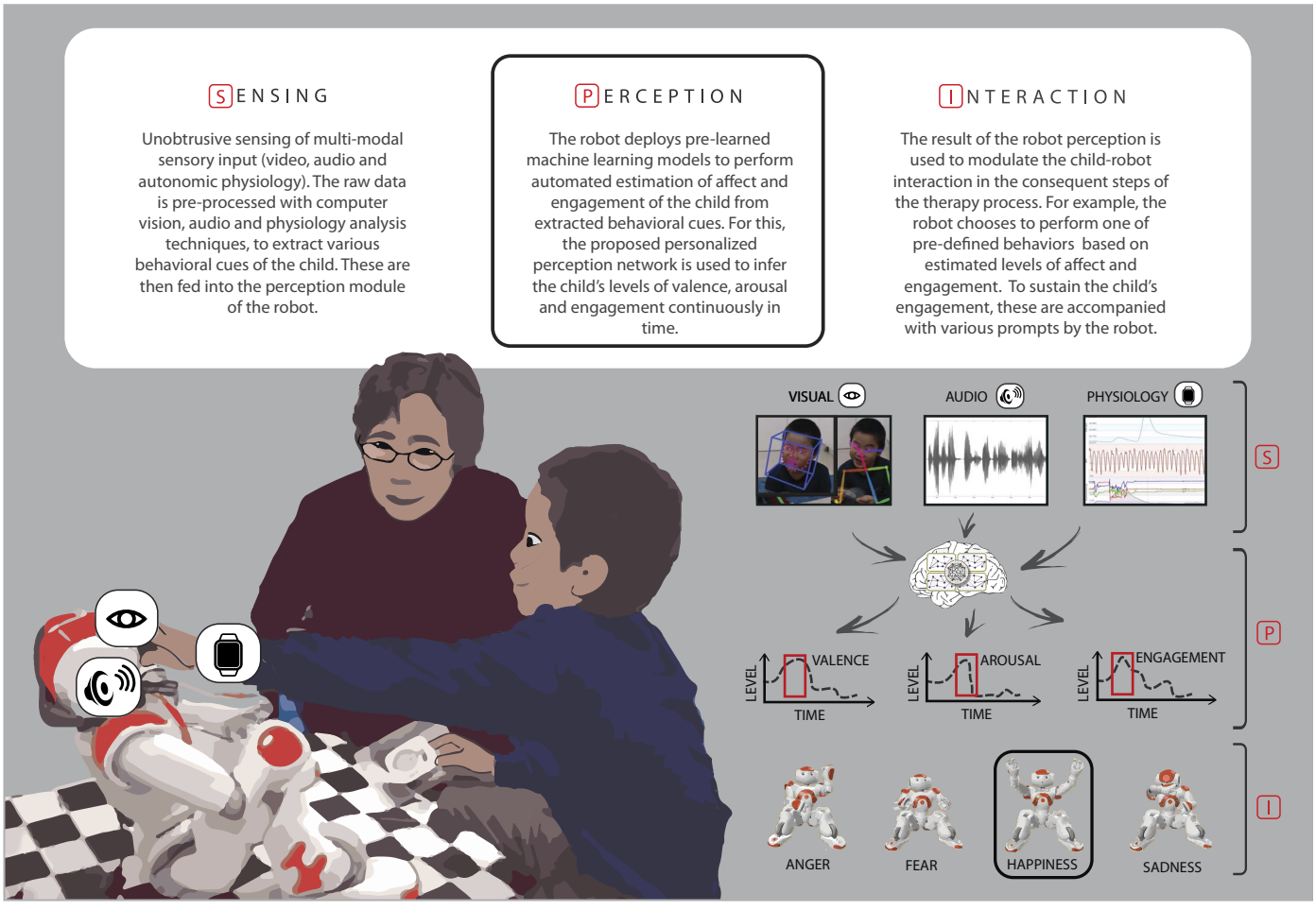

Fig. 1. Overview of the key stages (sensing, perception, and interaction) during robot-assisted autism therapy. Data from three modalities (audio, visual, and autonomic physiology) were recorded using unobtrusive audiovisual sensors and sensors worn on the child's wrist, providing the child's heart-rate, skin-conductance (EDA), body temperature, and accelerometer data. The focus of this work is the robot perception, for which we designed the personalized deep learning framework that can automatically estimate levels of the child's affective states and engagement. These can then be used to optimize the child-robot interaction and monitor the therapy progress (see Interpretability and utility). The images were obtained by using Softbank Robotics software for the NAO robot.

Robot sensing of outward and inward behavioral cues of the child used the open-source data processing tools: OpenFace (*29*), OpenPose (*30*), and openSMILE (*31*), together with tools that we built for processing audio-video and biosignals of the children (see note S3). The main focus of this work is the robot perception step, structured by us using deep neural networks (*6*). Specifically, we first trained the PPA-net using the processed features as input and the gold standard coding of affect and engagement as output. Our goal was to maximize the ICC agreement between the model estimates and gold standard coding to make the robot's perceived outputs closer to those of the human experts. The learned PPA-net is subsequently used to automatically estimate continuous levels of the child's valence, arousal, and engagement from new data (see System design). The estimated values can then be used to design and modulate more dynamic, engaging, and affect-sensitive interactions.

RESULTS

System design

Figure 2 shows the PPA-net that we designed and implemented, with (i) noisy and missing input features from the MDCA data set (feature layer); (ii) heterogeneous data (a famous adage says, "If you have met one person with autism, you have met one person with autism.") (context layer); and (iii) simultaneous and continuous estimation of children's affective states (valence and arousal) and engagement (inference layer).

In Fig. 3A, the diagonal elements show the fraction of data where the sensing extracted the features, whereas the off-diagonals show how often the two feature types were extracted simultaneously. About half of the data are missing at least one modality, for example, the child's face was not visible (the facial features were extracted successfully from 56% of image frames using the openFace tool) and/or a child did not wear the wristband (73% of children accepted to wear it). Figure 3B shows dependences of the gold standard labels obtained from human experts. As can be seen, these vary largely at the group level (cultures) and within children from each culture. Figure 3C shows the multimodal features in a two-dimensional space obtained using t-distributed stochastic neighbor embedding (t-SNE) (*32*), a popular technique for unsupervised dimensionality reduction and visualization. Even without using children's IDs, their data cluster in the projected space, revealing their high heterogeneity.

