Robotics: Science and Systems 2019
Freiburg im Breisgau, June 22-26, 2019

Predicting Human Interpretations of Affect and Valence in a Social Robot

David McNeill
Department of Computer Science
Boise State University
Boise Idaho, 83725
davidmcneill@u.boisestate.edu

Casey Kennington
Department of Computer Science
Boise State University
Boise Idaho, 83725
caseykennington@boisestate.edu

Abstract—**In this paper we seek to understand how people interpret a social robot's performance of an emotion, what we term 'affective display,' and the positive or negative valence of that affect. To this end, we tasked annotators with observing the Anki Cozmo robot perform its over 900 pre-scripted behaviors and labeling those behaviors with 16 possible affective display labels (e.g., interest, boredom, disgust, etc.). In our first experiment, we trained a neural network to predict annotated labels given multimodal information about the robot's movement, face, and audio. The results suggest that pairing affects to predict the valence between them is more informative, which we confirmed in a second experiment. Both experiments show that certain modalities are more useful for predicting displays of affect and valence. For our final experiment, we generated novel robot behaviors and tasked human raters with assigning scores to valence pairs instead of applying labels, then compared our model's predictions of valence between the affective pairs and compared the results to the human ratings. We conclude that some modalities have information that can be contributory or inhibitive when considered in conjunction with other modalities, depending on the emotional valence pair being considered.**

I. Introduction

As robots are employed in increasing numbers across industries, the expectation that these robots will interact naturally with people–lay people who will not understand how they work–is also increasing. However, people often assign anthropomorphic characteristics to robots, for example stereotypical gender [8], social categorizations [9], roles [32], age [26], as well as intelligence, interpretability, and sympathy [24] to the robots with which they interact. As noted in [23], affective expression and interpretation can be a highly effective method for coordinating actions between members of a team and, if correctly mapped to a robot's affordances, should reduce the ambiguity that exists in robot-human interactions [12] and facilitate shared tasks between robots and humans.

In this paper we explore how people perceive a robot's affective display based on multimodal information (i.e., how it looks, sounds, and behaves), without the context of a specific task, and despite what a designer might have intended for people to interpret from the programmed behaviors. For our work we used the Anki Cozmo robot. Cozmo is marketed as a toy robot for children, is small in size (see Figure 1) and inexpensive, and yet is useful as a research platform (e.g., [26]) because the SDK allows developers to access low-level information about the robot state and it has ample degrees of freedom. It includes animated eyes, a head and lift that can move up and down, track wheels that can be used to turn and move the robot forward or backward, and a speech synthesizer. Cozmo has a built-in camera and simple built-in object and facial detection software. Importantly for this work, Cozmo has 940 pre-scripted behaviors (termed *animations*) which, when invoked, produce movements, sounds, and facial animations that are easily observable by a person. Moreover, following [21] which notes that the more human-like a robot is, the more social engagement it receives, Cozmo's animated eyes, speech capabilities, lift (which resembles a human arm in some ways) and freedom of movement provide enough human-like qualities to permit an anthropomorphic interpretation of Cozmo's affective display.

In the sections that follow, we explain related work, how we tasked annotators with assigning 16 emotional labels to the 940 Cozmo animations, then compare the resulting annotations with (what we interpret to be) designer intent. We found that what designers intend is not what is perceived by others. We describe our data collection, analysis, and resulting dataset in Section III. We then perform three progressive experiments: in Experiment 1 (Section IV) given multimodal facial, movement, and audio information derived from the animations, we train a neural network to predict the annotated emotion labels, following on advances made in current state of the art human-robot interaction (HRI) research conducted by [22] and [11], both of which proved the efficacy of applying statistical models toward classifying social context and human emotions, respectively, by mapping a variety of multimodal low-level features collected from a robot to gold-standard label data. Our results show that treating this task as a standard labeling task is also potentially useful in predicting human interpretations of robot affective display.

In Experiment 2 (Section V) we follow the approach of [31], dividing our 16 emotion labels according to valence (i.e., positive or negative sentiment), and form 8 separate pairs of mutually exclusive positive and negative emotions. We then task a set of binary neural network classifiers to predict the valence of each pairing. The result was more informative and potentially useful in predicting human interpretations of robot behavior. This leads to Experiment 3, where we generated novel robot behaviors and tasked raters with assigning a scale to each valence pair and compared those ratings to our model predictions and conclude.

II. Related Work

Research in social robotics and human-robot interaction has explored how affect is displayed in human-robot interaction tasks; the focus, however, has largely been on human affect. For example, [17] explored how robots are perceived by different age groups, such as the elderly.

More relevant to our work are [24] and [26], both of which use facial expressions of the human participants to predict how the humans perceive robot intelligence and age, respectively. As in their work, we use multimodal features, but the features we focus on are derived from the robot and not the participants. Also related to our work is [15] which examined how empathy in a robot's speech can be interpreted by people; here we consider affects and modalities beyond speech. In particular, our work connects to multimodal aspects of human-robot interaction and learning, including grounded semantics [33, 16], engagement [3, 21], establishing common ground [5], interpreting intent behind robot movements [13], as well as learning verbal behaviors and action demonstrations [17]. We also relate to recent work [4] that focused on displaying curiosity in robots as a precursor to learning tasks, whereas here we extend our focus beyond just curiosity and look into many possible affective displays.

