HUMANOID ROBOTS

Human-robot facial coexpression

Yuhang Hu[1]*, Boyuan Chen[2,3,4], Jiong Lin[1], Yunzhe Wang[5], Yingke Wang[5], Cameron Mehlman[1], Hod Lipson[1,6]*

Large language models are enabling rapid progress in robotic verbal communication, but nonverbal communication is not keeping pace. Physical humanoid robots struggle to express and communicate using facial movement, relying primarily on voice. The challenge is twofold: First, the actuation of an expressively versatile robotic face is mechanically challenging. A second challenge is knowing what expression to generate so that the robot appears natural, timely, and genuine. Here, we propose that both barriers can be alleviated by training a robot to anticipate future facial expressions and execute them simultaneously with a human. Whereas delayed facial mimicry looks disingenuous, facial coexpression feels more genuine because it requires correct inference of the human's emotional state for timely execution. We found that a robot can learn to predict a forthcoming smile about 839 milliseconds before the human smiles and, using a learned inverse kinematic facial self-model, coexpress the smile simultaneously with the human. We demonstrated this ability using a robot face comprising 26 degrees of freedom. We believe that the ability to coexpress simultaneous facial expressions could improve human-robot interaction.

INTRODUCTION

Few gestures are more endearing than a smile. But when two people smile at each other simultaneously, the effect is amplified: Not only is the feeling mutual but also for both parties to execute the smile simultaneously, they likely are able to correctly infer each other's mental state in advance. This cognitive affirmation further cements the emotional bond by establishing that the two parties are "on the same wavelength" (*1*–*3*). Social alignment behaviors, such as simultaneous smiling, are important for successful social interactions because they indicate mutual understanding and shared emotions (*4*–*6*). Put simply, if a smile is simultaneous, then it is more likely to be genuine (*7*).

Facial expressions have been widely studied in various fields such as psychology, neuroscience, and robotics. For some facial gestures, observation of a facial movement of others will inadvertently generate spontaneous similar facial movements (*8*–*11*). For example, the atmosphere created by two people smiling at the same time can often reflect the harmony and sincerity of the communication (*12*). However, it is essential to note that this mirroring is not universal. In cases of social misalignment, contrasting facial reactions can emerge, such as responding to anger with fear (*13*). The imperceptible, subtle synchronization of some expressions is an ability that may carry substantial evolutionary advantage because it fosters social cohesion and mutual understanding—both crucial for group survival (*14*). In everyday interactions, if an individual displays a delayed smile while others smile in unison, then it might be perceived as disingenuous or submissive.

Across different ages and ethnic and cultural backgrounds, people frequently express similar states of mind through similar facial movements (*15*). However, note that there is substantial evidence showing cultural variance in both the display and the perception of facial expressions. Although facial movements might be universally recognized to some extent, the way they are expressed and interpreted can differ across cultures (*16*, *17*). In addition, cross-ethnic and cross-age differences can further influence the perception of these expressions. For instance, younger individuals might interpret certain facial cues differently than older individuals, and people from one ethnic background might perceive facial expressions differently when displayed by someone from another ethnic background (*18*, *19*). It is often posited that facial expressions frequently mirror inner emotions, leading to the near-simultaneous occurrence of expression formation and emotion experience (*20*, *21*). However, recent studies, including that of Barrett *et al.* (*22*, *23*), suggest that this relationship is not always straightforward and can vary on the basis of numerous factors. Therefore, we recognize that the following work only scrapes the surface of a very complex and powerful mode of human-robot interaction (*24*, *25*).

In the realm of human-robot interaction, we believe that the importance of anticipatory facial expressions is paramount. Currently, most robots can only perceive human emotions and respond after the human has finished making expressions (*26*). Such reactive expressions lack the authenticity and immediacy that come with anticipatory expressions. Robots that are limited to mimicking human expressions after they occur cannot fully integrate into human social environments because the delay in response is perceived as artificial and unrelatable.

For robots to be perceived as genuine and emotionally intelligent, they must be capable of anticipatory facial expressions. This is particularly critical for smiles, which play an outsized role in social bonding. As shown in Fig. 1C, there is a stark contrast between a mimicry smile and an anticipatory smile. An anticipatory smile, generated with the understanding and prediction of the other party's emotional state, is vital for creating genuine human-robot emotional bonds. Anticipatory models in robots can bring human-robot interactions closer to human-human interactions, bridging the gap in social communication and leading to more integrated and emotionally intelligent robotic systems.

[1]Creative Machines Laboratory, Department of Mechanical Engineering, Columbia University, New York, NY 10027, USA. [2]Mechanical Engineering and Materials Department, Duke University, Durham, NC 27708, USA. [3]Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. [4]Department of Computer Science, Duke University, Durham, NC 27708, USA. [5]Department of Computer Science, Columbia University, New York, NY 10027, USA. [6]Data Science Institute, Columbia University, New York, NY, 10027, USA.
*Corresponding author. Email: yuhang.hu@columbia.edu (Y.H.); hod.lipson@columbia.edu (H.L.)

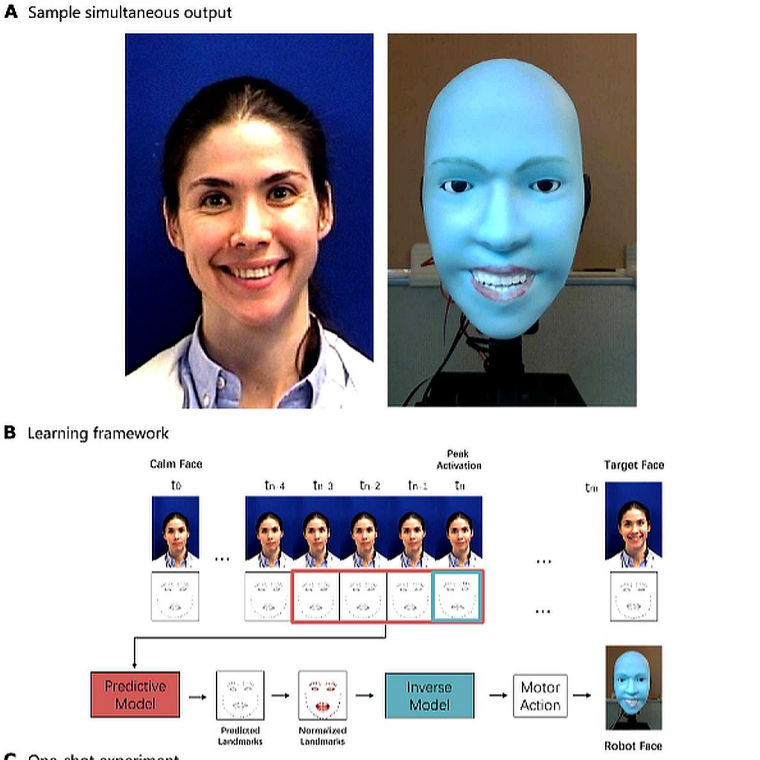

Fig. 1. Facial coexpression process. (A) Sample simultaneous output. **(B)** Depiction of the overall pipeline. The human face is in a calm (baseline) state at time t_0, and the expression at time t_n is when the facial expression changes with the greatest acceleration. The future target face (t_m) has the maximum difference from the calm face. After detecting the peak activation, the landmarks extracted from t_0 to t_n are concatenated as the input of the predictive model. The inverse model takes the normalized facial landmarks as input and outputs a set of motor commands for controller execution. **(C)** The top row illustrates the process of facial coexpression, where the robot uses anticipatory models to produce facial expressions simultaneously with the human participant. The bottom row shows the mimicry baseline, in which the robot generates a facial expression identical to the human's but with a noticeable delay. Each row contains a sequence of four snapshots capturing the progression of the expression from initiation to completion. This visual representation highlights the synchronization and authenticity achieved through anticipatory facial expressions as compared with the delayed response in the mimicry baseline. For a dynamic view of this interaction, please refer to the video demonstration in movie S1.