At the feature layer, we leveraged the complementary nature of the multimodal data to deal with missing and noisy features using auto-encoders (AEs) (*33*) and the fusion of different modalities (see Materials and Methods). These feature representations were subsequently augmented (context layer) using expert's inputs from the childhood autism rating scale (CARS) (*34*), providing 15 behavioral scores of the child's mental, motor, and verbal ability. The architecture was designed to nest the children on the basis of their demographics (culture and gender), followed by individual network layers for each child. Because the target outputs exhibit different dependence structures for each child

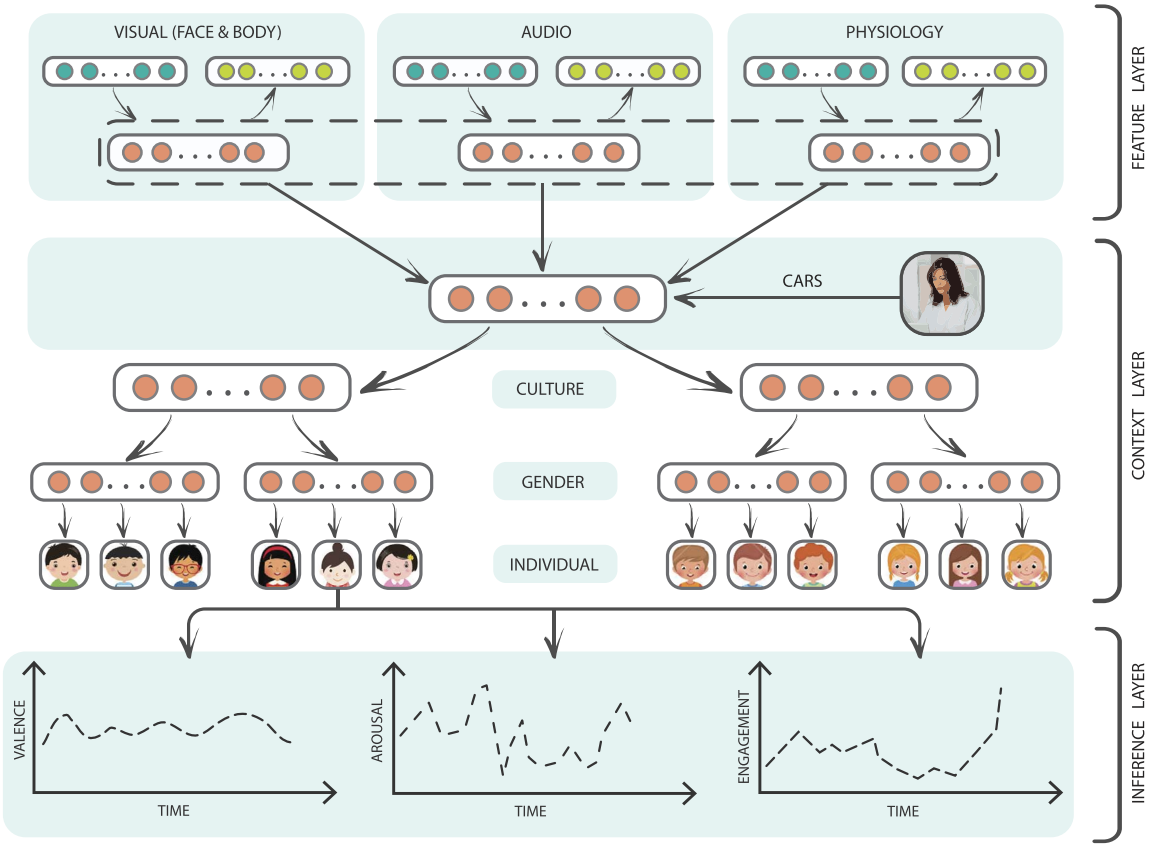

Fig. 2. PPA-net. The feature layer performs feature fusion using (supervised) auto-encoders designed to reduce noise and handle missing features. The inward and outward arrows depict the encoding (in orange)/decoding (in green) of the observed input features (in blue). At the context layer, behavioral scores of the child's mental, motor, and verbal ability are used to augment the input features using the expert knowledge [quantified by CARS (*34*)]. Personalization of the network is achieved using the demographic information (culture and gender), followed by individual network layers for each child. The inference layer performs the child-specific estimation of valence, arousal, and engagement levels. The activations of the hidden nodes (in orange) are learned during the network training (see Results).

