A Human-Agent Architecture for Explanation Formulation (An extended abstract)*

Yazan Mualla[1], Igor Tchappi[1], Timotheus Kampik[2], Amro Najjar[3], Davide Calvaresi[4], Abdeljalil Abbas-Turki[1], Stéphane Galland[1], and Christophe Nicolle[5]

[1] CIAD, Univ. Bourgogne Franche-Comté, UTBM, F-90010 Belfort, France
[2] Department of Computing Science, Umeå University, 90187 Umeå, Sweden
[3] AI-Robolab/ICR, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg
[4] University of Applied Sciences and Arts of Western Switzerland, Sierre, Switzerland
[5] CIAD UMR 7533, Univ. Bourgogne Franche-Comté, UB, F-21000 Dijon, France

1 Introduction

With the widespread use of AI systems, understanding the behavior of intelligent agents and robots is crucial to facilitate successful human-computer interaction (HCI) [3]. Recent studies have confirmed that explaining an agent's behavior to humans fosters the latter's acceptance of the agent [2, 4]. However, providing overwhelming or unnecessary information may also confuse humans and cause failure [15]. For these reasons, *parsimony* has been outlined as one of the key features of successful explanations in HCI [10, 9]; in this context, a parsimonious explanation is defined as the simplest explanation (*i.e.,* least complex) that describes the situation adequately (*i.e.,* descriptive adequacy) [9, 5]. While parsimony is receiving growing attention in the literature, most works are carried out on the conceptual front, and little research has been done from engineering and empirical HCI perspectives.

2 Contribution

This work proposes a mechanism for parsimonious eXplainable AI (XAI) [6, 7, 16]. In particular, it introduces the process of *explanation formulation* and proposes HAExA, a human-agent explainability architecture (Figure 1) allowing to make this formulation operational for remote robots. In HAExA, **remote robots** (right) are represented as agents that generate contrastive explanations[6] [12] to explain their behaviors based on the changes in the environment and their goals. **Assistant agents** (center) collect the remote agents' raw explanations to communicate filtered explanations to the **human** (left); the filtering helps prevent that humans get overwhelmed by the information the remote agents provide. Considering that the assistant agents have a global overview of the environment, they may post-process the raw explanations received from the remote agents to aggregate, update, and filter them; subsequently, they communicate the updated and filtered explanations to the human.

* This work has been accepted in the Journal of Artificial Intelligence on the 2nd of August 2021 [14]. DOI: https://doi.org/10.1016/j.artint.2021.103573

[6] Broadly speaking, contrastive explanations answer *why A and not B?* questions.

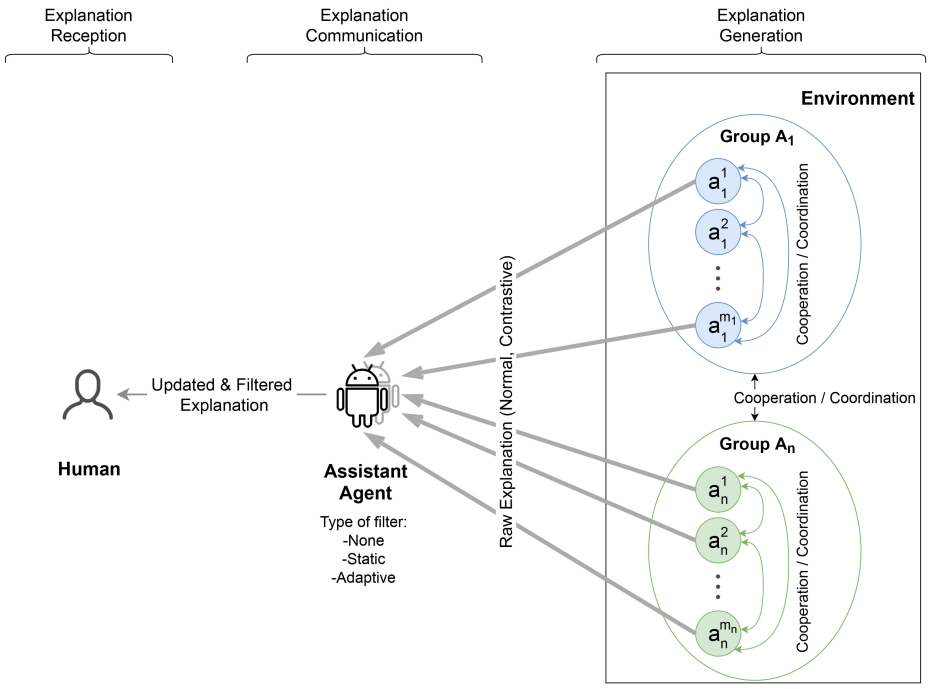

Fig. 1. Human-Agent Explainability Architecture (HAExA).