Humans can generate thousands of different facial expressions to convey countless and nuanced emotional states, and this ability is one of the most potent and effective interfaces in human social interaction (*27*). During the coronavirus disease 2019 epidemic, masks made social interaction awkward because they obscured facial expressions. At the same time, remote conferencing became more effective when video cameras were turned on (*28*–*30*). In a similar vein, once robots can reveal rich three-dimensional (3D) facial expressions, they can enhance their communication ability and be more conducive to building trust with humans.

Although robots have advanced notably in the past few years because of developments in artificial intelligence, there has been relatively little progress in the field of facial robots. Face animatronics require complex hardware and software design. Although past work has resulted in impressive humanlike face robots, these rely mostly on preprogrammed facial animations (*31*–*37*). These expressions are usually carefully preprogrammed, tuned, and choreographed rather than being spontaneous. Recent developments in facial robotics have focused on diversifying and improving dynamic facial expressions of emotion (*38*, *39*), which is a step toward creating more humanlike interactions.

Our previous robot platform Eva was an early example of a robot with the capability to self-model its own facial expressions (*1*). However, to achieve more convincing social interaction, the robot must learn to predict not just its facial expression but also that of the conversant agent with whom it communicates.

Here, we present an anthropomorphic facial robot called Emo, which features notable hardware improvements compared with Eva. Emo is equipped with 26 actuators (Fig. 2) that offer more degrees of freedom, allowing for asymmetric facial expressions, as opposed to the 10 actuators for Eva's face. One of the key differences in Emo's design is the use of direct-attached magnets to deform the replaceable face skin instead of the cable-driving mechanism (Bowden cable) used in Eva. This approach provides more precise control over facial expressions. Furthermore, Emo features embedded cameras in its eyes, enabling humanoid visual perception. These high-resolution RGB (red, green, blue) cameras, one inside the pupil of each eye, enhance the robot's ability to interact with its environment and better predict the conversant's facial expressions. In addition to these hardware upgrades, we introduce a learning framework consisting of two neural networks—one to predict Emo's own facial expression (the self-model) and another to predict the conversant's facial expression (the conversant model). Our soft-skinned human face robot has 23 motors dedicated to controlling facial expressions and three motors for neck movement. Overall, these improvements make Emo a substantially different and more advanced facial robot compared with its predecessor, Eva. We also propose an upgraded inverse model that can allow the robot to generate motor commands more than five times faster than the previous generation on the same computing hardware. We demonstrate a predictive model that can anticipate the target facial expressions of the conversant in real time. By combining the self-model and the anticipatory conversant model, the robot can perform coexpression. Our method generalizes more than 45 human participants. Last, we present how to use both models to achieve simultaneous human-robot expression on our physical robot.

RESULTS

A face robot design

Our results were achieved using our anthropomorphic facial robot with 26 actuators and interchangeable soft facial skin (Fig. 2). The entire facial skin was fabricated from silicone and attached to the robotic hardware using 30 magnets (Fig. 2A). The robotic facial skin can be replaced with alternative designs for a different appearance and for skin maintenance. For example, in Fig. 2B, we changed the

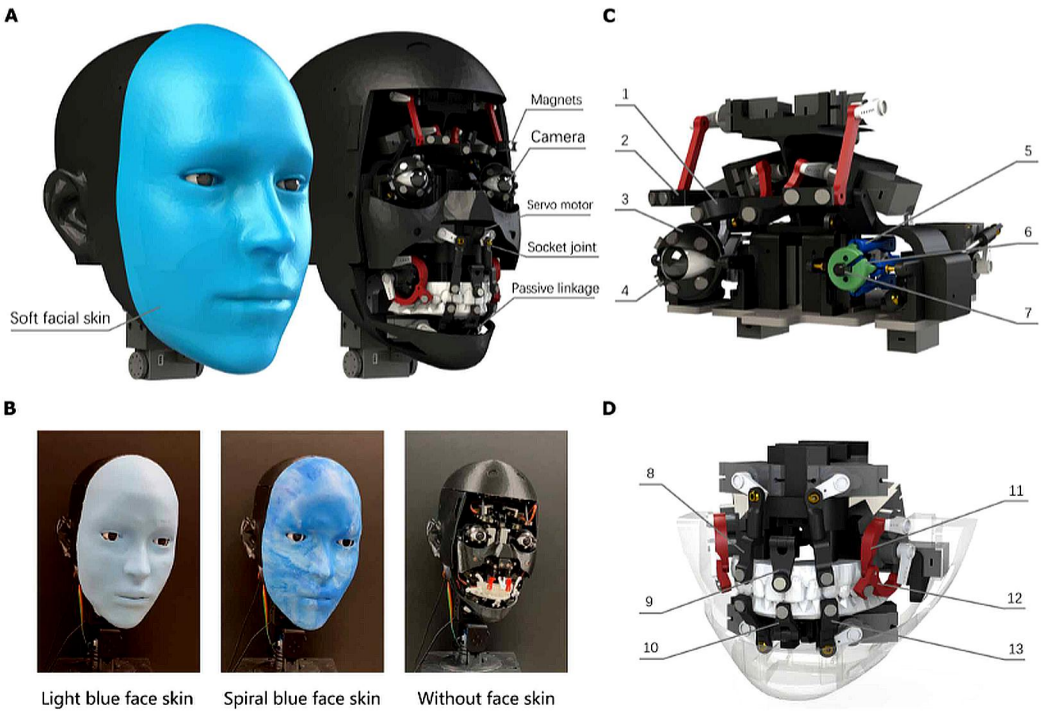

Fig. 2. Robot face platform. **(A)** Design overview. Our face robot contains 26 motors and uses position control. The soft face skin can easily connect to the hardware mechanism by magnets. Three motors control neck movements in three axes (roll, pitch, and yaw). Twelve motors control the upper face including eyeballs, eyelids, and eyebrows. Eleven motors control the mouth mechanism and the jaw. **(B)** The magnetic connection design allows the robot's facial skin to be easily replaced. **(C)** Eyes module. (1 and 2) Linkages with magnet connections control eyebrows. (3) Upper eyelid. (4) Lower eyelid. (5) Eyeball linkage. (6) Eyeball frame. (7) Camera. **(D)** Mouth module. (8 to 10 and 13) Mouth passive linkages. (11 and 12) Linkages of 2D five-bar mechanism.

robot face from a light blue face to a darker shade of blue with spiral marks. The robot consists of three subassembly modules: two eye modules, a mouth module, and a neck module.

The eye module controls the movements of eyeballs, eyebrows, and eyelids, as shown in Fig. 2C. Each eye frame is imbued with a high-resolution RGB camera. The eye frame is separately actuated by two motors through a parallelogram mechanism in two axes of pitch and yaw. The merit of this design is that it creates more space in the center of the eye frame, allowing us to mount the camera module in its natural location corresponding to the human pupil. This design facilitates more natural face-to-face interactions between robots and humans. It also allows for correct and natural gaze, a key aspect of nonverbal communication, especially at close range. A video demonstration of our robot with cameras in its eyes to track people's faces is provided in movie S2.

The movements of the mouth are complex. Whereas most animatronic robot faces typically exhibit only simple jaw motion, we aimed to replicate the complex motion of human lips through a mechanical structure. To solve this challenge, we designed several passive joints and linkages so that while the robot mouth moves, the soft skin can flex over the passive degrees of freedom of the mechanical structure to form complex yet natural-looking deformations. The mouth module contains nine kinematic chains shown in Fig. 2D. Six of them with passive joints control the upper and lower lips. Two five-bar linkages control the movements of the mouth corner, and the final linkage mechanism controls the movement of the jaw. We provide a video to demonstrate the movement of the mouth module and the robot's entire hardware in movie S3.

