The Importance of Word Choice for Trusting Robot Explanations

Gregory LeMasurier[1], Giordano Arcieri[2], and Holly A. Yanco[1]

Abstract— **It is especially important that robots are able to explain their failures in a manner that helps a person understand what went wrong and to appropriately align their trust with the system. In this work we updated our Generative explanation system to create explanations catered to novices. We also leverage context to help ground the language used in explanations with the environment. Through our study we highlight the importance of word choice on people's perception of a robot's trustworthiness.**

I. INTRODUCTION AND RELATED WORK

Robots must effectively communicate their failures to people who may be nearby. When explaining failures, a robot should do so in a manner that appropriately aligns the person's trust in the system.

Robots can use several trust repair strategies such as apologies, explanations, denials, and promises to help a person align their trust in the system [1]. These trust repair strategies can convey the same information through different language choices, which can differ in effectiveness and can influence a person's perception of the system.

Explanations and apologies have been found to be effective trust repair strategies [1]. Systems that explain their behavior or failures have been shown to increase people's trust [2], [3], [4], [5], [6]. Explanations can contain different types of information and can be structured in many different ways [7], [8], [9], [10], [11], [12]. Robot explanations should take the recipients' roles and experience into account [13] and should include sufficient details for non-experts to understand and provide appropriate assistance [14].

Explanations can be created using templates or by generating text [15]. Templated explanations are more controlled and can be more accurate compared to generative approaches. However, this comes with a trade-off, as templated explanations are not as fluent as generative explanations. This is especially important to consider as large language models are becoming popular tools for generating explanations of robot behavior [16], [17], [18], [19], [20]. Although accuracy has been argued to be more important than fluency [21], several participants in our prior work [22] commented on the importance of fluency on robot trust (e.g., "I lose a little faith in a supposed smart robot when its explanations aren't spoken in correct English"). This motivated us to compare Templated and Generative explanations, where we found that people surprisingly perceived the robot using Templated explanations to be similarly or more trustworthy compared to those using Generative explanations [17]. In this work we evaluate our new Generative explanations and further investigate these mixed results.

II. METHODOLOGY

To generate explanations when a robot fails, we utilize our Proactive explanation architecture and LLMs [17]. This architecture takes advantage of the hierarchical structure of Behavior Trees (BTs) [23] to automatically generate robot explanations [24]. The robot can leverage information from the BT by framing its internal states and actions into semantic sets: {goal, subgoals, steps, actions} [24]. We also incorporate Assumption Checkers (ACs) [25], [26], into the BT to track information about the robot's internal states or environmental conditions throughout task execution. By using robot and object profiles, this architecture can abstract out robot and object specific information leading to a more generalizable framework [27].

Through this framework, a robot can proactively detect and explain failures before they occur, resulting in better human perception and more understanding of the robot's failure [22]. The robot's explanations consist of information from the BTs, ACs, and the Object and Robot Profiles. These explanations can be created using templates or can be generated using LLMs [17] such as GPT-4o. To enable LLMs to generate accurate explanations we prompt it with information obtained from the BT, robot profile, object profiles to give it context for the scenario and the robot's functions. We also prompt the LLM with information obtained at the time the system predicted that a failure would occur including information from the BT and ACs that were violated.

In this work, we modified the prompt used in our previous work [17] with the goal of generating explanations that were better suited for novices. In addition, we add a condition where the prompt included an image from the robot's camera at the time of failure. This provided the LLM with context that it could use to help ground its language, such as references to objects, to the real world. Our complete prompt can be found on our GitHub repository[1].

We then conducted a user study to evaluate our explanation systems which used the same manufacturing scenario as

*This research has been funded by the Office of Naval Research (N00014-18-1-2503, N00014-23-1-2744, N00014-24-1-2634). The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Dept. of Defense or the U.S. Government.

[1]Richard A. Miner School of Computer & Information Sciences, University of Massachusetts Lowell, Lowell, MA, USA

[2]Manning College of Information & Computer Sciences, University of Massachusetts, Amherst, MA, USA

Corresponding author: gregory_lemasurier@student.uml.edu

[1] github.com/uml-robotics/GPTRobotFailureExplanationPrompts/

TABLE I

EXPLANATIONS USED FOR EACH CONDITION.

Explanation Type	Screw Bin Empty	Screw Bin Moved	Caddy Out of Reach
TEM	I do not see any screws on the table so I will not be able to pick screw.	Objects appear to have moved so I will not be able to pick screw.	My arm can not reach the caddy so I will not be able to place object into caddy.
GEN	I couldn't find any screws on the table so I can't pick one up.	Objects seem to have shifted, so I can't pick up the screw.	I couldn't find a way to reach the caddy, so I can't place the screw.
GEN+C	I don't see any screws in the blue bin, so I can't pick a screw.	The screws in the green bin have shifted, so I can't safely pick one up.	I can't reach the green caddy because it's too far, so I cannot place the screw.

our prior work [22], [17]. Participants observed videos of a Fetch mobile manipulator robot working on a kit assembly task with a worker (the experimenter) with the objective of evaluating its performance. The experimenter would imitate potential real-world failures by manipulating the environment (e.g., moving the screw bin while the robot was in the process of picking up a screw). When the robot proactively detected that a failure would likely occur based on violated ACs, it would then generate an explanation.

