DOI:
10.1109/RO-MAN53752.2022.9900589

Document Version
Peer reviewed version

Analysing Eye Gaze Patterns during Confusion and Errors in Human–Agent Collaborations

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

***Abstract*— As human–agent collaborations become more prevalent, it is increasingly important for an agent to be able to adapt to their collaborator and explain their own behavior. In order to do so, they need to be able to identify critical states during the interaction that call for proactive clarifications or behavioral adaptations. In this paper, we explore whether the agent could infer such states from the human's eye gaze for which we compare gaze patterns across different situations in a collaborative task. Our findings show that the human's gaze patterns significantly differ between times at which the user is confused about the task, times at which the agent makes an error, and times of normal workflow. During errors the amount of gaze towards the agent increases, while during confusion the amount towards the environment increases. We conclude that these signals could tell the agent what and when to explain.**

I. Introduction

Advances in artificial intelligence and robotics are progressively enabling humans and robots to collaborate in shared environments [1]. One of the key challenges in such scenarios is that an autonomous robot needs to be able to model individual users and predict their mental state in order to adapt to the user's needs and expectations [2]. Such a model of the user can be inferred from various signals, e.g., the user's past actions [3] or physiological and social cues [4]. One of those signals is a person's eye gaze, which has been previously used to study their focus of attention, intentions, or emotional state [5], and has been accredited with a communicative function in human–robot interactions (HRI) [6]. Therefore, it is important for a socially capable and aware robot to be able to interpret it.

In this study, we assess people's reactions to critical situations during a collaboration with an agent by analysing their gaze patterns. Previous research has investigated how gaze patterns change either when a robot makes an error [7], [8], [9] or when the human is confused and requires assistance to solve a task [10], however, it does not distinguish between the two. Moreover, the collaborative scenarios used so far are too simple to be representative of real-world interactions and mostly restrict the role of one of the team members to merely guide their partner through the task. We address this gap in literature, by analysing the user's gaze patterns in a complex cooperative task where both, human and agent, contribute equally towards a common goal by carrying out various interdependent sub-tasks. This environment allows us to elicit confusion in the user and includes multiple pre-programmed agent-errors.

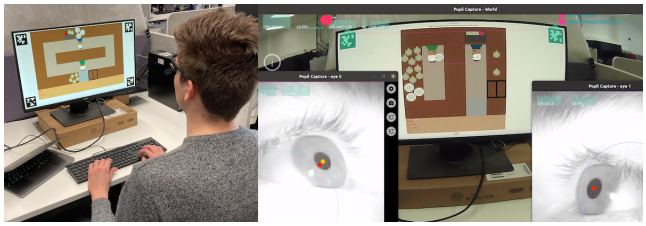

Fig. 1. The experiment setup with a participant wearing the eye tracker (left) and the recording setup (right).

Being able to infer human confusion and their own mistakes from gaze patterns will allow robots to: proactively adapt to their human partner; give correctly timed explanations to either help the user with their task or explain the robot's own motivations; correct their erroneous behavior; or ask the user for help and clarifications. Lastly, such gaze patterns might even help the robot during its training phase as — if proven to be indicative of critical situations — it could serve as an additional implicit reward signal.

In order to analyse gaze patterns during critical moments, we quantify how much people look at the agent, the task environment, and their own character playing the collaborative game Overcooked, as pictured in Figure 1. We compare the gaze patterns that were recorded during situations in which the agent makes errors with those recorded during moments of user-confusion and those recorded when a productive workflow is achieved. To better understand the potential of gaze as reward signal during training, we, furthermore, investigate the correlation between gaze patterns and the overall team performance as indicated by the scored points.

II. Related Work

Eye gaze plays an important role in information sharing and social interactions of humans and animals [11]. Moreover, sensitivity to gaze promotes cognitive development; e.g., following someone's gaze helps infants to learn the mappings between words and objects [12]. In addition, research shows how having direct access to another person's gaze targets during collaborations can improve human–human team performance [13]. For these reasons, we think it is

*: 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.

L. Wachowiak, P. Tisnikar, 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

worth to investigate the role of eye gaze in identifying critical states in HRI and eye gaze's potential as reward signal.