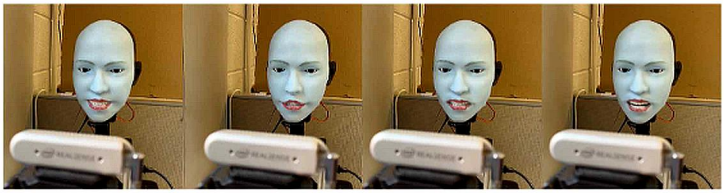

Fig. 3. Data collection. The robot looks at itself through a camera to learn to make facial expressions. We set up an RGB camera in front of the robot and random-sampled the motor actions to actuate the face robot. The motor actions are constrained to avoid self-collision or tearing of the soft face skin. Using this process, the robot learns the relationship between motor commands and facial appearance, without human supervision.

Inverse model for generating robot expressions

We propose a self-supervised learning process to train our face robot to generate human facial expressions without explicit choreography of the motion and without human labels. The traditional method of controlling robots relies on kinematic equations and simulations, but this only applies to rigid-body robots with known kinematics. Our robot has soft deformable skin and several passive mechanisms with four socket joints, so it is difficult to obtain the kinematic equation of the robot kinematics. We overcame this challenge by leveraging a visual-based self-supervised learning method in which the robot can learn the relationship between motor commands and the resulting facial expression by observing itself in a mirror.

The facial expressions of the robot are controlled by 19 motors, of which 18 are symmetrically distributed, and one motor controls the jaw movement. In our case, the expressions in the facial dataset are all symmetrical; therefore, the symmetrically distributed motors can share the same motor commands when controlling the robot. Consequently, the actual control commands only need 11 parameters normalized to the range [0, 1].

The facial inverse model was trained using the dataset generated by the robot itself (Fig. 3), which consisted of motor commands and the resulting facial landmarks. We collected the data using a process of random "motor babbling" in a self-supervised manner. Before sending the commands into the controller, the process automatically removes motor commands that may tear the face skin or cause self-collision. After the servo motors reached the target position defined by the commands, we captured robot face images with an RGB camera and extracted the robot's facial landmarks.

Given the dataset of motor commands and facial landmarks, we then aimed to train an inverse model that, given facial landmarks, can generate the corresponding motion commands. The inverse model is formed by several layers of multilayer perceptrons that implicitly capture the robot's face morphology, elasticity, and kinematics. Each datum tuple consists of entire face landmarks set, represented by a vector of 113×2 size and the corresponding 11 motor values. We collected 1000 data points, of which 200 were used for validation and the remaining 800 were used for training. Because the eye module movement in the upper half of the robot's face is relatively independent of the movement of the mouth module in the lower half, the overall training dataset can be separated into two parts. The training data can be augmented by extracting the upper half facial landmarks (52×2) and the lower half facial landmarks (61×2), respectively, from two separate sub-datasets and combining them together to form augmented data.

We evaluated the effectiveness of the inverse model by comparing our method with three baselines. The first baseline is to generate motor commands randomly, and the second is to sample and compare commands from the training dataset randomly. Both baselines use random selection but in different distributions because the commands that we used to generate the inverse model dataset are modified by constraint functions. The third baseline is the nearest neighbor. It compares the landmarks with the training dataset and directly uses the command of the closest landmark as the output. We used the L1 metric to measure the distance of motor commands normalized to [0, 1]. Figure 4A shows a boxplot of our inverse model evaluation. The inverse model generates motor commands that result in facial expressions that are more accurate than the three baselines. Our model successfully learns the relationship between the motor commands and the soft facial skin morphology and elasticity.

Predictive model for expression anticipation

For the robot to achieve authentic and timely facial expressions, it must anticipate facial expressions in advance, giving its mechanical apparatus sufficient time to actuate. To accomplish this, we developed a predictive facial expression model and trained it with a video dataset of humans making expressions. The model is able to predict the target expression that a person will produce on the basis of just initial and subtle changes in their face.

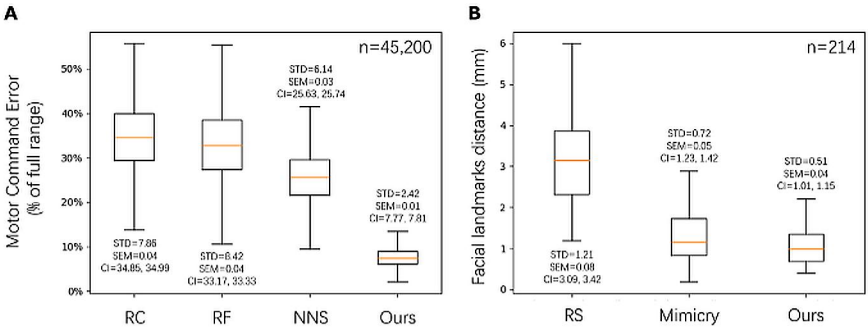

Fig. 4. Evaluations. **(A)** We compared the performance of the inverse model with other baselines using 45,200 samples. **(B)** Two baselines [random search (RS) and mimicry baseline] were used to evaluate the effect of our predictive model. We tested 214 different expressions and used facial landmarks to measure the errors. For both (A) and (B), we conducted a detailed statistical analysis, including the calculation of SD, SEM, and 95% confidence intervals (CI). The precise values in the boxplots and histogram plots for data distribution are shown in the Supplementary Materials.

First, we quantified the facial expression dynamics using the Euclidean distance between each set of facial landmarks and the facial landmarks of the initial ("resting") facial expression in each video. We defined the resting facial landmarks as the mean landmarks of the first five frames and the target facial landmarks as those that maximally differ from the resting facial landmark. The Euclidean distance between the resting facial landmarks and the landmark from other frames continuously changes and can be differentiated. Therefore, we can calculate the trend of expression change through the second derivative of the landmark's distance with respect to time. The details of data collection for the predictive model and training process are provided in the Supplementary Materials. We used the video frame at the moment when the expression change acceleration is the largest as the "peak activation."

To improve the accuracy and avoid overfitting, we augmented each datum by sampling the surrounding frames. Specifically, during the training procedure, the input of the predictive model is an arbitrary set of four frames from a total of nine frames before and after the peak activation. Similarly, the label is randomly sampled from the four frames after the target face. The dataset contains 45 human participants and 970 videos in total. Eighty percent of the data were used for training the model, and the rest was used for validation. We analyzed the entire dataset and obtained the average time that humans normally take to make a facial expression as 0.841 ± 0.713 s. The prediction model and inverse model (referring solely to the processing speed of the neural network models used in our paper) can run about 650 frames per second (fps) and 8000 fps, respectively, on a MacBook Pro 2019 without a GPU device. This frame rate does not include data capture or landmark extraction times. Our robot can successfully predict the target human facial expressions and generate the corresponding motor commands within 0.002 s. This timing leaves about 0.839 s to capture facial landmarks and execute the motor commands to produce the target facial expression on the physical robot face.

To quantitatively evaluate the accuracy of predictive facial expressions, we compared our methods with two baselines. The first baseline randomly selects one picture in the inverse model training dataset as the prediction. The dataset of this baseline contains a large number of pictures of robot expressions generated by motor babbling. The second baseline is a mimicry baseline that selects the facial landmarks at the peak activation as the predicted landmarks. If the peak activation is close to the target face, then the baseline can be very competitive with our method. However, the results of experiments demonstrate that our method is better than this baseline, showing that the predictive model successfully learns to predict the future target face by generalizing subtle changes in the face instead of simply copying the facial expression in the last input frames. Figure 4B shows the quantitative evaluation of the predictive model. We computed the mean absolute error between the predicted landmarks and the ground truth landmarks, which consist of human target facial landmarks with dimensions of 113×2. The table results (table S2) demonstrate that our method outperforms both baselines, exhibiting a smaller mean error and a narrower standard error.