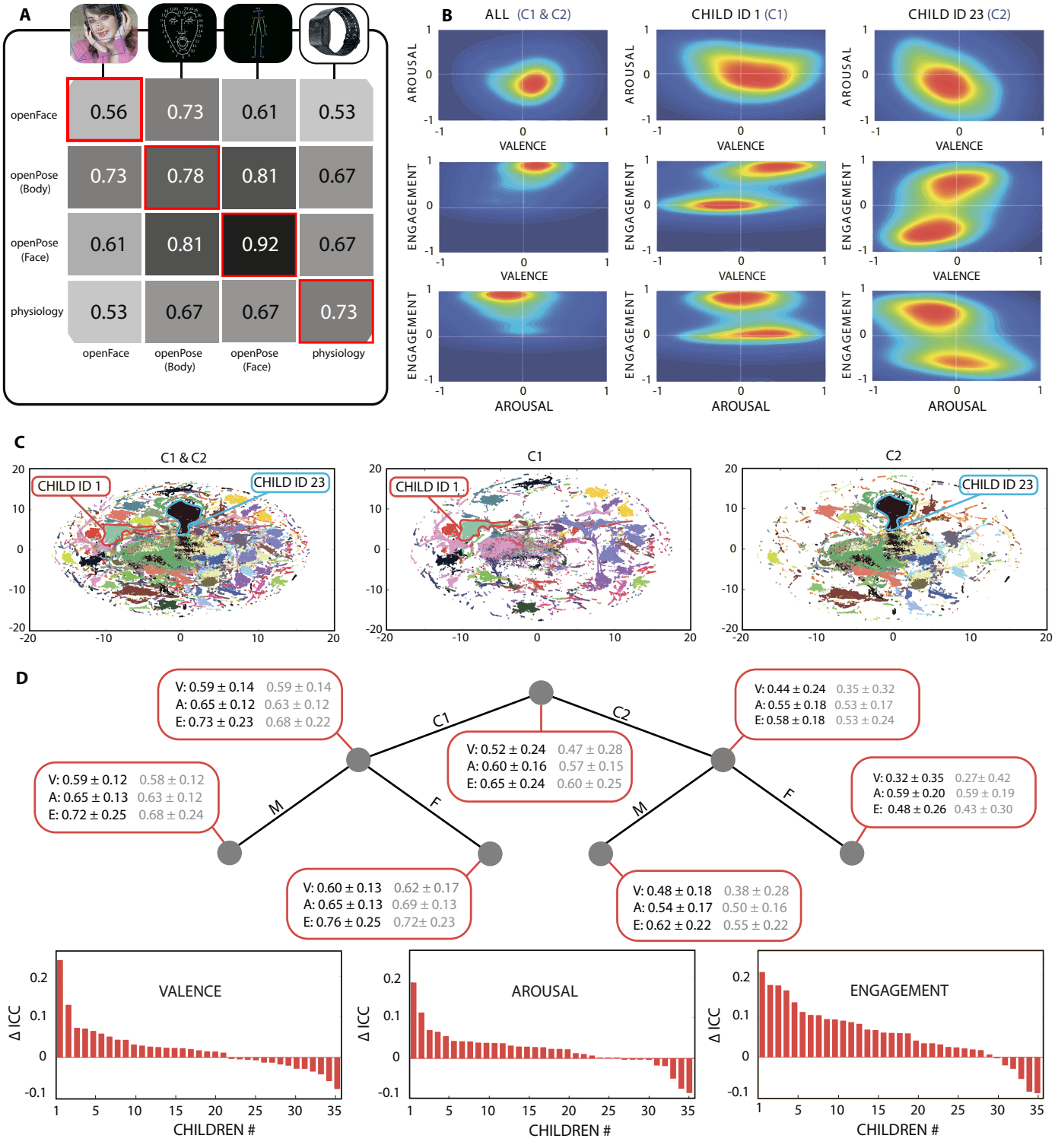

Fig. 3. Data analysis and results. (**A**) The fraction of data present across the different modalities both individually and concurrently. (**B**) The heat maps of the joint distributions for the valence, arousal, and engagement levels, coded by human experts. Large differences in these patterns are present at the culture and individual levels. (**C**) Clustering of the children from C1 and C2 using the t-SNE, an unsupervised dimensionality reduction technique, applied to the auto-encoded features (see Effects of model personalization). (**D**) ICC scores per child: C1 (n = 17) and C2 (n = 18) for valence (V), arousal (A), and engagement (E) estimation. Compared with the GPA-net (in gray), the performance of the PPA-net (in black) improved at all three levels (culture, gender, and individual) in the model hierarchy. Bottom: Sorted improvements in the ICC performance (ΔICC) between PPA-net and GPA-net for each child.

(Fig. 3B), we used multitask learning (*35*, *36*) to learn the child-specific inference layers for valence, arousal, and engagement estimation. To evaluate the PPA-net, we separated each child's data into disjoint training, validation, and testing data subsets (see Study design).

Effects of model personalization

The main premise of model personalization is that disentangling different sources of variance is expected to improve the individual estimation performance compared with the traditional one-size-fits-all approach. Figure 3D depicts the ICC scores computed at each level in the model hierarchy, comparing the PPA-net and group-level perception of affect network (GPA-net). The latter was trained with the same depth layers as PPA-net but shared across children (see Group-level network). Overall, the strength of the model personalization can be seen in the performance improvements at culture, gender, and individual levels in all (sub)groups of the children. A limitation can be seen in the adverse gender-level performance by the PPA-net (versus GPA-net) on the two females from C1 when estimating their valence (ICC = 0.60 versus 0.62) and arousal (ICC = 0.65 versus 0.69) levels. Here, the PPA-net overfits the data of these two children—a common bottleneck of ML algorithms trained on limited data (*37*). Nevertheless, it did not affect their engagement estimation (ICC = 0.76 versus 0.72) because the PPA-net successfully leveraged data from the children showing similar engagement behaviors.

The individual estimation performance by the two networks is compared at the bottom of Fig. 3D. The improvements in ICC due to the network personalization range from 5 to 20% per child. We note drops in the PPA-net performance on some children, which is common in multitask learning where the gain in performance on some tasks (here, "tasks" are children) comes at the expense of others. The PPA-net significantly outperformed GPA-net on 24 children for valence, 21 children for arousal, and 28 children for engagement, out of 35 children total. Thus, personalizing the PPA-net leads to overall better performance than using the group-level GPA-net.

Interpretability and utility

A barrier to adoption of deep learning is when interpretability is paramount. Understanding the behavioral features that lead to a particular output (e.g., engagement levels) builds trust with clinicians and therapists. To analyze the contribution of each behavioral modality and its features in the PPA-net, we applied DeepLift (Learning Important FeaTures) (*38*), an open-source method for computing importance scores in a neural network. Figure 4A shows the importance of features from different modalities in the engagement estimation task: Features with high positive scores are more important for estimating high levels of engagement and vice versa for negative scores. We note that the body and face modalities dominated when the scores were computed for both cultures together. At the culture level, the body features produced high scores in C1 and low scores in C2. This reveals that behaviors of high engagement in C1 were accompanied by large body movements, whereas in C2, this was the case when children disengaged—underscoring cultural differences in the two groups. The expert knowledge provided by CARS also affected estimation of engagement. CARS scores in our previous work (*27*) showed statistically significant differences between the two cultures (see table S2), which may explain the difference in the importance of the CARS features found here. Differences also occurred at the gender level but are difficult to interpret due to the male/female imbalance.

The PPA-net's utility can be demonstrated through visualization of several components of the robot sensing and perception. Figure 4B depicts the estimated affect and engagement levels along with the key autonomic physiology signals of a child undergoing the therapy (currently, these results are obtained by an off-line analysis of the recorded video). We observed that the PPA-net accurately detected changes in the child's engagement levels (e.g., during the disengagement segment) while providing estimates that were overall highly consistent with human coders. Figure 4C summarizes each phase of the therapy in terms of its valence, arousal, and engagement levels, along with their variability within each phase of the therapy. Compared with expert humans, the PPA-net produced these statistics accurately for engagement and arousal while overestimating the valence levels. However, as more target data per child become available, these can be improved by retraining his/her individual layer.

Alternative approaches

How much advantage does the new personalized deep learning approach provide over traditional ML? We compared the performance of the PPA-net with traditional multilayer perceptron (MLP) deep networks optimized using standard learning techniques without sequential nesting of the layers (Materials and Methods). The MLP layers were first optimized jointly in a child-dependent manner (CD-MLP), followed by fine-tuning (personalization) of the individual layers (P-MLP). We also compared traditional ML models: lasso regression (LR) (*5*), support vector regression (SVR) (*5*), and gradient boosted regression trees (GBRTs) (*39*). LR is usually considered a baseline; SVR has previously been used in engagement estimation; and GBRTs are often used in clinical tasks due to their ease of interpretation.

Table S1 shows that the personalized deep learning strategy (PPA-net) reaches an average performance of ICC = 59% for estimating valence, arousal, and engagement levels. The joint learning of all layers in the personalized MLP (P-MLP) results in a lack of discriminative power with an average drop of 6% in performance. Compared with unpersonalized models (GPA-net, CD-MLP, LR, SVR, and GBRTs), there is also a gap in performance. Whereas LR fails to account for highly nonlinear dependencies in the data (drop of 25%), the nonlinear kernel method (SVR) achieves it to some extent; however, SVR fails to reach the full performance attained by the PPA-net due to the absence of the hierarchical structure (drop of 9%). On the other hand, GBRTs are capable of discovering a hierarchy in the features, yet they lack a principled way of adapting to each child (drop of 10%). To provide insights into the importance of the child data for the estimation accuracy, we also included the results obtained by running a child-independent experiment (i.e., no target child data were used during training) using the MLP network (CI-MLP). This led to a large drop in performance (drop of 39%), evidencing high heterogeneity in the children data and the challenge it poses for ML when trying to generalize to new children with ASC.

Effects of different modalities

To assess the contribution of each behavioral modality, we evaluated the PPA-net using visual (face and body), audio, and autonomic physiology features both independently and together. In fig. S3 (note S1), we show the average results for both cultures and for children within each culture. As anticipated, the fusion approach outperforms the individual modalities across all outputs (valence, arousal, and engagement), confirming the complimentary nature of the modalities (*40*). Also, higher performance was achieved on C1 than C2 using the multimodal approach. Furthermore, the body features outperformed the other individual modalities, followed by the face and physiology features.

