ERR@HRI 3.0 Challenge: Multimodal Detection of Errors and Anticipation in Human-Robot Interactions

Maria Teresa Parreira
Cornell University
Ithaca, NY, USA

Micol Spitale
Politecnico di Milano
Milan, Italy

Maia Stiber
Microsoft Research
Redmond, WA, USA

Shiye Cao
Johns Hopkins University
Baltimore, MD, USA

Amama Mahmood
Johns Hopkins University
Baltimore, MD, USA

Chien-Ming Huang
Johns Hopkins University
Baltimore, MD, USA

Hatice Gunes
University of Cambridge
Cambridge, UK

Wendy Ju
Cornell Tech
New York, NY, USA

ABSTRACT

As robots become increasingly integrated into human environments, their ability to detect and respond to errors remains critical for maintaining user trust and interaction quality. While recent advances in machine learning have improved error detection capabilities, most approaches are limited to specific contexts, controlled settings, or pre-extracted features, limiting their generalizability and applicability to real-world conditions. To address this challenge, the third edition of the ERR@HRI Challenge (ERR@HRI 3.0) provides researchers with two complementary datasets that enable end-to-end innovation in methods for both detecting and preventing errors in human-robot interaction. The challenge offers raw, non-anonymized video data from naturalistic settings: (1) the Bystander Affect Detection (BAD) dataset, containing webcam recordings of 45 participants' spontaneous reactions to robot and human failure scenarios; and (2) the Bad Idea dataset, featuring 29 participants' anticipatory facial responses while predicting action outcomes *before* failures occur. Both datasets were collected via crowdsourcing, capturing the inherent variability of real-world conditions—diverse lighting, camera angles, participant positioning, and environmental contexts. This naturalistic variability, while challenging, provides an authentic testbed for developing robust error detection systems. Participants are invited to develop machine learning models that can generalize across these diverse contexts and temporal stages. Submissions will be evaluated on standard classification metrics (e.g. F1-score) with consideration for real-world deployment constraints. This challenge is a step toward developing robust error detection and prevention systems that can operate in the variable conditions of real-world human-robot collaboration.

ICMI '26, 5-9 October 2026, Napoli, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Robot Failure, Error Detection, Human-Robot Interaction, Multimodal Interaction, Benchmarking.

ACM Reference Format:
Maria Teresa Parreira, Micol Spitale, Maia Stiber, Shiye Cao, Amama Mahmood, Chien-Ming Huang, Hatice Gunes, and Wendy Ju. 2026. ERR@HRI 3.0 Challenge: Multimodal Detection of Errors and Anticipation in Human-Robot Interactions. In *Proceedings of 28th ACM International Conference on Multimodal Interaction (ICMI '26)*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Robot errors, deviations from expected or intended behavior [5], are not merely technical malfunctions but social events that can disrupt interaction flow, diminish user trust, and negatively impact the overall quality of human-robot collaboration [15]. While recent advances in machine learning and sensor technologies have enhanced robots' perceptual and decision-making capabilities, autonomous systems still struggle to reliably detect when they have made a mistake, particularly in dynamic, real-world settings.

The challenge of error detection in human-robot interaction is multifaceted. Errors manifest in diverse ways—from functional failures like navigation mistakes or object manipulation errors, to social missteps such as interrupting users or misunderstanding conversational intent [11, 20]. Traditional approaches often rely on task-specific metrics or domain knowledge that limit generalizability across different robots, tasks, and contexts [4]. Additionally, methods that depend on users to explicitly report errors introduce delays that hinder timely recovery.

An alternative approach leverages the multimodal behavioral signals that humans naturally exhibit in response to unexpected events. Just as people infer errors from bystanders' reactions—confusion, concern, or surprise signaling that something went wrong—robots might detect their own mistakes through observable social cues [1]. Recent work has demonstrated that user reactions, including facial expressions, speech patterns, and body language, contain information about interaction failures [8, 9, 11, 16, 18, 19]. However, most existing approaches have been constrained in two important ways: first, they rely on pre-extracted features rather than raw sensor data, limiting the space of applicable methods; and second, they focus on controlled laboratory settings that may not reflect the variability of real-world deployments.