Combining the self-model with the anticipatory model

The final step in the process involves using both the predictive model and the inverse model together to achieve human-robot facial simultaneous expressions. This task is not the same as face mimicry because the predictive model does not observe the target face, so the task is first to predict the facial expression and then to swiftly generate the predicted facial expression.

The overall pipeline described in Fig. 1B shows how the robot simultaneously generates the same facial expression as the human participant by predicting the human target facial expression first on the basis of intermediate frames and then producing action commands using the inverse model in the remaining time before the target expression appears.

We ran both models on a MacBook Pro 2019 (Intel Core i9) and sent the motor commands to the robot controller for execution. The entire pipeline runs at 25 Hz. We designed the models to be very lightweight, so our robot does not need to rely on GPU computing or high-performance servers. This allows extra computing power to be used for other functions in the future, such as listening, thinking, and speaking.

We conducted an experiment by running both our method and the mimicry baseline on our physical robot. The comparison of both

methods is shown in diagrams in Fig. 1C and table S2. In this experimental setup, the timeline begins at $t = 0$, which marks the starting point when both the robot and human face begin the expression process. When $t = n$, this represents the moment the robot detects the peak activation and begins to predict human facial expressions. The objective is to achieve simultaneous human-robot facial expressions when $t = m$, with m denoting the target time by which the robot aims to have matched the human's facial expression.

We conducted experiments on videos of various human participants shown in Fig. 5 with different facial expressions. The performance was calculated from the testing dataset used to train the predictive model. The columns with four sequential frames are the input frames observed by the predictive model. The target face column is the target facial expression that the robot does not perceive. Ground truth pictures are the normalized target facial landmarks directly put into the inverse model to produce the robot's target face. We actuated motor commands on our physical robot and took robot front-face pictures, as shown in the actual final robot face column, to demonstrate that our approach successfully allowed the robot to learn to predict the human target face using only the initial subtle changes in the face. The results also demonstrate that our learning framework generalizes across various human participants and diverse facial expressions.

Evaluation of facial expression prediction using a confusion matrix

To further evaluate the performance of our robot in predicting facial expressions, we constructed a confusion matrix on the basis of the prediction commands for the facial expressions. The primary task here was to predict the commands that would generate target facial expressions. Given that the commands are normalized between 0 and 1, representing the activation of muscles on the robot's face, we can classify each command as activated or not. Each facial expression is generated by a set of 11 motor commands, with each command representing the actuation of muscles on the robot's face. Our dataset comprised 214 test samples, resulting in a total population of 2354 commands.

We set a calm facial muscle as the reference point and defined positive samples as those where the L1 distance between the target face commands and the calm face commands is greater than 0.25. The choice of L1 distance and the threshold of 0.25 is informed by the normalization of the commands to a range of 0 to 1; thus, the calm face with a range of ±0.25 covers a half region, which is 0.5. Conversely, negative samples are defined as those where this distance is within 0.25. If the predicted commands fall in the same region as the target commands, then they are considered true; otherwise, they are classified as false.

Figure 6 visually illustrates the prediction process and comparison with ground truth in four representative cases. Figure 6A showcases a true positive case where the robot correctly predicted a big smile from a calm face. Figure 6B illustrates a false positive case where the robot inaccurately predicted a smile even though the ground truth showed calmer facial muscles. Figure 6C represents a false negative case where the robot failed to predict a facial expression that was present. Last, Fig. 6D shows a true negative case, where the robot accurately predicted that there was no substantial deviation from the calm face. Table 1 shows that our model correctly predicts the activation of facial muscles for expressions in approximately 72.2% of the cases. The high positive predictive value suggests that when our model predicts that a muscle will be activated, it is correct 80.5% of the time. However, the false omission rate of 0.462 and the false positive rate of 0.446 suggest that there is room for improvement in minimizing both the false negatives and false positives.

Given that the commands in our setup are normalized to a range between 0 and 1, we designated the calm face as the origin. In this normalized space, we used the L1 distance to measure the deviation between the target face's commands and the origin calm face's commands. The L1 distance was chosen because of its interpretability and sensitivity to changes in the command dimensions. To differentiate between positive and negative samples, we established a threshold for the L1 distance. We selected a threshold of 0.25, ensuring that the calm face region, which is considered within ±0.25 of the origin, constitutes half of the total normalized range (0.5). This number can easily classify notable changes as activated commands.

DISCUSSION

We present a face robot with soft anthropomorphic face skin and its controller capable of performing simultaneous expressions by predicting human facial expressions. The overall pipeline consists of two neural networks: the predictive model and the inverse model. We demonstrated the effectiveness of both models by quantitative evaluation in comparison with other baselines. The results show that our predictive model successfully anticipates diverse human target facial expressions and can generate the predicted facial expression long enough in advance to allow the mechanical apparatus sufficient time to actuate.

It is imperative to acknowledge that discretion must be exercised in the selection of facial expressions for the robot to mimic. Certain

Table 1. Confusion matrix of facial expression prediction. Confusion matrix summarizing the performance of our facial expression prediction model within a total population of 2354 instances. It quantifies the model's accuracy in predicting the activation of facial muscles during expressions, highlighting a success rate of approximately 72.2%. Acc., accuracy. The table includes several key metrics: positive predictive value (PPV), false omission rate (FOR), false positive rate (FPR), sensitivity (SEN), likelihood ratio–positive (LR^+).

		Predicted condition		
Total population = 2354		**Positive**	**Negative**	
Actual condition	Positive	1307	338	SEN = 0.795
	Negative	316	393	FPR = 0.446
	Acc. = 0.722	PPV = 0.805	FOR = 0.462	LR+ = 1.782

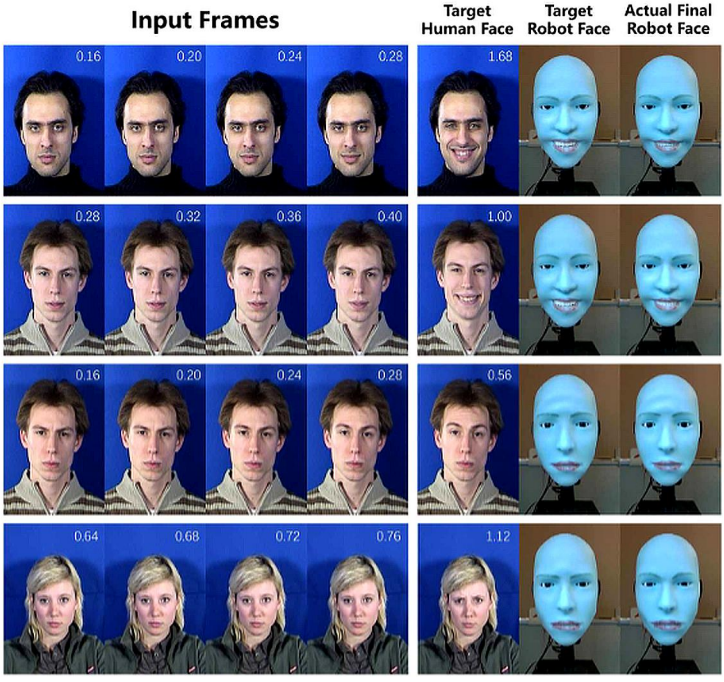

Fig. 5. Result visualization. The landmarks of the four consecutive frames on the left were input into the predictive model. The facial landmarks of target face frames are the labels of the predictive model. We derived the motor commands to produce the ground truth face by directly inputting the facial landmarks of the human target face into the inverse model. The number on the upper right in each picture is the time stamp. The prediction pictures are the results of the entire pipeline by observing only the four input frames. Additional examples of these results can be found in fig. S4.

facial gestures such as smiling, nodding, and maintaining eye contact are typically reciprocated naturally and are perceived positively in human communication (*40*, *41*). Conversely, the mimicking of expressions such as pouting or frowning should be approached with caution because these could potentially be misconstrued as mockery or convey unintended sentiments (*42*).

