“Bad Idea, Right?” Exploring Anticipatory Human Reactions for Outcome Prediction in HRI

Maria Teresa Parreira[1], Sukruth Gowdru Lingaraju[1], Adolfo Ramirez-Artistizabal[2], Alexandra Bremers[1], Manaswi Saha[2], Michael Kuniavsky[2], Wendy Ju[1]

Abstract— **Humans have the ability to anticipate what will happen in their environment based on perceived information. Their anticipation is often manifested as an externally observable behavioral reaction, which cues other people in the environment that something bad might happen. As robots become more prevalent in human spaces, robots can leverage these visible anticipatory responses to assess whether their own actions might be “a bad idea?” In this study, we delved into the potential of human anticipatory reaction recognition to predict outcomes. We conducted a user study wherein 30 participants watched videos of action scenarios and were asked about their anticipated outcome of the situation shown in each video (“good” or “bad”). We collected video and audio data of the participants reactions as they were watching these videos. We then carefully analyzed the participants’ behavioral anticipatory responses; this data was used to train machine learning models to predict anticipated outcomes based on human observable behavior. Reactions are multimodal, compound and diverse, and we find significant differences in facial reactions. Model performances are around 0.5-0.6 test accuracy, and increase notably when nonreactive participants are excluded from the dataset. We discuss the implications of these findings and future work. This research offers insights into improving the safety and efficiency of human-robot interactions, contributing to the evolving field of robotics and human-robot collaboration.**

Index Terms— **robot error; social signals; anticipation; error prevention; computer vision; human-AI collaboration**

I. Introduction

There is a social aspect to estimating the wisdom of our actions: when we are trying something new in the presence of others, we might watch the reaction of bystanders to judge whether our actions are risky. For example, if we see someone squinting and frowning at us as we are walking quickly across the sidewalk in winter, we might slow down, in case they know the ground is icy or otherwise hazardous. In doing so, we leverage the ability of other humans to perceive likely harms, and their tendency to manifest their perception as observable signals. This ability to sense and adapt to social signals given off by others makes use of the “wisdom of crowds”; people’s observable reactions form a rich source of data that, if effectively harnessed, could significantly enhance robotic systems’ efficiency and prevent error.

In the field of Human-Robot Interaction (HRI), there is a growing body of literature that explores human reactions to better robots’ functionality [3, 16, 6, 8, 7, 30]. In this line of research, many works have leveraged human reactions to failure [29, 26, 17, 4], but few delve into the human ability to observe situations and *anticipate* outcomes. There is untapped potential in using observable anticipatory behavior to raise robots’ social awareness and prevent robot failure.

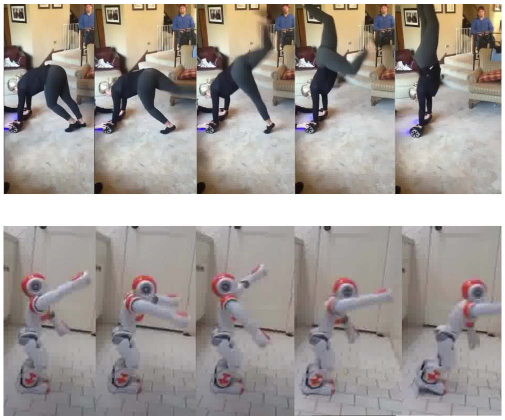


Fig. 1: In the online user study, participants were shown videos displaying a variety of scenarios featuring robots and humans. Videos ended before the outcome was displayed, and participants were asked about the anticipated outcome of each video.

In this work, we collected a dataset of human anticipatory reactions to videos displaying different human and robot scenarios. After observing the videos, participants were asked to predict the unshown outcome of the situation (“you think the situation ends...”, “well” or “poorly”). This anticipated outcome – the anticipated resolution of the setting shown in the video, based on the available context only – was then mapped to the behavioral reactions. We analyzed this dataset to understand anticipatory reactions and their potential in HRI. First, we categorized non-verbal anticipatory behaviors and extracted qualitative observations; we then explored the use of machine learning to develop systems that can predict errors before they happen. This research offers insights into improving the safety and efficiency of human-robot interactions, contributing to the evolving field of robotics and human-robot collaboration.

II. Related Work

Humans as sensors The idea of using *humans as sensors* [21] of the environment, namely for error detection, means grasping the ecological and behavioral patterns ingrained in human reactions. Human reactions to robot error are complex, diverse and multimodal. Common reactions include verbalization [18], body motion [18, 31, 10], gaze [18, 1] and facial expressions [18, 15, 27]. Mirnig et al. [22] analyzed a large video corpus of humans interacting with failing robots and found that social cues are common in reactions to robot

[1] Cornell Tech [2] Accenture Labs

error, but the specific reactions can differ with the type of error. Stiber and Huang [27] reported on the dynamic evolution of human reactions to robot failures (e.g. eyebrow raises evolving into verbal utterances).