This study was a mixed 3 (System Type: *Templated (TEM)*, *Generative (GEN)*, *Generative with Context (GEN+C)*) × 3 (Failure Type: *Screw Bin Empty*, *Screw Bin Moved*, *Caddy Out of Reach*) design online user study which was conducted on Prolific (N = 252). The first explanation generated by our Generative systems for each Failure Type were used in this study. The explanations used in each condition can be found in Table I. Nine total videos[2] of the robot interacting with a person were made, one for each of the combinations of system type × failure type. Each participant watched three videos and experienced all three explanation and all three failure types, one of each per video according to their randomly assigned configuration. The ordering of the scenarios was counterbalanced to reduce ordering effects.

We hypothesized that GEN+C would be perceived as more trustworthy compared to TEM and GEN, which do not have access to the environmental context (H1).

III. RESULTS

We evaluated people's perceived trustworthiness of each system through the MDMT V2 Capacity trust scale [28]. We ran Kruskall-Wallis tests and did not observe significant differences between our system types across each of our failure types: Screw Bin Empty ($\chi^2(2) = 2.27, p = 0.321$) TEM: ($M = 2.96, SD = 1.18$), GEN: ($M = 2.86, SD = 1.15$), GEN+C: ($M = 3.14, SD = 1.10$)), Screw Bin Moved ($\chi^2(2) = 1.53, p = 0.465$) TEM: ($M = 2.54, SD = 1.05$), GEN: ($M = 2.46, SD = 1.13$), GEN+C: ($M = 2.66, SD = 1.14$)), and Caddy Out of Reach ($\chi^2(2) = 0.65, p = 0.722$) TEM: ($M = 3.03, SD = 1.12$), GEN: ($M = 3.07, SD = 1.12$), GEN+C: ($M = 2.94, SD = 1.05$)).

IV. DISCUSSION

Ultimately, we did not observe support for H1. While providing additional context to help ground the explanations did not result in significant changes in perceived trustworthiness, this information could still be beneficial to a user's understanding. In future work we plan to further investigate users' responses to our understandability questions.

In this study we did not observe the same significant differences as we had in our prior work [17]. We had previously observed a significant difference in perceived trustworthiness between our Generative and Templated explanations, where the Templated explanation was perceived as more trustworthy in the Caddy Out of Reach Failure Condition. Both studies followed the same procedure, used the same videos with edited audio for the explanations, and our experimenters validated our Generative explanations to ensure that they were accurate and contained the same type of information. Since the content of the explanations were valid, we believe that the language used to communicate the failure may explain the differences we observed across studies.

In our previous work, we observed that language used by the LLM seemed more technical in some cases. This was especially true for the Caddy Out of Reach Failure Type's Generative explanation: "I couldn't generate a path to the caddy, so I'm unable to position the screw for placement in the caddy." The phrases "generate a path" and "position the screw for placement" do not describe the failure in language that a person would likely use when talking to another person. We believe this was because we had prompted the LLM with technical information, including a BT, leading it to generate more technical explanations. In this study, we modified our prompt to guide the LLM to generate explanations for novices and as a result our GEN and GEN+C explanations did not contain technical terminology. We hypothesize that this helped them perform on par with our handcrafted TEM explanations. This implies that it is very important for LLMs to be prompted to generate responses catered to the roles and experience of the recipient, as has also been suggested by existing literature [13].

Based on participants' responses to free response questions in our prior work, we anticipated that grammatically incorrect explanations from TEM would be perceived as less trustworthy than the grammatically correct GEN and GEN+C explanations. We did not observe significant differences between our explanation groups which suggest that grammar might not impact people's trust as much as the language used to describe the failures. This implies that the tradeoff between generalizability and consistency could be more important to consider than the tradeoff between fluency and accuracy.

[2] youtube.com/playlist?list=PLIwwT33Qq2HRE7w-XSjfr5FiwtR8lyifo

References

[1] C. Esterwood and L. P. Robert, "Repairing Trust in Robots?: A Meta-analysis of HRI Trust Repair Studies with A No-Repair Condition," in *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE Press, 2025, p. 410–419.

[2] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[3] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations," in *Proceedings of the 2016 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

[4] M. Kwon, S. H. Huang, and A. D. Dragan, "Expressing Robot Incapability," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2018, pp. 87–95.