It is worth noting, however, that in certain contexts, mimicking such expressions can be strategically used for humor or diffusion of tense situations (*43*–*45*). Moreover, in future work, it will be important to consider that the genuineness of smiles is not solely determined by their anticipatory nature but also involves specific facial movements, such as the Duchenne marker, and temporal features, like sustainment and decay rate (*46*).

Our key contribution is the development of both robot hardware and learning algorithms for enabling anticipatory facial expressions. Although the effectiveness of the proposed method has been validated quantitatively using standard facial tracking metrics, we recognize that the ultimate measure of success is how these expressions are perceived by human users. An essential future step is to validate the emotional effect of these expressions in real-world human-robot interactions in various contexts to determine their psychological validity. This is an area of focus for our upcoming research.

In addition, one of the limitations of the current study is the potential lack of cultural sensitivity in the model's predictions and mimicking of expressions. Different cultures may have varying norms and meanings associated with certain facial expressions (*47*). For instance, although a smile is generally considered a sign of happiness or friendliness in many cultures, it can also be a sign of embarrassment or uncertainty (*48*). Similarly, direct eye contact might be perceived as a sign of confidence and honesty in some cultures but can be considered rude or confrontational in other cultures (*49*). Future work could explore integrating cultural context into the model, possibly by incorporating datasets from various cultural backgrounds and incorporating an understanding of cultural norms in the algorithm.

We also acknowledge that facial mimicry alone, even when done simultaneously, is far from capturing the full range of facial communication ability of humans and may even feel distasteful when performed by an adult-looking robot. However, just like infants learn to imitate their parents before graduating to make independent facial expressions, we believe that robots must learn to anticipate and mimic human expressions as a first step before maturing to more spontaneous and self-driven expressive communication (*50*).

Potential Influence on other fields

The potential influence of this research extends beyond robotics to fields such as neuroscience and experimental psychology. In the field of neuroscience, the study of mirror neurons provides a relevant example. Mirror neurons are brain cells that fire both when an animal acts and when the animal observes the same action performed by another (*51*). These neurons have been implicated in understanding other people's actions, imitating behavior, and empathy (*52*, *53*). A robotic system that can predict and synchronize facial expressions can be used as a tool for studying the mirror neuron system (*54*). By interacting with participants while measuring brain activity, researchers can gain insights into the neural correlates of social interaction and communication (*55*).

In experimental psychology, understanding facial expressions is crucial, for instance, in education and therapy for individuals with autism spectrum disorder (ASD). People with ASD often have difficulty interpreting facial expressions (*56*). Robots with the ability to predict and synchronize facial expressions could be used as educational tools to help individuals with ASD develop better social communication skills. Studies have shown that robots can be effective in engaging children with ASD and promoting social interaction (*57*).

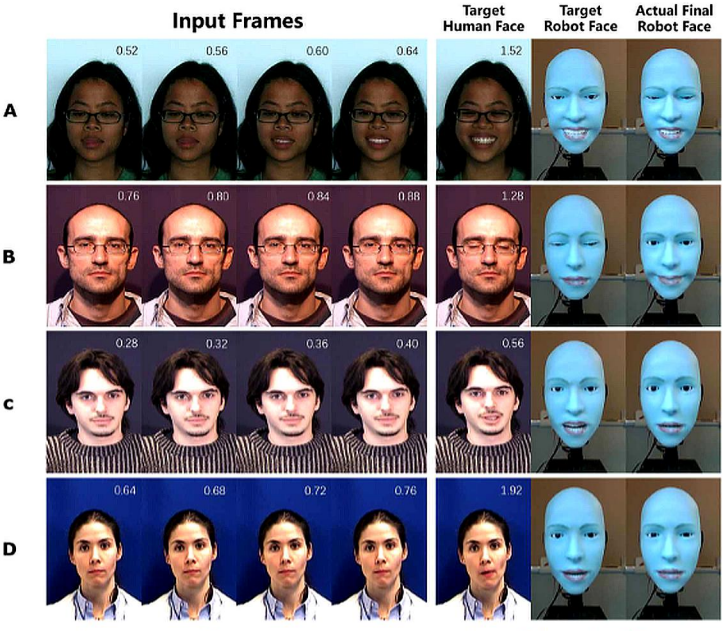

Fig. 6. Visual representation of four cases. (**A**) True positive—correct prediction of a big smile; (**B**) false positive—incorrect prediction of a smile; (**C**) false negative—failure to predict an actual facial expression; (**D**) true negative—correct prediction of no notable deviation from the calm face.

Being able to predict and recognize emotions through facial expressions is also crucial for empathy (*58*). Empathy, in turn, is a fundamental component of effective communication and maintaining social relationships (*59*). Thus, we believe that the ability to perceive people's emotions before their expressions is an essential first step toward building robots that are more socially adept (*60*). In this study, we focus on developing a robotic face capable of anticipatory facial expressions, laying the groundwork for more authentic human-robot interactions. Understanding and optimizing this interaction paves the way for potential applications in therapy, education, and everyday communication.

Ethical considerations

Last, as we reflect on the advancements in robotic facial expressions, we remain cognizant of the ethical dimensions associated with this technology. As robots evolve in their capacity to emulate humanlike behaviors, they gain the potential to forge stronger connections with humans (*61*). Although this capability heralds a plethora of positive applications, ranging from home assistants to educational aids, it is incumbent on developers and users to exercise prudence and ethical considerations (*62*). The potential for misuse of such technology, such as deception or manipulation, underscores the need for robust ethical frameworks and governance to ensure that these innovations are aligned with the values and well-being of society (*63*, *64*).

METHODS

Data representation

We extracted facial landmarks from pictures using Mediapipe (*65*) because landmarks had lower dimensions than raw images and were robust in describing facial features across genders and ethnicities. To realize the natural human-robot interactions, the robot needed to have a high-speed response at a rate similar to that of a human. Reducing the dimension of observation helped reduce the time required for computation and prevented people from losing patience and interest because of time-consuming data calculation. In our work, we selected 113 landmarks of the 468 original facial landmarks to represent facial expressions.

Inverse model training

The inverse model took in the robot landmarks or normalized human landmarks to generate motor commands. The movement command consisted of 11 numbers: two for the eyebrows, two for the eyelids, six for the mouth, and one for the jaw. We normalized all the motor action values to the [0, 1] range. The input landmark matrix of dimension 113×2 represented the positions of the 113 face points on the x and y axes. Given input landmarks, the inverse model outputted 11 motor values aimed at imitating those landmarks on the robot's face.

The inverse model had three fully connected layers (Fig. 7A). The activation function used in the first two layers is a bilinear rectified linear unit, and the last layer is Sigmoid. The model was optimized with a mean square error loss using the Adam optimizer (*66*) and a 10^{-6} learning rate.

During the data collection phase, the robot generated symmetrical facial expressions, which we thought could cover most of the situation and reduce the size of the model. We used an Intel RealSense D435i to capture RGB images and cropped them to 480 pixels by 320 pixels. We logged each motor command value and robot image to form a single data pair without any human labeling.

We used the setup described above to create two datasets, each consisting of 1000 robot facial expressions. In the first set, the robot only moved the eyes and eyebrows; in the second dataset, the robot only moved the mouth. This separation was conducted because the top half and bottom half of the robot face can be actuated independently, giving rise to 1,000,000 combinations in total. This augmentation method improved the efficiency of our training model process and also prevented the robot from overusing the motors and causing hardware problems. All the datasets to train the inverse model and the trained model are provided in the Supplementary Materials. We divided each group of 1000 expressions into 800 pairs for training and 200 for validation.