Other works also extensively characterize human social cues in response to robot error [18, 10]. However, these focus only on reactions to error *after* the event, disregarding the potential for leveraging anticipatory reactions. Notably, Stiber et al. [27, 26] report participants exhibiting reactions *before* error occurs, if the error seems predictable (i.e. if the outcome is obvious). Previous works [12, 9] reported how anticipation can improve human-robot interactions, but this is focused on the robot anticipating human intent. What we propose is analogous but complementary—the human anticipates the outcome of robots' actions, and the robot uses this data to prevent erroneous behavior.

Models for error detection using social cues While the use of social cues, namely human behavioral features, has long been studied, the literature has largely focused on the extrapolation of internal states (such as valence or arousal [32], engagement [20] or comfortability [19]). Recent works, however, have been exploring how to leverage humans' ability to (visibly) react to their environment and improve robotic systems [11, 5, 13, 8]. For example, Hwang et al. [15] and Cui et al. [8] both used human facial expressions as feedback input for robot task learning. Richter et al. [25] used participants' gaze and lip movement to improve dialogue systems.

In the field of failure in HRI, many works have investigated robot error prediction through human reactions. For example, Bremers et al. [4] used a corpus of bystander human reactions to failure to develop a Convolutional Neural Network (CNN)-based error detection system. Stiber et al. [28] used facial activation units (AUs) as input to a binary classifier to predict the timing of robot failure. Similar work [29] also made use of facial social cues to detect different types of error, using a deep neural network with 3 hidden layers. Kontogiorgos et al. [18] used a Recurrent Neural Network (RNN) and a Gradient Booster Tree in a multi-modal dataset to classify different error types from human reactions. Other approaches include zero-shot learning from human behavior [24]. While not directly in scope, there is extensive work on techniques for action anticipation [34], including in human-robot interaction and using social cues [14]. To the best of our knowledge, no works have explored human anticipatory reactions as they map to predicted outcomes.

Accordingly, we investigated human anticipatory reactions through a controlled online user study employing video stimuli. We wanted to analyze whether naturalistic reactions exhibit identifiable patterns that can be utilized to develop models for preventing robot errors. In sum, our contributions are: 1) collecting a dataset of anticipatory naturalistic human reactions to observed scenarios; 2) exploring the reaction patterns through extraction of facial activation features across different predicted outcomes, and 3) testing and benchmarking different model architectures for outcome prediction.

III. Study Design

To capture human reactions to anticipated outcomes, we designed an online user study to collect visual responses to videos displaying different scenarios. The protocol and dataset are described below.

A. Stimulus dataset

We selected a set of 30 stimulus videos, which include videos where humans and robots are featured, and there is a build-up in action to an outcome that can be positive or negative. Figure 1 shows a couple sequences from the stimulus dataset, and the full list of stimulus videos are available in an online repository [1]. The short-length videos (9.62 ± 2.77 s) were piloted and selected based on diversity of outcomes (good or bad resolution) and predictability (e.g. including bad outcomes that are predictably bad, as well as surprisingly bad).

B. Experimental procedure

We conducted an online crowd-sourced study to collect webcam reactions to stimulus videos from a global sample recruited through Prolific, in line with previous work[4]. After giving informed consent, participants provided demographic information (age, gender, and race/ethnicity). Following this, participants were shown a series of scenarios through short videos. The protocol included a “warm-up” round of 3 videos, followed by the main round of data collection (30 videos). Each stimulus video was shown twice. First, a shorter version of the video was shown; the video stops before resolution to the video action was reached (e.g., someone swinging from a rope, approaching a tree branch). After watching this video, participants were asked *“You think this situation ends...”* with the options “well” or “poorly”. After this, participants were able watch a longer version of the video, featuring the resolution of the video action. This two-stage video stimulus design was set up after we found in pilot studies that participants needed to be shown that both good and bad outcomes were possible for the videos, to have the full range of outcome anticipation response.

Participants would see each stimulus video while their laptop or computer webcam recorded their facial responses. Participants were not able to see their own image while the stimulus videos played and the order of stimulus videos was randomized. Compensation was provided at rate of USD 15/hour for participants that took less than 60 min to complete the study (watched all 30 videos). The full procedure took around 30 minutes to complete. This data was collected under Cornell University exempt IRB protocol #1609006604.