Eye Gaze in Error and Confusion Detection. Various studies have investigated how humans react to robots making errors [14]. Researchers have analyzed social and physiological reactions of different modalities, e.g., facial expressions [15], eye gaze [8], [7], [9], or posture [16], as well as the changes in the user's subjective perception of the agent [17], [18], [19]. Among other things, research shows how people react to robots making errors of different severity [15] and type [16], [9], e.g., comparing their reaction to social and technical errors. Moreover, there is research highlighting the usefulness of warning the user of possible errors before they happen, in addition to exploring various recovery strategies [17]. Eye gaze, specifically, has been shown to be focused on the robot for a longer duration when it makes errors as compared to when it works as planned [9]. This finding is reiterated by Kontogiorgos et al. [8], who show an increase in the number of gaze shifts towards the agent during error situations as well as an increase in the proportion of time spent looking at the robot. Additionally, they show that this increase is even more prominent if the robot is more human-like. This gaze pattern is also backed up by a case study indicating that deviations from normal eye gaze patterns during collaborative tasks hint at something unexpected having occurred [7].

Furthermore, eye gaze patterns have been studied with the goal of identifying situations in which the user requires help. One study, for instance, explores the user's gaze movements between task environment and robot in a scenario where they sort medication under the guidance of a robot [10]. In another study, the researchers aim to predict the user's confusion from their eye gaze during a task, in which the user has to find specific information on a website [20].

In comparison, the novelty of our study lies in analysing the user's gaze patterns in a fully collaborative task in which both, human and agent, have to contribute towards a common goal by fulfilling their own distinctive set of sub-tasks. In such a scenario, it is not only the agent that can make errors, but also the human participant that might not be aware of the next optimal action and might be confused about what to do. To the best of our knowledge, previous literature analyses either human errors and confusion or errors made by the agent. In addition, previous work does not consider scenarios in which both partners work together on complex team tasks with interdependent sub-tasks, which we argue are more representative of real-world collaborations.

Eye Gaze as Reward Signal. We are interested in the potential of human gaze as implicit reward signal from which an agent can learn. Saran and Short [21] show an effective robot learning framework in which gaze patterns augment the kinesthetic demonstrations of a human teaching a robot to perform manipulation tasks. Another approach for leveraging physiological and social cues as reward signal is presented by the EMPATHIC framework that allows agents to learn how to perform simple tasks based on mapping the facial expressions and head movements of a human observer to a reward function [22]. In this study, we correlate gaze patterns with the outcome of the collaboration to show the potential of eye gaze as reward signal during interactive learning with a human without access to explicit demonstrations.

III. Experiment Design

A. Hypotheses

Using a within-participants design[1], we test the following hypotheses in our experiment:

H1: The user's gaze will be proportionally more focused on the agent during time-frames in which the agent makes errors compared to those where the agent makes no errors.

H2: The user's gaze patterns during time-frames in which they do not know what to do next will differ from when they are sure what to do next.

H3: The less the user looks at the agent the better the overall team performance.

H1 is inspired by previous research on error situations during human–robot interactions showing an increase in gaze shifts towards the robot during errors [9]. In H2, we do not hypothesize a direction in which the gaze shift will differ as previous research has shown that gaze patterns during confusion can become complicated and are hard to manually interpret [20]. Lastly, in the case H3 holds, it would provide a simple mapping between an observable property of the interaction (success) and the gaze patterns.

B. Measurements

The experiment has a single independent variable with three conditions describing the different states the human–agent team can be in during the interaction. The conditions are: (1) *agent-error*, which encompasses all points in time during which the agent makes an error; (2) *user-confusion*, describing states where the user does not know what to do; and (3) *normal workflow*, where both team members work towards the common goal with neither agent-errors nor user-confusion being present. Each participant will experience all three states during the experiment, and participants who do not experience all three conditions will be excluded. Section IV-B describes the manual classification of each interaction's time points into one of these three conditions.

Our dependent variables for H1 and H2 are the participants' eye gaze patterns as indicated by an area-of-interest analysis, where we classify the recorded gaze location into one of three categories: *gaze on user*, *gaze on AI agent*, or *gaze on environment*. For H3, the dependent variable is the team performance as measured by the achieved game score.

C. Collaboration Scenario

Our experiments were conducted in a collaborative domain inspired by the video game Overcooked [23]. In Overcooked, multiple agents or players work together in a kitchen with the goal of preparing food according to incoming orders. This requires the coordination of various interdependent sub-tasks, e.g., preparing ingredients, cooking, and delivering dishes.

