Measuring, Predicting, and Leveraging Trust for Proactive Dialog in Human-AI Teams

MATTHIAS KRAUS, University of Augsburg, Germany

Effective communication, prediction of teammates' actions, and high-level coordination are essential for success in Human-AI teams. One important research topic concerning Human-AI teams is the question of how to model the AI teammate's proactive behavior and communication during collaboration. Proactivity is defined as an AI's self-initiating, anticipatory behavior for contributing to effective and efficient task completion. However, wrongful proactive behavior may have potentially dangerous consequences, which may ultimately sabotage effective teamwork between humans and AI. To prevent this, the design of adequate proactivity for AI-based systems in order to support humans is an important but challenging topic. In this work, we summarize our recent works for modeling proactive dialog in decision-making and assistance agents. This includes methods for measuring human-AI trust during collaboration, a framework for predicting the user's trust level during collaboration, and an approach to utilizing trust for adapting proactive dialog strategies. Additionally, we present an overview of our research results and evaluations, which demonstrate the importance of trust in forming effective Human-AI teams.

Additional Key Words and Phrases: human-computer interaction, dialog system, proactive conversational AI, reinforcement learning, trust, human-AI team

ACM Reference Format:
Matthias Kraus. 2023. Measuring, Predicting, and Leveraging Trust for Proactive Dialog in Human-AI Teams. 1, 1 (April 2023), 13 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

For decades, working together with artificial intelligence (AI) for solving specific tasks or complex problems has been intriguing to researchers, but also movie writers as well. For example, in a scene from the science-fiction movie "Her" by Spike Jonze, a very sophisticated artificial personal assistant named *Samantha* is able to help the main character in organizing his work by reviewing incredibly fast a large number of e-mails and providing helpful suggestions using natural conversation. However, *Samantha* was not only able to solve specific tasks, but could also socialize and form a personal bond with its user to build an effective team. Despite the availability of conversational assistants like Amazon Alexa, or Siri, or the recent hype about text-based assistants like ChatGPT, that are able to provide assistance for various tasks using natural language and other modalities, we are still not able to achieve such a level of Human-AI team (HAIT). Generally, HAITs require close coordination between humans and AI teammates, who work together towards a common goal [63]. Effective communication, prediction of teammates' actions, and high-level coordination are essential components of this collaborative effort. In order to achieve success, humans and AI teammates must share critical information through various forms of communication and work together to progress through tasks in a coordinated manner.

Author's address: Matthias Kraus, matthias.kraus@uni-a.de, University of Augsburg, Augsburg, Germany.

Manuscript submitted to ACM

For establishing effective HAITs, several research topics need to be addressed. For example, expectations of humans toward AI teammates need to be identified [63]. It needs to be determined which team roles the AI-powered system can take during collaboration [53] or how a congruent human's mental model of an AI's capabilities can be achieved for increasing team performance [3]. Another important research topic concerning HAITs is the question of how to model the AI teammate's proactive behavior and communication thereof during collaboration (e.g. see Horvitz et al. [19]). Proactivity can be defined as an AI's self-initiating, anticipatory behavior for contributing to effective and efficient task completion. The concept of proactivity has been shown to be essential for human teamwork as it leads to higher job and team performance and is associated with leadership and innovation [9]. In addition, it refines one's intrinsic motivation and self-regulation [12] while also creating coworker trust in work environments and positively contributing to socialisation [46]. However, the design of adequate proactivity for AI-based systems in order to support humans is still an open question and a challenging topic. Particularly, since proactive actions cause a shift of control towards the agent which may induce a loss of self-autonomy and wrongful proactive behavior may have potentially dangerous consequences. Therefore, a formation of trust for the user is required during collaboration, otherwise, the AI agent will possibly be rejected and becomes obsolete [51].

In that sense, it is essential to study the impact of proactive system actions on the human-agent trust relationship and how to use information about an AI agent's perceived trustworthiness in order to model appropriate proactive behavior for forming effective HAITs. In this paper, we summarize our recent works for addressing these issues. Foremost, we are interested in how to adequately communicate proactive behavior using natural language. For this, we present our methods for measuring human-AI trust during collaboration and present the results of the impact of different proactive dialog strategies on the user's perceived trust in the system. Further, we present our framework for predicting the user's trust level during collaboration with an AI assistant with a focus on decision-making tasks. Finally, we present our approach to utilizing trust for adapting proactive dialog strategies and report the results of our evaluation. To provide the necessary background on our work, we first provide some essential related work regarding human-computer trust (HCT) and proactive human-computer interaction (HCI).