[1] https://github.com/mteresaparreira/badidea.git

C. Feature extraction

The reaction videos feature participants' reactions to stimulus videos, as recorded through the participant's own web cameras. The reaction videos were collected at 30 fps, but vary in image resolution, lighting and camera angles (Fig. 2). Following an analysis of the collected data, and informed by prior work in the field of failure detection and affect computation [29, 4], we explored the space of human anticipatory reactions by extracting facial activation features (i.e., facial expressions). We made use of Openface [2]. We extracted only the features that regard gaze, head motion, and facial unit activation (AUs), in a total of 49 features. After qualitative analyses of the videos and model development (see below), only AUs were used, as gaze and body pose varied arbitrarily. Thus, the data used for training consisted of 35 features (the full feature list can be found in the online study repository[1]).

D. Model Development

To test the feasibility of automatically detecting anticipated outcomes, we made use of machine learning methods.

Problem Formulation Predicting anticipated outcomes based on human reactions was formulated as a sequential decision-making problem: at each time step t, the environment is captured as a state variable $s_t \in S$, and the model output is an outcome label $f_t \in F$, where F is a discrete variable that describes the **participant's predicted outcome** (and not the actual outcome of the video): 0 if *good*, 1 if *bad*. In a real-use HRI system, this binary classification would allow the robot to understand if it should proceed or not with its current action.

Action Space The reactions were labeled according to outcome predicted by the participants. In both *good* and *bad* anticipated outcomes, we extracted 3 seconds of data (at 30 fps) preceding the video cut-off moment. This time horizon was deemed to lead to better model performances.

Models tested We tested a wide range of deep learning model architectures, namely Recurrent Neural Networks (RNNs), which can capture temporal dependencies within the data. Long Short Term Memory (LSTMs) are commonly used, but tend to overfit. Alternatively, Gated Recurrent Units (GRUs) were also tested, which are preferred in smaller datasets, as well as Bidirectional LSTMs (BiLSTMs), which are made up of two hidden-layer LSTMs. We also tested a similar Deep Neural Network to that suggested by Stiber et al. [29]—3 hidden layers, with 64, 128 and 64 units (multi-layer DNN, ml-DNN). We used categorical cross-entropy as the loss function on all models.

Datasets We trained the models on subsets and variations of the data. We tested normalization, as well as Principal Component Analysis (PCA) for feature reduction. In order to evaluate the effect of data "cleaning"—i.e., removing nonreactive participants, that is participants with no visible facial reactions to the stimulus videos—we created a "short-list dataset" based on qualitative observations of the participants' reactions. We tested each model on both the *full* and the *short-list* dataset.

TABLE I: Samples (frames) per dataset per anticipated outcome.

Train Dataset	Good	Bad	Total
Full	41265	35909	77174
Short list	20000	17222	37222

Evaluation Metrics The models were evaluated based on the macro averages of the following metrics: accuracy, F1-score, precision and recall.

Model Training We performed hyperparameter tuning on an 70-20-10 train-val-test split, with 5 cross-validation folds with no overlapping participants. The best candidate model of each type was picked based on macro-accuracy and F1-score on the test set.

E. Participants

After exclusion due to video recording and feature extraction issues, our dataset is comprised of data from 29 participants. Ages range from 20-39 (27.10±4.98). 12 participants identify as female, 14 as male and 3 who self-identified. Racial/ethnical distribution includes 15 Caucasian/White or Asian/White, 9 African/African American/Black, 4 Hispanic/Latino and 1 Asian/Asian American participants. Participants took an average of $33m48s \pm 9m38s$ to complete the full data collection process.

The short-list dataset consisted of 14 participants, based on the salience of reactions as described above.

