Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots

Francisco J. Chiyah Garcia
Heriot-Watt University
Edinburgh, UK
fjc3@hw.ac.uk

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

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

Atanas Laskov
SeeByte Ltd
Edinburgh, UK
atanas.laskov@seebyte.com

Pedro Patron
SeeByte Ltd
Edinburgh, UK
pedro.patron@seebyte.com

Helen Hastie
Heriot-Watt University, MACS
Edinburgh, UK
h.hastie@hw.ac.uk

ABSTRACT

Autonomous systems in remote locations have a high degree of autonomy and there is a need to explain *what* they are doing and *why* in order to increase transparency and maintain trust. Here, we describe a natural language chat interface that enables vehicle behaviour to be queried by the user. We obtain an interpretable model of autonomy through having an expert 'speak out-loud' and provide explanations during a mission. This approach is agnostic to the type of autonomy model and as expert and operator are from the same user-group, we predict that these explanations will align well with the operator's mental model, increase transparency and assist with operator training.

KEYWORDS

Multimodal output, natural language generation, autonomous systems, trust, transparency, explainable AI

ACM Reference Format:
Francisco J. Chiyah Garcia, David A. Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots. In *Proceedings of Explainable Robotic Systems Workshop, ACM Human-Robot Interaction conference.* ACM, New York, NY, USA, 2 pages.

1 INTRODUCTION

Autonomous systems (AxV) now routinely operate in regions that are dangerous or impossible for humans to reach, such as the deep underwater environment. Typically, remote robots instil less trust than those co-located [1, 6]. This combined with high vulnerability in hazardous, high-stakes environments, such as that described in [7], means that the interface between operator and AxV is key in maintaining situation awareness and understanding. Specifically, AxVs need to be able to maintain a continuous communication with regards to what they are doing; and increase transparency through explaining their actions and behaviours.

Explanations can help formulate clear and accurate mental models of autonomous systems and robots. Mental models, in cognitive theory, provide one view on how humans reason either functionally (understanding what the robot does) or structurally (understanding how it works). Mental models are important as they strongly

Explainable Robotic Systems Workshop, ACM Human-Robot Interaction conference, March 2018, Chicago, IL USA
2018.

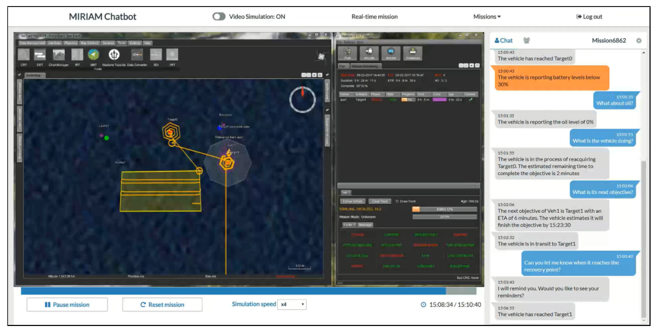

Figure 1: The MIRIAM multimodal interface with SeeTrack interface showing the predicted path of the vehicle on the left and the MIRIAM chat interface on the right

impact how and whether robots and systems are used. In previous work, explanations have been categorised as either explaining 1) machine learning as in [11] who showed that they can increase trust; 2) explaining plans [2, 13]; 3) verbalising robot [12] or agent rationalisation [3]. However, humans do not present a constant verbalisation of their actions but they do need to be able to provide information on-demand about *what* they are doing and *why* during a live mission.

We present here, MIRIAM, (Multimodal Intelligent inteRactIon for Autonomous systeMs), as seen in Figure 1. MIRIAM allows for these 'on-demand' queries for status and explanations of behaviour. MIRIAM interfaces with the Neptune autonomy software provided by SeeByte Ltd and runs alongside their SeeTrack interface. In this paper, we focus on explanations of behaviours and describe a method that is agnostic to the type of autonomy method. With respect to providing communication for monitoring, please refer to [5] for further details and an overview of the system.

2 EXPLANATIONS FOR REMOTE AUTONOMY

Types of explanations include *why* to provide a trace or reasoning and *why not* to elaborate on the system's control method or strategy [4]. Lim et al. (2009) [10] show that both *why* and *why not* explanations increase understanding but only *why* increases trust. We adopt here the 'speak-aloud' method whereby an expert provides rationalisation of the AxV behaviours while watching videos of missions on the SeeTrack software. This has the advantage of being

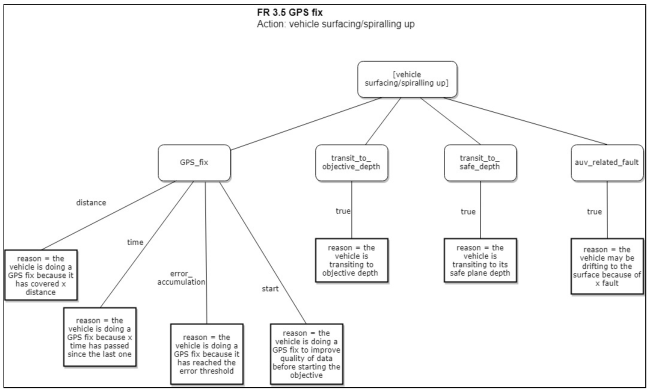

Figure 2: Autonomy model for vehicle surfacing

agnostic to the method of autonomy and could be used to describe rule-based autonomous behaviours but also complex deep models. Similar human-provided rationalisation has been used to generate explanations of deep neural models for game play [3].