2 RELATED WORK

2.1 Human-Computer Trust

In discussing the concept of trust between artificial agents and humans, we have adopted Lee and See's definition of trust as "the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [35]. This definition is frequently used and HCI we consider it to be very appropriate for explaining trust in HAITs. Lee and See also propose three factors that should be taken into account when modeling trust: the human, the autonomous partner, and the environment. Each of these factors has unique properties that influence the relationship between humans and autonomous partners. These properties include the gender and personality of the user or agent, the degree of automation, the anthropomorphism of the agent, and the type and difficulty of the task. For more detailed information on the impact of each of these factors, we recommend referring to Schaefer et al. [51]. In our work, we use trust-related properties as features to estimate the user's trust in a proactive AI agent during collaboration. How we predicted the user's trust in a proactive dialog system is described in Section 4.

In studies involving HCI, the user's trust in a system is typically measured subjectively using self-reported questionnaires (e.g. see [36, 37]. For our studies, we primarily used a questionnaire based on a hierarchical model of trust that differentiates between affect-based and cognition-based trust developed Madsen and Gregor [36]. Affect-based trust is

primarily concerned with attributes relevant to long-term relationships, whereas cognition-based trust encompasses features relevant to short-term interactions, such as perceived understandability, technical competence, and reliability. In short-term interactions, the functionality and usability of a system are particularly important, with the system's competence and reliability having a significant impact on the human-autonomous partner relationship [34, 44]. Additionally, we utilize a single-scale trust measurement applying a questionnaire developed by Jian et al. [21] and its German variant [24].

2.2 Proactivity in HCI

The concept of proactivity has been widely studied in the fields of HCI and AI [7, 19, 40, 50, 62], but there are varying definitions across different research areas. In recommendation systems [8], proactivity refers to suggesting specific items to simplify a user's navigation, while in open-domain dialog systems, it involves actively leading the conversation and changing the topic [57, 60, 61, 64]. In personal intelligent assistants [50, 62], proactivity is closely related to mixed-initiative interaction [19] and can be realized in multiple ways. In our works, we applied the definition of proactivity for such interaction. In mixed-initiative interactions, a proactive dialog is used to communicate and negotiate a system's decision-making process to minimize the risk of system failure and to solve tasks efficiently. The degree of proactivity can be modeled using different levels of autonomy (LoA) [52]. We transferred the concept of the LoA into the dialog domain similar to the proposal of an Interface-Proactivity Continuum by Isbell and Pierce [20]. For this, we defined proactive dialog actions [28], which include four levels of proactivity: *None*, *Notification*, *Suggestion*, and *Intervention*. These actions range from non-intrusive to very intrusive. As we considered HAITs for decision-making, the system aimed to provide helpful information and suggestions using natural language. With the reactive *None* action, the system waited for the user to explicitly request suggestions. The more cautious proactive actions, *Notification* and *Suggestion*, allowed the user to confirm the system's suggestions and differed only in their level of directness. The *Intervention* actions completely removed the responsibility from the user and autonomously selected an option.

For realizing proactive personal assistants, rule-based approaches are often used to ensure reliable and predictable behavior. Examples of proactive assistants include RADAR [11, 14] and CALO [62], which used task-specific dialog managers to determine the level and timing of proactive behavior based on a user's attention and interruption policies. However, the rules for proactive behavior may not be transferable to different scenarios and task domains, and they can be costly to develop and perceived as inflexible. For these reasons, we deem statistical methods for proactive dialog management beneficial. Especially, the reinforcement learning (RL) paradigm [54] seems to be predestined for proactive dialog management due to its nature of long-term planning and optimal decision-making under uncertainty. Therefore, we apply RL for deciding on optimal proactive system actions during mixed-initiative interaction in a HAIT consisting of one user and one agent. This will be described more in detail in Section 6. However, we first present our methods for measuring trust during proactive dialog and report a summary of our evaluation results.