IV. Exploring Anticipatory Reactions

A. Dataset

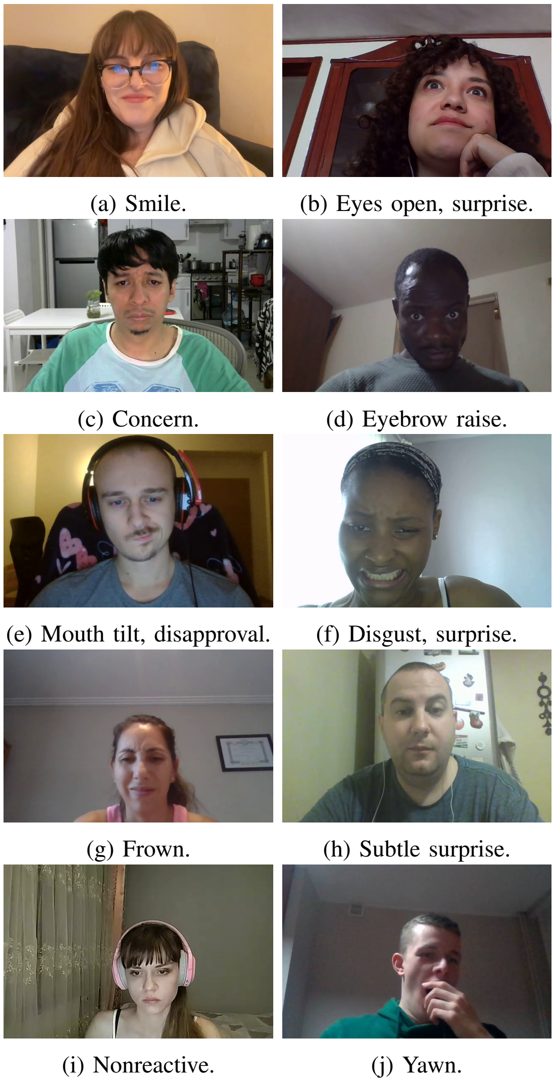
The dataset is composed of participants' visual responses to the shorter stimulus videos. We *mapped anticipatory reactions to self-reported outcome prediction*. Frames from videos where participants deemed the outcome to end "well" were labeled as 0 (*anticipated good outcome*). Frames from videos where participants deemed the outcome to end "poorly" were labeled as 1 (*anticipated bad outcome*).

We excluded all frames where feature prediction confidence (from Openface) was lower than 70% (total of 2680 frames, 3.47% of the dataset). One participant was excluded at this stage, as more than 50% of their videos were below this threshold. The final and short-list datasets are shown on Table I.

B. Describing Human Anticipatory Reactions

To better understand the space of human anticipatory behaviors as social signalling, we began by manually analyzing the dataset, collecting observations on the reactions and respective anticipated outcomes, based on behavioral analysis and ethnomethodology. Below, we list our findings.

a) ***Outcomes anticipated as bad generate more salient and diverse anticipatory reactions:*** Participants' reactions to outcomes anticipated as bad were generally more diverse and salient (described in detail below), whereas anticipated



(a) Smile.

(b) Eyes open, surprise.

(c) Concern.

(d) Eyebrow raise.

(e) Mouth tilt, disapproval.

(f) Disgust, surprise.

(g) Frown.

(h) Subtle surprise.

(i) Nonreactive.

(j) Yawn.

Fig. 2: Anticipatory behaviors displayed by participants. Participants provided explicit consent for reproduction.

good outcomes were preceded by mostly neutral facial expressions.

*b) **Anticipatory reactions are diverse, compound, multimodal, and evolving**:* Anticipatory behavior is rich in facial expressions, head motion and body pose changes, as well as multimodal, with vocal, verbal and non-verbal reactions accompanying gestures and expressions. Importantly, it evolves over time with compounding behaviors and changing emotional displays. In Figure 2, we show examples of these behaviors. Some examples include:

- chuckle into loud laughter;
- evolving surprise (eyebrows raise, into mouth opening, into vocal utterance).
- surprise (eyes widen, eyebrows raise) into disgust (lip corner depression, frowning);
- surprise into humor (smile, head shake, chuckle);
- confusion (eyes squint, lips purse, into shoulder shrug and vocal utterances of disbelief)
- disgusted humor (head nods, into chuckle and smile)
- disapproval (head shake, lip tightening)
- smirks (humorous if accompanied by eyebrow raise, disgusted if accompanied by eyebrow frown)

*c) **Reactions are person-dependent**:* Different participants behave differently – both in the diversity of behaviors shown and in their magnitude. While some participants are very visibly reactive (especially for anticipated bad outcomes), others display very subtle (Figure 2h) to no reactions at all (Figure 2i).

*d) **Datasets are noisy**:* Quantitative approaches to analyzing and leveraging anticipatory reactions must take into account displayed behaviors that are not part of these reactions. Examples are yawning (Figure 2j), sneezing, hands in face. Other noisy behaviors stemming from our data collection process in particular include distraction (looking away), camera device position changes (e.g. person moves their laptop), among others. Time and task fatigue are also a source of noise, as reactions tend to lose prominence with evolving trials. Another factor to consider is outcome anticipation mapping. While participants were asked to make a judgment based on the situation shown in each video, pilot studies indicated that participants might answer based on meta-cognition—thinking about what the expected answer might be, or based on past experience of watching or participating in situations similar to that displayed in the videos; hence, the reactions will not map to the intuitive judgment given the context shown but rather a rationalized outcome prediction.