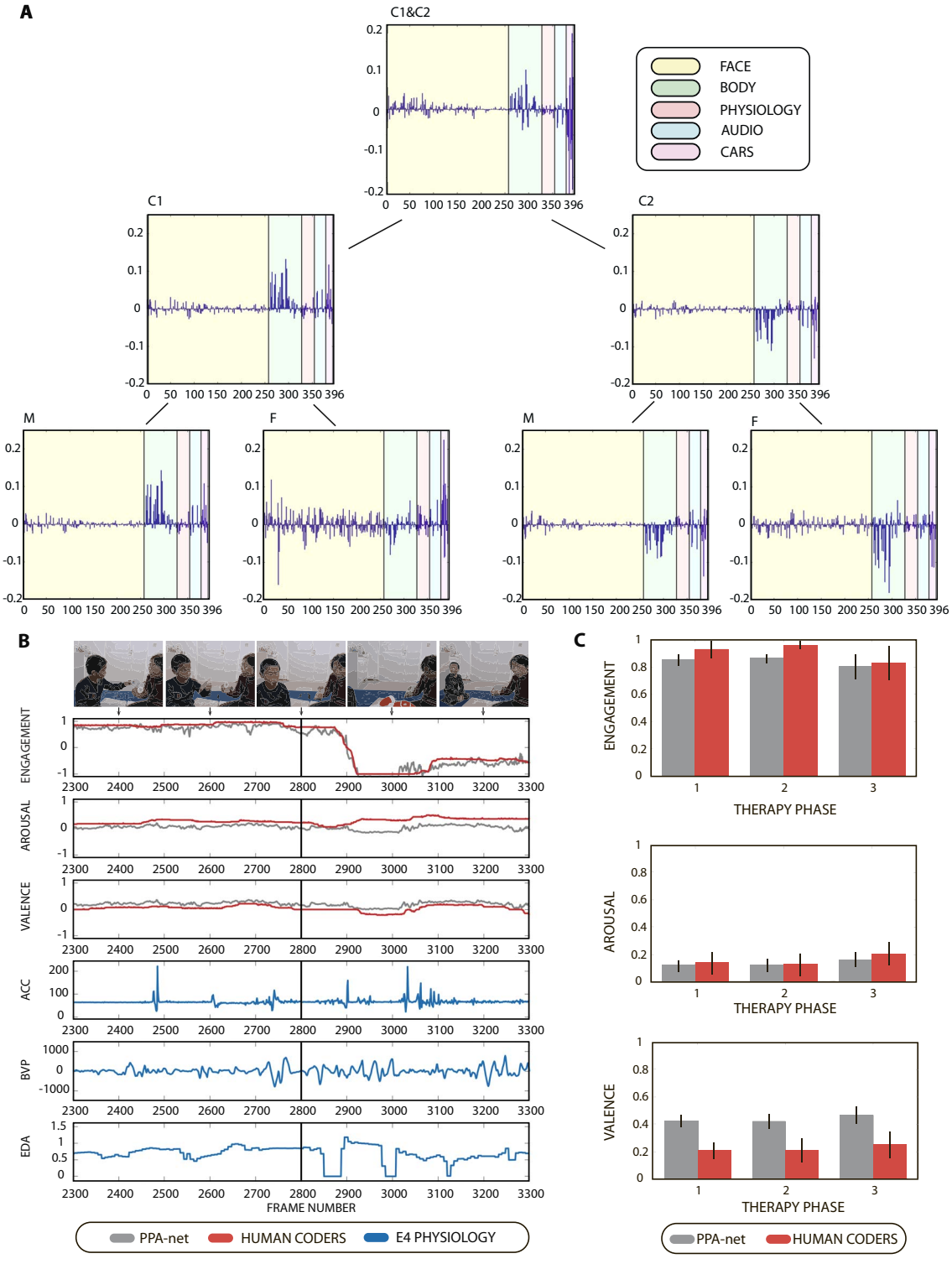

Fig. 4. Interpretability and utility. (**A**) Interpretability can be enhanced by looking at the influence of the input features on the output target. Here, the output is estimated engagement level. The relative importance scores (*y* axis) are shown for input features from each behavioral modality (*x* axis). These are obtained from the DeepLift (*38*) tool, which provides negative/positive values when the input feature drives the output toward −1/+1. (**B**) Estimated engagement, arousal, and valence levels of the child as the therapy progresses. These were obtained by applying the learned PPA-net to the held-out data of the target child. We also plotted the corresponding signals measured from the child's wrist: the movement intensity derived from accelerometer readings (ACC), blood-volume pulse (BVP), and EDA. (**C**) Summary of the therapy in terms of the average ± SD levels of affect and engagement within each phase of the therapy: (1) pairing, (2) recognition, and (3) imitation.

We attribute the low performance of the audio modality to a high level of background noise, which is difficult to control in real-world settings (*41*). The performance of the autonomic physiology features was comparable with the best performing individual modality (i.e., body) in C1, but this was not true in C2. We noticed that children from C1 moved their hands less, resulting in more accurate sensory readings of their biosignals and thus better quality features for the PPA-net, which may explain the difference in their performance across the two cultures.

DISCUSSION

This work demonstrated the technical feasibility of using ML to enable robot perception of affect and engagement in children undergoing autism therapy. This is driven by the societal need for new technologies that can facilitate and improve existing therapies for a growing number of children with ASC (*7*). To this end, we developed a personalized ML framework, PPA-net, that adapts to a child's affective states and engagement across different cultures and individuals. This framework builds on state-of-the-art deep learning techniques (*6*, *37*) that have shown great success in many tasks [e.g., object and scene recognition (*42*)] but have not previously been explored for robot perception in autism therapy.

The state of the art for recognition of arousal and valence benchmarked on the largest public data set of human facial expressions [Affect-Net (*43*)] claims a human labeler agreement of 60.7%, with automated estimation reaching 60.2% for valence and 53.9% for arousal. These results are on images of typical individuals, whereas coding and automatic estimation of affect and engagement of children with ASC are considered far more challenging. On our autism data set, the average ICC of human experts was 55.3%, whereas the PPA-net achieved an average ICC of 59%, thus providing a consistent estimator of the affective states and engagement. To accomplish this, we personalized robot perception by allowing the deep networks to share data at each level in the model hierarchy, and using network operators (learn, nest, and clone) and specifically designed fine-tuning strategies (Materials and Methods) based on the notions of network nesting (*44*) and deeply supervised nets (*45*). Overall, this led to statistically significant improvements over the GPA-net (see Effects of model personalization). From experiments on interpretability, we found that leveraging contextual information, such as demographics and expert knowledge, helped disentangle different sources of variance arising from cultures and individuals (see Interpretability and utility). Traditional ML, such as the SVMs used in previous attempts of robot perception, and the ensemble of regression trees, GBRT, do not offer this flexibility. By contrast, the PPA-net brings together interpretability, design flexibility, and overall improved performance.

Faced with the real-world problem of noisy and missing data from children with ASC, we showed that the fusion of outward (audiovisual) and inward (physiological) expressions of affect and engagement helped mitigate the effects of these artifacts (see Effects of different modalities). Our experiments also revealed that the body and face modalities played a central role, and by leveraging the autonomic physiology data, the estimation performance was further enhanced (*25*). Although the audio modality underperformed in our experiments, future background noise reduction, speaker diarization, and better audio descriptors may improve its usefulness (*41*). We also found that CARS scores largely influence the estimation performance (see Interpretability and utility), suggesting that the expert knowledge plays an important role in facilitating the robot perception of affect and engagement.