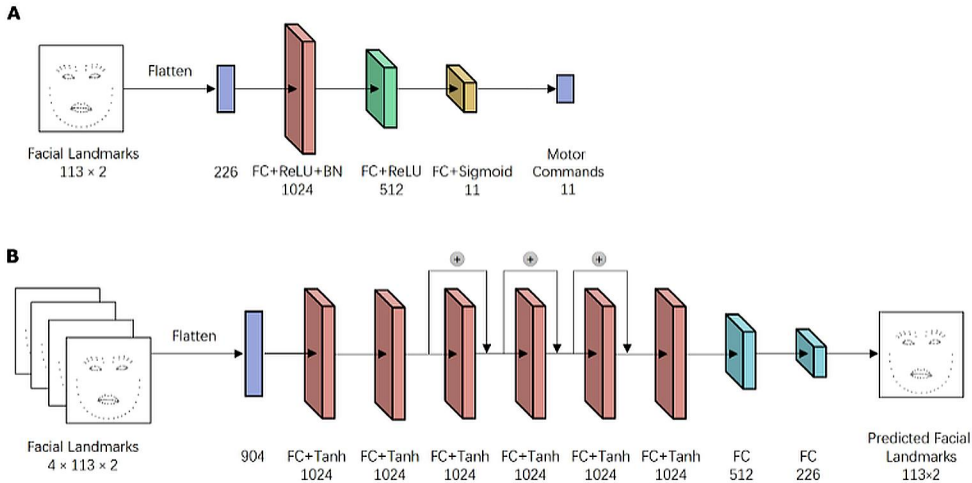

Fig. 7. Model architecture. (**A**) Inverse model architecture. A flattened input datum from the facial landmarks in pixel coordinates was transmitted through three fully connected layers (FCs). The first FC output was fed into the rectified linear unit (ReLU) and batch normalization (BN) processing before continuing to transmit to the next FC (*72*). The last FC layer used the Sigmoid activation function to map the resulting values between 0 and 1. (**B**) Predictive model architecture. The input was flattened data with a size of 4 × 113 × 2 consisting of four groups of facial landmarks. The model had eight FC layers and outputted 226 size data that could be reshaped to a group of facial landmarks. The output of the first six FCs entered the Tanh activation function before entering the next layer. The FC layer between 4 and 6 has a skip connection, which could amplify the output of the previous layer to preserve the information of the original data.

Predictive model training

The predictive model generated predicted target landmarks on the basis of a sequence of landmarks after the peak activation. The predictive model was a residual neural network with eight fully connected layers optimized by a mean square error loss and Adam optimizer (Fig. 7B). We used a learning rate of 10^{-5} and a batch size of 128 during the training procedure.

We constructed our dataset to train a predictive model using the MMI Facial Expression database (*67*, *68*). This database contained 2900 videos of 75 human participants ranging in age from 19 to 62 making 79 series of expressions. The human participants were of European, Asian, or Hispanic/Latino ethnicity. It is important to note that although this dataset provides a range of facial expressions and some diversity in participant ethnicity, it does not encompass a comprehensive representation of all global ethnicities. The selection of the training dataset was constrained by the capabilities of our robot hardware. For instance, facial expressions like pouting, sticking the tongue out, and puffing out the cheeks were not achievable by our robots, and we manually removed such data to form a dataset that was more representative of our robot's capabilities.

Among the 970 videos we selected, 756 videos were used for training and 214 videos were used for testing. We chose to split the videos from each participant in an 80:20 ratio for training and validation because of the varied number of videos provided by each participant (for instance, participant #18 provided 83 videos, but participant #25 provided only 2). This method allowed us to ensure a balanced and representative distribution of data for training and validation, resulting in more stable and reliable model performance. In our Supplementary Materials, we performed two additional fivefold cross-validation tests using different data-splitting methods to assess the performance of our model. After splitting the data into training and testing sets, we extracted landmarks from the expressions of four frames before and after the peak activation. This resulted in a single input datum of size 9 × 113 × 2, representing the concatenated landmarks from multiple frames. Each label datum with a size of 4 × 113 × 2 was composed of the landmarks extracted from the target face frame and the subsequent three frames. When training the predictive model, we sampled four sets of landmark data from each input data and one set of landmark data from label data to form a data pair. With this data augmentation method, we could theoretically construct 1,629,600 pairs of data. The inverse model and the predictive model were neural networks implemented with Pytorch (*69*). We provide all the details regarding training for the predictive model including the dataset and the trained model in the Supplementary Materials.

Predictive model data generation

We trained our predictive model with human facial expression videos. In this section, we describe how to generate the data for training the model automatically in detail. Because we used facial landmarks to represent the human face, we could quantify the change in the human face by calculating the distance between facial landmarks in each frame with the resting facial landmarks. In Fig. 7A, we leveraged a Savitzky-Golay filter (*70*) to smooth the raw data curve. The frame corresponding to the peak of this smoothed curve was identified as the target face, which had the maximum deviation from the resting face. Then, we calculated the second

derivative of the processed curve to depict the acceleration of face change. The maximum value of this new curve was the peak activation, as shown in Fig. 8B. To improve the data efficiency and make the performance more robust, we sampled the data around peak activation as input data and target face as label data.

Normalization algorithm

The motion space of the robot was different from that of the human, so to make the model trained by the robot face data meet the input data of the human face, we needed a normalization process that maps human facial landmarks to robot facial landmarks as shown in fig. S1. This was necessary because the range of motion of the person might exceed that of the robot. Furthermore, the normalization process could yield more general results because the human facial motion space was different among humans. We could obtain landmarks in robot space L_R by normalizing human landmarks L_H using the following equations

$$\text{if } L_H - H_s > 0: \quad L_R = \text{Min}((R_{\max} - R_s), (L_H - H_s)) \times K + R_s \quad (1)$$

$$\text{if } L_H - H_s > 0: \quad L_R = \text{Min}((R_{\max} - R_s), (L_H - H_s)) \times K + R_s \quad (2)$$

where H_s, R_s, $R_{\max}$, and $R_{\min}$ represent the human and robot resting facial expression and the robot value range of the spatial position of the landmark in the inverse model dataset. K is the scaling factor that adjusts the proportionality of the landmark mapping. This normalization approach differs from our previous work (*71*), where the calculations of human moving range and robot moving range are unnecessary, so the robot can directly do normalization without collecting such priors (human facial moving range) from human participants. It will be more accurate for robots to learn human expressions because the expressions of robots and humans will be on the same scale.

Ethics approval and consent to participate

This study involved the use of identifiable images of the first author, Y.H., who has explicitly consented to the publication of these images without anonymization. Because the study did not involve additional human participants directly, the requirement for broader ethics approval was not applicable to this work.

Regarding the use of the MMI Facial Expression database, we complied with the requirements set forth by the database's end user license agreement (EULA) (*68*). The database is freely available to the academic scientific community for noncommercial use under the condition that users register and agree to the EULA. This agreement explicitly allows for the academic use of imagery contained within the database, including publication and presentation, provided that the participants depicted have granted their permission for such uses. The database can be accessed at http://mmifacedb.com, and it requires users to register and agree to the EULA before gaining access to the materials.

By exclusively using images for which explicit consent has been obtained (from the first author) and relying on an open-source database compliant with ethical research standards, this study upholds the principles of consent and ethics in academic research.

