Flexible Robot Error Detection Using Natural Human Responses for Effective HRI

Maia Stiber
mstiber@jhu.edu
Johns Hopkins University
Baltimore, Maryland, USA

ABSTRACT

Robot errors during human-robot interaction are inescapable; they can occur during any task and do not necessarily fit human expectations. When left unmanaged, robot errors harm task performance and user trust, resulting in user unwillingness to work with a robot. Existing error detection techniques often specialize in specific tasks or error types, using task or error specific information for robust management and so may lack the versatility to appropriately address robot errors across tasks and error types. To achieve flexible error detection, my work leverages natural human responses to robot errors in physical HRI for error detection across task, scenario, and error type in support of effective robot error management.

CCS CONCEPTS

• **Computer systems organization → Robotics**.

KEYWORDS

robot error, social signals, error detection

ACM Reference Format:
Maia Stiber. 2024. Flexible Robot Error Detection Using Natural Human Responses for Effective HRI. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3610978.3638365

1 INTRODUCTION

Deployed robots are estimated to make significant errors every few hours [8]. In HRI, where effective interactions are built upon shared trust and task performance [11, 15, 26], unexpected robot errors (robot actions that stray from a user's expectations or mental model of the task and the robot [29]) damage user trust and hurt safety and task performance [6, 32], depending on their quantity and severity [27, 29]. Ignoring errors outright is not viable and if not properly managed, errors will require more recovery time and users can become disinclined to collaboration [16–18, 24, 32].

Timely and appropriate error management can address this problem [18, 37], but necessitates an understanding of errors and their impact. The basis of *error management*—adapted from aviation teamwork—has four main aspects: (1) detection, (2) classification, (3) mitigation, and (4) recovery [18, 19].

Prior work on error detection uses domain-specific knowledge [4, 9, 14, 25] and are not adaptable to new error types, scenarios, or tasks [20]. Their rigidity is due in part to the general information framework required to make a robotic system error-aware: the information needed limits detection to certain task structures or renders it ineffective when applied across scenarios. In addition, these techniques do not account for factors such as a participant's mental model of the task and robot, even though how a robot's behavior strays from a user's expectations determines whether the user considers the robot action an error [30]. Therefore, these methods cannot be easily used across person, task, or error [10].

However, error detection flexibility is important because errors are unexpected and do not fit humans' preconceptions. Moreover, the nature of the task can affect whether a robot action is an error and the situation and team members can dictate the error's impact on the collaboration [12, 30]. Thus, my overarching question: *How can we develop an automatic error detection method for unexpected robot errors during human-robot interaction that is flexible enough to be robust across task, user, and error type?*

Errors' unexpectedness is likely to generate social signals from human collaborators [13]. Through social signals, humans reflexively impart information about error, task, and their own mental model of the robot to the robot in a format common across all users because people exhibit more behaviors during error situations than error-free ones [7]. Gaze [1, 21, 22, 31], facial expressions [36], verbalizations [22], and body movements [4, 13, 22] have already been shown to effectively signal errors, including their severity and type [5, 22, 23]. However, this prior research used social scenarios with social interactions and humanoid robots. Error detection using social signals in physical HRI, where *object manipulation* is central (as opposed to *communication* in social-based interaction), has yet to be explored. The Media Equation theory [28] offers that people behave toward computers as they do in human-human social relationships and respond socially to technical robot actions [13], suggesting this is a valid area of exploration.

By developing an understanding of how we can use instinctive human reactions to robot actions, we can create a model of the user, transfer social and contextual information to the robot, and develop complementary situational awareness and knowledge about user intent. This would allow more timely error detection and effective recovery. Only one prior study has illustrated the feasibility of this modality for detecting conversational failures and that study had different social signals indicative of errors in different scenarios [22].

Through my research, I will show that *in physical HRI, social signals exhibited in response to robot errors are good indicators and enable flexible automatic error detection.*

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3638365

2 OVERALL GOALS

I focus on close-proximity, dyadic, physical HRI. My approach is two-fold: (1) understand and analyze social signals (facial action units, or AUs) in response to unexpected robot errors and (2) use these for flexible automatic error detection. I target application across task, person, and error type and execution in real-time—these factors effect error detection efficacy.

3 UNDERSTANDING SOCIAL SIGNALS IN RESPONSE TO UNEXPECTED ERRORS

Completed: AUs Across Person, Task, and Error Type. I ran two Programming by Demonstration (PbD) studies to explore users' natural responses to unexpected robot errors across person, task, and error type. The first study (N=23), PbD grocery unpacking, showed that users consistently exhibit AUs in response to unexpected errors and AUs hold discriminitive power to detect such errors (physical error type) during physical interactions in a timely manner across people, despite great variability in participant reactions and small training set [34]. This was done with a ML based error detection algorithm (trained on my study data) that used AUs at each time step to output error detection time step. *No* inputs contained task-specific information.

To explore different tasks and error types, I ran a study (N=5) evaluating AUs on PbD pipe sorting and physical, conceptual, and generalization errors types [34]. Evaluation showed that modeling AUs (with model only trained on prior study data) may be useful in error detection across PbD tasks and error types. The algorithm performed similarly with physical and conceptual errors but was delayed when considering generalization errors.