[1]We pre-registered the hypotheses, study design, and analysis plan at https://doi.org/10.17605/OSF.IO/PT296

Environments based on Overcooked were used in recent reinforcement learning research [24], [25] as they provide a collaborative domain in which multiple agents have common goals as opposed to the more broadly studied domains in which agents act alone or in competition. Another line of research uses Overcooked to give agents a theory of mind that allows them to infer the user's current intentions and, thus, helps the agents to adapt their actions to a particular user [26] or to generate user-centered explanations [27].

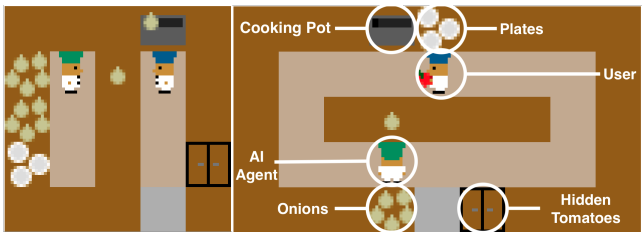

Fig. 2. The two Overcooked levels used in the study.

D. Task and Agent Errors

We base our Overcooked environment on the implementation by Carroll et al. [24]. Our extended version is available online[2]. In our version, each participant plays two levels together with a rule-based agent. A level lasts for 90 seconds, which ensures that there is enough time for the user to experience confusion and encounter the agent's error. In both levels, the human and agent work together to cook as many soups as possible by: putting two onions and one tomato into a cooking pot, turning the pot on, putting the cooked soup on an empty plate, and delivering the plate to a serving station. The user acts in the environment by pressing the arrow keys to move and using the spacebar to interact with objects.

In the first level, which can be seen on the left of Figure 2, the kitchen is split into two detached parts with the agent working on the left and the participant on the right. The agent has to hand over the plates and onions to the participant, who then has to cook them, put them on a plate, and deliver them. Additionally, the user needs to take a tomato from their side of the kitchen and place it into the cooking pot with the onions. To ensure some confusion during the interaction, we hide the tomatoes in a cupboard placed on the bottom right of the level. When reading the instructions beforehand, the user is only presented with a sprite of the tomatoes and is not told about the cupboard. In order to prevent the user getting stuck due to not finding the tomatoes, we display a hint if the user fails to deliver a soup after 15 seconds. Additionally, we pre-programmed the agent to make errors after every two soups that were delivered successfully by giving the user a plate too early when they would have still required another onion or an onion when they would have required a plate.

The second level is shown on the right of Figure 2. Here, both cooks work in a shared environment where the onions are separated from the pot by a counter in the middle. Therefore, the most efficient strategy is for one cook to provide the ingredients over the counter and for the other cook to take them from the other side and put them in the pot. The confusion in this level is mainly generated by initially having to figure out this strategy as, if the participant tries to do everything on their own, their path will be blocked by the agent. In the error condition, we programmed the agent to suddenly change strategies by walking to the pot and placing the onion in the pot instead of handing them over as before. After the error, they return to the optimal strategy.

[2]https://github.com/lwachowiak/overcooked-demo

E. Study Protocol

Firstly, participants filled out a questionnaire about their previous experience with HRI and Overcooked. They then read instructions indicating what the task environment looks like, how they interact in it, and what their goal is. Next, we calibrated the eye tracker. Afterwards, the participants played the two levels of Overcooked described in Section III-D. After playing, we conducted a retrospective think-aloud session [28], during which the participants were asked to re-watch a recording of their task performance while narrating their thoughts during that task retrospectively.

F. Participants

We recruited 35 participants from a pool of computer science students and university staff. We had to exclude 5 participants as they did not produce enough soups to reach the error condition. Subsequently, we ended up with 9 female and 21 male participants within an age range of 19 to 66 with an average age of 25.6 (SD=8.2). On a Likert scale from 1 to 5, the average experience with playing Overcooked was 2.1 (SD=1.6) and the average familiarity with HRI research was 2.9 (SD=1.2). The study was approved by the KCL Ethics Committee (Ref.: MRSP-21/22-29071) and participants were given an information sheet and signed a consent form.

