Towards Autonomous Collaborative Robots that Adapt and Explain

Peter Tisnikar*, Lennart Wachowiak*, Gerard Canal, Andrew Coles, Matteo Leonetti, and Oya Celiktutan

***Abstract*— As we will increasingly encounter robots in our everyday lives, engage with them in social interactions, and collaborate on common tasks, it is important we endow them with the capability of adapting to our abilities and preferences. Moreover, we want them to be able to explain the decisions they make in collaborative tasks to maximise rapport and build trust and acceptance. Current explanations are missing a focus on the individual user's needs, which is why we want to learn when and what to explain in a collaboration. This paper proposes the roadmap that we plan to implement with the goal of inferring the user's mental state from physiological and social signals in order to inform explanation generation and robot adaptation. We present preliminary results utilizing eye gaze and a planned framework that allows a collaborative robot to adapt to its human collaborator and tailor its explanations to them in order to minimise the confusion in an interaction.**

I. Introduction

Social robots can help us tackle societal challenges in a variety of areas, especially, in education or healthcare [1]. One of their applications, for instance, is in home assistance, where robots will extend the autonomy of the elderly by providing support with tasks that become too physically taxing for the person. However, the environments and interactions in which these robots will participate bring a host of new challenges. The humans a robot interacts with will differ in many aspects such as age, cultural background, physical ability, or personal preferences.

We are interested in giving robots the ability to understand their human collaborators and interaction partners in order to tailor their behavior to a specific person. Every human will have their own approach to solving a certain task, and the robot should recognize these approaches as valid plans instead of errors in the collaborator's policy. It should also be able to justify its choices and help the user when they are unsure of the robot's actions or confused by the task itself. In the example of caring for an elderly person, the robot should be able to recognize that different people have different preferences, for example, one person might be fine with a robot coming in contact with them during assisted dressing while another might not. Furthermore, the robot should be able to explain its actions, for instance, why it tries to give the care recipient a certain type of medication. By doing that, the robots in our homes will make a step forward from being machines that need constant supervision to partners capable of doing things the way we want them to do them, without us explicitly guiding them every step of the way.

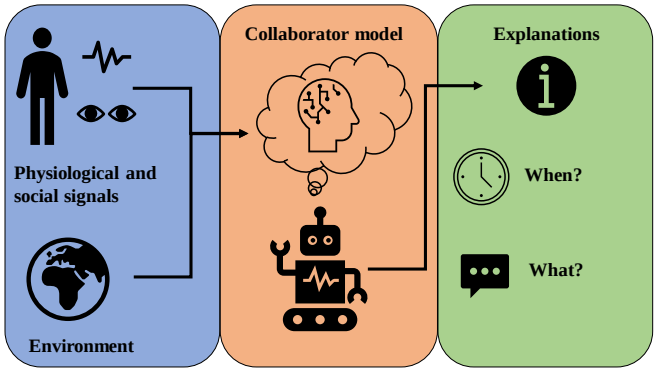

Fig. 1. The proposed explanation framework takes into account the physiological and social signals that it observes from the human collaborator, the task-specific information, and creates a mental model of the collaborator for providing explanations.

In this work, we outline the approach we will be taking to address some of these questions, namely the adaptation to a user's plan and explanations given to the user when confusion arises. We base our approach on the preliminary results of a study of eye gaze in a collaborative task, discussed in Section III. We aim to contribute a novel framework, which incorporates both of these properties and we plan to implement this on a physical robot, testing it on an interactive collaborative task. We will base our framework on the use of physiological and social cues that a person expresses during their interaction with the robot: eye gaze, heart rate, electrodermal activity, posture, and facial expressions (as seen in Figure 1). We argue that these provide a rich source of information, which can be used to generate explanations with appropriate detail at the right time and shape the robot's behaviour in a way that maximises the quality of the interaction by minimising confusion of the user.

II. Related Work

The need to design a framework which allows robots to become user-aware collaborators is well known. Lemaignan et al. [2] propose a cognitive deliberative architecture for a robot that is capable of human–robot interaction (HRI) and outline the necessary capabilities of such robot: a mental model of the collaborator, which is used when the robot plans its actions. A survey by Tabrez et al. [3] shows that most

*: Authors have contributed equally to this work.

This work was supported by UK Research and Innovation (EP/S023356/1), in the UKRI CDT in Safe and Trusted AI. This work was also supported by the CHIST-ERA project COHERENT (EP/V062506/1) and the EPSRC project LISI (EP/V010875/1). Gerard Canal was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK IC Postdoctoral Research Fellowship programme.