2 ERR@HRI 3.0 CHALLENGE

2.1 Previous ERR@HRI Editions

The ERR@HRI initiative was established as a platform for benchmarking both datasets and machine learning models designed to detect robot errors from multimodal behavioral signals. The inaugural ERR@HRI challenge [17], held in conjunction with ICMI'24, focused on detecting interaction ruptures during longitudinal robotic well-being coaching sessions, providing a dataset of 89 sessions across 23 participants. The challenge tracks included 1) detection of robot mistakes (e.g., interrupting or not responding), 2) detection of user awkwardness (e.g., when the coachee feels uncomfortable interacting with the robot) and 3) detection of interaction ruptures (IR) (disjunction of 1 and 2), and a total of 3 teams participated. Building on this foundation, ERR@HRI 2.0 [3], held at ACM MM'25, expanded the scope to include conversational errors from LLM-powered robots across two distinct embodiments (social robot [2] and voice assistant[10]) and five collaborative tasks, while introducing separate detection challenges for system-level errors versus user-perceived errors, and a total of 2 teams submitted results. Together, these challenges have demonstrated both the feasibility and the challenges of developing generalizable error detection models from human behavioral responses, with important contributions to the field in the form of participants' reports [6, 7, 13, 14, 21].

2.2 ERR@HRI 3.0

Building on prior editions, ERR@HRI 3.0 addresses two challenges in existing error detection research. First, we provide raw, non-anonymized video data rather than pre-extracted features, enabling participants to explore end-to-end learning approaches, modern computer vision techniques, and novel feature representations. Second, we focus on naturalistic data collected via crowdsourcing, which inherently contains the variability—in lighting, camera positioning, participant behavior, and environmental context—characteristic of real-world deployments.

The specific goals of ERR@HRI 3.0 are to:

- Advance multimodal error detection by enabling end-to-end learning from raw visual data
- Promote development of robust models that handle naturalistic data variability
- Explore the relationship between anticipatory and reactive error responses
- Establish benchmarks for generalizable error detection across contexts and temporal stages
- Foster community development of open, reproducible error detection methods

This challenge contributes to the broader goal of creating robots that can operate more safely and effectively in human environments by better understanding when and how their actions deviate from expectations.

Specifically, this challenge offers:

(1) **The Bystander Affect Detection (BAD) Dataset [1]**: Audiovisual webcam recordings of 45 participants' facial reactions to 46 robot and human failure scenarios, capturing observer responses to errors. This non-anonymized dataset enables fine-grained analysis of how people react when witnessing mistakes, providing a foundation for detecting errors through bystander behavior.

(2) **The Bad Idea Dataset [12]**: Audiovisual webcam recordings of 29 participants' anticipatory reactions while predicting whether 30 action scenarios would end well or poorly, captured *before* the outcomes were revealed. This dataset uniquely focuses on anticipatory responses, enabling research into how observable human behavior might signal impending errors before they fully occur.

The BAD and Bad Idea datasets introduce unique challenges beyond traditional generalization concerns. Both datasets were collected via crowdsourcing [1], rather than controlled laboratory conditions, meaning they inherently contain high variability in data quality. While such variability presents modeling challenges, it also offers a more realistic testbed for developing robust error detection systems that must function across diverse deployment conditions. Furthermore, these datasets are provided in non-anonymized form, preserving the original visual information rather than extracting only pre-computed features. This design decision expands the space of applicable methods, enabling researchers to apply modern computer vision techniques such as convolutional neural networks (CNNs) directly on image data, explore end-to-end learning approaches, or develop novel feature representations tailored to error detection tasks.

Participants are invited to develop machine learning models that can leverage one or more of these datasets to advance our understanding of multimodal error detection across diverse HRI contexts.

