Explanation Styles for Trustworthy Autonomous Systems

Extended Abstract

David A. Robb
Heriot-Watt University
Edinburgh, United Kingdom
d.a.robb@hw.ac.uk

Xingkun Liu
Heriot-Watt University
Edinburgh, United Kingdom
x.liu@hw.ac.uk

Helen Hastie
Heriot-Watt University
Edinburgh, United Kingdom
h.hastie@hw.ac.uk

ABSTRACT

We present a study that explores the formulation of natural language explanations for managing the appropriate amount of trust in a remote autonomous system that fails to complete its mission. Online crowd-sourced participants were shown video vignettes of robots performing an inspection task. We measured participants' mental models, their confidence in their understanding of the robot behaviour and their trust in the robot. We found that including history in the explanation increases trust and confidence, and helps maintain an accurate mental model, but only if context is also included. In addition, our study exposes that some explanation formulations lacking in context can lead to misplaced participant confidence.

KEYWORDS

Explanations; transparency; trust; robot faults; mental models

ACM Reference Format:
David A. Robb, Xingkun Liu, and Helen Hastie. 2023. Explanation Styles for Trustworthy Autonomous Systems: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

As robots and autonomous systems (RAS) are becoming increasingly deployed, it is important to understand the reasoning behind their decisions. However, the behaviour of these robots can seem opaque. Transparency is closely linked to trust [8, 12] and explainability can be a means to facilitate this transparency [11]. It is key that the user has the correct mental model of what the system can and cannot do, so that the system is not undertrusted or overtrusted [5]. This will increase adoption of these systems going forward. Das *et al.* [2] explored explanation styles in the context of failing robots that were co-located with the users (e.g. in homes and hospitals). When users are remote from their robot or autonomous systems (e.g. systems for inspecting offshore wind farms or nuclear energy plants) trust has been shown to be generally lower [1, 4, 6]. Therefore, the influence of explanations on trust for remote robots is even more important to understand. Here, we have explored the use of such robots with a commercial organisation, for use in offshore inspection, developing a natural language interface for use by a remote operator to garner situation awareness of the remote robot [9]. Trust was recognised as being key in the acceptance of these new inspection robots. Our research questions (RQ) are as follows:

- RQ1: How do different types of explanations affect a user's mental model of a remote mobile robot, in terms of their ability to correctly identify the cause of robot failure, and their confidence in the accuracy of their failure identification?
- RQ2: How do different types of failure explanation affect a user's trust of a remote mobile robot?

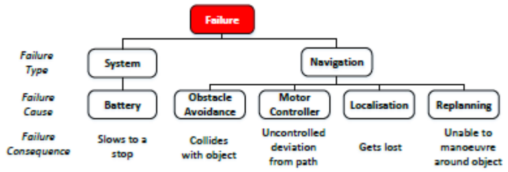

Figure 1: Fault Taxonomy

2 STUDY DESIGN & RESULTS

We developed a remote robot fault taxonomy (Fig. 1) and handcrafted explanations, adapted from Das *et al.*, [2021] and based on insights expressed by our industry partner. Video vignettes[1] were produced of a simulated remote mobile inspection robot moving in a gas station displaying fault behaviours as per Fig.1. Participants watched these videos followed by text of the robot's explanation of its behaviour. Participants were recruited via the online platform, Prolific. We used a between-groups design with the style of explanations varying between conditions (Table 1). The control condition (CTL) was an explanation containing no new information and simply restated the robot's goal. Participants were assigned to a condition randomly, completed a pre-task robot attitude questionnaire to rule out between group bias, did their fault identification task of 7 trials and, finally, completed a post-task trust questionnaire (the 7-item Letras-G Trust scale used by Lim et al [2022]). Each task trial consisted of watching a video, identifying what it showed (by multiple choice response), reading the robot's text explanation of its behaviour (Table 1), and, lastly, identifying again what had occurred, in the light of the explanation. Two of the trials were for familiarisation showing successful robot missions and the other five failing missions were randomly ordered. Mental Model (MM) Correctness (DeltaCor (-1.0 to 1.0)) and MM Confidence (DeltaConf (-100 to 100)) were gathered by assessing the difference between the correctness of their failure identification (0 or 1), and their confidence in that identification (on a scale 0 to 100) [10], before and after reading the robot's explanation (averaged over all the failure trials). 238 sets of responses were analysed.