3 Evaluation and Results

To evaluate HAExA, several research hypotheses are investigated in an HCI study using an agent-based simulation based on a scenario of package delivery in smart cities (*see* our demo paper [13]). The study relies on well-established XAI metrics [8] to estimate how understandable the explanations are to the human participants. The study investigates the impact of the different techniques of explanation formulation (static filter, adaptive filter, and adaptive filter with contrastive explanations) on humans. The participants' responses are collected using a 5-Likert scale [1]. The significance of these responses is statistically analyzed and presented using statistical testing: Non-parametric (Kruskal-Wallis), Parametric (ANOVA), and Cronbach's alpha.

Based on the analysis of *subjective* and *objective understandability*, we gathered evidence that adaptively filtered and contrastive explanations improve human understandability compared to statically filtered explanations (*i.e.,* non-adaptive to the environment). Our insights indicate that contrastive explanations can be used without risking a detrimental effect on understandability. Our study could not confirm the same effect on *trust* (which remains a challenge identified in many other works in the literature [11, 8]). Nevertheless, the results provide empirical insights on human-multiagent system explainability as a starting point that future research on XAI could expand.

References

1. Albaum, G.: The likert scale revisited. Market Research Society. Journal. **39**(2), 1–21 (1997)
2. Anjomshoae, S., Najjar, A., Calvaresi, D., Främling, K.: Explainable agents and robots: Results from a systematic literature review. In: Proc. of 18th Int. Conf. on Autonomous Agents and MultiAgent Systems. pp. 1078–1088. Int. Foundation for Autonomous Agents and Multiagent Systems (2019)
3. Bainbridge, W.A., Hart, J., Kim, E.S., Scassellati, B.: The effect of presence on human-robot interaction. In: RO-MAN 17th IEEE Int. Symposium on Robot and Human Interactive Communication. pp. 701–706 (2008)
4. Calvaresi, D., Mualla, Y., Najjar, A., Galland, S., Schumacher, M.: Explainable multi-agent systems through blockchain technology. In: Calvaresi, D., Najjar, A., Schumacher, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 41–58. Springer International Publishing, Cham (2019)
5. Contreras, H.: Simplicity, descriptive adequacy, and binary features. Language pp. 1–8 (1969)
6. Gunning, D.: Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), nd Web (2017)
7. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI—Explainable Artificial Intelligence. Science Robotics (2019)
8. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608 (2018)
9. Krizek, G.C.: Ockham's razor and the interpretations of quantum mechanics (2017)
10. Laird, J.: The law of parsimony. The Monist **29**(3), 321–344 (1919), http://www.jstor.org/stable/27900747
11. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: Explainable reinforcement learning through a causal lens. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 2493–2500 (2020)
12. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
13. Mualla, Y., Kampik, T., Tchappi, I.H., Najjar, A., Galland, S., Nicolle, C.: Explainable agents as static web pages: Uav simulation example. In: Calvaresi, D., Najjar, A., Winikoff, M., Främling, K. (eds.) Explainable, Transparent Autonomous Agents and Multi-Agent Systems. pp. 149–154. Springer International Publishing, Cham (2020)
14. Mualla, Y., Tchappi, I., Kampik, T., Najjar, A., Calvaresi, D., Abbas-Turki, A., Galland, S., Nicolle, C.: The quest of parsimonious xai: a human-agent architecture for explanation formulation. Artificial Intelligence p. 103573 (2021)
15. Mualla., Y., Tchappi., I., Najjar., A., Kampik., T., Galland., S., Nicolle., C.: Human-agent explainability: An experimental case study on the filtering of explanations. In: Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: HAMT,. pp. 378–385. INSTICC, SciTePress (2020). https://doi.org/10.5220/0009382903780385
16. Ras, G., van Gerven, M., Haselager, P.: Explanation methods in deep learning: Users, values, concerns and challenges. In: Explainable and Interpretable Models in Computer Vision and Machine Learning, pp. 19–36. Springer (2018)