Statistical analysis

We used Python (version 3.9) for statistical analysis. We compared the performance of our inverse model against three baselines: random commands, random face, and nearest neighbor search,

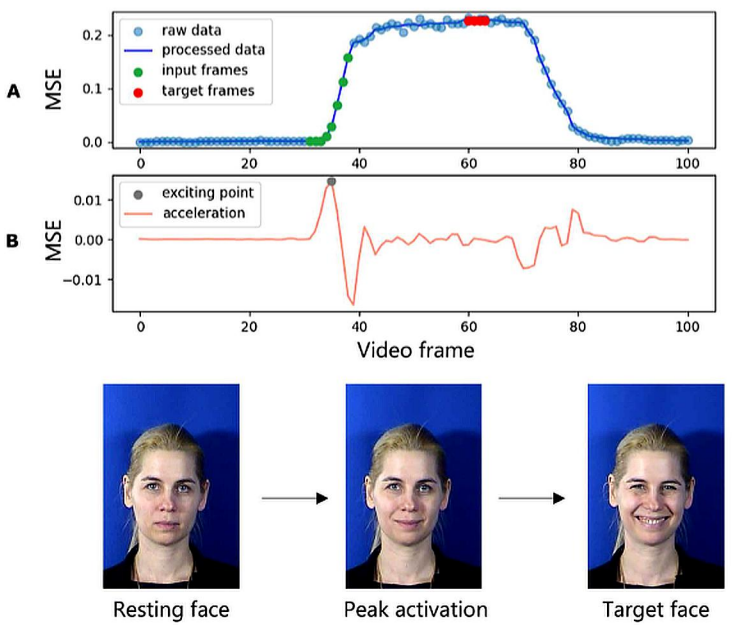

Fig. 8. Data for training predictive model. (**A**) The raw data consisted of the distances between facial landmarks in each frame and those of the resting face, calculated using the mean square error (MSE) method. The facial landmarks of the resting face were the mean of the five initial frames. After smoothing the raw data, we got processed data as a blue curve. When training the predictive model, we sampled the input data from the input frames shown with green points on the curve and sampled one of the target frames as the label. (**B**) The acceleration of face change. The peak activation at the gray point was the maximum value of this curve.

using a dataset comprising 45,200 samples. Similarly, we assessed the effect of our predictive model in comparison with two baselines, random search and mimicry, using 214 distinct expressions measured against facial landmarks. For both scenarios, we detailed the statistical analysis by calculating the SD, SEM, and 95% confidence intervals to ensure a robust evaluation of model performance. We conducted a *t* test to compare the mean prediction errors between our method and the mimicry baseline. We used a significance level set at 0.05 to determine whether there was a statistically significant difference between the means of these two independent groups. To assess the generalizability and robustness of our predictive model, we used fivefold cross-validation tests executed in two distinct manners: based on video samples and based on participants. In addition, we constructed a confusion matrix to assess the model's ability to predict facial expression commands, using L1 distance to classify the activation of facial muscles.

Supplementary Materials

This PDF file includes:
Supplementary Methods
Figs. S1 to S4
Tables S1 to S5
References (*73*, *74*)

Other Supplementary Material for this manuscript includes the following:
Movies S1 to S4
MDAR Reproducibility Checklist

REFERENCES AND NOTES

1. C. Frith, Role of facial expressions in social interactions. *Philos. Trans. R. Soc. B Biol. Sci.* **364**, 3453–3458 (2009).
2. T. L. Chartrand, J. L. Lakin, The antecedents and consequences of human behavioral mimicry. *Annu. Rev. Psychol.* **64**, 285–308 (2013).
3. U. Hess, A. Fischer, Emotional mimicry as social regulation. *Pers. Soc. Psychol. Rev.* **17**, 142–157 (2013).
4. S. G. Shamay-Tsoory, J. Aharon-Peretz, D. Perry, Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. *Brain* **132**, 617–627 (2009).
5. C. D. Frith, U. Frith, The neural basis of mentalizing. *Neuron* **50**, 531–534 (2006).
6. S. Garrod, M. J. Pickering, Joint action, interactive alignment, and dialog. *Top. Cogn. Sci.* **1**, 292–304 (2009).
7. E. G. Krumhuber, A. S. R. Manstead, Can Duchenne smiles be feigned? New evidence on felt and false smiles. *Emotion* **9**, 807–820 (2009).
8. E. J. Moody, D. N. McIntosh, L. J. Mann, K. R. Weisser, More than mere mimicry? The influence of emotion on rapid facial reactions to faces. *Emotion* **7**, 447–457 (2007).
9. L. K. Bush, C. L. Barr, G. J. McHugo, J. T. Lanzetta, The effects of facial control and facial mimicry on subjective reactions to comedy routines. *Motiv. Emot.* **13**, 31–52 (1989).
10. U. Dimberg, Facial reactions to facial expressions. *Psychophysiology* **19**, 643–647 (1982).
11. D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, J. L. Wilbarger, When the social mirror breaks: Deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Dev. Sci.* **9**, 295–302 (2006).
12. T. L. Chartrand, J. A. Bargh, The chameleon effect: The perception–behavior link and social interaction. *J. Pers. Soc. Psychol.* **76**, 893–910 (1999).
13. A. Moors, P. C. Ellsworth, K. R. Scherer, N. H. Frijda, Appraisal theories of emotion: State of the art and future development. *Emot. Rev.* **5**, 119–124 (2013).
14. J. L. Lakin, V. E. Jefferis, C. M. Cheng, T. L. Chartrand, The chameleon effect as social glue: Evidence for the evolutionary significance of nonconscious mimicry. *J. Nonverbal Behav.* **27**, 145–162 (2003).
15. C. Darwin, P. Prodger, *The Expression of the Emotions in Man and Animals* (Oxford Univ. Press, 1998).
16. R. E. Jack, Culture and facial expressions of emotion. *Vis. Cogn.* **21**, 1248–1286 (2013).
17. R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, P. G. Schyns, Facial expressions of emotion are not culturally universal. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 7241–7244 (2012).
18. N. C. Ebner, M. Riediger, U. Lindenberger, FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behav. Res. Methods* **42**, 351–362 (2010).
19. R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, R. Caldara, Cultural confusions show that facial expressions are not universal. *Curr. Biol.* **19**, 1543–1548 (2009).
20. P. Ekman, Facial expression and emotion. *Am. Psychol.* **48**, 384–392 (1993).
21. A. J. Fridlund, *Human Facial Expression: An Evolutionary View* (Academic Press, 2014).
22. L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, S. D. Pollak, Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychol. Sci. Public Interest* **20**, 1–68 (2019).
23. E. Krumhuber, A. Kappas, Moving smiles: The role of dynamic components for the perception of the genuineness of smiles. *J. Nonverbal Behav.* **29**, 3–24 (2005).
24. C. L. Breazeal, "Sociable machines: Expressive social exchange between humans and robots," thesis, Massachusetts Institute of Technology, Cambridge, MA (2000).
25. L. F. Barrett, M. Lewis, J. M. Haviland-Jones, *Handbook of Emotions* (Guilford Publications, 2016).
26. X. Feng, Y. Wei, X. Pan, L. Qiu, Y. Ma, Academic emotion classification and recognition method for large-scale online learning environment—Based on A-CNN and LSTM-ATT deep learning pipeline method. *Int. J. Environ. Res. Public Health* **17**, 1941 (2020).
27. R. L. Birdwhistell, *Kinesics and Context: Essays on Body Motion Communication*, (University of Pennsylvania Press, 2010).
28. J. N. Bailenson, Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technol. Mind Behav.* **2**, 10.1037/tmb0000030 (2021).
29. E. Peper, V. Wilson, M. Martin, E. Rosegard, R. Harvey, Avoid Zoom fatigue, be present and learn. *NeuroRegulation* **8**, 47–56 (2021).
30. K. A. Karl, J. V. Peluchette, N. Aghakhani, Virtual work meetings during the COVID-19 pandemic: The good, bad, and ugly. *Small Group Res.* **53**, 343–365 (2022).
31. H. S. Ahn, D.-W. Lee, D. Choi, D.-Y. Lee, M. Hur, H. Lee, Designing of android head system by applying facial muscle mechanism of humans, in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)* (IEEE, 2012), pp. 799–804.
32. W. T. Asheber, C.-Y. Lin, S. H. Yen, Humanoid head face mechanism with expandable facial expressions. *Int. J. Adv. Robot.* **13**, 29 (2016).
33. H. Kobayashi, F. Hara, Facial interaction between animated 3D face robot and human beings, in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation* (IEEE, 1997), pp. 3732–3737.
34. D. Loza, S. Marcos Pablos, E. Zalama Casanova, J. Gómez García-Bermejo, J. L. González, Application of the FACS in the design and construction of a mechatronic head with realistic appearance. *J. Phys. Agents* **7**, 31–38 (2013).
35. K. Berns, J. Hirth, Control of facial expressions of the humanoid robot head ROMAN, in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems* (IEEE, 2006).
36. T. Hashimoto, S. Hitramatsu, T. Tsuji, H. Kobayashi, Development of the face robot SAYA for rich facial expressions, in *2006 SICE-ICASE International Joint Conference* (Elsevier, 2006), pp. 3119–3124.
37. T. Hashimoto, S. Hiramatsu, H. Kobayashi, Dynamic display of facial expressions on the face robot made by using a life mask, in *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots* (IEEE, 2008), pp. 521–526.
38. C. Chen, O. G. B. Garrod, J. Zhan, J. Beskow, P. G. Schyns, R. E. Jack, Reverse engineering psychologically valid facial expressions of emotion into social robots, in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (IEEE, 2018).
39. C. Chen, L. B. Hensel, Y. Duan, R. A. A. Ince, O. G. B. Garrod, J. Beskow, R. E. Jack, P. G. Schyns, Equipping social robots with culturally-sensitive facial expressions of emotion using data-driven methods, in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)* (IEEE, 2019), pp. 1–8.
40. J. J. Gross, Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology* **39**, 281–291 (2002).
41. M. Stel, R. Vonk, Mimicry in social interaction: Benefits for mimickers, mimickees, and their interaction. *Br. J. Psychol.* **101**, 311–323 (2010).
42. N. H. Frijda, *The Emotions* (Cambridge Univ. Press, 1986).
43. A. Niculescu, B. van Dijk, A. Nijholt, H. Li, S. L. See, Making social robots more attractive: The effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* **5**, 171–191 (2013).
44. J. L. Lakin, T. L. Chartrand, Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychol. Sci.* **14**, 334–339 (2003).
45. R. A. Martin and T. Ford, *The Psychology of Humor: An Integrative Approach* (Academic Press, 2018).
46. R. Song, H. Over, M. Carpenter, Young children discriminate genuine from fake smiles and expect people displaying genuine smiles to be more prosocial. *Evol. Hum. Behav.* **37**, 490–501 (2016).
47. H. A. Elfenbein, N. Ambady, On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychol. Bull.* **128**, 203–235 (2002).
48. D. Matsumoto, T. Kudoh, American-Japanese cultural differences in attributions of personality based on smiles. *J. Nonverbal Behav.* **17**, 231–243 (1993).
49. M. Argyle, L. Lefebvre, M. Cook, The meaning of five patterns of gaze. *Eur. J. Soc. Psychol.* **4**, 125–136 (1974).
50. A. N. Meltzoff, M. K. Moore, Explaining facial imitation: A theoretical model. *Infant Child Devel.* **6**, 179–192 (1997).