Mental Model Correctness Results. With regards RQ1, an Independent-Samples Kruskal-Wallis test on participants' DeltaCor

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

[1] Available as a playlist: https://bit.ly/robotexpstudyplaylist

Table 1: Formulation of the conditions, showing the elements included in each condition for one example video vignette depicting system failure: "Low battery level" in Figure 1. Column heads, a_t, a_{t-1}, and c_t indicate the *current action*, *prior action(s)* and *context/circumstances* elements respectively, with t indicating *time*.

Abbr	*Condition*	a_t	a_{t-1}	c_t	*Example robot explanation*
CTL	Control Condition				*My goal was to move from Point A to Point B.*
AB	Action Based	✓			*I could not navigate on my planned path to my destination.*
CB	Context Based	✓		✓	*I could not navigate a path to my destination because I had been working hard previously and was low on power.*
ABH	Action Based History	✓	✓		*I departed from my start point and was navigating a path to my destination, Then I slowed to a stop and I could not continue on my planned path to my destination.*
CBH	Context Based History	✓	✓	✓	*I departed from my start point and was navigating a path to my destination. Then I slowed to a stop because I had been working hard previously and was low on power and so I could not continue on my planned path to my destination.*

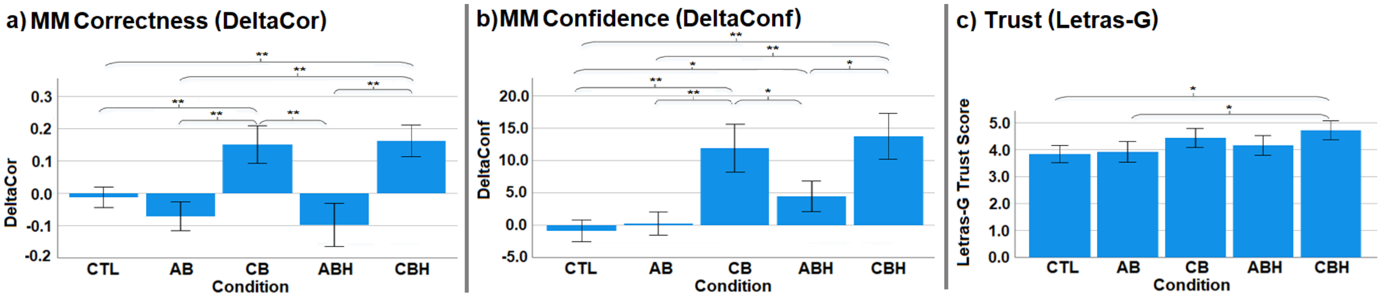

Figure 2: Mental model (MM) and Trust measurement charts showing the means and 95% confidence limits by condition group. Statistically significant pairwise differences are marked: * p<.05, ** p<.001

showed that there is a statistically significant difference between the DeltaCor scores of the condition groups ($\chi^2(4)$ = 70.530, p<.001). See Fig 2a. We can see that, in the control condition, correctness changes little between viewing the video and reading the explanation. However, reading the explanations in AB and ABH conditions leads to a negative (non-significant) effect on correctness compared to CTL. Reading the explanations in both CB and CBH conditions leads to statistically significantly improved correctness (positive DeltaCor) compared to CTL, AB and ABH.

Mental Model Confidence Results. The same test was conducted on participants' DeltaConf score, showing a statistically significant effect of Condition on participants' DeltaConf scores ($\chi^2(4)$ = 75.384, p<.001). Fig. 2b shows that, in CTL, confidence changes little between viewing the video and reading the explanation; likewise with AB. However, we see that reading the explanations in CB, ABH, and CBH conditions leads to statistically significantly increased confidence (positive DeltaConf) compared to CTL.

Trust Measurement Results. With regards RQ2, Fig. 2c shows that CBH elicits the highest trust, and CTL the lowest. An Independent-Samples Kruskal-Wallis test on these Letras-G scale scores showed a statistically significant difference between the trust scores of the condition groups ($\chi^2(4)$ = 16.911, p=.002). Pairwise comparisons, using Bonferroni correction, show statistically significant differences between CBH and CTL and between CBH and AB conditions.