C. Statistical Analysis of Behavioral Patterns

To understand the patterns of behavior in anticipatory reactions, we analyzed the extracted features through statistical methods. We focused only on intensity of facial unit activation features. This is because both gaze and pose vary widely across participants (different camera angles, positions), hence we did not want to foster sporadic findings of statistical significance that do not pertain to actual anticipatory behavior. The following analysis is based on 17 activation unit features.

We checked if different facial unit activations can identify with specific anticipated outcomes by comparing their variances using Welsh t-test on the entire dataset of each facial unit following previous work [29]. We corrected p-values using Bonferroni. For all features but 5, the feature was deemed a significant predictor of the anticipated outcome. Table II shows the features and respective corrected p-values. Figure 3 shows again the features deemed significant and the mean of their normalized value. It can be seen that for anticipated bad outcomes there is more intensity of activation, which is consistent with observational findings described in Section IV-B. The full results from the statistical analysis can be found in the online repository.

D. Predicting Outcomes based on Anticipation

We benchmarked different model architectures through the methods described in Section III. Table III describes the best-performing models for each model type, both in the full

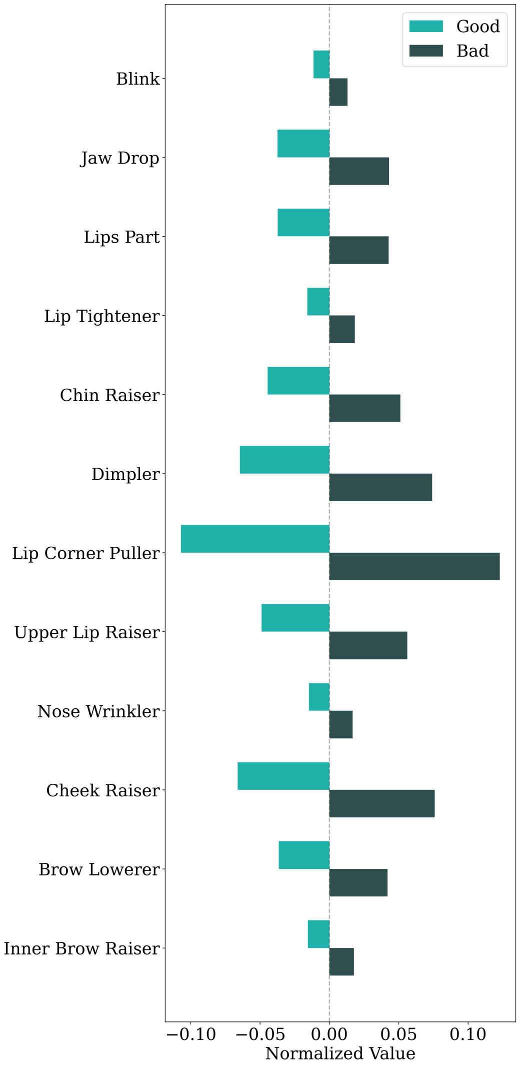


Fig. 3: Significant AUs, normalized mean value across the dataset for each anticipated outcome class (light blue: good outcome; dark blue: bad outcome). Standard deviation not shown as it is a large value, consistent with observations of the wide reaction range.

dataset and in the short-list dataset. In all cases, the dataset used consisted of frames taken at a $30fps$ frame rate, with a 3 second time window. These were deemed to be the dataset criteria that lead to the best results. These performances are discussed in Section V.

V. Discussion

In this work, we explored the visual behavioral patterns in anticipatory human reactions. We collected a dataset of naturalistic anticipatory reactions from people, through online and remote data collection. While this data collection method presents a challenge for data handling, due to high environmental diversity,it also enables us to collect data from more participants, and under circumstances that might be more ecologically valid for future applications.

In an implemented scenario, robots operating directly or within shared spaces with humans would achieve more seamless and efficient functioning by avoiding actions that are deemed as leading to poor outcomes. Based on past experiences and domain expertise, humans are able to analyze scenarios and predict outcomes; analogously, externalized social signals such as anticipatory reactions to events can be used as sensors to the environment for “preprocessing” contextual information. Robots can and should make use of this ability in human-robot interaction scenarios in order to operate more efficiently, by avoiding actions anticipated as erroneous (i.e. avoiding robot failure).

TABLE II: Facial Unit Activation features with significantly different intensity across anticipated outcomes. AU – action unit. p-values: * – p <0.05; *** — p <0.001.