From [10] we took the importance of including face and movement information, which they showed made "a robot more compelling to work with." In [29] they showed that audio data was the most significant feature in a robot's ability to model "contingency" (the ability to detect an effect on the environment from its own actions). [31] also employed a multimodal approach in predicting between seven basic emotions, as well as overall valence and arousal in a robot. Their findings showed that movement was a better predictor than the LED light-strip which had been placed on the robot. Moreover, we approached our classification task by attempting more simple and powerful classification methods (e.g., see [22]), but found in our results that neural networks worked best to classify our data, as in the findings of [11], which calls for a more diverse and complex model of human emotion, which we attempt to address, at least in part here.

III. Data

In this section, we explain the data we collected and offer some analysis of that data. Our goal in this data collection is to better understand how people interpret the affective display of Cozmo as it performs its pre-scripted animations, and how those interpretations differ from what we understood to be the animation designer's intent.

A. Data Collection

For each of Cozmo's 940 available, pre-scripted animations, we recorded video and audio of the robot's behavior. For each recording, we position Cozmo in a starting position where it faced the camera, then initiated the animation. We kept the camera as close to Cozmo as possible while still recording the animations from within a single camera position (i.e., for some animations, Cozmo moved around, requiring wider camera coverage). An example of three frames derived from one of these video recordings is in Figure 1: though Cozmo does not appear to move, its eyes have the appearance of looking around and portraying boredom.

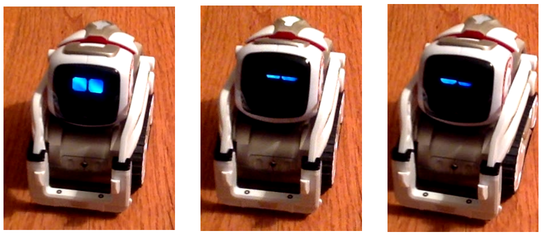

Fig. 1. Three example frames of a video recording of Cozmo for a *bored* animation.

We then posted these recordings (i.e., containing the audio and video) on Amazon Mechanical Turk with the following instructions for the workers: You will be shown a video of a small robot. Please describe what the robot is doing in the video, and provide a selection of the emotions that you think the robot is displaying (in this paper we only focus on the resulting emotion labels). Following [28], we used the following 16 emotions: *interest, alarm, confusion, understanding, frustration, relief, sorrow, joy, anger, gratitude, fear, hope, boredom, surprise, disgust, desire*. Taking note from [1] that there is no mutual exclusivity between emotions, we allowed workers to be able to select any number of these emotionss, thereby not constraining the number of emotions they could assign. However, we did not give them a free-form input so as to keep the task within reasonable constraints.[1] Each worker was paid $1.00 to describe and label 10 randomly assigned videos and could repeat the process for another set, if they desired. The emotion check boxes were arranged randomly.[2] Each animation recording was labeled by two workers.

As explored in [7] and [18], culture and language can influence the valence of equivalent emotions and may have skewed our workers' labels. However, one barrier to accepting our task was that workers had to demonstrate fluency in written English; also, according to [14], on the days we collected labels, the majority of our workers were based in the United States.[3] Additionally, studies demonstrating a cultural influence on emotional experience and representation are not conclusive, as noted in [20].

We gathered 1,870 labeled recordings (to ensure that each

[1] Though see [27] which showed that non-linguistic speech utterances can be interpreted categorically.

[2] Due to an oversight, about half of the tasks ended up not being randomly arranged, but the distribution of labels for those versus the ones that were randomized did not show any significant difference.

[3] On December 17, 2018, 71% of workers were in the United States, 22.92% were in India, and 5.21% were from other countries. On January 16, 2019, 75% of workers were in the United States, 12.50% in India, and 12.50% from other countries.

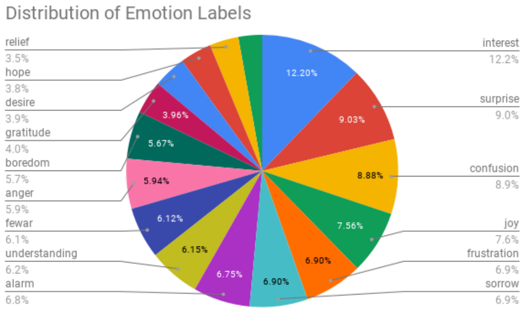

Fig. 2. Distribution of emotion labels as assigned by the workers.

worker received the same number of labels, some were labeled 3 times). Figure 2 shows the distribution over the labels. The most common label is *interest* at 12.2%, the least common is *disgust* at 2.82%, with a fairly uniform distribution over each of the 16 labels. We take this to mean that no single label was either over- or under-favored by the workers.

Figure 3 shows a count of the number of labels for each animation. For example, if an animation has a recording that the worker labeled as *surprise* and *disgust*, then that animation received a count of 2. Of the 1,870 labeling tasks, 1008 only received one label, 486 received 2, 165 received 3, and we found smaller counts for higher numbers of labels. From this we infer something important: while more than half of the workers assigned a single label to recordings, nearly half received more than one label. This is the first evidence of ambiguity in interpreting the affective display of a particular behavior.

To further measure the challenge that people have in interpreting the robot's affective display, we calculated inter-annotator agreement using Cohen's Kappa statistic [6]. As each recording received labels from two different workers, we treated the two workers as two different annotators with one important proviso: when two workers agreed on at least one label, we marked the two annotations as agreed upon. This resulted in a Kappa score of 0.26, which is considered in the "fair agreement" range. That this value is not below or equal to zero tells us that there is some agreement in how people perceive a robot's affective display.