3 MEASURING TRUST FOR PROACTIVE DIALOG

For measuring the impact of proactive dialog on the trust relationship in HAITs, we conducted three experiments with full-functioning prototypes and compared different proactive strategies according to the proactive dialog act taxonomy mentioned earlier. Here, our method was as follows: depending on the type of experiment we measured trust using the scale by Jian et al. [21] and its sub-components using the scales provided by Madsen and Gregor [36] during and/or after the experiment. As trust is a subjective variable that differs between individuals, we measured a trust baseline before the experiments for each study participant using the propensity to trust autonomous systems by Merritt et al.

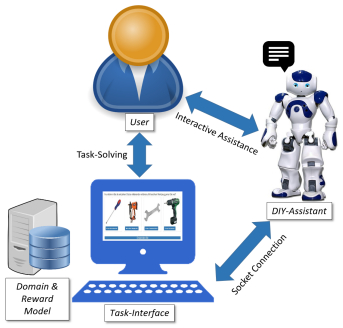

Fig. 1. Depiction of the study apparatus. Taken from [28].

[39]. In the following, we shortly describe each experiment and summarize our findings regarding the impact of each strategy on trust and its sub-components as well as the trust trajectories during the experiments.

3.1 The Planning Agent Experiment

3.1.1 Overview. In this experiment [29], we investigated the impact of proactive dialog strategies on trust dependent on task difficulty. For assessing the difficulty of a task, we made use of the cognitive load measurements [56] which were explicitly rated by users during evaluation using a scale by Klepsch et al. [23]. We hypothesized that dependent on the task difficulty, different proactive levels were more trustworthy than others according to varying needs of assistance dependent on task difficulty [15, 47]. Further, it was tested if there exist general differences between the proactive dialog act concepts and the trust relationship in HAITs. For evaluation, we implemented a prototype that assisted with decision-making by initiating proactive behavior dependent on recognized user uncertainty. Generally, during decision-making users have to make choices using their knowledge and the available information. Here, we expected the need for proactivity if users had problems with making a profound decision. A sign of arising problematic situations could be user hesitation to select a decision option. To prevent such problematic situations, proactive behavior may be beneficial. It was assumed that using this trigger mechanism might increase trust.

As a scenario, we selected a planning task, in which users had to make step-by-step decisions on how to plan the building of a DIY project while being assisted by an AI agent. We selected a NAO robot as an embodied AI agent. The agent could act according to one of the four proactive dialog actions, which was kept constant during the experiments and treated as a between-subject variable, while task difficulty and the triggers were used as within-subject variables. We compared two different planning task conditions and two-timing triggers - an insecurity-based trigger and a random trigger. The study apparatus is described in Fig. 1.

3.1.2 Findings. Our study found that low- and medium-level proactive actions were more trusted than reactive actions in the easier task condition, but that there were no significant differences in overall trust between proactive and reactive dialog actions.

We also found that the *None* proactive action received low trust ratings in the easier task condition, but was trusted similarly to medium-level proactive actions in the more difficult task condition. Among the proactive strategies, the *Notification* action was found to have the most impact on conveying the competence and reliability of the system, while the *Intervention* action was found to be less trusted than more conservative strategies. The study also found significant

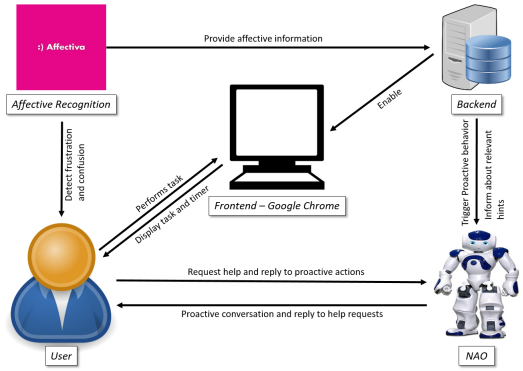

Fig. 2. System architecture. Taken from [25], licensed under CC BY 4.0 (https://creativecommons.org/ licenses/by/4.0).

gender differences in the impact of proactive system behavior on cognitive-based trust, reliability, and competence. The interplay between gender and trust is a common phenomenon in engineering and science [26, 33, 58]. Female study participants were less experienced with conversational agents and DIY tasks than male subjects. This suggested the first evidence, that the perception of proactive system behavior as trustworthy is crucially affected by the user's experience with the task and technology. Furthermore, we found significant differences between the genders regarding the big five personality traits as females rated themselves higher for neuroticism, conscientiousness, and to some degree openness toward new experiences.