Feature	AU	p-value
AU01_r	Inner Brow Raiser	***
AU04_r	Brow Lowerer	***
AU06_r	Cheek Raiser	***
AU09_r	Nose Wrinkler	***
AU10_r	Upper Lip Raiser	***
AU12_r	Lip Corner Puller	***
AU14_r	Dimpler	***
AU17_r	Chin Raiser	***
AU23_r	Lip Tightener	***
AU25_r	Lips Part	***
AU26_r	Jaw Drop	***
AU45_r	Blink	*

The collected reactions to the scenarios shown were first analyzed statistically, revealing complex, compound, and diverse reactions, especially to outcomes anticipated as “bad”. Stiber and Huang [27] had previously reported that reactions are different in their salience and evolve with time when humans are in the presence of failing robots. While these variations unveil the richness of human reactions and our ability to anticipate outcomes, they also pose a challenge for automatic detection of the predicted outcomes, since patterns in the data are not found trivially.

In our statistical analysis of facial units, we identified 12 action units with significantly different activation intensities when outcomes are predicted as good and bad. These concern AUs related to mouth motion (e.g. mouth opening in surprise reaction) and smiling/laughing (lips, cheeks), as well as eyebrow motion (e.g. eyebrows frowning in confused reaction). For all AUs deemed significantly different, the intensity of activation was lower for outcomes anticipated as “good” than for “bad” outcomes (Fig. 3). These results are promising for the envisioned use case—where robots are able to prevent actions deemed as leading to bad outcomes by tracking (bystander) human reactions.

Following these analyses, we tested and benchmarked a set of machine learning models, exploring the potential for a use-case scenario of error prevention systems. Using a non-overlapping participant approach for training and testing the models, and in spite of extensive model development testing, model performances are only slightly above chance for the full dataset approach. Importantly, by manually curating the dataset—removing nonreactive participants—we obtain noteworthy performance improvements of up to 7 percentage points for accuracy. These performances are not surprising given the above-mentioned variability of reactions in the

TABLE III: Best performing models (test performance across 5 test folds, $M \pm SD$ for 7000 epochs). Grey rows represent models trained on the short-list dataset. *norm* indicates dataset was normalized. SL: sequence length.

Model	**HyperP**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
GRU	SL=10, Units=64, Dropout=0.0 Act: sigmoid , Opt: Adadelta	0.527 ± 0.045	0.556 ± 0.033	0.544 ± 0.025	0.511 ± 0.062
GRU	*short-list, norm* SL=10, Units=128, Dropout=0.2 Act: sigmoid , Opt: Adadelta	0.583 ± 0.031	0.578 ± 0.020	0.574 ± 0.017	0.568 ± 0.021
LSTM	SL=5, Units=128, Dropout=0.2 Act: sigmoid , Opt: Adadelta	0.544 ± 0.037	0.561 ± 0.018	0.557 ± 0.018	0.538 ± 0.039
LSTM	*short-list* SL=10, Units=128, Dropout=0.0 Act: sigmoid , Opt: Adadelta	0.602 ± 0.039	0.588 ± 0.045	0.578 ± 0.039	0.575 ± 0.038
BiLSTM	SL=10, Units=32, Dropout=0.6 Act: sigmoid , Opt: Adadelta	0.545 ± 0.046	0.568 ± 0.018	0.561 ± 0.02	0.537 ± 0.054
BiLSTM	*short-list, norm* SL=10, Units=128, Dropout=0.6 Act: sigmoid , Opt: Adadelta	0.577 ± 0.046	0.565 ± 0.043	0.558 ± 0.036	0.556 ± 0.034
ml-DNN [29]	*norm* Dropout=0.0 Act: sigmoid , Opt: SGD	0.512 ± 0.014	0.51 ± 0.019	0.509 ± 0.018	0.504 ± 0.017
ml-DNN [29]	*short-list* Dropout=0.0 Act: sigmoid , Opt: SGD	0.586 ± 0.036	0.581 ± 0.029	0.577 ± 0.029	0.569 ± 0.033

dataset. Better performance should be achieved by increasing the dataset size, to ensure that more samples of the different reactions and reaction patterns are represented. Another option would be to implement a single-user system, where model generalization is conceptualized as the minimum amount of data needed to generalize to other samples from the same user (e.g. through a calibration round). In sum, there is potential for using anticipatory reactions, but larger datasets or different data collection methods (e.g. in-person data collection) should be explored.

The growing body of literature that leverages human reactions to failure for robot error detection [4, 18, 29, 23] reveals the potential of exploring these cues in human-robot interactions. Making use of anticipatory reactions would help *preventing* robot error, thus allowing for better alignment of robotic perception into the human-robot social sphere. On a final note, the ethical use of such systems should be considered and discussed. User consent and understanding of the system is pivotal and system designers must uphold privacy and security as priorities.

