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:
Ruben S. Verhagen, Mark A. Neerincx, Can Parlar, Marin Vogel, and Myrthe L. Tielman. 2023. Personalized Agent Explanations for Human-Agent Teamwork: Adapting Explanations to User Trust, Workload, and Performance: 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 & 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 interdependencies 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].

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.

Figure 1: Cropped image of the world used during our study.

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 = 0.38) 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).

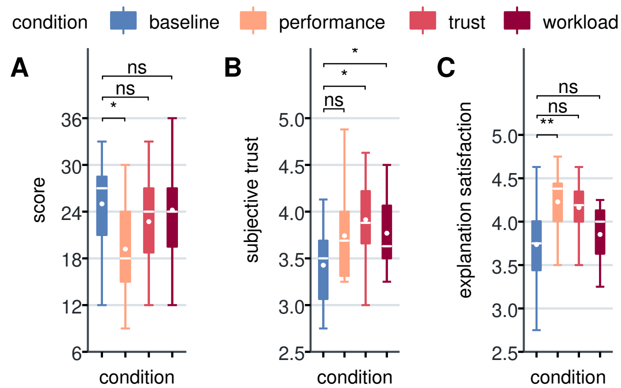

Figure 2: Boxplots of score (A), trust (B), and explanation satisfaction (C), *p<0.05. **p<0.01.

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.

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

[1] Sule Anjomshoae, Amro Najjar, Davide Calvaresi, and Kary Främling. 2019. Explainable agents and robots: Results from a systematic literature review. In *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019*. International Foundation for Autonomous Agents and Multiagent Systems, 1078–1088.

[2] Rita Borgo, Michael Cashmore, and Daniele Magazzeni. 2018. Towards providing explanations for AI planner decisions. *arXiv preprint arXiv:1810.06338* (2018).

[3] Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. 2017. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *arXiv preprint arXiv:1701.08317* (2017).

[4] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. 2018. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science* 19, 3 (2018), 259–282.

[5] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web* 2, 2 (2017), 1.

[6] Maaike Harbers, Reyhan Aydogan, Catholijn M Jonker, and Mark A Neerincx. 2014. Sharing information in teams: giving up privacy or compromising on team performance?. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 413–420.

[7] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and evaluation of explainable BDI agents. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. IEEE, 125–132.

[8] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[9] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[10] Matthew Johnson, Jeffrey M Bradshaw, Paul J Feltovich, Catholijn M Jonker, M Birna Van Riemsdijk, and Maarten Sierhuis. 2014. Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction* 3, 1 (2014), 43–69.

[11] Matthew Johnson and Alonso Vera. 2019. No AI is an island: the case for teaming intelligence. *AI magazine* 40, 1 (2019), 16–28.

[12] Frank Kaptein, Joost Broekens, Koen Hindriks, and Mark Neerincx. 2017. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 676–682.

[13] Glen Klein, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman, and Paul J Feltovich. 2004. Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems* 19, 6 (2004), 91–95.

[14] Bertram F Malle. 2004. How the mind explains behavior. *Folk Explanation, Meaning and Social Interaction. Massachusetts: MIT-Press* (2004).

[15] Bertram F Malle. 2011. Attribution theories: How people make sense of behavior. *Theories in social psychology* 23 (2011), 72–95.

[16] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[17] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[18] Yazan Mualla, Igor Tchappi, Timotheus Kampik, Amro Najjar, Davide Calvaresi, Abdeljalil Abbas-Turki, Stéphane Galland, and Christophe Nicolle. 2022. The quest of parsimonious XAI: A human-agent architecture for explanation formulation. *Artificial Intelligence* 302 (2022), 103573.

[19] Diane Nahl. 2005. Affective and cognitive information behavior: Interaction effects in Internet use. *Proceedings of the American Society for Information Science and Technology* 42, 1 (2005).

[20] Mark A Neerincx et al. 2003. Cognitive task load analysis: allocating tasks and designing support. *Handbook of cognitive task design* 2003 (2003), 283–305.

[21] Mark A Neerincx, Maaike Harbers, Dustin Lim, and Veerle van der Tas. 2014. Automatic feedback on cognitive load and emotional state of traffic controllers. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 42–49.

[22] Mark A Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. 2018. Using perceptual and cognitive explanations for enhanced human-agent team performance. In *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer, 204–214.

[23] Mayada Oudah, Talal Rahwan, Tawna Crandall, and Jacob Crandall. 2018. How AI wins friends and influences people in repeated games with cheap talk. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[24] Raja Parasuraman, Thomas B Sheridan, and Christopher D Wickens. 2008. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of cognitive engineering and decision making* 2, 2 (2008), 140–160.

[25] David V Pynadath, Ning Wang, Ericka Rovira, and Michael J Barnes. 2018. *Clustering behavior to recognize subjective beliefs in human-agent teams.* Technical Report. UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES.

[26] Lara Quijano-Sanchez, Christian Sauer, Juan A Recio-Garcia, and Belen Diaz-Agudo. 2017. Make it personal: a social explanation system applied to group recommendations. *Expert Systems with Applications* 76 (2017), 36–48.

[27] Eduardo Salas, Dana E Sims, and C Shawn Burke. 2005. Is there a "big five" in teamwork? *Small group research* 36, 5 (2005), 555–599.

[28] Lindsay Sanneman and Julie A Shah. 2022. The Situation Awareness Framework for Explainable AI (SAFE-AI) and Human Factors Considerations for XAI Systems. *International Journal of Human–Computer Interaction* (2022), 1–17.

[29] Anthony R Selkowitz, Shan G Lakhmani, and Jessie YC Chen. 2017. Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research* 46 (2017), 13–25.

[30] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[31] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. 2021. Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI* 8 (2021), 640647.

[32] Jurriaan van Diggelen, JS Barnhoorn, Marieke MM Peeters, Wessel van Staal, ML Stolk, Bob van der Vecht, Jasper van der Waa, and Jan Maarten Schraagen. 2019. Pluggable social artificial intelligence for enabling human-agent teaming. *arXiv preprint arXiv:1909.04492* (2019).

[33] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2021. A two-dimensional explanation framework to classify ai as incomprehensible, interpretable, or understandable. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*. Springer, 119–138.

[34] Ruben S Verhagen, Mark A Neerincx, and Myrthe L Tielman. 2022. The influence of interdependence and a transparent or explainable communication style on human-robot teamwork. *Frontiers in Robotics and AI* 9 (2022), 243.