2.3 Materials

A challenge website and GitHub repository will be set up with a commitment to be maintained for at least the next 3 years. For the BAD and Bad Idea datasets, which contain non-anonymized visual data, participants must agree to strict usage terms, including: no right to redistribute; datasets can only be used for the challenge; no right to use datasets to defame human participants; and proper data security measures. Access to the datasets will be provided through signing EULA agreements.

2.4 Datasets Overview

The ERR@HRI 3.0 Challenge provides two complementary datasets, each capturing different aspects of error detection in human-robot interaction.

2.4.1 BAD (Bystander Affect Detection) Dataset. **Data format:** Raw video files (.mp4) containing participants' faces as they watched failure scenarios (average 17.84 s); Video dataset (46 videos) used as stimuli (.mp4).

Labels: Videos in the BAD dataset are labeled according to the stimulus video content: *Failure* (human or robot) or *Control* (no failure) (6 videos).

[1] https://www.prolific.com/

Table 1: Comparison of the datasets in ERR@HRI 3.0

Characteristic	BAD Dataset	Bad Idea
Participants	45	29
Total recordings	2,054 videos	951 videos
Total time (s)	36650.58	1850.54
Temporal focus	During failure	Before failure
Error types	Observed failures	Anticipated outcomes
Setting	Crowdsourced online	
Data format	Video and audio (.mp4)	
Anonymized	No	

2.4.2 Bad Idea Dataset. **Data format:** Raw video files (.mp4) of participants watching and reacting to scenarios, along with their outcome predictions (average 1.95 s). Video dataset (30 videos) used as stimuli (.mp4).

Labels: The Bad Idea dataset includes annotations for the **participant's prediction** of whether the scenario will end *Well* (scenario shown ends in good outcome, e.g., a human on a bike does not fall) or *Poorly* (scenario ends in a bad outcome, e.g. robot jumping fails the landing). Importantly, these labels are based on *what the participant predicted was going to happen* after watching the video with the outcome cut-off, not what really happens. Nonetheless, data is mostly balanced (1.15 good-to-bad outcome prediction ratio).

Table 1 provides an overview of the two datasets.

2.5 Challenge Tasks

The ERR@HRI 3.0 Challenge consists of multiple sub-challenges organized around the two datasets:

2.5.1 Track 1: Bystander Reaction Detection (BAD Dataset). Binary classification of whether the participant is observing a failure (vs. control scenario).

2.5.2 Track 2: Anticipatory Response Prediction (Bad Idea Dataset). Binary classification of participants' predicted outcome (well/poorly) from their anticipatory behavior.

2.5.3 Cross-Dataset Generalization (Optional). Participants are encouraged to explore transfer learning and generalization across datasets, e.g. train on BAD dataset and test on Bad Idea dataset.

2.6 Evaluation Metrics

Models will be evaluated using:

- **Offline metrics**: F1-score, Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC)
- **Windowed predictions with video-level labels**: Use fixed-length sliding windows, and consider a video correctly classified if any window in the video predicts the correct label.
- **Earliest Detection Time**: For correctly classified videos, report the percentage of video elapsed before the first correct prediction (lower is better).
- **False Negative Rate per video**: Count how many windows "miss" positive predictions in error/bad outcome videos.

3 CHALLENGE LOGISTICS

3.1 Challenge Format

The ERR@HRI 3.0 Challenge will be held as a half-day event. The challenge will include:

- Presentation of challenge results and winning submissions
- Invited talk on error detection in human-robot interaction
- Panel discussion on future directions for generalizable error detection systems
- Presentations from participating teams
- Winners announcement
- Networking session for participants and organizers

3.2 Soliciting Participation

We will recruit challenge participants through multiple channels:

- Email advertisements to relevant mailing lists (e.g., robotics-worldwide, HRI-announcements, SIGCHI announcements)
- Announcements on social media platforms (X, LinkedIn, research-focused communities)
- Direct outreach to research groups working in human-robot interaction, affective computing, and multimodal machine learning
- Promotion through affiliated workshops and conferences (ACM/IEEE HRI, IEEE ICRA, IEEE/RSJ IROS, IEEE ACII)
- Challenge website with registration form and comprehensive information

