Initial Metrics for Evaluating Communication of a Robot's Self-Assessed Proficiency to Humans

Adam Norton, Holly Yanco
adam_norton,holly_yanco@uml.edu
University of Massachusetts Lowell

Jacob Crandall, Mike Goodrich
crandall,mike@cs.byu.edu
Brigham Young University

Aaron Steinfeld
steinfeld@cmu.edu
Carnegie Mellon University

1 INTRODUCTION

Advanced robots may be able to self-assess their proficiency at performing a task and communicate that proficiency to a human teammate before, during, or after a task is performed. Effective communication of robot proficiency can result in more fluent HRI across a human-robot team. Metrics are needed to evaluate the communication of self-assessed proficiency; concepts from related fields can be leveraged, including those from human-robot interaction [3] and explainable artificial intelligence [2].

2 METRICS

Several prospective metrics are defined below pertaining to the qualities of the communication conveyed by the robot. These can be used to characterize a given communication or tuned as part of a test method in order to compare communication methods.

- Succinctness: a function of the length, clarity, and precision of a communication (e.g., conveying a proficiency measure of 95% vs. a thorough explanation of how 95% proficiency was determined).
- Latency: timeliness of the communication as a function of when the information is conveyed, when the human receives it, when the human understands it, and if there are particular thresholds for this timing (e.g., the robot needs help to complete the task vs. the robot reports after the fact).
- Abstraction: the level of detail of the communication; factors include task component specificity (e.g., this object vs. an object), robot's confidence in its proficiency measures (e.g., 90% confidence vs. high confidence), and behavior performance (e.g., specific code module failure vs. object detection algorithm failure).
- Uncertainty: in addition to the confidence value of the robot's self-assessed proficiency, there will be an additional confidence value associated with the communication of that proficiency to the human (e.g., past communications of proficiency in X manner have shown to be understandable).

Metrics are also proposed to measure the human agent's interpretation of the communication. These can be measured in order to assess the effectiveness of the communication, either through subjective means (e.g., surveys) or objective responses from the human agent (e.g., if the human effectively understood the robot, then they respond in a manner that demonstrates this).

- Congruity: intersection of mental models between robot and human (e.g., robot's understanding of task progression is different than that of the human's).
- Perceptibility: awareness of the human that the robot is conveying information about its proficiency (e.g., messages could be missed due to high workload demand on the human, distraction, or lack of effective communication methods by the robot).
- Understandability: degree of understanding of the information communicated to the human, demonstrated by human reaction (i.e., participating in the task) or by in-situ questionnaires.
- Processing difficulty: the cognitive workload required from the human to process the communication; a function of all of the other metrics listed above.

Many of these metrics are dependent on or correlated with others. For example, understandability may be affected by the latency if enough time has passed between the execution of performance related to the proficiency measure and the communication of that proficiency. If the information being communicated must be responded to in a timely manner in order to continue task execution, then understandability will decay over time.

The accuracy of the proficiency being communicated must also be considered. Due to the nature of learning systems and artificial intelligence, it may be difficult to produce a ground truth to compare the robot's self-assessed proficiency to. However, it may be possible to intentionally induce proficiency assessments at prescribed levels (e.g., low, medium, high) by varying task parameters. For example, if a robot is stacking blocks and one block is intentionally positioned such that the robot will believe it is difficult to grasp (e.g., obstructed by another object), it can communicate this to the human agent. If the communication was understood by the human then they can clear the obstruction for the robot.

The proposed metrics in this paper will continue to be developed and will be exercised in upcoming evaluations of a prototype robot system that is able to self-assess its proficiency [1]. The structure of the evaluations will be used to inform the development of test method concepts that are more broadly applicable to evaluating the accuracy and effectiveness of explanations provided by robots.

3 ACKNOWLEDGEMENTS

This work has been supported in part by the Office of Naval Research (N00014-18-1-2503).

REFERENCES

[1] Zhao Han, Jordan Allspaw, Adam Norton, and Holly A Yanco. 2019. Towards A Robot Explanation System: A Survey and Our Approach to State Summarization, Storage and Querying, and Human Interface. *arXiv preprint arXiv:1909.06418* (2019).

[2] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[3] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. 33–40.