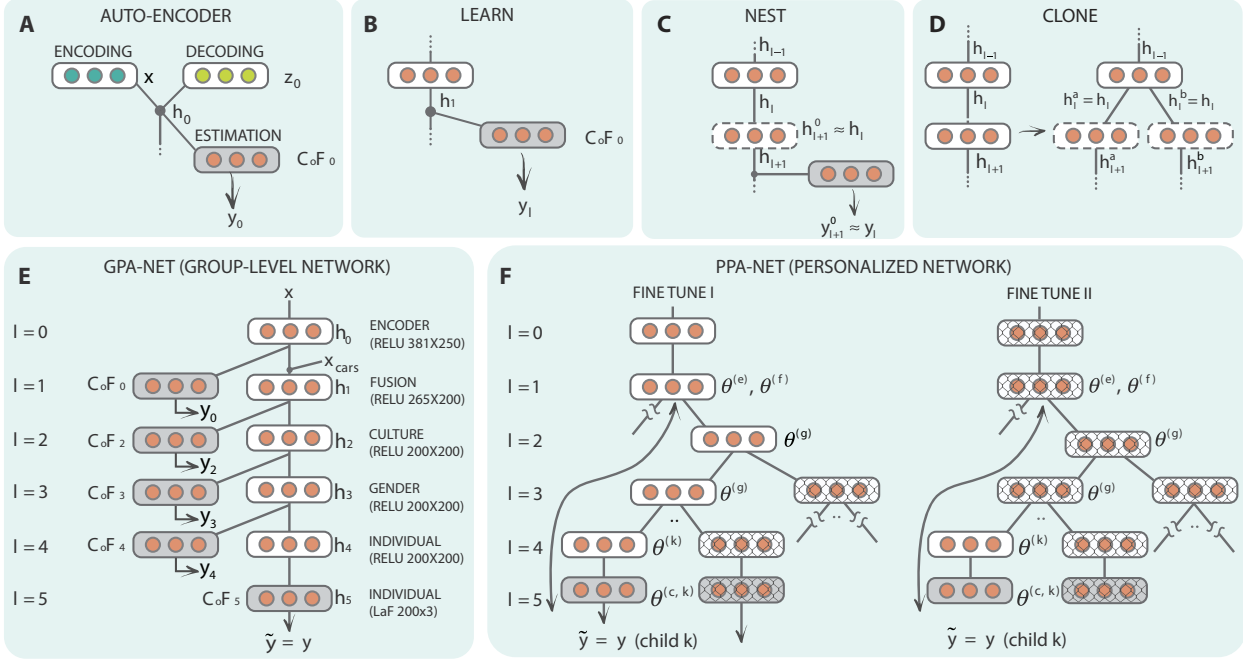

Fig. 5. The learning of the PPA-net. (**A**) The supervised AE performs the feature smoothing by dealing with missing values and noise in the input while preserving the discriminative information in the subspace h_0—constrained by the CoF_0. The learning operators in the PPA-net—(**B**) learn, (**C**) nest, and (**D**) clone—are used for the layer-wise supervised learning, learning of the subsequent vertical layers, and horizontal expansion of the network, respectively. (**E**) The group-level GPA-net is first learned by sequentially increasing the network depth using learn and nest and then used to initialize the personalized PPA-net weights at the culture, gender, and individual level (using clone). (**F**) The network personalization is then accomplished via the fine-tuning steps I and II (Materials and Methods).

Implications

How can this technology help support therapists and clinicians working with children with ASC, given that today's framework is not yet real time? Whereas a future real-time version will likely be used to automatically adapt robot interactions or to assist a therapist with real-time insights, for example, how internal physiology may give early indications before outward changes become visible, there are also uses today based on viewing offline data, as shown in Fig. 4. The data can help a therapist see idiosyncratic behavioral patterns of the interacting child and track how these change over multiple therapy sessions. The output of the robot perception can easily summarize each phase of the therapy (Fig. 4C). Although people are generally not good at detecting or tracking small behavioral changes, a robot partner could excel at this, giving insights that help a therapist better see when even small progress is made.

Limitations and future work

This work has several limitations and opportunities for future improvement. In the PPA-net, we split the children based on demographics; a hybrid adaptive robot perception could combine previous knowledge (e.g., demographics) with a data-driven approach to automatically learn the network structure. Also, our current framework is static, whereas the data are inherently dynamic. By incorporating temporal information, different network parameters may be learned for each phase. Whereas the multimodal data set used in our work contains a limited one session per child, 25 min at 25 frames/s, randomly sampled and split into nonoverlapping training, validation, and test sets, ideally, the system would have access to multiple therapy sessions. This would allow the robot to actively personalize the PPA-net as the therapy progresses, giving enough sessions so that data samples that are further apart in time and less likely to be correlated could be drawn for training the models. For this, ML frameworks such as active learning (*46*) and reinforcement learning (*47*) are a good fit.

Another constraint of the current solution is that the video comes from a fixed (in position) background camera/microphone. This provides a stable camera, allowing us to better focus on discrimination of affective states and engagement levels. Although the "active vision" view from the robot's (moving) perspective would enable more naturalistic interactions, it poses a number of stabilization and multiview challenges (*48*). Furthermore, this framework needs to be extended and optimized to handle previously unseen children. In this context, one of the most important avenues for future research on robot perception for autism therapy is to focus on its utility and deployment within everyday autism therapies. Only in this way can the robot perception and the learning of children with ASC be mutually enhanced.

MATERIALS AND METHODS

Study design

Experimentals were preapproved by the Institutional Review Boards of Chubu University (Japan), Massachusetts Institute of Technology (United States), and Mental Health Institute (Serbia), with informed consent in writing from the children's parents. We used data from 35 children (30 males and 5 females, ages 3 to 13) diagnosed with autism, being part of our MDCA data set (*27*). The data of one male child (C2) from this data set were excluded from analysis due to the low-quality recordings (see note S2). Human experts who coded the affect and engagement in the data performed blinded assessment of these outcomes. Evaluations in Results were made after randomly splitting each child's data into three disjoint sets: 40% training, 20% validation (to select model configuration), and 40% testing (to evaluate generalization). To minimize bias in the data selection, we repeated this process 10 times. The input features were z-normalized, and the models' performance was reported as average ± SD of ICCs across the children.

The overall objective of this study was to show the benefits of the personalized robot perception over a traditional one-size-fits-all ML approach. For this, we implemented the personalized deep learning architecture using feed-forward multilayer neural networks (*6*). Each layer receives the output of the layer above as its input, producing higher-level feature representations. We began with the GPA-net, with all layers shared among the children. Network personalization was then achieved by (i) replicating the layers to construct the architecture in Fig. 2 and (ii) applying the fine-tuning strategies to optimize the network performance on each child. The last network layers were then used to make individual estimations of affect and engagement. As shown previously, this approach shows benefits over nonpersonalized ML solutions, measured by the ICC between the estimated and human coding of affect and engagement.