An interpretable model of autonomy was then derived from the expert, as partially shown in Figure 2. If a *why* request is made, the decision tree is checked against the current mission status and history and the possible reasons are determined, along with a probability. As we can see from example outputs in Figure 3A, there may be multiple reasons with varying levels of certainty depending on the information available at a given point in the mission. Hence in this example, when the same *why* question is asked at a later point then only one higher confidence answer is given.

In the example scenario given in Figure 3B, the operator is able to observe in the SeeTrack interface that the vehicle has not done a GPS fix for some time. The user asks why it is not doing a GPS fix and the answer explains the relevant constraints on the vehicle, as captured in the interpretable autonomy model. The surface representations of the explanations are generated using template-based NLG. The wording of the output reflects the certainty on three levels: above 80% (high), 80% to 40% (medium) and below 40% (low).

3 DISCUSSION AND FUTURE WORK

Future work includes conducting user evaluations to examine the trade-off between providing all of the information, even if one is not 100% sure ('completeness') versus providing only those statements with very high confidence ('soundness'). The former is shown in Figure 3. This trade-off will vary between personnel with different information needs and expertise and domains as discussed in [9]. Verbal indicators (e.g. "It is likely/probable") have been used in weather reporting to reflect levels of certainty [8]. However, informal feedback from users indicate that the use of such verbal indicators may reduce confidence in the reporting system and therefore may not be suitable for highly critical, high risk situations. Exactly how these should be expressed is the subject of future work.

We present here a method for explaining behaviours of remote AxV, which is agnostic to the autonomy model. Fortunately, in this domain, it is appropriate for the expert to be from the same pool of end-users (i.e. operators) and therefore explanations are likely to align with their mental models and assumptions about the

Chat A: Why	Chat B: Why not
User: Why is the vehicle spiralling down? **System:** The vehicle is spiralling down to Suvery0's depth (16m) (high confidence) **User:** Why is the vehicle spiralling up? **System:** This could be due to the following reasons: -The vehicle is performing a GPS fix (medium confidence) -The vehicle is transiting to its safe plan depth (5m) (medium confidence) .... **User:** Can you confirm now why it is spiralling up? **System:** The vehicle is performing a GPS fix (high confidence) **User:** Why is it doing a GPS fix? **System:** The vehicle is doing a GPS fix at the start of Survey0 to improve the quality of the data gathered (medium confidence)	**User:** What is the vehicle doing now? **System:** The vehicle is in transit to Point0. The estimated remaining time to complete the objective is 15 minutes **User:** Why is it spiralling? **System:** The vehicle is transiting to its safe plane depth (5m) (high confidence) **User:** Why is the vehicle not doing a GPS fix? **System:** The vehicle is not allowed to do GPS fixes during its transit to Point0 (high confidence)

Figure 3: A: MIRIAM chat showing examples of *why* explanations; B showing examples *why not* explanations

system. This will not always be the case, as described in [2], e.g. for in-home help robots, where users and planning experts have disparate mental models. Future work will involve the evaluation of explanations with respect to mental models.

ACKNOWLEDGMENTS

This research was funded by DSTL (REGIME 0913-C2-Ph2-009 and MIRIAM/ACC101939); EPSRC (EP/R026173/1, 2017-2021); RAEng/ Leverhulme Trust Senior Research Fellowship Scheme (Hastie/ LT-SRF1617/13/37). Thanks to Prof. Y Petillot and Dr Z Wang.

REFERENCES

[1] W. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati. 2008. The Effect of Presence on Human-Robot Interaction. *Proc. IEEE Int. Symp. on Robot and Human Interactive Communication* (2008), 701–706.
[2] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proc. of IJCAI'17.* 156–163. arXiv:1701.08317
[3] U. Ehsan, B. Harrison, L. Chan, and M. O. Riedl. 2017. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. arXiv:1702.07826
[4] S. Gregor and I. Benbasat. 1999. Explanations From Intelligent Systems: Theoretical Foundations and Implications for Practice. *MIS Quarterly vol 23 No. 4* (1999), 497–530.
[5] H. Hastie, F. J. Chiyah Garcia, P. Patron, A. Laskov, and D. A. Robb. 2017. MIRIAM: A Multimodal Chat-Based Interface for Autonomous Systems. In *Proc. of ICMI'17.*
[6] H. Hastie, X. Liu, and P. Patron. 2017. Trust Triggers for Multimodal Command and Control Interfaces. In *Proc. ICMI'17.*
[7] H. Hastie, K. Lohan, M.J. Chantler, D.A. Robb, S. Ramamoorthy, R. Petrick, S. Vijayakumar, and D. Lane. 2018. The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. In *Proc. of Explainable Robotic Systems Workshop, HRI'18.*
[8] H. Kootval. 2008. Guidelines on Communicating Forecast Uncertainty. (2008).
[9] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symp. on Visual Languages and Human Centric Computing.* 3–10.
[10] B. Y. Lim, A. K. Dey, and D. Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proc. of CHI'09.* 2119–2129. https://doi.org/10.1145/1518701.1519023
[11] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (2016), 1135–1144.
[12] S. Rosenthal, S. P Selvaraj, and M. Veloso. 2016. Verbalization: Narration of Autonomous Robot Experience. In *IJCAI.* 862–868.
[13] N. Tintarev and R. Kutlak. 2014. SAsSy - Making Decisions Transparent with Argumentation and Natural Language Generation. (2014), 1–4.