P. Tisnikar, L. Wachowiak, G. Canal, A. Coles, and M. Leonetti are with the Department of Informatics, King's College London, WC2R 2LS, United Kingdom. O. Celiktutan is with the Department of Engineering, King's College London, WC2R 2LS, United Kingdom. {name.surname}@kcl.ac.uk

collaborative or interactive robots use some form of mental modelling of the collaborator by use of inverse reinforcement or inverse planning algorithms.

Others have focused on the inference of mental states using physiological [4] and social cues [5] expressed by the human during the interaction. Eye gaze, for instance, is well known to reflect cognitive processes [6] and can, therefore, also provide information about the mental state of the human during interaction with a robot. Due to this property, eye gaze can be used as a reward which is used to shape a robot's behaviour, as seen in [7], where a neural network is trained to map human facial expressions to performance and use them to improve an agent's policy. Furthermore, eye gaze signals can be used as an enhancement to other types of teaching. Saran et al. [8] enhance kinesthetic teaching with eye gaze fixations, as they show that eye gaze reveals information that might not always be directly observable from the teacher's actions alone. Lastly, eye gaze is known to indicate task intent in collaborations, which showcases the usefulness of being aware of your collaborators gaze when wanting to adapt to them [9].

Beyond its use as learning signal, eye gaze has the potential to be used in explainability research. Use of eye gaze and physiological signals in general has, to the best of our knowledge, barely been explored in the explainable agents literature. However, such cues could play a role in overcoming one of the major shortcomings of current explanations — not being user-centered [10], [11]. Some areas of research focus on automatically detecting critical moments of an interaction, e.g., a robot making errors or a human not knowing how to proceed in a task. Such moments can be interpreted as warranting an explanation given by the agent: the agent explaining why they engaged in the erroneous behavior and the agent explaining the task to the user. Kurylo and Wilson [12], for instance, try to identify when a user requires assistance based on gaze patterns. Kontogiorgos et al. [13], on the other hand, investigate users' reactions to agent errors, which in the future could also be used as an indicator for when to provide an explanation. In addition, Trung et al. attempt to automatically detect errors made by a robot only using a persons head and shoulder movements [14]. Lastly, Das et al. present a method on how to generate explanations that help users understand why a robot failed [15].

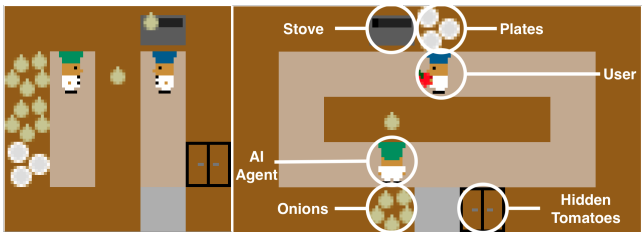

Fig. 2. The two video game levels used in the study.

III. Eye Gaze Patterns during Human–Agent Collaborations

To extend the body of work on using physiological and social signals to optimize HRI, we conducted a study investigating how eye gaze patterns differ at various phases of an interaction [16]. In the study, we asked participants to play a collaborative video game in which they had to coordinate with an AI agent to maximise their game score (Figure 2). This task was based on an implementation of the game Overcooked provided by Carroll et al. [17]. Overcooked requires the human–agent team to cook soups by accomplishing interdependent sub-goals and was previously used as testbed in explainability research [18].

From our preliminary analysis, we conclude that eye gaze patterns significantly differ between the investigated conditions: normal workflow, confusion by task, and agent error. In our future work, we want to exploit such patterns in order to adapt to the human collaborator and generate user-centered explanations. For example, if the agent could infer that it is making an error based on the user's gaze signal the agent would know that it has to stop this erroneous behavior or initiate a dialogue with the user in order to discuss whether to continue or change its behavior.

IV. Roadmap

In order to develop and implement the framework, we propose three distinct phases through which we aim to first extend the study of social signal into an interaction with a physical robot, then enable the robot to adapt to the human collaborator, and provide explanations based on the inferred mental model.

A. Data Collection with a Physical Robot

We wish to extend this study into a collaborative task with a physical robot. The test task will recreate one of the levels from the collaborative task of our initial study, as seen on the left-hand side of Figure 2. We will extend the study by adding other modalities such as heart rate, electrodermal activity, posture, and facial expressions. A collaborative task with forced collaboration gives us greater amount of control over the space of strategies and possible errors, whilst guaranteeing collaboration between partners. We hope to extend the findings of our initial study, which can then be used to provide policy shaping signals for adaptation, and underpin the timing and content of explanations.