This work was extended beyond PbD to human-robot collaboration by creating and analyzing an open-source dataset from three HRI studies [35]. The dataset contains 7hrs 37min of interaction video across 73 participants, with calculated facial AUs. The first scenario was human-robot collaboration (HRC): participants completed an assembly task alongside the autonomous robotic system. The second scenario was also HRC: cooking tasks where the robot provided ingredients and the human cooked. The third scenario was the same as the prior PbD studies. Analysis showed that social signals are widespread and prompt across error, task, and scenario, potentially enabling earlier error detection. I postulate that social signals are a pivotal input source for flexible, timely error detection.

Completed: Understanding AUs and Error Context. My preliminary study (N=7) exploring users' responses to errors of varying severity (context) found that more severe errors caused faster, more intense reactions; these behaviors escalated over time [33]. This work used a subset of the data collected in [35], where I showed that I can model and distinguish between social and non-social contexts using AUs; therefore, one should potentially consider including contextualization in order to effectively employ social signals.

Ongoing: AUs Across Machine Embodiments and Interaction Paradigms. My prior explorations have had users interact with a non-anthropomorphic robot arm in both uni-directional (user tells robot to do something) and bi-directional (collaboration between user and robot) interactions; I am extending this across embodiments for error detection. I am assembling a corpus of data, including the previous dataset [35] and data collected from laboratory studies with organic and inorganic errors during user interactions with smart speakers. The corpus varies across embodiments (robot arm, social robot, smart speaker), interaction paradigms (uni- and bi-directional), and AUs. I will initially use inferential statistics to examine embodiment effects, explore AU modeling to determine if these models can be used across embodiments, and determine via modeling if error detection performance improves by incorporating interaction paradigm information to provide context for AUs.

Proposed: AUs in the Wild. All previous studies were conducted in laboratory settings and so I am planning to deploy a robot arm that makes coffee (service robot) and collect video and audio data to explore what social signals are exhibited during natural user interactions. At the end of each interaction, participants will be surveyed to collect information about their perception of that robot and its services. The study will encourage repeat interactions to explore how social signals change over time by providing punch cards for free coffee.

4 FLEXIBLE AUTOMATIC ERROR DETECTION

Completed: Framework for Flexible Error-Aware HRI. I have shown that natural AUs can be used to detect and temporally localize errors with reasonable accuracy and timeliness across different tasks, error types, and people. However, I also observed that social signals may not be sufficient as not everyone exhibits social reactions and others may overreact [34]. In addition, social signals may need to be contextualized for appropriate error detection [33].

To improve my method's robustness and traditional error detection's flexibility, I introduced a three-layer conceptual error-aware framework: explicit indicators (*e.g.,* human manual reporting [14]), implicit domain-specific indicators (*e.g.,* task tracking [4]), and implicit social signal indicators (my modification [35]). This final layer provides information shown in prior studies to potentially offer flexibility, capturing unexpectedness and variability in HRI.

Ongoing: Proactive Flexible Error Detection System and Validation. I am using the above framework—as an integrated robot error detection system [35]—to explore the benefits of proactive error detection using AUs across tasks. I created a robotic system, built on Microsoft's Platform for Situated Intelligence (\psi) [3], that inputs two synchronized video streams, calculates AUs (using OpenFace [2]) per stream, selects the AUs from the more confident facial detection, and logs them. This process was used for all of the data collection in section 3. Additionally, using speech-to-text and LLMs (allowing more fluid verbal interaction), potential errors can also be detected via speech and non-lexical utterances. The \psi interface communicates with a robot to control its actions, feeds AUs in real time to the error detection model described in section 3, and handles these speech cues. Upon error detection, the robot will automatically recover based on pre-programmed recovery behavior. I will run a between-subjects study to compare the effects of this method with a reactive one on perceived trust and teamwork metrics. The goal is to validate that proactive error detection using social signals promotes effective HRI.

ACKNOWLEDGEMENTS

This work was supported by the Johns Hopkins Institute for Assured Autonomy and National Science Foundation award #2143704.

REFERENCES

[1] Reuben M. Aronson and Henny Admoni. 2018. Gaze for Error Detection During Human-Robot Shared Manipulation. In *RSS Workshop: Towards a Framework for Joint Action*.

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*.

[3] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Patrick Sweeney, Anne Loomis Thompson, and Eric Horvitz. 2021. Platform for Situated Intelligence. arXiv:2103.15975 [cs.AI]

[4] Riccardo Bovo, Nicola Binetti, Duncan P. Brumby, and Simon Julier. 2020. Detecting Errors in Pick and Place Procedures. In *ACM Conference on Intelligent User Interfaces*.

[5] Alexandra Bremers, Alexandria Pabst, Maria Teresa Parreira, and Wendy Ju. 2023. Using Social Cues to Recognize Task Failures for HRI: A Review of Current Research and Future Directions. *arXiv preprint arXiv:2301.11972* (2023).