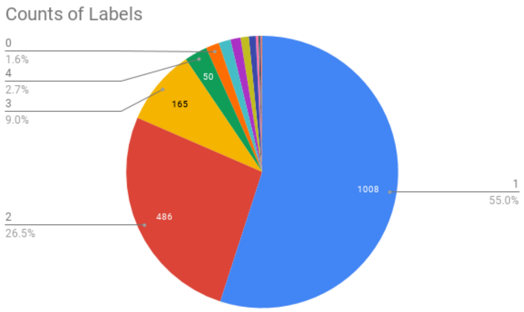

Fig. 3. Total counts of labels per task.

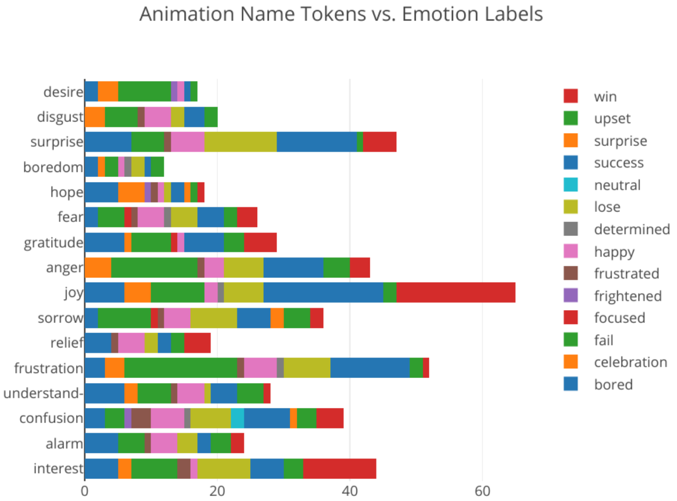

Fig. 4. Common designer name tokens compared with annotator emotion labels; lower tokens map to counts on the left side of the bars.

B. Data Analysis

Though it is not possible to fully recover the intent of the designers who created the animations, we can estimate the intended affective display of Cozmo from the designer-written animation names. Below are some examples of these animation names:

- `bored_event`
- `greeting_happy`
- `explorer_driving01_loop_01_head_angle`
- `rollblock_fail`

Note that some names have words that denote affective displays, while others only focus on the function of the animation and not how it might be interpreted as an emotion or affect. By taking the individual word tokens (i.e., between the underscores) we identified the common words that we interpreted to denote affective displays: *bored, celebration, fail, focused, frightened, frustrated, happy, determined, lose, neutral, success, surprise, upset, win*; 145 of the 940 animation names had at least one of these tokens in them.

We compared the annotator-labeled affects with the affective tokens from their corresponding animations. This comparison is shown in Figure 4, where the labels are on the y-axis and the count of tokens of intended affect interpretation is represented in the bars. In some cases there are clear analogs to the emotion list we used from [28], i.e., *bored=boredom*, *frightened=fear*, *frustrated=frustration*, *happy=joy*, and *surprise=surprise*, but even for those pairings, affect was interpreted in many different ways. The label for *surprise*, for example, was used to identify animations with *bored, fail, frustrated, happy, lose, success, upset*, and *win* tokens. In this case, *surprise* as a token in an animation name was never actually interpreted as *surprise* by the annotators. On the other hand, the token *win* was interpreted by workers as nearly every affective display (see the red/rightmost items in each bar).

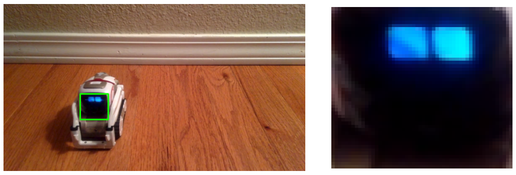

Fig. 5. Example of face tracking and the corresponding extracted face frame.

C. Data Modalities

In this paper, we explore three sources of information that are available to people when observing and interpreting the robot animations: (1) Cozmo's produced sounds, (2) facial animations, (3) and movements. For the experiments below, we obtained representations of each of these modalities. Obtaining Cozmo's produced sounds was straightforward: we extracted the audio from the recordings for each animation. The other two modalities required additional steps which we explain in the following subsections.

1) Movement via Internal State: The Cozmo SDK allows developers to obtain the internal state of the robot at any state change update event. Some examples are itemized below; the entire set of 47 state variables is listed in the Appendix:[4]

- `left_wheel_speed`
- `lift_position_height`
- `accelerometer_x`
- `gyro_x`

On average, animations had 73 state change updates with *sorrow*-labeled animations being the longest (92 on average), and *surprise*-labeled animations being the shortest (58 on average). For each change in the state of the robot, we recorded the entire state of the robot resulting in a sequence of state changes for each animation, which we used to represent movement over time.

2) Face Animations: The internal state updates do not include information about the state of the face. Cozmo's face display is an OLED (organic light-emitting diode) where the facial animations are pre-defined and inaccessible through the SDK. To obtain facial information for each animation, we passed the video recordings through a computer vision processing script that located the eyes by color (which was unique to the scenes in the recordings) and created a bounding box around them. An example of what this looked liked for a single frame is depicted in Figure 5; Cozmo is facing the camera, the script located the face (i.e., the eyes) and formed a bounding box, then extracted the contents of that bounding box into an individual face image. Processing each animation recording in this manner resulted in a sequence of face images, one for each frame where the face was found in the frame (i.e., there were some frames where Cozmo was not facing the camera, and therefore no face images were extracted).