3.2.1 Evaluation Protocol.

1. **Training data**: Participants receive training and validation splits for each dataset. Data sets will be subject-independent to evaluate generalizations to unseen participants
2. **Test data**: Test sets (without labels) will be released one month before submission deadline
3. **Submission limits**: Each team may submit up to 3 times on the test set
4. **Ranking**: Teams will be ranked separately for each track
5. **Paper submission**: Participating teams must submit a description paper detailing their approach
6. **Code availability**: Teams are strongly encouraged to release code repositories. Code will be requested from participants for reproducibility

Teams will be ranked separately for each track based on F1-score. Winners will be selected for:

- Track 1: Bystander Reaction Detection
- Track 2: Anticipatory Response Prediction
- Best Overall Performance (across tracks)
- Best Cross-Dataset Generalization (optional track)

3.2.2 Review Process. All submitted papers will undergo peer review by members of the Technical Program Committee. Proposed reviewers include Rutherford Agbeshi Patamia (Deakin University), Xun Jiang (University of Electronic Science and Technology of China, Tongji University), Peter Tisnikar (King's College London), Lennart Wachowiak (King's College London), Pradip Pramanick (University of Naples Federico II), Ruben Janssens (Ghent University).

3.3 Important Dates

The proposed timeline for the challenge is as follows:

- **March 1, 2026**: Challenge announcement and call for participation
- **March 15, 2026**: Registration opens; training and validation data released
- **May 1, 2026**: Baseline models and evaluation scripts released
- **June 1, 2026**: Test data released (without labels)
- **June 15, 2026**: Submission deadline for predictions
- **June 22, 2026**: Paper submission deadline
- **July 15, 2026**: Notification of acceptance
- **August 6, 2026**: Camera-ready papers due
- **October 5, 2026**: Challenge workshop at ICMI 2026

3.4 Funding

Currently, this challenge does not have dedicated funding. Organizers are affiliated with institutions that support their participation in this event. We will provide certificates and a symbolic award to the challenge winners.

3.5 Organizers

Maria Teresa Parreira
PhD Candidate at the Information Science Department at Cornell University, USA
Short bio: Maria Teresa Parreira is a PhD candidate at the Information Science Department at Cornell University, supervised by Professors Wendy Ju and Malte Jung. Her research focuses on building tools for socially competent agents through direct interaction with the environment, leveraging multimodality and human-in-the-loop sensing. She was an organizer of the 2024 and 2025 ERR@HRI challenges.

Micol Spitale
Assistant Professor, DEIB, Politecnico di Milano, Italy
Short bio: Micol Spitale is an Assistant Professor (tenure-track) at the Department of Electronics, Information and Bioengineering at the Politecnico di Milano. In recent years, her research has been focused on the field of Social Signal Processing, Human-Robot Interaction, and Affective Computing, exploring ways to develop robots that are socio-emotionally adaptive in different contexts with the goal of having a positive impact on the society. She was an organizer of the 2024 and 2025 ERR@HRI challenges.

Maia Stiber
Senior Researcher at Microsoft Research, USA
Short bio: Maia Stiber is a Senior Researcher at Microsoft Research. Her research focuses on leveraging multimodal behavioral signals to develop robots that are human-aware to assist human and support well-being. She received her Ph.D. in Computer Science from Johns Hopkins University. She was an organizer of the 2024 and 2025 ERR@HRI challenges.

Shiye Cao
PhD Candidate at Johns Hopkins University, USA
Short bio: Shiye Cao is a PhD candidate in the Intuitive Computing Lab at Johns Hopkins University. Her research focuses on enhancing the social capabilities of robots to more closely emulate natural human-human interactions and empower stakeholders to co-design personalized and meaningful interactions. She was an organizer of the 2025 ERR@HRI challenge.