IV. Data Collection and Annotation

A. Data Collection and Processing

To record the user's gaze we used the wearable eye tracker Pupil Core [29]. The setup is depicted in Figure 1. Additionally, we captured the computer screen to record the task performance. The retrospective think-aloud session was recorded with a microphone and transcribed afterwards.

We processed the collected eye gaze data to allow for an area-of-interest analysis. In order to achieve this, we computed the coordinates of the agent and the human-controlled character in the video recorded with the eye tracker. To do so, we extracted the pixel-coordinates based on the color of their hats. This information allowed us to directly compare the coordinates of the eye gaze with the coordinates of the two characters. Hence, we computed for each frame whether the gaze was focused on the agent, the human-controlled character, or the remaining task environment. The gaze data had to be downsampled to 12 samples per second by computing medians in order to be aligned with the video data. The resulting dataset contains 61,951 frames of synchronized eye gaze and agent location data, with statistics based on condition and gaze target available in Table I.

TABLE I

STATISTICS FOR THE RESULTING DATASET OF ALL PARTICIPANTS

	Gaze fixated on		
	Agent	User-controlled character	Environment
Frames – Lvl 1	806 (2.6%)	7,117 (22.7%)	23,432 (74.7%)
Frames – Lvl 2	2,192 (7.6%)	6,964 (22.8%)	21,440 (70.1%)
	Frames labeled as		
	User-confusion	Agent-error	Normal Workflow
Frames – Lvl 1	12,672 (40.4%)	1,872 (6.0%)	16,811 (53.6%)
Frames – Lvl 2	3,373 (11.0%)	2,115 (6.9%)	25,108 (82.1%)

During most frames, the gaze is directed at the environment, followed by the user-controlled character, and the agent.

B. Annotation

We annotated the collected game play and eye gaze data so that each point in time was labeled either as the agent making an error, the user being confused about what to do, or as normal workflow without any agent-error or user-confusion. As in the work by Trung et al. [16], the label for *agent-error* starts at the point in time when the agent first deviates from its normal behavior and ends when the agent returns to the correct routine. A point in time is labeled as *user-confusion* if the user does not execute any productive actions for some time or indicates during the think-aloud session that they did not know what to do. Initially, two annotators labeled the first three participants and subsequently revised the annotation rules. Labeling the remaining 27 participants resulted in the two annotators agreeing on 91% of the data points. Computing Cohen's Kappa, a chance-corrected coefficient of agreement, we obtain a value of 0.75 for the first level, 0.85 for the second level, and 0.80 overall, which corresponds to "substantial agreement" according to Landis et al. [30]. The agreement is higher for the second level as the annotators encountered less user-confusion with the users already accustomed to the mechanics of the task. Conflicts were resolved by going through the data once more together and discussing all disagreements. This process resulted in a dataset with most frames being labeled as normal workflow, followed by user-confusion, and agent-error (see Table I).

V. RESULTS

We present the results as follows: (1) a comparative analysis of gaze patterns across conditions; (2) a replication of the analysis but with redefined areas-of-interest where the areas for each agent are defined more leniently; and (3) the correlations between score and eye gaze patterns.

A. Results Overall

Figure 3 shows the area-of-interest analysis grouped by our independent variable, splitting the interaction into three conditions: user-confusion, agent-error, and normal workflow. During all three conditions, people look at the environment the most. However, this value is the highest during user-confusion. Compared to the other conditions, people look at the agent the most during agent-errors. During normal workflow, people look at themselves more than otherwise.

Firstly, in order to infer whether there is a difference in the combined dependent variables between at least two of

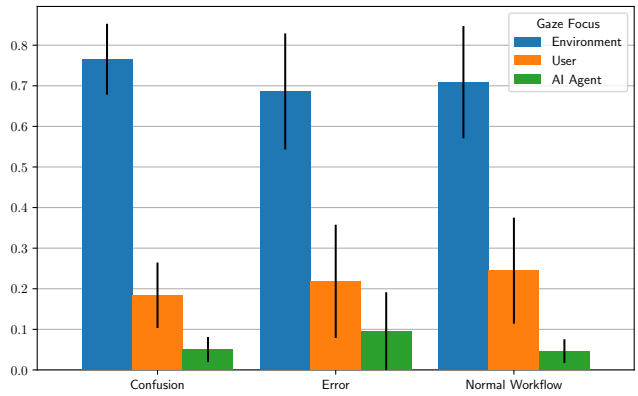

Fig. 3. Proportional amount of gaze towards different areas-of-interest with respect to the three conditions: Confusion, Error, and Normal Workflow.