Feature fusion and autoencoding

We applied fusion to the face (x_f), body (x_b), audio (x_a), and autonomic physiology (x_p) features of each child encoded as a vector of real-valued numbers as $x = [x_f; x_b; x_a; x_p] \in \mathcal{R}^{D_x \times 1}$, where $D_x = 396$ is total number of the input features. The continuous labels for valence (y_v), arousal (y_a), and engagement (y_e) for each child were stored as $y = [y_v; y_a; y_e] \in \mathcal{R}^{D_y \times 1}$, where $D_y = 3$. To reduce problems from partially observed and noisy features in the input x (fig. S3A), we used an AE (*49*) in the first layer of the PPA-net. The AE transforms x to a hidden representation h_0 (with an encoder) through a deterministic mapping using a linear activation function (LaF),

$$h_0 = f_{\theta_0^{(e)}}(x) = W_0^{(e)}x + b_0^{(e)}, \quad \theta_0^{(e)} = \left\{W_0^{(e)}, b_0^{(e)}\right\} \qquad (1)$$

parametrized by $\theta_0{}^{(e)}$, where e designates the parameters on the encoder side, W is the weight coefficient matrix, and b is the bias vector. This hidden representation is then mapped back to the input, producing the features

$$z_0 = f_{\theta_0^{(d)}}(h_0) = W_0^{(d)}h_0 + b_0^{(d)}, \quad \theta_0^{(d)} = \left\{W_0^{(d)}, b_0^{(d)}\right\} \qquad (2)$$

where d designates the parameters of the decoder, and $W_0^{(d)} = W_0^{(e)T}$ are the tied weights used for the inverse mapping of the encoded features (decoder). Here, T represents the matrix transpose operation. In this way, the input data were transformed to lower-dimensional and less-noisy representations (the "encoding"). The encoded subspace also integrates the correlations among the modalities, rendering more robust features for subsequent network layers.

We augmented the encoding process by introducing a companion objective function (CoF) for each hidden layer (*45*). The CoF acts as a regularizer on the network weights, enabling the outputs of each layer to pass the most discriminative features to the next layer. Using the CoF, the AE also reconstructs target outputs y_0 (in addition to z_0) as

$$y_0 = f_{\theta_0^{(c)}}(h_0) = W_0^{(c)}h_0 + b_0^{(c)}, \quad \theta_0^{(c)} = \left\{W_0^{(c)}, b_0^{(c)}\right\} \qquad (3)$$

where c designates the parameters of the CoF. The AE parameters $\omega_0 = \left\{\theta_0^{(e)}, \theta_0^{(d)}, \theta_0^{(c)}\right\}$ were then optimized over the training data

set to minimize the mean squared error (MSE) loss, defined as $\alpha(a,b) = \sum_{i=1}^{d} ||a_i - b_i||^2$, for both the decoding (α_d) and output (α_c) estimates:

$$\omega_0^* = \underset{\omega_0}{\operatorname{argmin}}\, \alpha(x,y) = \underset{\omega_0}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\left((1-\lambda)\alpha_d\left(x^{(i)}, z_0^{(i)}\right) + \lambda\alpha_c\left(y_0^{(i)}, y^{(i)}\right)\right) \quad (4)$$

where N is the number of training data points from all children. The parameter $\lambda \in (0, 1)$ was chosen to balance the network's generative power (feature decoding) and discriminative power (output estimation) and was optimized using validation data (yielding $\lambda = 0.8$). The learned $f_{\theta_0^{(e)}}(\cdot)$ was applied to the input features x, and the resulting code h_0 was then combined with each child's CARS ($x_{\text{cars}} \in \mathcal{R}^{15\times1}$) as $h_1 = [h_0; x_{\text{cars}}]$. This new feature representation was used as input to the subsequent layers in the PPA-net.

Group-level network

We first trained the GPA-net, where all network layers are shared among the children (Fig. 5E). GPA-net weights were used to initialize the PPA-net, followed by the proposed fine-tuning strategies to personalize it (network personalization). The former step is important because each layer below the culture level in the PPA-net uses only a relevant subset of the data (e.g., in C1, data of two females are present below the gender layer), resulting in fewer data to train these layers. This, in turn, could easily lead to overfitting of the PPA-net, especially of its child-specific layers, if only the data of a single child were used to learn their weights. To this end, we used a supervised layer-wise learning strategy, similar to that proposed in recent deep learning works (*44*, *45*). The central idea is to train the layers sequentially and in a supervised fashion by optimizing two layers at a time: the target hidden layer and its CoF.

We defined two operators in our learning strategy: learn and nest (Fig. 5). The learn operator is called when simultaneously learning the hidden and CoF layers. For the hidden layers, we used the rectified linear unit (ReLU) (*6*), defined as $h_l = \max(0, W_l h_{l-1} + b_l)$, where $l = 1,\ldots,4$ and $\theta_l = \{W_l, b_l\}$. ReLU is the most popular activation function that provides a constant derivative, resulting in fast learning and preventing vanishing gradients in deep neural networks (*6*). The combined AE output and CARS (h_1) were fed into the fusion ($l = 1$) layer, followed by the culture ($l = 2$), gender ($l = 3$), and individual ($l = 4, 5$) layers, as depicted in Fig. 5, where each CoF is a fully connected LaF with parameters $\theta_l^{(c)} = \{W_l^{(c)}, b_l^{(c)}\}$. The optimal parameters of the lth layer $\omega_l = \{\theta_l, \theta_l^{(c)}\}$ were found by minimizing the loss:

$$\omega_l^* = \underset{\omega_l}{\operatorname{argmin}}\, \alpha_c(h_l, y) = \underset{\omega_l}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\alpha_c\left(y_{l+1}^{(i)}, y^{(i)}\right) \quad (5)$$

where α_c is the MSE loss (feature fusion and autoencoding) computed between the output of the ReLU layer (h_{l+1}) passed through the LaF layer of the CoF (y_{l+1}) and the true outputs (y).

When training the subsequent layer in the network, we used the nest operator (Fig. 5) in a similar fashion as in (*44*) to initialize the parameters as

$$\theta_{l+1} = \left\{W_{l+1} \leftarrow I + \epsilon, \quad b_{l+1} \leftarrow 0\right\}$$
$$\theta_{l+1}^{(c)} = \left\{W_{l+1}^{(c)} \leftarrow W_l^{(c)}, \quad b_{l+1}^{(c)} \leftarrow b_l^{(c)}\right\} \quad (6)$$

where the weight matrix W_{l+1} of the ReLU was set to an identity matrix (I). To avoid the network being trapped in a local minimum of the previous layer, we added a low Gaussian noise [$\epsilon_{i,j} = \mathcal{N}(0, \sigma^2), \sigma = 0.01$] to the elements of I. We set the parameters of the supervised linear layer using the weights of the CoF above, which assured that the network achieved similar performance after nesting of the new ReLU layer. Before we started training the nested layer, we "froze" all the layers above by setting the gradients of their weights to zero—a common approach in a layer-wise training of deep models (*44*). This allowed the network to learn the best weights for the target layer (at this stage). The steps learn and nest were applied sequentially to all subsequent layers in the network. Then, the fine-tuning of the network hidden layers and the last CoF was done jointly. We initially set the number of epochs to 500 with early stopping [i.e., training until the error on a validation set reaches a clear minimum (*6*, *50*) (~100 epochs)].