D. Final Dataset

After the post-processing described above, the resulting dataset contains the following:

- audio/video recordings of each animation
- two text descriptions for each animation
- two sets of emotion labels for each animation
- extracted internal state update changes
- extracted eye/face images for each frame of each animation recording

IV. Experiment 1: Predicting Perceptions of Affective Display

In this experiment, we use the data we collected to train and evaluate a neural network model to predict what people perceive about the affective display of Cozmo's animations. This amounts to a standard classification task, a common approach for automation: the goal is to determine how well a classifier can predict at least one of the possible affect labels.

A. Task & Procedure

Given features that represent the animation (i.e., movement, audio, and facial features as described above), predict one or more of the 16 affect labels. We split our data into train, development, and evaluation sets as follows: for the data, we replicated animations according to the number of labels they had been assigned, assigning one label from that animation's label set to each of its replicates to allow for a one-to-one classification. For the development and evaluation sets, we removed any animations that were in the training data so any development or evaluation animation was never observed in training (recall that there were two sets of labels for each animation). This resulted in 1008 instances for training, 209 instances for development and 150 instances for evaluation.

1) Modalities & Features: To represent the animations, we use features from three different modalities:

a) movement: to represent the movement of Cozmo over time, we used the internal state updates. Each state update recorded a vector of 47 continuous feature values. The number of state updates varied for each animation; i.e., the vector of 47 values could change as little as once to as many as over a thousand event updates.

b) audio: to represent the sounds coming from Cozmo, we used the audio recordings of the animations and applied a short-time Fourier transform, chroma gram, mel-frequency cepstral coefficients, mel-scaled spectogram, spectral contrast, and tonal centroid features, using the librosa python library [19] taking the average of each. This resulted in 119 audio features for each animation.

c) face: to represent the face (i.e., animated eyes), we padded the extracted face images with black to match the required input size for the pre-trained VGG19 convoluational neural network which takes in an image at the bottom layer and outputs a softmax distribution over 1000 possible classes [30] and used that vector as a representation of the face. The VGG19 was trained on the ILSVRC-2012 data set which contains 1.3 million images grouped into 1000 classes (i.e., the images depicted individual entities such as an animal or an object). We used the development data to empirically determine parameters, including which pre-trained network and which layer of the model that we should use. This resulted

[4] We only considered variables that did not remain constant across all animations.

in a vector of 1000 dimensions for each face image for each animation (the shortest animation only had 85 face images; the longest had 8,129). Due to the limitations of our hardware, we were only able to use the first 50 frames (we also attempted to average over all the face frames, but it did not perform well on our development set). Our results show below that this was not a limitation.

It is important to note that fusing the three modalities frame by frame was not possible because the sampling rate of the audio, the movements, and the faces were not temporally aligned. This means that our model will only be able to predict perception of affective display after an animation is complete. We leave early multimodal fusion of the three modalities and incremental prediction of affect for future work.

B. Model & Training

We assumed that, because this was a sequential task, a recurrent neural network architecture would work well for our data and this task. However, early testing showed that the recurrent models were not learning the mapping between input features and labels, likely due to the size of our dataset. We opted for a more parsimonious architecture which worked well on our development set: a neural network with an input layer, a single dense hidden layer with 2000 nodes, and an output layer distribution over the 16 possible labels. We used the *tanh* activation function for the hidden layer and softmax for the output layer. We applied categorical cross entropy for the loss function, the adam optimizer, a batch size of 32, and trained for 1000 epochs. We did not find that additional regularization (e.g., dropout) made any difference in our model, and comparisons between the training and development accuracies showed that the model did not overfit our data. For the input features, we flattened and padded each modality and concatenated them together.

C. Metrics

We perform ablation tests for each of the three modalities of movement, audio, and face using the same model architecture (we tried architectures with fewer hidden layer nodes when fewer features were used, e.g., audio only, but found that our model architecture worked the best for all variants). We report three metrics to give an overall understanding of how our model performs this classification task in decreasing degrees of constraint:

accuracy′: Standard accuracy would take the argmax of the output distribution and compare that to a single label. Because the affective displays could correspond to more than one emotion, if a user checked several boxes in their rating (users assigned between 2 and 15 possible labels; the average number of emotions assigned was 1.8 overall), we opted for a less constrained accuracy score (hence, *accuracy′*): if the argmax of the output distribution matched *any* of the labels that a user had assigned to that affective display, then that counted towards the accuracy′ score. This inflates the accuracy slightly, but gives the model a chance to get at least one of the assigned labels correct.[5]

average accuracy: Because any given animation could have multiple labels, we take the prediction distribution and for every affect that received a probability of more than 0.1 in that distribution (0.1 was determined using the development set), we counted that as a positive guess for that affect. We compared this to the labels (i.e., by comparing two binary vectors) and compute the accuracy for each animation, then take the average. This too will seem inflated as many zeros in the binary vectors will increase the accuracy, but it allows our model to predict multiple labels, which better follows what people do when interpreting emotions.

average fscore: Similar to how we compute average accuracy, we take the binary vectors (i.e., affects with a probability higher than 0.1 are considered as guesses) and calculate the f1 metric (i.e., harmonic mean of precision and recall) for each animation and average the fscores. This metric gives a better idea of how well the model is performing overall; it is less constrained than accuracy and allows for multiple label guesses for a single animation.

The baseline we are comparing against is the most common baseline of 12.2%, which would have been obtained if the model were to only predict the most common label, interest. We report these three metrics for all variations of the modalities, except when considering the movement and face modalities in isolation as these did not perform better than the most common baseline. For this final evaluation, we used the evaluation set of 150 animations.

