Joel Greenyer, Malte Lochau, and Thomas Vogel (Hrsg.): Explainable Software for Cyber-Physical Systems (ES4CPS), GI-Dagstuhl Seminar 19023, January 2019,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2019 11

Explainable Autonomy through Natural Language

Francisco Javier Chiyah Garcia[1] David A. Robb[1] Helen Hastie[1]

Robots and autonomous systems are increasingly being deployed remotely in hazardous environments such as in the nuclear or energy sector domains [Ha18, Li17, KY12, Na13, SK16, Wo17]. Typically, these remote robots instil less trust than those co-located [Ba08, HLP17, Li15] and thus it is important to maintain a high level of transparency regarding their behaviour. This is particularly important for robots in remote locations where they cannot be directly observed.

The obscure nature of autonomous systems makes understanding them a difficult task for non-experts. The lack of knowledge and transparency about how the systems operate is reflected in decreased trust and understanding, which ultimately has negative effects on the human-machine cooperation [Dz03]. Therefore, the interface between the user and the system is key to maintaining situation awareness and understanding between the system and the human operator [Ro18].

The interfaces used to monitor autonomous systems often combine graphical elements such as the systems' updates and location so operators can understand what is happening. The complexity of these interfaces usually requires previous training and an operator overseeing the system's actions. As we move towards unmanned systems without continuous supervision, it is vital that these systems are able to communicate their current status and explain the reasoning behind their actions in a clear manner.

Explainability is an important facet of a transparent system [WTB17] as it can provide the user with a high fidelity mental model, along with increased confidence and performance [Le18, LDA09]. 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) [JL80]. Mental models are important as they strongly impact how and whether robots and autonomous systems are used.

A better mental model of a system, in terms of *what* the system can do and *why* it is doing certain actions, can provide the user with the knowledge to understand and predict the system's behaviours. A system that is easier to predict for users is more likely to be adopted over time and a clearer mental model can also help to avoid wrong assumptions and misuse of the system.

[1] Heriot-Watt University, School of Mathematics and Computer Sciences, Edinburgh, United Kingdom
{fjc3,d.a.robb,h.hastie}@hw.ac.uk

Explanations can contribute to building accurate mental models of a system. For example, [LDA09] showed that explaining "*why*" a system behaved in a certain way increased both understanding and trust, whilst "*why not*" explanations showed only an increase in understanding, and thus both are important with regards the user's mental model.

However, the way these explanations are given to the user can affect their benefits. As [GB99] shows, users will only take the time to process explanations if the benefits are perceived to be worth it, and do not adversely add to cognitive load [Me16]. [Ku13] explored how the amount of information that the explanations give, in terms of *soundness* (the level of detail) and *completeness* (the amount of reasons to give), can affect the user's mental model. Thus it is important to control both what to explain and how to present it to the user.

Interaction through natural language provides an intuitive means of keeping the user informed and requires little training. Conversational agents are becoming widely used to perform tasks such as asking for the weather or restaurant recommendations. Their convenience and ability to give on-demand instant responses makes them appealing for a range of users and devices. Previous works have investigated natural language to provide explanations of a variety of systems including: deep learning models [RSG16]; planning systems [TK14, Ch17]; and verbalising robot [RSV16] or agent [HER17] rationalisation. However, autonomous systems work in a continually changing environment with dynamic goals, which cannot always be known ahead of time and preprogrammed.

The data that autonomous systems gather from the environment to decide their next action offers an insight into how they behave. Our most recent works studied how this data can be used to increase trust and understanding of an autonomous system, as well as derive explanations about its behaviour [Ro18, Ch18a, Ch18b, Ha17]. We showed that a conversational agent running alongside the system's interface can increase the user's situational awareness and that it requires little training to understand and use. We also found that explanations with a high completeness (multiple explanations about the autonomy model) performed the best in increasing the user's understanding, whereas explanations with high soundness (high level of detail) were not as necessary.

Future work includes how to expand the information that the Conversational Agent can process and understand, as well as deeper integration with the autonomous system. Uncertainty is a challenging issue when working with autonomous systems and therefore, understanding the best way to handle it and present it to the user is essential to increase the system's transparency. Finally, another important goal that we are working towards is the generalisation of the agent, so it can be applied to other systems regardless of the domain.

References

[Ba08] Bainbridge, Wilma A.; Hart, Justin; Kim, Elizabeth S.; Scassellati, Brian: The Effect of Presence on Human-Robot Interaction. In: 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Munich, Germany, pp. 701–706, 2008.

[Ch17] Chakraborti, Tathagata; Sreedharan, Sarath; Zhang, Yu; Kambhampati, Subbarao: Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. IJCAI'17, Melbourne, Australia, pp. 156–163, 2017.

[Ch18a] Chiyah Garcia, Francisco J.; Robb, Dave A.; Liu, X.; Laskov, Atanas; Patron, Patron; Hastie, Helen: Explain Yourself: A Natural Language Interface for Scrutable Autonomous Robots. In: Proceedings of Explainable Robotic Systems Workshop. HRI'18, Chicago, IL, USA, 2018.

[Ch18b] Chiyah Garcia, Francisco J.; Robb, David A.; Laskov, Atanas; Liu, Xingkun; Patron, Pedro; Hastie, Helen: Explainable Autonomy: A Study of Explanation Styles for Building Clear Mental Models through a Multimodal Interface. In: Proceedings of The 11th International Natural Language Generation Conference. INLG'18, Tilburg, The Netherlands, p. 99–108, 2018.

