Personalized Agent Explanations for Human-Agent Teamwork: Adapting Explanations to User Trust, Workload, and Performance

Extended Abstract

Ruben S. Verhagen  
Delft University of Technology  
Delft, the Netherlands  
R.S.Verhagen@tudelft.nl

Mark A. Neerincx  
Delft University of Technology  
Delft, the Netherlands  
M.A.Neerincx@tudelft.nl

Can Parlar  
Delft University of Technology  
Delft, the Netherlands  
C.Parlar@student.tudelft.nl

Marin Vogel  
Delft University of Technology  
Delft, the Netherlands  
M.Vogel-3@student.tudelft.nl

Myrthe L. Tielman  
Delft University of Technology  
Delft, the Netherlands  
M.L.Tielman@tudelft.nl

ABSTRACT

For human-agent teams to be successful, agent explanations are crucial. These explanations should ideally be personalized by adapting them to intended human users. So far, little work has been conducted on personalized agent explanations during human-agent teamwork. Therefore, an online experiment (n = 60) was conducted to compare personalized agent explanations against a baseline of non-personalized explanations. We implemented four agents who adapted their explanations during a search and rescue task randomly, or based on human workload, performance, or trust. Results show that personalized explanations can increase explanation satisfaction and trust in the agent, but also decrease performance. Therefore, we conclude that personalized agent explanations can be beneficial to human-agent teamwork, but that user modelling and personalization techniques should be carefully considered.

KEYWORDS

Explainable AI; Human-Agent Teamwork; Personalization

ACM Reference Format:


1 INTRODUCTION & BACKGROUND

Humans and autonomous intelligent agents are increasingly working together in human-agent teams [11, 13, 32]. Mutual understanding is crucial within these teams, but the behavior of agents is often hard to understand [1, 10, 11, 13, 17, 27, 33, 34]. Fortunately, Explainable Artificial Intelligence (XAI) methods can make agents understandable to humans, for example by accompanying decisions with explanations [1, 5, 14, 15]. Three explanation phases can be distinguished during human-agent collaboration: explanation generation, communication, and reception [22]. Various explanation types can be generated, such as confidence explanations, feature attributions, and counterfactual explanations [17, 30, 31]. These can be communicated in different forms like textually, verbally, or combining both [17]. The reception of such explanations concerns how well humans understand them, which requires user studies on their effectiveness in realistic human-agent settings [18, 22].

One of the main goals within XAI community is the development of user-aware agents able to adapt their explanations according to intended user [1, 17]. This could be achieved by maintaining a user model and using that model to personalize agent explanations by adapting them to the specifics of the model [1, 17]. So far, little work has been conducted on such personalized agent explanations during human-agent teamwork [1]. However, several studies highlight the importance of these explanations [2–4, 7, 12, 23, 25, 26, 31]. In summary, these studies often include approaches for modelling a human user and/or generating personalized explanations. However, only few works include user studies validating such approaches, and none of these involve user studies during human-agent teamwork. Our study will fill this gap by implementing and comparing three types of personalized agent explanations against non-personalized explanations during human-agent teamwork.