D. Results

modalities	accuracy′	avg accuracy	avg f1
audio only	0.35	0.71	0.55
movement+face	0.28	0.67	0.52
movement+audio	0.27	0.66	0.5
face+audio	0.37	0.71	0.55
all modalities	0.34	0.67	0.51

TABLE I
EXPERIMENT 1 RESULTS: ABLATION TESTS FOR AUDIO, MOVEMENT, AND FACE MODALITIES PREDICTING 16 LABELS.

The results are shown in Table I. Overall, the model works well above baseline and yields respectable results, though more work is needed to produce a model that could predict reliably which of these 16 possible affects a robot is displaying. The results tell us that though the movement and face modalities are useful together (on their own they only worked as well as baseline), the audio modality was picked up by the model as being the most useful. The best performing variant across the three modalities is face+audio. We interpret this to mean that the facial features and sounds produced by Cozmo

[5] Of course, if a worker had selected all labels for the test animations, then with this approach accuracy′ would reach 100%, but it was very uncommon for works to give more than 6 labels to any single animation (a mean of 1.8 labels were given to each animation).

play a role in how people perceive affective display, even if designers intended for another affect or no affect at all.

That our model performs well above the baseline is clear, but this standard classification task that attempts to learn a mapping from the input features to emotion labels may not be the ideal approach to this task. In the following section, we explain an experiment that alters this standard classification approach used in many automation tasks and follows from prior work in treating the affective displays as valence pairs.

V. Experiment 2: Predicting Perceptions of Valence

This experiment is motivated by the fact that though interpretation of affective display is not mutually exclusive, as shown in Section III and in Experiment 1, certain affects can be treated as opposites. We therefore break apart the task of classification of the 16 possible affective displays into 8 binary classifiers for valence pairings. We follow [28] and pair the emotions into positive and negative valence pairs shown in in Table II. We hypothesize that doing so will allow us to consider which modalities influence which affects and valence pairs more directly. We can then make use of the individual binary classifiers to make graded predictions about what affect a human would interpret.

positive valence	**negative valence**
interest	alarm
understanding	confusion
relief	frustration
joy	sorrow
gratitude	anger
hope	fear
surprise	boredom
desire	disgust

TABLE II
Valence parings of 16 specific affects.

A. Task & Procedure

The task for this experiment is to evaluate a set of binary classifiers that can determine positive or negative valence based on the pairs in Table II. We used the same training, development, and evaluation sets as in Experiment 1. However, to train and evaluate the individual binary classifiers, we only considered training examples where either of the two affects were labeled. This resulted in much smaller training sets where there was a large skew in how many training examples were represented for each affect. To mitigate this skew, we randomly sampled from the more common affect in the pair so the resulting training set for each pair was balanced. This resulted in much smaller training sets (the smallest training set was only 46 examples, the largest was 137, average was 88).

a) Modalities & Features: The smaller training sets for this task meant we needed to simplify the features from the those used in Experiment 1 for the face modality: instead of using some of the extracted faces, we averaged over all faces in an animation. That resulted in a vector of 1000 features for each animation.

B. Model, Training, Metrics

As in Experiment 1, we applied simpler models appropriate to the small amount of training data for each valence pair. After using our development dataset to tune the hyper parameters, we applied the final model to all variants of our data, a neural network with a dense hidden layer of 128 nodes with the *tanh* activation function. The top layer was binary yet was represented as two possible classes; as in Experiment 1, we used the softmax activation. We applied categorical cross entropy for the loss function, the adam optimizer with a batch size of 32, and trained for 500 epochs with early stopping, optimizing for validation set accuracy with a patience of 25 (training rarely went beyond 100 epochs with this setup). We then used the best performing model for evaluation. We note here that we evaluated linear classifiers such as logistic regression and decision trees, but the neural network, even with this fairly small, simpler network when compared to others in the literature, worked better than the linear classifiers on our development set. As in Experiment 1, we concatenated the input features and performed ablation tests on the modalities.

As this was a binary classification task, we only calculated the accuracy of each valence pair. Having sampled the data to reach an even split of each valence pair, the baseline for this task is 50%.

C. Results

The results are displayed in Table III. These results add nuance to the story from Experiment 1: audio and face modalities together perform the best for 5 of the 8 valence pairs. Digging deeper into the results, however, we find other interesting patterns: for *understand-confusion*, face alone worked as well as face+audio, telling us that unlike the results in Experiment 1, the facial features are most informative for this valence pair. Facial features are also the most informative for the *hope-fear* valence pair. We interpret this to mean that when determining if a robot is showing understanding vs. confusion or hope vs. fear, showing some kind of affective display in a face (even if it's only animated eyes) plays an important role. The same can be said of the audio modality in interpreting the valence of *relief* or *frustration* (e.g., an easy way to display relief is through a sigh; frustration, a growl). As in Experiment 1, the movement modality was not a big contributor to any valence pair (in fact, in many cases the accuracy was less than baseline), except for *desire-disgust*, where the resulting accuracy tied with the other two modalities considered together–clearly, some modalities have information that can be contributory or inhibitive when considered in conjunction with other modalities. We take this to mean that movement alone can predict the perceived valence of *desire* vs. *disgust*, likely due to the fact that movement towards something or away from something is an easy way to display desire or disgust.