The network parameters were trained using the standard back-propagation algorithm (*6*), which learns how the model should change its parameters used to compute the representation in each layer from the representation in the previous layer. The loss of the AE layer and each pair of the ReLU/LaF(CoF) layers was minimized using the Adadelta gradient descent algorithm with learning rate $lr = 1$, 200 epochs, and a batch size of 100. The optimal network configuration had 396×250 and 250×3 hidden neurons (h) in the AE and its CoF layers, respectively. Likewise, the size of the fusion ReLU was 265, $(250 + 15) \times 200$, and 200×200 for all subsequent ReLU layers. The size of their CoF layers was 200×3.

Network personalization

To personalize the GPA-net, we devised a learning strategy that consists of three steps: the network initialization followed by two fine-tuning steps. For the former, we introduced a new operator, named clone, which widens the network to produce the architecture depicted in Fig. 2. Specifically, the AE ($l = 0$) and fusion ($l = 1$) layers were configured as in the GPA-net (using the same parameters). The clone operator was then applied to generate the culture, gender, and individual layers, with parameters initialized as

$$\begin{aligned} l = 2 &: \theta_l^{(q)} \leftarrow \theta_l^0, \quad q = \{C_1, C_2\} \\ l = 3 &: \theta_l^{(g)} \leftarrow \theta_l^0, \quad g = \{\text{male}, \text{female}\} \\ l = 4 &: \theta_l^{(k)} \leftarrow \theta_l^0, \quad k = \{1,..,K\} \\ l = 5 &: \theta_{l-1}^{(c,k)} \leftarrow \theta_{l-1}^{(c,0)}, \quad k = \{1,..,K\} \end{aligned} \quad (7)$$

As part of the clone procedure, the culture and gender layers were shared among the children, whereas the individual layers were child-specific.

To adapt the network parameters to each child, we experimented with different fine-tuning strategies. We report here a two-step fine-tuning strategy that performed the best. First, we updated the network parameters along the path to a target child while freezing the layers not intersecting with that particular path. For instance, for child k and demographics $\{C_1, \text{female}\}$, the following updates were made: $\omega_{l=1:5,k}^* = \left\{\theta_0, \theta_1, \theta_2^{(C_1)}, \theta_3^{(\text{female})}, \theta_4^{(k)}, \theta_5^{(c,k)}\right\}$. Practically, this was achieved by using a data batch of 100 random samples of the target child to compute the network gradients along that child path. In this way, the network gradients were accumulated across all the children and then back-propagated (1 epoch). This was repeated for 50 epochs, and the stochastic

gradient descent (SGD) algorithm (lr = 0.03) was used to update the network parameters. At this step, SGD produced better parameters than Adadelta. Specifically, because of its adaptive lr, Adadelta quickly altered the initial network parameters, overfitting the parameters of deeper layers for the reasons mentioned above. This, in turn, diminished the shared knowledge provided by the GPA-net. On the other hand, the SGD with the low and fixed lr made small updates to the network parameters at each epoch, allowing the network to better fit each child while preserving the shared knowledge. This was followed by the second fine-tuning step, where the child-specific layers $\left(\omega_k^* = \left\{\theta_3^{(k)}, \theta_3^{(c,k)}\right\}\right)$ were further optimized. For this, we used again Adadelta (lr = 1), one child at a time, and 200 epochs. We implemented the GPA-net/PPA-net using the Keras API (*51*) with a TensorFlow backend (*52*), on a Dell Precision workstation (T7910), with the support of two graphics processing units (GPUs) (NVIDIA GF GTX 1080 Ti). The paired t tests with unequal variances were performed to estimate ICC differences between PPA-net and GPA-net from 10 repetitions across random splits of the child data (see Effects of model personalization). The significance level was set to 0.05.

SUPPLEMENTARY MATERIALS

robotics.sciencemag.org/cgi/content/full/3/19/eaao6760/DC1
Note S1. Details on model training and alternative approaches.
Note S2. Data set.
Note S3. Feature processing.
Note S4. Data coding.
Fig. S1. Empirical cumulative distribution function of ICC and MSE.
Fig. S2. The learning of the networks.
Fig. S3. PPA-net: The performance of the visual (face and body), audio, and physiology features.
Table S1. Comparisons with alternative approaches.
Table S2. Summary of the child participants.
Table S3. Summary of the features.
Table S4. The coding criteria.
References (*53*–*56*)