2 METHOD

We conducted a one-way between subjects experiment (n = 60) to compare three user-aware agents providing personalized explanations against a baseline providing non-personalized explanations. Using the MATRX software (https://matrx-software.com/), we built a two-dimensional grid world consisting of 14 areas, 26 collectable objects, 12 obstacles, and one drop zone (Figure 1). Next, we created three victims (critically injured, mildly injured, and healthy) and added obstacles in front of areas (boulder, tree, or stone). Finally, we added a human and an artificial agent (called RescueBot) to our world, which had to collaborate during a search and rescue task. The objective of this task was to find the target victims and carry them to the drop zone. We implemented several soft and hard dependencies between human and agent, such as carrying critically injured victims jointly. Participants had eight minutes to complete the task, received six points for rescuing critical victims, and three points for rescuing mild victims. Finally, we objectively measured task completeness and score, while subjectively measuring trust in the agent [9], workload [8], and explanation satisfaction [9].
Whenever RescueBot found an obstacle or victim, it provided decision support using suggestions and explanations based on crowd sourced data (Figure 1). More specifically, nine people were shown our environment, confronted with task dilemmas, and asked to make decisions and which features contributed most to these decisions. We used this data to generate one suggestion and confidence explanations, feature attributions, and counterfactual explanations. Next, we manipulated communication of these explanations by implementing non-user-aware, trust-aware, performance-aware, and workload-aware agents. The non-user-aware agent did not model the human user and for each decision randomly adapted its provided reasoning information. The trust-aware agent modelled user trust in the agent based on the number of followed and rejected agent suggestions. This agent increased its provided reasoning information when predicted trust decreased (and vice versa) [4, 16, 29]. Next, the performance-aware agent modelled user performance based on the difference between the predicted and real-time score of the task. This agent increased its provided reasoning information when predicted performance decreased (and vice versa) [4, 16, 29]. Finally, the workload-aware agent modelled user workload during the task based on cognitive and affective load [6, 19–21]. This agent increased its provided reasoning information when predicted workload decreased (and vice versa) [4, 24, 28].

3 RESULTS

Compared to the baseline, we expect the personalized explanations to increase each of their respective user factors used for adapting the explanations. Therefore, we conducted either independent samples t-tests or Mann-Whitney U tests to compare personalized explanations against the baseline. Here, we only report significant results, everything not reported was not found statistically significant. Task score was statistically significantly higher for participants receiving non-user-aware explanations ($M = 25.00$, $SD = 6.58$) than participants receiving performance-aware explanations ($M = 19.20$, $SD = 6.88$), $t(28) = 2.36$, $p = 0.025$, $d = .86$ (Figure 2A). In addition, trust scores of participants receiving trust-aware explanations (mean rank = 36.47) were statistically significantly higher than trust scores of participants receiving non-user-aware explanations (mean rank = 21.93), $W = 63.00$, $p = 0.041$ (Figure 2B). Trust was also statistically significantly higher for participants receiving workload-aware explanations ($M = 3.77$, $SD = 3.38$) than participants receiving non-user-aware explanations ($M = 3.43$, $SD = 0.43$), $t(28) = -2.31$, $p = 0.028$, $d = 0.84$ (Figure 2B). Finally, explanation satisfaction was statistically significantly higher for performance-aware explanations ($M = 4.23$, $SD = 0.36$) than non-user-aware agent explanations ($M = 3.74$, $SD = 0.54$), $t(28) = -2.95$, $p = 0.0063$, $d = 1.08$ (Figure 2C).

4 DISCUSSION AND CONCLUSION

As expected, our results show that people receiving explanations adapted to their trust in the agent, have significantly higher trust in the agent than people receiving non-personalized explanations. The results demonstrate how people receiving explanations adapted to their workload, also have significantly higher trust in the agent than people receiving non-personalized explanations. Combining these results, it seems that providing personalized agent explanations is particularly beneficial to trust in the agent, irrespective of the user factor used for adapting. Our results further show that people receiving personalized explanations based on their performance, perform worse as a team than people receiving non-personalized agent explanations. On the other hand, people receiving these personalized explanations are still more satisfied with them than people receiving non-personalized explanations. The worse performance is actually the opposite of the goal of the performance-aware agent explanations. However, since the worst performing participants received the explanations with most reasoning information, reading these took more time and likely resulted in the worse performance.

All in all, our study shows that personalized agent explanations can result in a higher explanation satisfaction and trust in the agent than non-personalized explanations. This highlights the benefits of personalized agent explanations for human-agent teamwork. However, our findings also show that personalized agent explanations using a sub-optimal adaptation strategy can result in a worse team performance than non-personalized explanations. This demonstrates the importance of carefully considering and comparing different user modelling and explanation adaptation strategies.
Poster Session I

AAMAS 2023, May 29–June 2, 2023, London, United Kingdom

ACKNOWLEDGMENTS

This work is part of the research lab AI’MAN of Delft University of Technology. The authors want to thank Zhiqiang Lei for his contribution during research design and execution.

REFERENCES