It is important to note that though the training data was evenly distributed, the test set was not. This doesn't mean that the classifier simply got lucky and learned to predict the class that happened to be more represented in the test set. As seen

positive affect:	**interest**	**understand**	**relief**	**joy**	**gratitude**	**hope**	**surprise**	**desire**
negative affect:	**alarm**	**confusion**	**frustration**	**sorrow**	**anger**	**fear**	**boredom**	**disgust**
movement only	55.70	52.73	48.98	52.83	49.09	45.65	64.91	**64.29**
face only	62.03	**58.18**	61.22	60.38	**65.45**	**69.57**	52.63	50.00
audio only	62.03	56.36	69.39	60.38	**65.45**	65.22	63.16	46.43
movement+face	60.76	52.73	**75.51**	56.60	54.55	58.70	59.65	57.14
movement+audio	45.57	50.91	48.98	62.26	58.18	63.04	75.44	39.29
face+audio	**67.09**	**58.18**	63.27	**64.15**	63.64	65.22	**89.47**	**64.29**
all modalities	55.70	40.09	42.86	52.83	61.82	58.70	87.72	60.71

TABLE III
EXPERIMENT 2 RESULTS: ABLATION TESTS FOR AUDIO, MOVEMENT, AND FACE MODALITIES FOR 8 BINARY CLASSIFIERS. BEST PERFORMING ABLATIONS FOR EACH VALENCE PAIR (I.E., COLUMN) ARE MARKED IN BOLDFACE.

in Table III, the models learned something more generalizable, and in all cases no classifier accurately guessed all class labels. For example, in the *desire-disgust* valence pair, there was a 46/54% split in the test data, but for most of the ablation tests the classifier was able to perform better than the 54% most common baseline, though the training data was split evenly.

We find these results more informative than the results of Experiment 1 for several reasons: different pairings of affect make use of different information sources, and each pairing requires only a minimal amount of training data to be effective. Moreover, though it would seem like the pairings forced mutual exclusivity between the two affects, e.g., the model cannot predict *joy* and *sorrow* at the same time, the model's probability can be used to interpret the *degree* that it predicts a certain affect in the pair. This allows for a more graded approach to predicting affective display. In the following experiment, we test this idea with an altered labeling task to better match what we hypothesize to be a more informative approach to automating the interpretation of affect.

VI. EXPERIMENT 3: GENERATING AND EVALUATING NOVEL BEHAVIORS

A. Generating Novel Behaviors

To test the lessons from Experiment 2, we compared a new batch of graded human perceptions against our model. We generated our own custom animations and tasked participants with assigning a graded valence label to those animations. We explain below how we generated novel animations and used a Likert-style labeling process, and make a comparison between those labels and the predictions of the model we trained in Experiment 2.

a) Generating actions: For generating actions, we made use of the head and lift positions, left and right wheel distance traveled, as well as left and right wheel direction (i.e., opposing directions would result in spinning). Performing actions using any of these movement abilities requires a parameter to invoke that ability (e.g., a positive integer value for the lift signals to the robot to move the lift up) and a duration parameter of how long the robot should take to change its state to match that parameter. We constrained the possible parameter values for each of these abilities to fall within the acceptable range of values as calculated from the existing 940 animations. We randomly sampled values using a normal distribution (i.e., more emphasis on inner-quartile range of possible values), except for the forward/backward signals, which were randomly chosen binary values.

b) Generating faces: We gave Cozmo a repertoire of 13 possible face images, chosen from the selection of faces we extracted (see Section III) for variety and possible affective interpretation. For each animation, we randomly chose one of these images to statically display on Cozmo's OLED screen for the duration of the animation.

c) Generating audio: For audio, we constructed a list of five vowel sound approximations stated as either a question or exclamation (e.g., *aa!*, *aa?*, *ee!*, *ee?*, etc.), for a total of ten total utterance options. For each animation, we randomly selected one of these ten sounds and passed it through Cozmo's built-in speech synthesizer to play for the duration of the animation.

d) Final animations: Taking the action, face, and speech generation procedures together, we generated 20 novel animations for this experiment. On average, the duration of the animations was 4.7 seconds. We followed a similar procedure as the data collection in Section III on these animations: we recorded the video and audio of each novel animation with a camera, then extracted the face images, audio, and internal state updates from the robot for movement.

B. Procedure

We recruited 9 participants from Boise State University to observe each of the 20 animations, and after each animation we asked them to fill out a Likert-style questionnaire using the 8 valence pairs in Table II, where, following the Godspeed Questionnaire [2], we displayed the negative valence on the left and the positive valence on the right with a 5-point scale between them. We did not explicitly state that the middle value (i.e., 3) was neutral or none–we left it open for the participants to interpret (which was not allowed in the original data collection–participants always chose at least one affect). The participants were required to score all of the valence pairs (i.e., 8 possible scores for each questionnaire) for each of the 20 animations. We threw out cases where participants did not fill in a value for each of the valence pairings, which resulted in ten of the twenty with 9 scores, seven with 8 scores, and three with 7 scores. Importantly for this experiment, this rating approach allows us to take the variability of affective display into account [1] and make a direct comparison to our binary models from Experiment 2. We then used the full version of our model (i.e., all three modalities) from Experiment 2 to predict values for each valence pair for each animation. We explain below how we compared the questionnaire scores with our model's predicted values.

C. Comparison of Model Predictions and Questionnaires

We calculated the average assigned score for each valence pair for each animation and normalized the scores by dividing by the total possible values (i.e., five). This resulted in a value between zero and one. We were able to directly compare that with predictions from our model, which were also probability values between zero and one. The value represented the degree of belief that the model was predicting the positive valence. A value of 0.5 would mean the model is unable to choose between the two, and a value below 0.5 means the model is predicting the negative valence for the pair.