[5] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill, "Is It My Looks? Or Something I Said? The Impact of Explanations, Embodiment, and Expectations on Trust and Performance in Human-Robot Teams," in *International Conference on Persuasive Technology*. Springer, 2018, pp. 56–69.

[6] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, "A tale of two explanations: Enhancing human trust by explaining robot behavior," *Science Robotics*, vol. 4, no. 37, 2019.

[7] S. Stange and S. Kopp, "Effects of a Social Robot's Self-Explanations on How Humans Understand and Evaluate Its Behavior," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2020, pp. 619–627.

[8] D. Das, S. Banerjee, and S. Chernova, "Explainable AI for Robot Failures: Generating Explanations that Improve User Assistance in Fault Recovery," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2021, pp. 351–360.

[9] R. Thielstrom, A. Roque, M. Chita-Tegmark, and M. Scheutz, "Generating Explanations of Action Failures in a Cognitive Robotic Architecture," in *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*, 2020, pp. 67–72.

[10] L. Wachowiak, A. Fenn, H. Kamran, A. Coles, O. Celiktutan, and G. Canal, "When Do People Want an Explanation from a Robot?" in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. ACM, 2024.

[11] P. Khanna, E. Yadollahi, M. Björkman, I. Leite, and C. Smith, "Effects of Explanation Strategies to Resolve Failures in Human-Robot Collaboration," in *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2023, pp. 1829–1836.

[12] A. Rossi and S. Rossi, "On the Way to a Transparent HRI," in *Adjunct Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization (UMAP)*, 2024, pp. 215–219.

[13] M. Ribera and À. Lapedriza García, "Can we do better explanations? A proposal of User-Centered Explainable AI," in *Joint Proceedings of the ACM IUI 2019 Workshops*. CEUR Workshop Proceedings, 2019.

[14] Z. Han, E. Phillips, and H. A. Yanco, "The Need for Verbal Robot Explanations and How People Would Like a Robot to Explain Itself," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–42, 2021.

[15] A. Gatt and E. Krahmer, "Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[16] Z. Liu, A. Bahety, and S. Song, "REFLECT: Summarizing Robot Experiences for Failure Explanation and CorrecTion," *7th Conference on Robot Learning (CoRL 2023)*, 2023.

[17] G. LeMasurier, C. Tagliamonte, J. Breen, D. Maccaline, and H. A. Yanco, "Templated vs. Generative: Explaining Robot Failures," in *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2024, pp. 1346–1353.

[18] C. P. Lee, P. Praveena, and B. Mutlu, "REX: Designing User-centered Repair and Explanations to Address Robot Failures," in *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024, pp. 2911–2925.

[19] D. Sobrín-Hidalgo, M. A. González-Santamarta, Á. M. Guerrero-Higueras, F. J. Rodríguez-Lera, and V. Matellán-Olivera, "Explaining Autonomy: Enhancing Human-Robot Interaction through Explanation Generation with Large Language Models," *arXiv preprint arXiv:2402.04206*, 2024.

[20] F. Gebellí, L. Hriscu, R. Ros, S. Lemaignan, A. Sanfeliu, and A. Garrell, "Personalised Explainable Robots Using LLMs," in *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE Press, 2025, p. 1304–1308.

[21] E. Reiter. (2019) Generated Texts Must Be Accurate! (Accessed on: 2024-06-06). [Online]. Available: https://ehudreiter.com/2019/09/26/generated-texts-must-be-accurate/

[22] G. LeMasurier, A. Gautam, Z. Han, J. W. Crandall, and H. A. Yanco, "Reactive or Proactive? How Robots Should Explain Failures," in *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2024, pp. 413–422.

[23] M. Colledanchise and P. Ögren, *Behavior Trees in Robotics and AI: An Introduction*. CRC Press, 2018.

[24] Z. Han, D. Giger, J. Allspaw, M. S. Lee, H. Admoni, and H. A. Yanco, "Building The Foundation of Robot Explanation Generation Using Behavior Trees," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 3, pp. 1–31, 2021.

[25] A. Gautam, T. Whiting, X. Cao, M. A. Goodrich, and J. W. Crandall, "A Method for Designing Autonomous Robots that Know Their Limits," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 121–127.

[26] X. Cao, A. Gautam, T. Whiting, S. Smith, M. A. Goodrich, and J. W. Crandall, "Robot Proficiency Self-Assessment Using Assumption-Alignment Tracking," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 3279–3298, 2023.

[27] C. Tagliamonte, D. Maccaline, G. LeMasurier, and H. A. Yanco, "A Generalizable Architecture for Explaining Robot Failures Using Behavior Trees and Large Language Models," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2024, pp. 1038–1042.

[28] D. Ullman and B. F. Malle, "MDMT: Multi-Dimensional Measure of Trust," 2019.