Amama Mahmood
Postdoctoral Fellow at Johns Hopkins University, USA
Short bio: Amama Mahmood is a Postdoctoral Fellow at Malone Center for Engineering in Healthcare at Johns Hopkins University. Her research focuses on the design, development, evaluation, and integration of AI assistants that support health and well-being. She has been awarded Malone Postdoctoral Fellowship, JHU Computer Science Department Fellowship, and Creel Family Fellowship. She is a Fulbright Scholar. She was an organizer of the 2025 ERR@HRI challenge.

Chien-Ming Huang
Associate Professor at Johns Hopkins University, USA
Short bio: Chien-Ming Huang is the John C. Malone Associate Professor in the Department of Computer Science at the Johns Hopkins University. His research focuses on designing interactive AI aimed to assist and collaborate with people. His research has received media coverage from MIT Technology Review, Tech Insider, and Science Nation. Huang completed his postdoctoral training at Yale University and received his Ph.D. in Computer Science at the University of Wisconsin–Madison. He is a recipient of the NSF CAREER award. He was an organizer of the 2024 ERR@HRI and 2025 ERR@HRI challenges.

Hatice Gunes
Full Professor of Affective Intelligence & Robotics at the University of Cambridge, UK
Short bio: Hatice Gunes is a Full Professor at the University of Cambridge's Department of Computer Science and Technology, where she directs the Cambridge Affective Intelligence and Robotics Lab (AFAR Lab) and leads award-winning research on multimodal, social, and affective intelligence for AI systems. Her work spans Machine Learning, Affective Computing, Social Signal Processing, and Robotics. She has received numerous awards, including best paper awards at prestigious conferences such as IEEE ACII and IEEE FG. She has also played key and leading roles on such events. Her research on creating robotic coaches for assessing/promoting mental wellbeing received extensive media coverage with >1,700 global reports in The Guardian, BBC News, Medical News Today, Science Daily, Telegraph, Sky News, ITV News, Bloomberg, New York Post etc. Most recently, she was named a Finalist for Sony Women in Tech. Award with Nature 2025. She was an organizer of the 2024 and 2025 ERR@HRI challenges.

Wendy Ju
Associate Professor of Information Science at the Jacobs Technion-Cornell Institute at Cornell Tech, USA
Short bio: Wendy Ju is an Associate Professor of Information Science at the Jacobs Technion-Cornell Institute at Cornell Tech and inaugural faculty in Cornell's new multi-college Design Tech department. Prof. Ju has innovated numerous methods for early-stage

prototyping of automated systems to understand how people will respond to systems before the systems are built. In her research, Prof. Ju has worked closely with industrial partners such as Toyota, Spotify, Intel, Ford, Bosch, Renault, Fiat Chrysler, Panasonic, Volvo, Nissan and Mitsubishi. She has a PhD in Mechanical Engineering from Stanford, and a Master's in Media Arts and Sciences from MIT. Her monograph on The Design of Implicit Interactions was published in 2015. She was an organizer of the 2024 and 2025 ERR@HRI challenges.

3.6 Call for Participation

The ERR@HRI 3.0 challenge addresses the problem of error detection in human-robot interaction (HRI) by providing the community with complementary datasets that span the temporal spectrum of error management, from anticipatory responses before failures occur to reactive responses during observed errors.

Upon acceptance of the ERR@HRI 3.0 terms and conditions and signing the End User Licence Agreement (EULA), we will share two datasets containing raw video recordings: (1) the Bystander Affect Detection (BAD) dataset with webcam recordings of 45 participants' reactions to 46 robot and human failure scenarios, and (2) the Bad Idea dataset with webcam recordings of 29 participants' anticipatory reactions while predicting action outcomes before failures occur. Both datasets are provided as raw video files to enable end-to-end learning approaches.

The ERR@HRI 3.0 challenge tasks are:

- **Track 1**: Bystander Reaction Detection (BAD Dataset) - Binary classification of whether the participant is observing a failure vs. control scenario
- **Track 2**: Anticipatory Response Prediction (Bad Idea Dataset) - Binary classification of participants' predicted outcome (well/poorly) from their anticipatory behavior
- **Cross-Dataset Generalization** (Optional): Transfer learning across datasets

We invite participants to collaborate in teams to submit their multi-modal ML models for evaluation, which will be benchmarked based on various performance metrics, including F1-score, and early detection metrics. Teams can participate in all tracks and they will be evaluated separately.