We directly compared the model predictions with the questionnaire averages using Pearson's correlation coefficient. We found moderate correlations between three valence pairings: *relief-frustration* (0.44), *joy-sorrow* (0.36), and *surprise-boredom* (0.34); all others were below 0.3. The surprising yet welcome result in this experiment is that the highest and most consistent correlations came from the model that used all three modalities, whereas in prior experiments the *audio* and *face* modalities tended to perform best either together or, in some cases, separately. We take this to mean that all three modalities are informative when compared with more realistic ratings from human users.

Figure 6 shows the model predictions and average human ratings for one of the animations, which illustrates the general pattern over all animations that specific pairings generally lie fairly close to each other. Two pairs: *interest-alarm* and *understanding-confusion* had very low correlations (i.e,. <0.1) which tells us that our chosen model confuses those two particular pairings (though an application of the *face*-only model resulted in a 0.24 correlation for the latter pair).

These results show that our classifiers can yield reliable predictions when compared to humans for novel animations. Those classifiers could be useful in specific tasks, e.g., *relief-disgust* and *surprise-boredom* in tasks where a robot's minimal level of social engagement with a person is important–a model that predicts more interpretations of boredom would need to alter its behavior to elicit more engagement with a user.

VII. Conclusion

In this paper, we examined how people perceive the affective display of the Anki Cozmo robot. Our data collection and analyses confirmed prior research which has shown that affective labels are not mutually exclusive; an entity such as a person or a robot can experience, for example, joy and desire at the same time. Moreover, the perceived affect is often not what the designer intended–even if the designer did not intend any affect at all. This is an ongoing research problem which arguably does not need more experiments to be validated; however, this paper's novel contribution to HRI is in its increasingly granular approach to classification, both in terms of the robot modalities being considered in Experiment 1, which are checked against individual emotional valences in Experiment 2, along with the creation of a novel dataset, of which there are few existing corpora for the reasons stated in [22]. We showed in Experiment 1 that movement, facial, and audio modalities supply useful information as features in a classification task, particularly the facial and audio modalities. As noted in [28], while the affective displays may not be mutually exclusive they can be grouped into valence pairings, which we explored in Experiment 2. Our set of binary classifiers further the research of [31], in finding that different features carry different weight in communicating emotional valence. Our experiments also confirm claims made in [26] and [22] that speech is an important modality for natural interaction between people and robots, yet robotics researchers often avoid audio as a medium of communication between robots and people [25].

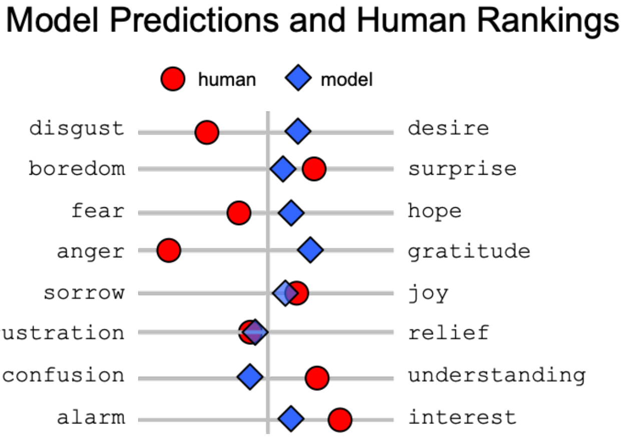

Fig. 6. Comparison of model prediction and average questionnaire results for one of the novel animations.

Our experiments and resulting models in this paper can inform future robotics research by automatically predicting affective display and valence between the affects. We also assert that though we showed how three modalities–movement, face, and audio–are useful when predicting perceptions of affect and valence, we were only able to obtain the face and audio modalities by external sensors (i.e., a camera and a microphone). Future work in robotics that takes affect and valence into account should be able to observe all three modalities directly from the robot itself, echoing the findings of [29] on the importance of immediacy and integration of a robot's "perceptual" feedback.

This paper's primary contributions are: (1) a novel dataset explained in Section III that we will make publicly available, (2) a multimodal computational model that could be used in live interactive tasks and analysis of how those modalities play into that model's predictions of perception, and (3) motivation for robotics researchers to evaluate how real people interpret robot behaviors.

Acknowledgements We would like to thank the anonymous reviewers for their informative comments and the Anki developers for answering questions in a timely manner. This work was approved by the SB-IRB, Protocol Number 126SB18242.

References

[1] Lisa Feldman Barrett. Variety is the spice of life: A psychological construction approach to understanding variability in emotion. *Cognition & Emotion*, 23(7): 1284–1306, 2009. ISSN 0269-9931.

[2] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[3] Dan Bohus and Eric Horvitz. Models for Multiparty Engagement in Open-World Dialog. In *Computational Linguistics*, number September, pages 225–234, London, UK, sep 2009. Association for Computational Linguistics.

[4] Jessy Ceha, Joslin Goh, Corina McDonald, Dana Kulić, Edith Law, Nalin Chhibber, and Pierre-Yves Oudeyer. Expression of Curiosity in Social Robots: Design, Perception, and Effects on Behaviour. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, Glasgow, UK, 2019.

[5] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 33–40, Bielefeld, Germany, 2014.

[6] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 1960.

[7] Thi Le Quyen Dang, Nguyen Tan Viet Tuyen, Sungmoon Jeong, and Nak Young Chong. Encoding Cultures in Robot Emotion Representation. *26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Lisbon, Portugal(Aug 28 - Sept 1): 547–552, 2017.