A. Limitations and Future Work

This work explores how anticipatory behavior could be used to predict action outcomes and prevent error. This is preliminary work that analyses these reactions, leaving opportunities for future work that dives into implementing these systems.

Regarding data collection, our method of online crowdsourcing provides a dataset of naturalistic bystander reactions, but it comes with limitations. In this protocol, we map reaction behavior to the humans' predicted outcome. This means that we cannot control for data balance in our ground-truth labels, since they are generated by each participant as they observe the stimulus videos. Nonetheless, we piloted the dataset to prevent major data imbalances. This could also be a source of lower model performance, as some of the non-mixed participant folds might not be balanced. Another factor to consider is that we cannot control how participants choose to respond to their predicted outcome of the situation shown. For example, one participant might see a video of someone about to jump from a high building, anticipate a bad outcome, but answer “good outcome” given that a video was filmed and made available for that moment. This means that the reactions will be ill-mapped to the predicted outcome. This observation emerged from pilot testing and collecting participants' feedback. To prevent this, we explain at multiple points of the study protocol that participants should answer intuitively based only on what is shown in the video. We additionally did not control for whether the participant was familiar with any of the video stimuli, which might impact their behavioral response.

In this work, we focus only on the visual modality of the dataset, in spite of the multimodality of anticipatory responses described in Section IV (e.g. laughter). This is because the dataset contains background noise that could lead to erroneous model training. Future work should make use of multimodal data inputs. Other avenues for future developments include collecting and exploring in-person anticipation reactions, and making use of other machine learning techniques for automatic outcome prediction, such as complementing the system with zero-shot large language models [33].

VI. Conclusion

This study has shed light on the potential of harnessing human anticipatory reaction recognition to inform the predictive capabilities of robots in diverse environments. Through our user study, we revealed the multifaceted nature of human responses to various action scenarios, with significant differences in facial expressions that serve as valuable indicators of anticipated outcomes. Notably, the exclusion of nonreactive participants led to improvements in model performance, highlighting the importance of capturing salient anticipatory responses for accurate prediction. By leveraging human anticipatory reactions, robots can navigate complex social landscapes with greater adaptability. Further research endeavors hold the potential to refine predictive models, explore novel modalities of reaction recognition, and translate these insights into real-world applications across diverse domains.

References

[1] R. M. Aronson. Gaze for error detection during human-robot shared manipulation. In *RSS Workshop: Towards a Framework for Joint Action*, 2018.

[2] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, 2018. doi: 10.1109/FG.2018.00019.

[3] A. Bremers, A. Pabst, M. T. Parreira, and W. Ju. Using social cues to recognize task failures for hri: A review of current research and future directions. In *(under review) arXiv:2301.11972*, 2023. doi: https://doi.org/10.48550/arXiv.2301.11972.

[4] A. Bremers, M. T. Parreira, X. Fang, N. Friedman, A. Ramirez-Aristizabal, A. Pabst, M. Spasojevic, M. Kuniavsky, and W. Ju. The bystander affect detection (bad) dataset for failure detection in hri, 2023.

[5] J. Broekens. Emotion and reinforcement: affective facial expressions facilitate robot learning. In *Artifical intelligence for human computing*, pages 113–132. Springer, Cham, 2007.

[6] L. Cohen, M. Khoramshahi, R. N. Salesse, C. Bortolon, P. Słowiński, C. Zhai, K. Tsaneva-Atanasova, M. Di Bernardo, D. Capdevielle, L. Marin, et al. Influence of facial feedback during a cooperative human-robot task in schizophrenia. *Scientific reports*, 7(1):1–10, 2017.

[7] A. Cuadra, H. Lee, J. Cho, and W. Ju. Look at me when i talk to you: A video dataset to enable voice assistants to recognize errors. *arXiv preprint arXiv:2104.07153*, 2021.

[8] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. Knox. The empathic framework for task learning from implicit human feedback. In *Conference on Robot Learning*, 2020.

[9] P. F. Dominey, G. Metta, F. Nori, and L. Natale. Anticipation and initiative in human-humanoid interaction. In *Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots*, pages 693–699, 2008. doi: 10.1109/ICHR.2008.4755974.

[10] M. Giuliani, N. Mirnig, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi. Systematic analysis of video data from different human–robot interaction studies: a categorization of social signals during error situations. *Frontiers in psychology*, 6:931, 2015.

[11] C. J. Hayes, M. Moosaei, and L. D. Riek. Exploring implicit human responses to robot mistakes in a learning from demonstration task. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 246–252, New York, NY, USA, 2016. IEEE. doi: 10.1109/ROMAN.2016.7745138.

[12] G. Hoffman. Anticipation in human-robot interaction. In *2010 AAAI Spring Symposium*, pages Technical Report SS–10–06, 2010.