For more information about the challenge, check our website [PLACEHOLDER: URL]. To register and access the challenge data, please fill the registration form you will find on the website. For each task, training and validation data will be made available to participants. All submissions will be evaluated on a held-out test dataset to ensure a fair comparison. Participants will also be encouraged to submit a conference-style paper describing their proposed approach for tackling the challenge task(s) as well as the results obtained following the template of the ICMI 2026 main conference papers. All accepted challenge papers will be published in the main ACM ICMI 2026 proceedings.

Challenge timeline:

- **March 1, 2026**: Challenge announcement and call for participation
- **March 15, 2026**: Registration opens; training and validation data released
- **May 1, 2026**: Baseline models and evaluation scripts released
- **June 1, 2026**: Test data released (without labels)
- **June 15, 2026**: Submission deadline for predictions
- **June 22, 2026**: Paper submission deadline
- **July 15, 2026**: Notification of acceptance
- **August 6, 2026**: Camera-ready papers due
- **October 5, 2026**: Challenge workshop at ICMI 2026

The Organisers:

- Maria Teresa Parreira, Cornell University, USA
- Prof. Micol Spitale, Politecnico di Milano, Italy
- Dr. Maia Stiber, Microsoft Research, USA
- Shiye Cao, Johns Hopkins University, USA
- Dr. Amama Mahmood, Johns Hopkins University, USA
- Prof. Chien-Ming Huang, Johns Hopkins University, USA
- Prof. Hatice Gunes, University of Cambridge, UK
- Prof. Wendy Ju, Cornell University, USA

Contact us via email: [PLACEHOLDER]

4 CONCLUSION

The ERR@HRI 3.0 Challenge addresses critical gaps in robot error detection research by providing two complementary datasets that span the temporal spectrum of error management—from anticipatory responses before failures occur to reactive responses during observed errors. By offering raw, non-anonymized video data collected in naturalistic crowdsourced settings, the challenge enables end-to-end innovation in error detection methods while exposing models to the variability characteristic of real-world deployments.

This challenge represents a step toward developing more adaptive, context-aware robots that can operate safely in human environments. The techniques developed through this challenge have potential to advance our understanding of multimodal behavioral analysis, improve robot error awareness, and ultimately enhance the quality and safety of human-robot collaboration in real-world applications.

We invite researchers from the robotics, computer vision, affective computing, and human-robot interaction communities to participate in this challenge. We look forward to the innovative approaches that participants will develop and the insights this challenge will generate about multimodal error detection in human-robot interaction.

REFERENCES

[1] Alexandra Bremers, Maria Teresa Parreira, Xuanyu Fang, Natalie Friedman, Adolfo Ramirez-Aristizabal, Alexandria Pabst, Mirjana Spasojevic, Michael Kuniavsky, and Wendy Ju. 2023. The Bystander Affect Detection (BAD) Dataset for Failure Detection in HRI. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11443–11450.

[2] Shiye Cao, Jiwon Moon, Amama Mahmood, Victor Nikhil Antony, Ziang Xiao, Anqi Liu, and Chien-Ming Huang. 2025. Interruption handling for conversational robots. *Robotics: Science and Systems* (2025).

[3] Shiye Cao, Maia Stiber, Amama Mahmood, Maria Teresa Parreira, Wendy Ju, Micol Spitale, Hatice Gunes, and Chien-Ming Huang. 2025. ERR@HRI 2.0 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Conversations. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 14130–14135. doi:10.1145/3746027.3762073

[4] Jennifer Carlson and Robin R Murphy. 2005. How UGVs physically fail in the field. *IEEE Transactions on robotics* 21, 3 (2005), 423–437.