the conditions, we conducted a one-way repeated-measures MANOVA in SPSS. The multivariate test showed a significant difference with $F(4, 114) = 4.962$, $p = 0.001$, $Wilk's\ \Lambda = 0.725$, and $partial\ \eta^2 = 0.148$. Secondly, post-hoc univariate tests, which tell us which of the dependent variables differ, showed significance for gaze on agent and gaze on environment. As sphericity could not be assumed for the gaze on user and agent, we report their Greenhouse-Geisser corrected values: gaze on user ($F(1.63, 47.283) = 3.089$, $p = 0.065$), gaze on agent ($F(1.204, 34.907) = 6.078$, $p = 0.014$), gaze on environment ($F(2, 58) = 5.648$, $p = 0.006$). Lastly, we report the results of Bonferroni-corrected post-hoc pairwise comparisons in Table II, which tell us precisely where the differences are. The amount of time looking at the user is significantly higher during normal workflow compared to user-confusion. Moreover, the amount of time looking at the agent is significantly higher during agent-error as compared to normal workflow. Lastly, the amount of time looking at the environment is significantly higher during confusion as compared to both other conditions.

Thus, **H1** can be only partially confirmed as the difference between the proportional amount of gaze towards the agent during agent-errors is only significantly different from the amount during normal workflow, but not from the amount during user-confusion. **H2** can be confirmed as the gaze differs in multiple statistically significant ways during user-confusion.

TABLE II

PAIRWISE COMPARISONS FOR EACH DEPENDENT VARIABLE BETWEEN ALL COMBINATIONS OF CONDITIONS; *: $p < 0.05$

Gaze on	Condition A	Condition B	Diff.	Std. Err.	Sig.
User	Workflow	Confusion	0.60*	0.019	0.010
	Workflow	Error	0.026	0.029	1.000
	Confusion	Error	-0.034	0.024	0.507
Agent	Workflow	Confusion	-0.004	0.007	1.000
	Workflow	Error	-0.049*	0.018	0.032
	Confusion	Error	-0.045	0.019	0.074
Env.	Workflow	Confusion	-0.057*	0.020	0.022
	Workflow	Error	0.023	0.027	1.000
	Confusion	Error	0.080*	0.025	0.012

B. Impact of Area-of-Interest Bounds

The results presented in Section V-A were computed for when the gaze is on the user-controlled character or the agent directly. Figure 4 shows the result for the same analysis but with a more lenient classification of whether the gaze is on one of the cooks, thus, also considering glances in the direction of the agent or user-character. This is important since people do not always look directly at the cooks, especially when those are moving, and the tracker can sometimes be off by a few pixels. In order to realize this analysis, we consider a gaze as being on the agent if the distance to the agent's center is 80 pixels as compared to 40 pixels in the analysis of Section V-A. Although the general trends are the same, the differences between the conditions are more pronounced, which is directly reflected in the significance tests. In contrast to before, all three univariate tests are significant ($p < 0.001$). Furthermore, the pairwise comparisons show significance for the same pairings as before, as well as for the proportional gaze amount on the user between the conditions normal workflow and agent-error ($p < 0.01$), and the proportional gaze amount on the agent between the conditions agent-error and user-confusion ($p < 0.001$).

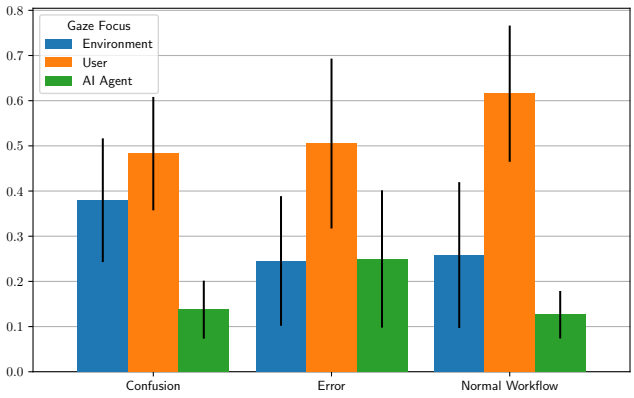

Fig. 4. Area-of-interest analysis with lenient cook detection bounds

C. Correlating Eye Gaze with Team Performance

We computed the Pearson correlations between the scores, i.e., number of delivered soup, of each participant and the respective proportional amount of gaze focused on the different areas of interest throughout the task. For each resulting correlation coefficient we computed p-values via permutation tests, i.e., recomputing the coefficient 10,000 times with permuted lists. None of the correlations was statistically significant. We, thus, cannot confirm **H3**.