[Dz03] Dzindolet, Mary T.; Peterson, Scott A.; Pomranky, Regina A.; Pierce, Linda G.; Beck, Hall P.: The role of trust in automation reliance. International Journal of Human - Computer Studies, 58(6):697–718, 2003.

[GB99] Gregor, Shirley; Benbasat, Izak: Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice. MIS Quarterly, 23(4):497–530, December 1999.

[Ha17] Hastie, Helen; Chiyah Garcia, Francisco J.; Robb, David A.; Patron, Pedro; Laskov, Atanas: MIRIAM: A Multimodal Chat-Based Interface for Autonomous Systems. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17. ACM, Glasgow, UK, pp. 495–496, 2017.

[Ha18] Hastie, Helen; Lohan, Katrin Solveig; Chantler, Mike J.; Robb, David A.; Ramamoorthy, Subramanian; Petrick, Ron; Vijayakumar, Sethu; Lane, David: The ORCA Hub: Explainable Offshore Robotics through Intelligent Interfaces. In: Proceedings of Explainable Robotic Systems Workshop, HRI'18. Chicago, IL, USA, 2018.

[HER17] Harrison, Brent; Ehsan, Upol; Riedl, Mark O.: Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. 2017.

[HLP17] Hastie, Helen; Liu, Xingkun; Patron, Pedro: Trust Triggers for Multimodal Command and Control Interfaces. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction, ICMI'17. ACM, Glasgow, UK, pp. 261–268, 2017.

[JL80] Johnson-Laird, Philip Nicholas: Mental models in cognitive science. Cognitive science, 4(1):71–115, 1980.

[Ku13] Kulesza, Todd; Stumpf, Simone; Burnett, Margaret; Yang, Sherry; Kwan, Irwin; Wong, Weng-Keen: Too much, too little, or just right? Ways explanations impact end users' mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing. San Jose, CA, USA, pp. 3–10, Sept 2013.

[KY12] Kwon, Young-Sik; Yi, Byung-Ju: Design and motion planning of a two-module collaborative indoor pipeline inspection robot. IEEE Transactions on Robotics, 28(3):681–696, June 2012.

[LDA09] Lim, Brian Y.; Dey, Anind K.; Avrahami, Daniel: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '09, pp. 2119–2129, 2009.

[Le18] Le Bras, Pierre; Robb, David A.; Methven, Thomas S.; Padilla, Stefano; Chantler, Mike J.: Improving User Confidence in Concept Maps: Exploring Data Driven Explanations. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, pp. 1–13, 2018.

[Li15] Li, Jamy: The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. International Journal of Human-Computer Studies, 77:23–37, 2015.

[Li17] Li, Jinke; Wu, Xinyu; Xu, Tiantian; Guo, Huiwen; Sun, Jianquan; Gao, Qingshi: A novel inspection robot for nuclear station steam generator secondary side with self-localization. Robotics and Biomimetics, 4(1):26, 2017.

[Me16] Mercado, Joseph E.; Rupp, Michael A.; Chen, Jessie YC.; Barnes, Michael J.; Barber, Daniel; Procci, Katelyn: Intelligent agent transparency in human–agent teaming for Multi-UxV management. Human Factors: The Journal of Human Factors and Ergonomics Society, 58(3):401–415, 2016.

[Na13] Nagatani, Keiji; Kiribayashi, Seiga; Okada, Yoshito; Otake, Kazuki; Yoshida, Kazuya; Tadokoro, Satoshi; Nishimura, Takeshi; Yoshida, Tomoaki; Koyanagi, Eiji; Fukushima, Mineo: Emergency response to the nuclear accident at the Fukushima Daiichi Nuclear Power Plants using mobile rescue robots. Journal of Field Robotics, 30(1):44–63, 2013.

[Ro18] Robb, David A.; Chiyah Garcia, Francisco J.; Laskov, Atanas; Liu, Xingkun; Patron, Pedro; Hastie, Helen: Keep Me in the Loop: Increasing Operator Situation Awareness through a Conversational Multimodal Interface. In: Proceedings of the 20th ACM International Conference on Multimodal Interaction. ICMI'18, ACM, Boulder, Colorado, USA, 2018.

[RSG16] Ribeiro, Marco Tulio; Singh, Sameer; Guestrin, Carlos: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'16, ACM, New York, NY, USA, pp. 1135–1144, 2016.

[RSV16] Rosenthal, Stephanie; Selvaraj, Sai P.; Veloso, Manuela: Verbalization: Narration of Autonomous Robot Experience. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence. IJCAI'16, AAAI Press, New York, NY, USA, pp. 862–868, 2016.

[SK16] Shukla, Amit; Karki, Hamad: Application of robotics in onshore oil and gas industry—-A review part I. Robotics and Autonomous Systems, 75:490–507, 2016.

[TK14] Tintarev, Nava; Kutlak, Roman: SAsSy– Making Decisions Transparent with Argumentation and Natural Language Generation. Proceedings of IUI 2014 Workshop on Interacting with Smart Objects, pp. 1–4, 2014.

[Wo17] Wong, Cuebong; Yang, Erfu; Yan, Xiu-Tian T.; Gu, Dongbing: An overview of robotics and autonomous systems for harsh environments. In: 2017 23rd International Conference on Automation and Computing (ICAC). IEEE, Huddersfield, UK, pp. 1–6, 2017.

[WTB17] Wortham, Robert H.; Theodorou, Andreas; Bryson, Joanna J.: Robot transparency: Improving understanding of intelligent behaviour for designers and users. In (Gao, Yang; Fallah, Saber; Jin, Yaochu; Lakakou, Constantina, eds): Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017. Lecture Notes in Artificial Intelligence, Springer, Guildford, UK, pp. 274–289, 7 2017.