[5] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[6] Ruben Janssens, Eva Verhelst, and Mathieu De Coster. 2024. Predicting Errors and Failures in Human-Robot Interaction from Multi-Modal Temporal Data. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 671–676. doi:10.1145/3678957.3688388

[7] Xun Jiang, Shuangle Li, Chong Liu, and Xing Xu. 2025. Multimodal Time Series Alignment for Error Detection in Human Robot Interactions. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 14143–14149. doi:10.1145/3746027.3762075

[8] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[9] Dimosthenis Kontogiorgos, Minh Tran, Joakim Gustafson, and Mohammad Soleymani. 2021. A systematic cross-corpus analysis of human reactions to robot conversational failures. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 112–120.

[10] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies* 195 (2025), 103406.

[11] Nicole Mirnig, Manuel Giuliani, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Impact of Robot Actions on Social Signals and Reaction Times in HRI Error Situations. In *Social Robotics*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Springer International Publishing, Cham, 461–471.

[12] Maria Teresa Parreira, Sukruth Gowdru Lingaraju, Adolfo Ramirez-Artistizabal, Alexandra Bremers, Manaswi Saha, Michael Kuniavsky, and Wendy Ju. 2024. "Bad Idea, Right?" Exploring Anticipatory Human Reactions for Outcome Prediction in HRI. In *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN)*. 2072–2078. doi:10.1109/RO-MAN60168.2024.10731310

[13] Rutherford Agbeshi Patamia, Ha Pham Thien Dinh, Ming Liu, and Akansel Cosgun. 2025. Beyond Technical Failures: Multimodal Time-Series Modelling for Detecting Social Breakdowns and User Repair Attempts in Human-Robot Interaction. In *Proceedings of the 33rd ACM International Conference on Multimedia* (Dublin, Ireland) *(MM '25)*. Association for Computing Machinery, New York, NY, USA, 14136–14142. doi:10.1145/3746027.3762074

[14] Pradip Pramanick and Silvia Rossi. 2024. PRISCA at ERR@HRI 2024: Multimodal Representation Learning for Detecting Interaction Ruptures in HRI. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 666–670. doi:10.1145/3678957.3688387

[15] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (Portland, Oregon, USA) *(HRI '15)*. Association for Computing Machinery, New York, NY, USA, 141–148. doi:10.1145/2696454.2696497

[16] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. VITA: A Multi-modal LLM-based System for Longitudinal, Autonomous, and Adaptive Robotic Mental Well-being Coaching. *arXiv preprint arXiv:2312.09740* (2023).

[17] Micol Spitale, Maria Teresa Parreira, Maia Stiber, Minja Axelsson, Neval Kara, Garima Kankariya, Chien-Ming Huang, Malte Jung, Wendy Ju, and Hatice Gunes. 2024. ERR@HRI 2024 Challenge: Multimodal Detection of Errors and Failures in Human-Robot Interactions. arXiv:2407.06094 [cs.RO] https://arxiv.org/abs/2407.06094

[18] Maia Stiber. 2022. Effective Human-Robot Collaboration via Generalized Robot Error Management Using Natural Human Responses. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (Bengaluru, India) *(ICMI '22)*. Association for Computing Machinery, New York, NY, USA, 673–678. doi:10.1145/3536221.3557028

[19] Maia Stiber, Russell H Taylor, and Chien-Ming Huang. 2023. On using social signals to enable flexible error-aware hri. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 222–230.

[20] Leimin Tian and Sharon Oviatt. 2021. A Taxonomy of Social Errors in Human-Robot Interaction. *J. Hum.-Robot Interact.* 10, 2, Article 13 (feb 2021), 32 pages. doi:10.1145/3439720

[21] Lennart Wachowiak, Peter Tisnikar, Andrew Coles, Gerard Canal, and Oya Celiktutan. 2024. A Time Series Classification Pipeline for Detecting Interaction Ruptures in HRI Based on User Reactions. In *Proceedings of the 26th International Conference on Multimodal Interaction* (San Jose, Costa Rica) *(ICMI '24)*. Association for Computing Machinery, New York, NY, USA, 657–665. doi:10.1145/3678957.3688386