[8] Friedericke Eyssel and Frank Hegel. She's Got the Look: Gender Stereotyping of Robots. *Journal of Applied Social Psychology*, 42(9):2213–2230, 2012.

[9] Friederike Eyssel and Dieta Kuchenbrandt. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology*, 51(4):724–731, 2012.

[10] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. A survey of socially interactive robots. In *Robotics and Autonomous Systems*, 2003.

[11] Mehdi Ghayoumi and Arvind K. Bansal. Emotion in Robots Using Convolutional Neural Networks. pages 285–295. Springer, Cham, 2016.

[12] Louisa Hall. How We Feel About Robots That Feel. *MIT Technology Review*, pages 1–11, oct 2017.

[13] Julian Hough and David Schlangen. Investigating Fluidity for Human-Robot Interaction with Real-time, Real-world Grounding Strategies. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 288–298, Los Angeles, sep 2016. Association for Computational Linguistics.

[14] Panagiotis G. Ipeirotis and Panagiotis G. Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16, dec 2010.

[15] Jesin James, Catherine Inez Watson, and Bruce MacDonald. Artificial Empathy in Social Robots: An analysis of Emotions in Speech. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 632–637. IEEE, aug 2018.

[16] Casey Kennington and David Schlangen. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China, 2015. Association for Computational Linguistics.

[17] Douwe Kiela, Luana Bulat, and Stephen Clark. Grounding Semantics in Olfactory Perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China, jul 2015. Association for Computational Linguistics.

[18] Barbara Lewandowska-Tomaszczyk and Paul A. Wilson. Compassion, empathy and sympathy expression features in affective robotics. In *2016 7th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000065–000070. IEEE, oct 2016.

[19] Brian Mcfee, Colin Raffel, Dawen Liang, Daniel P W Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python science conference (SCIPY 2015)*, pages 18–25, 2015.

[20] Peter E. McKenna, Ayan Ghosh, Ruth Aylett, Frank Broz, and Gnanathusharan Rajendran. Cultural Social Signal Interplay with an Expressive Robot. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents - IVA '20*, pages 211–218, New York, New York, USA, 2018. ACM Press.

[21] Lilia Moshkina, Susan B Trickett, and J G Trafton. Social Engagement in Public Places: A Tale of One Robot. In *ACM/IEEE International Conference on Human-robot Interaction*, 2014.

[22] Aastha Nigam and Laurel D. Riek. Social context perception for mobile robots. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3621–3627. IEEE, sep 2015.

[23] Jekaterina Novikova, Leon Watts, and Tetsunari Inamura.

Emotionally expressive robot behavior improves human-robot collaboration. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 7–12. IEEE, aug 2015.

[24] Jekaterina Novikova, Christian Dondrup, Ioannis Papaioannou, and Oliver Lemon. Sympathy Begins with a Smile, Intelligence Begins with a Word: Use of Multimodal Features in Spoken Human-Robot Interaction. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 86–94, 2017.

[25] Julia Peltason. *Modeling Human-Robot-Interaction based on generic Interaction Patterns.* PhD thesis, Bielefeld University, 2014.

[26] Sarah Plane, Ariel Marvasti, Tyler Egan, and Casey Kennington. Predicting Perceived Age: Both Language Ability and Appearance are Important. In *Proceedings of SigDial*, 2018.

[27] Robin Read and Tony Belpaeme. People Interpret Robotic Non-linguistic Utterances Categorically. *International Journal of Social Robotics*, 8(1):31–50, mar 2016.

[28] David L. Robinson. Brain function, emotional experience and personality. *Netherlands Journal of Psychology*, 2008.

[29] Elaine Schaertl Short, Mai Lee Chang, and Andrea Thomaz. Detecting Contingency for HRI in Open-World Environments. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, pages 425–433, 2018.

[30] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. sep 2014.

[31] Sichao Song and Seiji Yamada. Designing Expressive Lights and In-Situ Motions for Robots to Express Emotions. In *Proceedings of the 6th International Conference on Human-Agent Interaction - HAI '18*, pages 222–228, New York, New York, USA, 2018. ACM Press.

[32] Benedict Tay, Younbo Jung, and Taezoon Park. When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, 2014.

[33] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. Learning MultiModal Grounded Linguistic Semantics by Playing ” I Spy ”. In *Proceedings of IJCAI*, pages 3477—-3483, 2016.

Appendix

```
left_wheel_speed
right_wheel_speed
battery_voltage
time
nav_memory_map_sizes
nav_memory_map_x
nav_memory_map_y
pose_0
pose_1
pose_4
pose_5
pose_10
pose_12
pose_13
pose_14
is_moving
is_picked_up
is_animating
lift_in_pos
head_in_pos
are_wheels_moving
is_localized
pose_angle_rads
pose_angle_degs
pose_angle_abs_rads
pose_angle_abs_degs
pose_pitch_rads
pose_pitch_degs
pose_pitch_abs_rads
pose_pitch_abs_degs
head_angle_rads
head_angle_degs
head_angle_abs_rads
head_angle_abs_degs
lift_position_height
lift_position_ratio
lift_position_angle_rads
lift_position_angle_degs
lift_position_angle_abs_rads
lift_position_angle_abs_degs
dispatcher_has_in_progress_action
accelerometer_x
accelerometer_y
accelerometer_z
gyro_x
gyro_y
gyro_z
```