B. Adaptation

As physiological and social signals indicate the user's mental state, we can optimise for favourable patterns by shaping the agent's behaviour. For example, a robot's policy can be tailored to a person who is left-handed and prefers all interaction with the robot to happen on their left-hand side. If the robot insists on interaction taking place on their right-hand side, the person might become confused and frustrated by the robot's behaviour, as observed in [19]. Knox and Stone [20] have shown that policy shaping can be done with explicit rewards from the human teacher. However, they

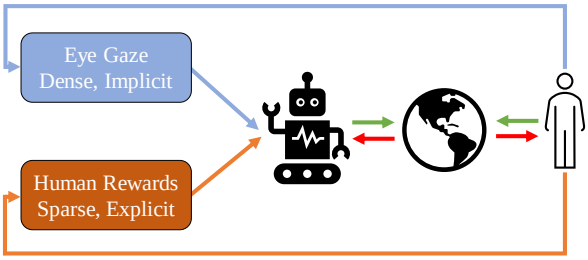

Fig. 3. The classic reinforcement learning paradigm is extended to include the human collaborator, who gives two types of rewards: Implicit reward based on physiological and social signals, and explicit reward based on their preference. The agent and the human act in the environment, and receive observations from it.

acknowledge that credit assignment difficulties, which are typical in problems with sparse rewards, remain in their approach. We argue that the continuous nature of implicit rewards alleviates some of these issues, as the reward function becomes denser, thus, improving the learning performance. However, we also anticipate that there will be conflicts between explicit and implicit signals, and we are interested in finding an arbitration scheme which resolves this conflict. We are therefore interested in comparing adaptation based on explicit rewards with adaptation based on implicit rewards alone, as well as a mix between explicit and implicit rewards, all in a collaborative scenario. The relations between those two type of rewards and the agent are depicted in Figure 3.

C. Explanations

Moreover, we want to use the physiological and social signals in order for the agent to learn *when* and *what* to explain during a collaboration, as shown in Figure 1. To achieve that, we will, firstly, train a classifier that predicts aspects of the user's mental state such as level of confusion and source of confusion. The classifier will be trained in a supervised manner using data labeled by multiple independent annotators. In order to decide whether a participant was confused for a given frame, the annotators will consider recordings of the interaction as well as (retrospective) think-aloud interviews [21] from the participants. For now, the only sources of confusion we consider are agent errors and difficulties in understanding the environment. In the future, a finer-grained taxonomy of such sources could be considered, for instance, including robot errors, social norm violations by the robot, e.g., [22], the robot's inability to execute a task, e.g., [23], difficult to grasp strategies, e.g., [24], and missing task or environment information. Secondly, we will integrate this information into a planner. This planner could model *giving an explanation* as action, while integrating the information regarding the user's confusion as preconditions that decide whether explanatory or productive actions are being executed, thus, deciding the timing of an explanation. Additionally, the explanation's content should target the source of confusion. For instance, when the user sees the agent making an error, the user might want to know why the agent did that in order to understand the faulty beliefs of the agent. If, on the other hand, the user is confused about what they have to do in a task, the agent might provide helpful advice regarding how to proceed. We expect some of the physiological and social signals to work better as strong signals of when and what to explain than others. Gaze targets might need to be redefined for different domains, whereas signals such as facial expressions or heart rate variability are expected to work across domains.

V. Conclusion

In this paper we presented a roadmap to a framework in which a robot is able to improve the interaction quality in a collaborative task with a human through proactive adaptation and user-centered explanations. We plan to achieve that by interpreting physiological signals such as heart rate and electrodermal activity as well as social signals such as eye gaze, facial expressions, and posture in order to inform the robot of the user's mental state. The robot will then be able to use this information to shape its policy in order to minimise the undesired mental states of the collaborator. We hope that by augmenting or completely removing explicit rewards, the collaboration will become more intuitive and personalised to each individual collaborator. Lastly, being able to detect when a user is confused as well as what they are confused about will allow the agent to know when and what to explain.

References

[1] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots: Concepts, design and applications," 2002.

[2] S. Lemaignan, M. Warnier, E. A. Sisbot, A. Clodic, and R. Alami, "Artificial cognition for social human–robot interaction: An implementation," *Artificial Intelligence*, vol. 247, pp. 45–69, Jun. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0004370216300790

[3] A. Tabrez, M. B. Luebbers, and B. Hayes, "A Survey of Mental Modeling Techniques in Human–Robot Teaming," *Current Robotics Reports*, vol. 1, no. 4, pp. 259–267, Dec. 2020. [Online]. Available: https://link.springer.com/10.1007/s43154-020-00019-0

[4] A. Kothig, J. Munoz, S. A. Akgun, A. M. Aroyo, and K. Dautenhahn, "Connecting humans and robots using physiological signals–closing-the-loop in hri," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 735–742.