[6] Daniel J. Brooks, Momotaz Begum, and Holly A. Yanco. 2016. Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 487–492.

[7] Dito Eka Cahya, Rahul Ramakrishnan, and Manuel Giuliani. 2019. Static and Temporal Differences in Social Signals Between Error-Free and Erroneous Situations in Human-Robot Collaboration. In *International Conference on Social Robotics*. 189–199.

[8] Jennifer Carlson and Robin R. Murphy. 2005. How UGVs Physically Fail in the Field. *IEEE TRANSACTIONS ON ROBOTICS* 21, 3 (2005), 423–437.

[9] Greg Chance, Antonella Camilleri, Benjamin Winstone, Praminda Caleb-Solly, and Sanja Dogramadzi. 2016. An assistive robot to support dressing - strategies for planning and error handling. In *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*.

[10] Filipa Correia, Carla Guerra, Samuel Mascarenhas, Francisco S Melo, and Ana Paiva. 2018. Exploring the impact of fault justification in human-robot trust. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*.

[11] Julie V Dinh and Eduardo Salas. 2017. Factors that influence teamwork. *The Wiley Blackwell handbook of the psychology of team working and collaborative processes* (2017).

[12] Romi Gideoni, Shanee Honig, and Tal Oron-Gilad. 2022. Is it personal? The impact of personally relevant robotic failures (PeRFs) on humans' trust, likeability, and willingness to use the robot. *arXiv preprint arXiv:2201.05322* (2022).

[13] Manuel Giuliani, Nicole Mirnig, Gerald Stollnberger, Susanne Stadler, Roland Buchner, and Manfred Tscheligi. 2015. Systematic Analysis of Video Data from Different Human-Robot Interaction Studies: A Categorisation of Social Signals During Error Situations. *Frontiers in Psychology* 6 (2015).

[14] Dylan F Glas, Satoru Satake, Florent Ferreri, Takayuki Kanda, Norihiro Hagita, and Hiroshi Ishiguro. 2012. The network robot system: enabling social human-robot interaction in public spaces. *Journal of Human-Robot Interaction* 1, 2 (2012), 5–32.

[15] Victoria Groom and Clifford Nass. 2007. Can robots be teammates?: Benchmarks in human–robot teams. *Interaction studies* (2007).

[16] Svyatoslav Guznov, J Lyons, Marc Pfahler, A Heironimus, Montana Woolley, Jeremy Friedman, and A Neimeier. 2020. Robot transparency and team orientation effects on human–robot teaming. *International Journal of Human–Computer Interaction* (2020).

[17] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* (2011).

[18] Robert L Helmreich. 2000. On error management: lessons from aviation. *Bmj* 320, 7237 (2000).

[19] Robert L Helmreich, Ashleigh C Merritt, and John A Wilhelm. 2017. The evolution of crew resource management training in commercial aviation. In *Human Error in Aviation*.

[20] Shanee Honig and Tal Oron-Gilad. 2021. Expect the Unexpected: Leveraging the Human-Robot Ecosystem to Handle Unexpected Robot Failures. *Frontiers in Robotics and AI* 8 (2021).

[21] 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*. Association for Computing Machinery, 53–62.

[22] 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*.

[23] Dimosthenis Kontogiorgos, Sanne van Waveren, Olle Wallberg, Andre Pereira, Iolanda Leite, and Joakim Gustafson. 2020. Embodiment Effects in Interactions with Failing Robots. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.

[24] John Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992).

[25] Zongyu Li, Kay Hutchinson, and Homa Alemzadeh. 2022. Runtime Detection of Executional Errors in Robot-Assisted Surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[26] John Mathieu, M Travis Maynard, Tammy Rapp, and Lucy Gilson. 2008. Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of management* (2008).

[27] Bonnie M Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996).

[28] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Cambridge University Press.

[29] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario. In *International Conference on Social Robotics*. Springer, 42–52.

[30] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. 2017. Human Perceptions of the Severity of Domestic Robot Errors. In *Social Robotics*. Springer International Publishing, 647–656.

[31] Rachid Riad Saboundji and Róbert Adrian Rill. 2020. Predicting Human Errors from Gaze and Cursor Movements. In *2020 International Joint Conference on Neural Networks (IJCNN)*.

[32] 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*. 141–148.

[33] Maia Stiber and Chien-Ming Huang. 2020. Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 97–101.

[34] Maia Stiber, Russell H. Taylor, and Chien-Ming Huang. 2022. Modeling Human Response to Robot Errors for Timely Error Detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*.

[35] 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.

[36] Pauline Trung, Manuel Giuliani, Markus Miksch, Gerald Stollnberger, Susanne Stadler, Nicole Mirnig, and Manfred Tscheligi. 2017. Head and Shoulders: Automatic Error Detection in Human-Robot Interaction. In *ACM International Conference on Multimodal Interaction*.

[37] Hiroyuki Yasuda and Mitsuharu Matsumoto. 2013. Psychological impact on human when a robot makes mistakes. In *2013 IEEE/SICE International Symposium on System Integration*.