Enhancing Transparency in Human-autonomy Teaming via the Option-centric Rationale Display

Ruikun Luo[1,2], Na Du[2], Kevin Y. Huang[2], and X. Jessie Yang[1,2]

[1]Robotics Institute, University of Michigan, Ann Arbor, MI
[2]Department of Industrial and Operations Engineering, University of Michigan, Ann Arbor, MI

Human-autonomy teaming is a major emphasis in the ongoing transformation of future work space wherein human agents and autonomous agents are expected to work as a team. While the increasing complexity in algorithms empowers autonomous systems, one major concern arises from the human factors perspective: Human agents have difficulty deciphering autonomy-generated solutions and increasingly perceive autonomy as a mysterious black box. The lack of transparency could lead to the lack of trust in autonomy and sub-optimal team performance (Chen and Barnes, 2014; Endsley, 2017; Lyons and Havig, 2014; de Visser et al., 2018; Yang et al., 2017).

In response to this concern, researchers have investigated ways to enhance autonomy transparency. Existing human factors research on autonomy transparency has largely concentrated on conveying automation reliability or likelihood/(un)certainty information (Beller et al., 2013; McGuirl and Sarter, 2006; Wang et al., 2009; Neyedli et al., 2011). Providing explanations of automation's behaviors is another way to increase transparency, which leads to higher performance and trust (Dzindolet et al., 2003; Mercado et al., 2016). Specifically, in the context of automated vehicles, studies have showed that informing the drivers of the reasons for the action of automated vehicles decreased drivers' anxiety, increased their sense of control, preference and acceptance (Koo et al., 2014, 2016; Forster et al., 2017).

However, the studies mentioned above largely focused on conveying simple likelihood information or used hand-drafted explanations, with only few exceptions (e.g.(Mercado et al., 2016)). Further research is needed to examine potential design structures of transparency autonomy.

In the present study, we wish to propose an option-centric explanation approach, inspired by the research on design rationale. Design rationale is an area of design science focusing on the "representation for explicitly documenting the reasoning and argumentation that make sense of a specific artifact (MacLean et al., 1991)". The theoretical underpinning for design rationale is that for designers what is important is not just the specific artifact itself but its other possibilities – why an artifact is designed in a particular way compared to how it might otherwise be. We aim to evaluate the effectiveness of the option-centric explanation approach on trust, dependence and team performance.

We conducted a human-in-the-loop experiment with 34 participants (Age: Mean = 23.7 years, *SD* = 2.88 years). We developed a simulated game *Treasure Hunter*, where participants and an intelligent assistant worked together to uncover a map for treasures. The intelligent assistant's ability, intent and decision-making rationale was conveyed in the *option-centric rationale* display. The experiment used a between-subject design with an independent variable – whether the option-centric rationale explanation was provided. The participants were randomly assigned to either of the two explanation conditions. Participants' trust to the intelligent assistant, confidence of accomplishing the experiment without the intelligent assistant, and workload for the whole session were collected, as well as their scores for each map.

The results showed that by conveying the intelligent assistant's ability, intent and decision-making rationale in the option-centric rationale display, participants had higher task performance. With the display of all the options, participants had a better understanding and overview of the system. Therefore, they could utilize the intelligent assistant more appropriately and earned a higher score. It is notable that every participant only played 10 maps during the whole session. The advantages of option-centric rationale display might be more apparent if more rounds are played in the experiment session. Although not significant at the .05 level, there seems to be a trend suggesting lower levels of workload when the rationale explanation displayed.

Our study contributes to the study of human-autonomy teaming by considering the important role of explanation display. It can help human operators build appropriate trust and improve the human-autonomy team performance.

REFERENCE

Beller, J., Heesen, M., and Vollrath, M. (2013). Improving the Driver–Automation Interaction: An Approach Using Automation Uncertainty. *Human Factors*.

Chen, J. Y. C. and Barnes, M. J. (2014). Human - Agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*.

de Visser, E. J., Pak, R., and Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned

From Human–Automation Research. *Human Factors*.

Forster, Y., Naujoks, F., Neukum, A., and Huestegge, L. (2017). Driver compliance to take-over requests with different auditory outputs in conditional automation. *Accident Analysis and Prevention*.

Koo, J., Kwac, J., Ju, W., Steinert, M., Leifer, L., and Nass, C. (2014). Why did my car just do that? Explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing*.

Koo, J., Shin, D., Steinert, M., and Leifer, L. (2016). Understanding driver responses to voice alerts of autonomous car operations. *International Journal of Vehicle Design*.

Lyons, J. B. and Havig, P. R. (2014). Transparency in a Human-Machine Context: Approaches for Fostering Shared Awareness/Intent. In Shumaker, R. and Lackey, S., editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*. Springer International Publishing.

MacLean, A., Young, R. M., Bellotti, V. M. E., and Moran, T. P. (1991). Questions, options, and criteria: Elements of design space analysis. *Humman-Computer Interaction*.

McGuirl, J. M. and Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors*.

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., and Procci, K. (2016). Intelligent agent transparency in human–agent teaming for multi-uxv management. *Human Factors*.

Neyedli, H. F., Hollands, J. G., and Jamieson, G. A. (2011). Beyond identity: Incorporating system reliability information into an automated combat identification system. *Human Factors*.

Wang, L., Jamieson, G. A., and Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors*.

Yang, X. J., Unhelkar, V. V., Li, K., and Shah, J. A. (2017). Evaluating effects of user experience and system transparency on trust in automation. In *HRI '17*.