3 DISCUSSION AND FUTURE WORK

Regarding RQ1, we clearly show that providing an explanation can increase the fidelity of a mental model and confidence in that mental model, with CBH explanations performing the best (confirming findings from Das *et al.*). Both AB and ABH seem to have a negative effect on fault identification. Interestingly, ABH led to a statistically significant *increase* in confidence (positive DeltaConf) compared to CTL. This unexpected combination of negative DeltaCor with positive DeltaConf is evidence that including history without context might lead to users having misplaced confidence in the quality of their MM. This could affect decision making and have serious consequences in the real world. However, participants in this study were lay people and expert operators might find these too verbose and prefer shorter explanations [3], to be explored in future work.

Regarding RQ2, CBH seems to be the best explanation method for increased trust, significantly higher than the CTL and AB (both of which just state facts). With regards the adoption of robots, it is key to manage trust so that it is appropriate. Given these findings, we can confidently state that including an explanation can help maintain a certain level of trust and, importantly, help rebuild the trust after an error. We have shown that providing explanations gives the user more understanding of the robot actions and thus more reassurance of the robot's autonomous capability, increasing trust over baselines that provide no or little explanations.

ACKNOWLEDGMENTS

We thank our industrial partner, Total Energies, and our funder, UKRI, through grants: HUman-machine teaming for Maritime Environments (HUME) EP/V05676X/1, Trustworthy Autonomous Systems Node in Trust EP/V026682/1, and UK Robotics and Artificial Intelligence Hub for Offshore Energy Asset Integrity Management (The Orca Hub) EP/R026173/1.

REFERENCES

[1] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *Proceedings of RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication.* IEEE, 701–706.

[2] Devleena Das, Siddhartha Banerjee, and Sonia Chernova. 2021. 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.* 351–360.

[3] Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference. on Natural Language Generation.* 99–108.

[4] Helen Hastie, Xingkun Liu, and Pedro Patron. 2017. Trust triggers for multimodal command and control interfaces. In *Proceedings of the 19th ACM International Conference. on Multimodal Interaction.* 261–268.

[5] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.

[6] Jinke Li, Xinyu Wu, Tiantian Xu, Huiwen Guo, Jianquan Sun, and Qingshi Gao. 2017. A novel inspection robot for nuclear station steam generator secondary side with self-localization. *Robotics and biomimetics* 4, 1 (2017), 1–9.

[7] Mei Yii Lim, José David Aguas Lopes, David A Robb, Bruce W Wilson, Meriam Moujahid, Emanuele De Pellegrin, and Helen Hastie. 2022. We are all Individuals: The Role of Robot Personality and Human Traits in Trustworthy Interaction. In *Proceedings of the 2022 31st IEEE International Conference. on Robot and Human Interactive Communication (RO-MAN).* IEEE, 538–545.

[8] Birthe Nesset, David A Robb, José Lopes, and Helen Hastie. 2021. Transparency in hri: Trust and decision making in the face of robot errors. In *Proceedings Companion of the 2021 ACM/IEEE International Conference. on Human-Robot Interaction.* 313–317.

[9] David A Robb, José Lopes, Stefano Padilla, Atanas Laskov, Francisco J Chiyah Garcia, Xingkun Liu, Jonatan Scharff Willners, Nicolas Valeyrie, Katrin Lohan, David Lane, et al. 2019. Exploring interaction with remote autonomous systems using conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference.* 1543–1556.

[10] Marta Romeo, Peter E McKenna, David A Robb, Gnanathusharan Rajendran, Birthe Nesset, Angelo Cangelosi, and Helen Hastie. 2022. Exploring Theory of Mind for Human-Robot Collaboration. In *Proceedings of the 2022 31st IEEE International Conference. on Robot and Human Interactive Communication (RO-MAN).* IEEE, 461–468.

[11] Alan FT Winfield, Serena Booth, Louise A Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I Muttram, Joanna I Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, et al. 2021. IEEE P7001: A proposed standard on transparency. *Frontiers in Robotics and AI* (2021), 225.

[12] Robert H Wortham. 2020. *Transparency for Robots and Autonomous Systems: Fundamentals, technologies and applications.* Institution of Engineering and Technology.