51. G. Rizzolatti, L. Craighero, The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–192 (2004).
52. M. A. Arbib, From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics. *Behav. Brain Sci.* **28**, 105–124 (2005).
53. R. Mukamel, A. D. Ekstrom, J. Kaplan, M. Iacoboni, I. Fried, Single-neuron responses in humans during execution and observation of actions. *Curr. Biol.* **20**, 750–756 (2010).
54. D. Floreano, A. J. Ijspeert, S. Schaal, Robotics and neuroscience. *Curr. Biol.* **24**, R910–R920 (2014).
55. L. M. Oberman, V. S. Ramachandran, The simulating social mind: The role of the mirror neuron system and simulation in the social and communicative deficits of autism spectrum disorders. *Psychol. Bull.* **133**, 310–327 (2007).
56. M. B. Harms, A. Martin, G. L. Wallace, Facial emotion recognition in autism spectrum disorders: A review of behavioral and neuroimaging studies. *Neuropsychol. Rev.* **20**, 290–322 (2010).
57. P. Pennisi, A. Tonacci, G. Tartarisco, L. Billeci, L. Ruta, S. Gangemi, G. Pioggia, Autism and social robotics: A systematic review. *Autism Res.* **9**, 165–183 (2016).
58. S. Baron-Cohen, S. Wheelwright, The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J. Autism Dev. Disord.* **34**, 163–175 (2004).
59. J. Decety, P. L. Jackson, The functional architecture of human empathy. *Behav. Cogn. Neurosci. Rev.* **3**, 71–100 (2004).
60. C. Breazeal, Emotion and sociable humanoid robots. *Int. J. Hum. Comput. Stud.* **59**, 119–155 (2003).
61. P. H. Kahn Jr., T. Kanda, H. Ishiguro, N. G. Freier, R. L. Severson, B. T. Gill, J. H. Ruckert, S. Shen, "Robovie, you'll have to go into the closet now": Children's social and moral relationships with a humanoid robot. *Dev. Psychol.* **48**, 303–314 (2012).
62. R. R. Murphy, D. D. Woods, Beyond Asimov: The three laws of responsible robotics, in *Machine Ethics and Robot Ethics*, W. Wallach, P. Asaro, Eds. (Routledge, 2020), pp. 405–411.
63. R. Calo, Artificial intelligence policy: A roadmap. *UCDL Rev.* **51**, 399 (2017).
64. A. F. T. Winfield, M. Jirotka, Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Phil. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **376**, 20180085 (2018).
65. C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, M. Grundmann, Mediapipe: A framework for building perception pipelines. arXiv:1906.08172 [cs.DC] (2019).
66. D. P. Kingma, J. Ba, Adam: A method for stochastic optimization. arXiv:1412.6980 [cs.LG] (2014).
67. M. Pantic, M. Valstar, R. Rademaker and L. Maat, Web-based database for facial expression analysis, in *2005 IEEE International Conference on Multimedia and Expo* (IEEE, 2005), p. 5.
68. M. Valstar, M. Pantic, Induced disgust, happiness and surprise: An addition to the mmi facial expression database, in *Proceedings of the 3rd International Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect* (ELRA, 2010), pp. 65–70.
69. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 1 (2019).
70. A. Savitzky, M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
71. B. Chen, Y. Hu, L. Li, S. Cummings, H. Lipson, Smile like you mean it: Driving animatronic robotic face with learned models, in *2021 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2021), pp. 2739–2746.
72. A. F. Agarap, Deep learning using rectified linear units (relu). arXiv:1803.08375 [cs.NE] (2018).
73. M. Mori, K. F. MacDorman, N. Kageki, The uncanny valley [from the field]. *IEEE Robot. Autom. Mag.* **19**, 98–100 (2012).
74. L. I. Labrecque, G. R. Milne, Exciting red and competent blue: The importance of color in marketing. *J. Acad. Market. Sci.* **40**, 711–727 (2012).

Acknowledgments

Funding: This work was supported by the US National Science Foundation (NSF) AI Institute for Dynamical Systems (DynamicsAI.org) under grant 2112085 and Amazon grant through the Columbia Center of AI Technology (CAIT). **Author contributions:** H.L., Y.H., and B.C. proposed the research. Y.H. designed the robots. Y.H., H.L., B.C., Yunzhe Wang, and Yingke Wang designed the algorithms. Y.H., J.L., and C.M. fabricated the robots and conducted the physical experiments. Y.H. and J.L. performed the numerical experiments. Y.H., B.C., and H.L. analyzed the data. Y.H. and H.L. wrote the paper. All authors provided feedback. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to support the conclusions of this manuscript are included in the main text or Supplementary Materials. Contact X.X. for materials. We used part of the data from a public database: MMI Facial Expression Dataset (https://mmifacedb.eu/). The codebase and dataset of the work can be found at https://doi.org/10.5061/dryad.gxd2547t7.

Submitted 27 April 2023
Accepted 27 February 2024
Published 27 March 2024
10.1126/scirobotics.adi4724