VI. Discussion

Studying the participants' eye gaze during user-confusion, agent-error, and productive workflow, we were able to identify multiple statistically significant differences. In light of these results, we discuss whether the differences are sufficient for an agent to identify these critical situations. In addition, we discuss qualitative insights, shortcomings, and how to transfer the study to physical robots.

Predicting Errors and Confusion. As the human's gaze patterns during critical moments such as user-confusion and agent-error differ from those during normal workflow, we conclude that being sensitive to the human's eye gaze can help an agent to identify their own errors and determine situations in which to help their human collaborator. However, these area-of-interest patterns are only observable over multiple seconds and a simple glance at the agent, for example, does not indicate much by itself. For instance, we observed that participants regularly looked at the agent when they needed another ingredient an were waiting for the agent to give it to them. In order to quickly and effectively identify user-confusion and agent-errors, further gaze-based signals, such as number of saccades and fixations, or additional modalities, e.g., facial emotions, should be included.

Gaze as Reward Signal. Although we could not show a correlation between score and gaze patterns, we believe that gaze can play a role as implicit reward signal since increased looks at the environment indicate confusion and increased looks at the agent indicate agent-errors, thus, states the agent tries to avoid. One reason why the patterns did not correlate with the score might be that people who scored highly also encountered more agent-errors in their playthrough.

Insights from the think-aloud sessions. The retrospective think-aloud protocols offer qualitative support for our analysis. During productive workflow, for instance, one participant mentioned that they "didn't really think about what they [the agent] were doing particularly or worrying about what was happening. I was just thinking about myself.", which fits well with the small amount of time people tended to look at the agent during normal workflow. The protocols also highlight different sources of confusion, e.g., confusion about starting the experiment ("At first it was very confusing") or about task mechanics ("I am not sure how to start the soup").

Shortcomings. In our analysis, a time frame is labeled as agent-error independently from whether the user notices it. However, in the retrospective think-aloud sessions only 44% of users that reached the error condition in the first level mentioned it explicitly compared to 86% who mentioned it for the second level. It would, thus, be interesting to see if an analysis of only those errors that were consciously noticed gives the same results. Moreover, our task caused more confusion than we anticipated. For instance, two participants got themselves into a deadlock by picking up a plate after putting an object on every counter, whereupon they could not proceed anymore as there was no way to get rid of that plate. In addition, multiple participants were confused by having to manually start the cooking process and found it hard to identify when the soup was ready. While on the one hand, these additional sources of confusion make for new interesting data points, they also make the annotation of time frames a lot harder and create a strong reliance on the think-aloud data. For follow-up experiments, we, therefore, plan to include a tutorial level instead of only providing written instructions. Furthermore, agent-error and user-confusion can be hard to keep apart, can interact in various ways, and can be influenced by the user's potentially faulty beliefs: users

can be confused by an agent's erroneous actions leaving them not knowing how to act; users can interpret correct agent behavior as erroneous if they themselves assume wrong goals; and users can engage in a completely flawed chain of actions but still be confident in their actions, believing to do the right thing. All these scenarios might warrant different interventions by the agent, but require a sophisticated theory of mind that infers the partners goals, beliefs, and emotions.

Transferring the experiment to physical robots. Since a task as complex as presented here is difficult to implement, a first challenge is to find an appropriate collaborative task that robots can achieve but that still incorporates both partners working towards a common goal and is demanding enough to elicit confusion. Secondly, in the physical world, participants do not usually look at themselves, which is why new task-related areas-of-interests have to be defined.

VII. Conclusion

In this study, we investigated people's gaze patterns during human–agent collaborations and showed how they differ across situations during which a normal workflow is given, the user is confused, or the robot is making an error. This marks an important step towards developing robots that can understand their collaborators and proactively adapt to them.