[5] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[6] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends in cognitive sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[7] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox, "The EMPATHIC Framework for Task Learning from Implicit Human Feedback," *arXiv:2009.13649 [cs]*, Dec. 2020, arXiv: 2009.13649. [Online]. Available: http://arxiv.org/abs/2009.13649

[8] A. Saran and E. S. Short, "Understanding Teacher Gaze Patterns for Robot Learning," p. 12.

[9] C.-M. Huang, S. Andrist, A. Sauppé, and B. Mutlu, "Using gaze patterns to predict task intent in collaboration," *Frontiers in Psychology*, vol. 6, Jul. 2015. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fpsyg.2015.01049/abstract

[10] M. Ribera and A. Lapedriza, "Can we do better explanations? a proposal of user-centered explainable ai." in *IUI Workshops*, vol. 2327, 2019, p. 38.

[11] S. Kambhampati, "Challenges of human-aware ai systems: Aaai presidential address," *AI Magazine*, vol. 41, no. 3, pp. 3–17, Sep. 2020. [Online]. Available: https://ojs.aaai.org/index.php/aimagazine/article/view/5257

[12] U. Kurylo and J. R. Wilson, "Using Human Eye Gaze Patterns as Indicators of Need for Assistance from a Socially Assistive Robot," in *Social Robotics*, M. A. Salichs, S. S. Ge, E. I. Barakova, J.-J. Cabibihan, A. R. Wagner, Á. Castro-González, and H. He, Eds. Cham: Springer International Publishing, 2019, vol. 11876, pp. 200–210.

[13] D. Kontogiorgos, S. van Waveren, O. Wallberg, A. Pereira, I. Leite, and J. Gustafson, "Embodiment Effects in Interactions with Failing Robots," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, Apr. 2020, pp. 1–14.

[14] P. Trung, M. Giuliani, M. Miksch, G. Stollnberger, S. Stadler, N. Mirnig, and M. Tscheligi, "Head and shoulders: automatic error detection in human-robot interaction," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. Glasgow UK: ACM, Nov. 2017, pp. 181–188. [Online]. Available: https://dl.acm.org/doi/10.1145/3136755.3136785

[15] D. Das, S. Banerjee, and S. Chernova, "Explainable ai for robot failures: Generating explanations that improve user assistance in fault recovery," in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 2021, pp. 351–360.

[16] L. Wachowiak, P. Tisnikar, G. Canal, A. Coles, M. Leonetti, and O. Celiktutan, "Analysing eye gaze patterns during confusion and errors in human–agent collaborations," in *2022 31th IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 2022, forthcoming.

[17] M. Carroll, T. L. Griffiths, R. Shah, M. K. Ho, S. A. Seshia, P. Abbeel, and A. Dragan, "On the Utility of Learning about Humans for Human-AI Coordination," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[18] X. Gao, R. Gong, Y. Zhao, S. Wang, T. Shu, and S.-C. Zhu, "Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. Naples, Italy: IEEE, Aug. 2020, pp. 1119–1126. [Online]. Available: https://ieeexplore.ieee.org/document/9223595/

[19] G. Canal, C. Torras, and G. Alenyà, "Are Preferences Useful for Better Assistance?: A Physically Assistive Robotics User Study," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–19, 2021.

[20] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: the TAMER framework," in *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*. Redondo Beach, California, USA: ACM Press, 2009, p. 9. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1597735.1597738

[21] K. A. Ericsson and H. A. Simon, *Protocol Analysis: Verbal Reports as Data.* MIT Press, 1984.

[22] N. Mirnig, M. Giuliani, G. Stollnberger, S. Stadler, R. Buchner, and M. Tscheligi, "Impact of robot actions on social signals and reaction times in hri error situations," in *International Conference on Social Robotics*. Springer, 2015, pp. 461–471.

[23] Z. Han, E. Phillips, and H. A. Yanco, "The need for verbal robot explanations and how people would like a robot to explain itself," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 4, pp. 1–42, 2021.

[24] A. Raymond, H. Gunes, and A. Prorok, "Culture-based explainable human-agent deconfliction," *arXiv preprint arXiv:1911.10098*, 2019.