[13] C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90, New York, NY, USA, 2016. IEEE. doi: 10.1109/HRI.2016.7451737.

[14] C.-M. Huang and B. Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90, 2016. doi: 10.1109/HRI.2016.7451737.

[15] K.-S. Hwang, J. Ling, Y.-Y. Chen, and W.-H. Wang. Reward shaping for reinforcement learning by emotion expressions. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1288–1293, New York, NY, USA, 2014. IEEE, IEEE.

[16] H. Jiang, E. A. Croft, and M. G. Burke. Social cue detection and analysis using transfer entropy. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, page 323–332, New York, NY, USA, 2024. doi: 10.1145/3610977.3634933.

[17] D. Kontogiorgos, A. Pereira, B. Sahindal, S. van Waveren, and J. Gustafson. Behavioural responses to robot conversational failures. In *2020 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 53–62, New York, NY, USA, 2020. doi: 10.1145/3319502.3374782.

[18] D. Kontogiorgos, M. Tran, J. Gustafson, and M. Soleymani. A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 112–120, 2021.

[19] M. E. Lechuga Redondo, R. Niewiadomski, R. Francesco, and A. Sciutti. Comfortability recognition from visual non-verbal cues. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, page 207–216, New York, NY, USA, 2022. doi: 10.1145/3536221.3556631.

[20] D. W. Lee, Y. Kim, R. Picard, C. Breazeal, and H. W. Park. Multipart: Multiparty-transformer for capturing contingent behaviors in group conversations, 2023.

[21] M. Lewis, H. Wang, P. Velagapudi, P. Scerri, and K. Sycara. Using humans as sensors in robotic search. In *2009 12th International Conference on Information Fusion*, pages 1249–1256, New York, NY, USA, 2009. IEEE, IEEE.

[22] N. Mirnig, M. Giuliani, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi. Impact of robot actions on social signals and reaction times in hri error situations. In A. Tapus, E. André, J.-C. Martin, F. Ferland, and M. Ammi, editors, *Social Robotics*, pages 461–471, Cham, 2015. Springer International Publishing.

[23] M. T. Parreira, S. G. Lingaraju, A. Ramirez-Aristizabal, M. Saha, M. Kuniavsky, and W. Ju. A study on domain generalization for failure detection through human reactions in hri. *arXiv preprint arXiv:2403.06315*, 2024.

[24] J. Ravishankar, M. Doering, and T. Kanda. Zero-shot learning to enable error awareness in data-driven hri. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 592–601, New York, NY, USA, 2024. doi: 10.1145/3610977.3634940.

[25] V. Richter, B. Carlmeyer, F. Lier, S. Meyer zu Borgsen, D. Schlangen, F. Kummert, S. Wachsmuth, and B. Wrede. Are you talking to me? improving the robustness of dialogue systems in a multi party hri scenario by incorporating gaze direction and lip movement of attendees. In *Proceedings of the Fourth International Conference on Human Agent Interaction*, HAI '16, page 43–50, 2016. doi: 10.1145/2974804.2974823.

[26] M. Stiber. Effective human-robot collaboration via generalized robot error management using natural human responses. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 673–678, 2022. doi: 10.1145/3536221.3557028.

[27] M. Stiber and C.-M. Huang. Not all errors are created equal: Exploring human responses to robot errors with varying severity. In *Companion Publication of the 2020 ICMI*, ICMI '20 Companion, page 97–101, 2021. doi: 10.1145/3395035.3425245.

[28] M. Stiber, R. Taylor, and C.-M. Huang. Modeling human response to robot errors for timely error detection. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 676–683. IEEE, 2022. doi: 10.1109/IROS47612.2022.9981726.

[29] M. Stiber, R. H. Taylor, and C.-M. Huang. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 222–230, 2023. doi: 10.1145/3568162.3576990.

[30] M. Stiber, M. Spitale, H. Gunes, and C.-M. Huang. Social signal modeling in human-robot interaction. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 1358–1360, 2024. doi: 10.1145/3610978.3638163.

[31] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi. Head and shoulders: Automatic error detection in human-robot interaction. ACM, 2017.

[32] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image Vision Comput.*, 31(2):153–163, feb 2013. doi: 10.1016/j.imavis.2012.03.001.

[33] B. Zhang and H. Soh. Large language models as zero-shot human models for human-robot interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7961–7968, 2023. doi: 10.1109/IROS55552.2023.10341488.

[34] Z. Zhong, M. Martin, M. Voit, J. Gall, and J. Beyerer. A survey on deep learning techniques for action anticipation, 2023.