REFERENCES AND NOTES

1. T. Kanda, H. Ishiguro, *Human-Robot Interaction in Social Robotics* (CRC Press, 2017).
2. D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot, R. el Kaliouby, AFFDEX SDK: A cross-platform real-time multi-face expression recognition toolkit, in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* [Association for Computing Machinery (ACM), 2016], pp. 3723–3726.
3. T. Fukuda, P. Dario, G.-Z. Yang, Humanoid robotics—History, current state of the art, and challenges. *Sci. Robot.* **2**, eaar4043 (2017).
4. L. D. Riek, Healthcare robotics. *Commun. ACM* **60**, 68–78 (2017).
5. C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, 2006).
6. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436–444 (2015).
7. M. J. Matarić, Socially assistive robotics: Human augmentation versus automation. *Sci. Robot.* **2**, eaam5410 (2017).
8. 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. *J. Behav. Robot.* **8**, 18–38 (2017).
9. D. Freeman, S. Reeve, A. Robinson, A. Ehlers, D. Clark, B. Spanlang, M. Slater, Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychol. Med.* 1–8 (2017).
10. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Association, 2013).
11. D. L. Christensen, D. A. Bilder, W. Zahorodny, S. Pettygrove, M. S. Durkin, R. T. Fitzgerald, C. Rice, M. Kurzius-Spencer, J. Baio, M. Yeargin-Allsopp, Prevalence and characteristics of autism spectrum disorder among 4-year-old children in the autism and developmental disabilities monitoring network. *J. Dev. Behav. Pediatr.* **37**, 1–8 (2016).
12. H. I. Krebs, J. J. Palazzolo, L. Dipietro, M. Ferraro, J. Krol, K. Rannekleiv, B. T. Volpe, N. Hogan, Rehabilitation robotics: Performance-based progressive robot-assisted therapy. *Auton. Robots* **15**, 7–20 (2003).
13. W. A. Bainbridge, J. W. Hart, E. S. Kim, B. Scassellati, The benefits of interactions with physically present robots over video-displayed agents. *Int. J. Soc. Robot.* **3**, 41–52 (2011).
14. S. Baron-Cohen, A. M. Leslie, U. Frith, Does the autistic child have a "theory of mind"? *Cognition* **21**, 37–46 (1985).
15. S. Harker, Applied behavior analysis (ABA), in *Encyclopedia of Child Behavior and Development* (Springer, 2011), pp. 135–138.
16. J. J. Diehl, L. M. Schmitt, M. Villano, C. R. Crowell, The clinical use of robots for individuals with autism spectrum disorders: A critical review. *Res. Autism Spectr. Disord.* **6**, 249–262 (2012).
17. B. Scassellati, H. Admoni, M. Matarić, Robots for use in autism research. *Annu. Rev. Biomed. Eng.* **14**, 275–294 (2012).
18. E. S. Kim, R. Paul, F. Shic, B. Scassellati, Bridging the research gap: Making HRI useful to individuals with autism. *J. Hum. Robot Interact.* **1** (2012).
19. K. Dautenhahn, I. Werry, Towards interactive robots in autism therapy: Background, motivation and challenges. *Pragmat. Cogn.* **12**, 1–35 (2004).
20. T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, R. Looije, M. Neerincx, Y. Demiris, R. Ros-Espinoza, A. Beck, L. Cañamero, A. Hiolle, M. Lewis, I. Baroni, M. Nalin, P. Cosi, G. Paci, F. Tesser, G. Sommavilla, R. Humbert, Multimodal child-robot interaction: Building social bonds. *J. Hum. Robot Interact.* **1**, 33–53 (2012).
21. J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, A. Paiva, Automatic analysis of affective postures and body motion to detect engagement with a game companion, in *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE, 2011), pp. 305–311.
22. 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 *14th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)* (IEEE, 2017), pp. 39–44.
23. S. M. Anzalone, S. Boucenna, S. Ivaldi, M. Chetouani, Evaluating the engagement with social robots. *Int. J. Soc. Robot.* **7**, 465–478 (2015).
24. M. B. Colton, D. J. Ricks, M. A. Goodrich, B. Dariush, K. Fujimura, M. Fujiki, Toward therapist-in-the-loop assistive robotics for children with autism and specific language impairment. *Autism* **24**, 25 (2009).
25. J. Hernandez, I. Riobo, A. Rozga, G. D. Abowd, R. W. Picard, Using electrodermal activity to recognize ease of engagement in children during social interactions, in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (ACM, 2014), pp. 307–317.
26. M. E. Hoque, Analysis of speech properties of neurotypicals and individuals diagnosed with autism and down, in *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (ACM, 2008), pp. 311–312.
27. O. Rudovic, J. Lee, L. Mascarell-Maricic, B. W. Schuller, R. W. Picard, Measuring engagement in robot-assisted autism therapy: A cross-cultural study. *Front. Robot. AI* **4**, 36 (2017).
28. P. E. Shrout, J. L. Fleiss, Intraclass correlations: Uses in assessing rater reliability. *Psychol. Bull.* **86**, 420–428 (1979).
29. T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: An open source facial behavior analysis toolkit, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (IEEE, 2016), pp. 1–10.
30. Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2017), pp. 1302–1310.
31. F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in openSMILE, the Munich open-source multimedia feature extractor, in *Proceedings of the 21st ACM international conference on Multimedia* (ACM, 2013), pp. 835–838.
32. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
33. Y. Bengio, L. Yao, G. Alain, P. Vincent, Generalized denoising auto-encoders as generative models. *Adv. Neural Inf. Process. Syst.* **26**, 899–907 (2013).
34. E. Schopler, M. E. Van Bourgondien, G. J. Wellman, S. R. Love, *The Childhood Autism Rating Scale–2 (CARS-2)* (Western Psychological Services, ed. 2, 2010).
35. Y. Zhang, Q. Yang, An overview of multi-task learning. *Nat. Sci. Rev.* **5**, 30–43 (2017).
36. S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, R. Picard, Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Trans. Affect. Comput.* (2017).
37. M. I. Jordan, T. M. Mitchell, Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
38. A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not just a black box: Learning important features through propagating activation differences. http://arxiv.org/abs/1605.01713 (2016).
39. J. H. Friedman, Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001).
40. R. Picard, M. Goodwin, Developing innovative technology for future personalized autism research and treatment. *Autism Advocate* **50**, 32–39 (2008).

41. B. W. Schuller, *Intelligent Audio Analysis* (Springer, 2013).
42. R. Salakhutdinov, J. B. Tenenbaum, A. Torralba, Learning with hierarchical-deep models. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1958–1971 (2013).
43. A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE. Trans. Affect. Comput.* (2017).
44. T. Chen, I. Goodfellow, J. Shlens, Net2Net: Accelerating learning via knowledge transfer, paper presented at International Conference on Learning Representations (ICLR 2016), Caribe Hilton, San Juan, Puerto Rico, 2 to 4 May 2016.
45. C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, Z. Tu, Deeply-supervised nets. *Artif. Intell. Stat.* 562–570 (2015).
46. B. Settles, Active learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **6**, 1–114 (2012).
47. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
48. S. Chen, Y. Li, N. M. Kwok, Active vision in robotic systems: A survey of recent developments. *Int. J. Rob. Res.* **30**, 1343–1377 (2011).
49. G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
50. After this, we also tried fine-tuning the last layer only; however, this did not affect the network performance.
51. F. Chollet, Keras (2015); https://github.com/fchollet/keras.
52. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A system for large-scale machine learning, in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* (USENIX, 2016), pp. 265–283.
53. 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).
54. J. A. Hadwin, P. Howlin, S. Baron-Cohen, *Teaching Children with Autism to Mind-Read: Workbook* (John Wiley & Sons, 2015).
55. E. Paul, *Facial Expressions* (John Wiley & Sons Ltd., 2005).
56. J. F. Cohn, F. De la Torre, Automated face analysis for affective computing, in *The Oxford Handbook of Affective Computing* (Oxford Univ. Press, 2015).

Acknowledgments: We thank J. Hernandez and M. Schmitt for help with processing of physiological/audio data and MIT students J. Busche and S. Malladi for their help with running the experiments. We also thank A. Lazarevic for assistance in preparing the figures. Our special thanks go to the children who participated in this study and their parents and to Serbian Autism Society and Autism Society of Japan for their assistance in data collection. **Funding:** This work was supported by Grant-in-Aid for Young Scientists B, grant no. 16K16106, Chubu University grant no. 27IS04I, and EU HORIZON 2020 grant nos. 701236 (EngageME) and 688835 (DE-ENIGMA). **Author contributions:** O.R. and R.W.P. conceived the personalized ML framework. O.R. derived the proposed deep learning method. M.D. and O.R. implemented the method and conducted the experiments. J.L. supported the data collection, processing, and analysis of the results. B.S. provided insights into the method and audio data processing. All authors contributed to the writing of the paper. **Competing interests:** R.P. is a cofounder of and owns shares in Affectiva and Empatica Inc., which make affect-sensing technologies (E4 sensor). B.S. is a cofounder of and owns shares in Audeering Inc., which makes audio-sensing technologies (openSMILE). The other authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions are included in the paper or the Supplementary Materials. Please contact O.R. for data and other materials.

Submitted 3 February 2018
Accepted 6 June 2018
Published 27 June 2018
10.1126/scirobotics.aao6760

Citation: 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).

Science Robotics

Personalized machine learning for robot perception of affect and engagement in autism therapy

Ognjen RudovicJaeryoung LeeMiles DaiBjörn SchullerRosalind W. Picard

Sci. Robot., 3 (19), eaao6760. • DOI: 10.1126/scirobotics.aao6760

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.