In the future, we want to explore such adaptations in the form of explanations provided by the agent, which either help the user resolve their confusion or explain the agent's own actions during errors. Moreover, the agent might adapt its behavior directly or use gaze as an additional reward signal during learning. The goal of such user-centered, socially aware behavior is to benefit people's perception of robots in their life, facilitate trust, and increase team performance. In order to model a user with high accuracy we also plan to include more modalities into the predictive system, e.g., facial emotions or posture. Furthermore, we plan to extend our study using physical robots and to experiment with a wider variety of error types and collaborative domains.

Acknowledgments

The authors would like to thank all participants.

References

[1] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, "Open Problems in Cooperative AI," *arXiv:2012.08630 [cs]*, Dec. 2020.

[2] 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.

[3] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, 2018.

[4] S. Rossi, F. Ferland, and A. Tapus, "User profiling and behavioral adaptation for HRI: A survey," *Pattern Recognition Letters*, vol. 99, pp. 3–12, Nov. 2017.

[5] 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.

[6] H. Admoni and B. Scassellati, "Social Eye Gaze in Human-Robot Interaction: A Review," *Journal of Human-Robot Interaction*, vol. 6, no. 1, p. 25, Mar. 2017.

[7] R. M. Aronson and H. Admoni, "Gaze for Error Detection During Human-Robot Shared Manipulation," *Fundamentals of Joint Action workshop, Robotics: Science and Systems*, 2018.

[8] 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*. ACM, Apr. 2020, pp. 1–14.

[9] D. E. Cahya, R. Ramakrishnan, and M. Giuliani, "Static and Temporal Differences in Social Signals Between Error-Free and Erroneous Situations in Human-Robot Collaboration," in *International Conference on Social Robotics*. Springer International Publishing, 2019, vol. 11876.

[10] U. Kurylo and J. R. Wilson, "Using Human Eye Gaze Patterns as Indicators of Need for Assistance from a Socially Assistive Robot," in *International Conference on Social Robotics*, 2019, vol. 11876.

[11] G. L. Davidson and N. S. Clayton, "New perspectives in gaze sensitivity research," *Learning & Behavior*, vol. 44, no. 1, Mar. 2016.

[12] M. Çetinçelik, C. F. Rowland, and T. M. Snijders, "Do the Eyes Have It? A Systematic Review on the Role of Eye Gaze in Infant Language Development," *Frontiers in Psychology*, vol. 11, p. 589096, Jan. 2021.

[13] O. Špakov, H. Istance, K.-J. Räihä, T. Viitanen, and H. Siirtola, "Eye gaze and head gaze in collaborative games," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, 2019.

[14] S. Honig and T. Oron-Gilad, "Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development," *Frontiers in Psychology*, vol. 9, p. 861, Jun. 2018.

[15] M. Stiber and C.-M. Huang, "Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity," in *Companion Publication of the 2020 International Conference on Multimodal Interaction*. ACM, Oct. 2020, pp. 97–101.

[16] 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*, Nov. 2017.

[17] M. K. Lee, S. Kiesler, J. Forlizzi, S. Srinivasa, and P. Rybski, "Gracefully mitigating breakdowns in robotic services," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Osaka, Japan: IEEE, Mar. 2010, pp. 203–210.

[18] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, "To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot," *Frontiers in Robotics and AI*, vol. 4, 2017.

[19] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 719–735, Jun. 2003.

[20] J. Salminen, M. Nagpal, H. Kwak, J. An, S.-g. Jung, and B. J. Jansen, "Confusion Prediction from Eye-Tracking Data: Experiments with Machine Learning," in *Proceedings of the 9th International Conference on Information Systems and Technologies*, Mar. 2019.

[21] A. Saran and E. S. Short, "Understanding Teacher Gaze Patterns for Robot Learning," in *Conference on Robot Learning*, 2020, p. 12.

[22] 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.

[23] "Overcooked," Ghost Town Games, 2016.

[24] 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.

[25] D. Strouse, K. R. McKee, M. Botvinick, E. Hughes, and R. Everett, "Collaborating with Humans without Human Data," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021, p. 28.

[26] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner, "Too Many Cooks: Bayesian Inference for Coordinating Multi-Agent Collaboration," *Topics in Cognitive Science*, vol. 13, no. 2, pp. 414–432, 2021.

[27] 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 *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020, pp. 1119–1126.

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

[29] M. Kassner, W. Patera, and A. Bulling, "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, Sep. 2014.

[30] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, 1977.