Overall, the findings suggest that proactive system behavior should be subtle and give the user a feeling of involvement in the task and that the user's experience with the task and technology, as well as their personality traits, can affect their perception of proactive system behavior as trustworthy. Regarding the timing triggers, no significant differences were found.

3.2 The Tutoring Agent Experiment

3.2.1 Overview. In this study [25], we also aimed to explore how different proactive dialog actions affected trust in HAITs. Specifically, we wanted to investigate whether negative cognitive-affective states could be used as an indicator for the user's need for conversational assistance during decision-making tasks. The attentional control theory suggested that user state anxiety indicated a need for assistance [10], but previous research had expanded to include other negative cognitive-affective states, such as boredom, frustration, and confusion [17]. To measure these states, we used a high-resolution camera and the Affectiva software [38]. We focused on the negative states of confusion and frustration according to the suggestion by Friemel et al.[13] and conducted a user study to examine the effects of proactive dialog on the user's perceived trust during a concept learning task. During the task, they interacted with a Nao robot as an AI agent similar to the previous experiment. The AI agent provided help in either a reactive or proactive manner, and different timing strategies were tested to determine whether there were general differences in perceived trust based on the user's current state. The triggers started either after a random time interval or after the detection of frustration or confusion. The trigger conditions were used as within-subject conditions, while the individual proactive dialog actions were used as between-subject variables. The study apparatus is described in Fig. 2.

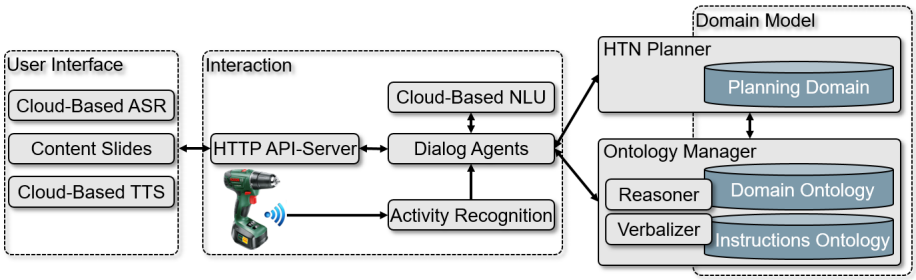

Fig. 3. Overview of Robert's architecture. A user interacted with the assistant's interface capable of multimodal input recognition. User input was forwarded to a server-based dialogue manager that mediated the interaction with the HTN planner and the ontology manager. In addition, the system was able to track the user's activity with a connected electric drill for proactive dialogue initiation. Taken from [27], reprinted according to author rights of ACM.

3.2.2 Findings. We found that low to medium levels of proactive dialog, such as *None*, *Notification*, and *Suggestion*, were associated with higher perceived trust ratings than a high level of proactivity (*Intervention*). This finding is consistent with previous research that has also shown a negative effect of high levels of proactivity on trust [49]. We also examined the trust trajectories throughout the experiment and found that the *None*, *Notification*, and *Suggestion*, strategies led to an increase in perceived user trust, while the *Intervention* strategy resulted in a decrease in trust.

We also looked at the effect of timing strategies on proactive dialog and found that high system proactivity triggered after the detection of user frustration or confusion had a significant negative effect on the perceived understandability of the AI agent. A possible explanation was, that users were focused on resolving their state of frustration or confusion and perceived the system interruption as disruptive or obstructive. Baker et a. [1] stated that frustration and confusion may be a natural aspect of the experience of learning when dealing with difficult material. Furthermore, Grasesser et al. [16] noted that confusion, although being considered a negative state, positively contributes to a user's learning experience. We concluded that acting upon recognized user confusion or frustration is not a reliable trigger mechanism for determining the need for proactive dialog behavior and that additional information, such as context or user features, may be necessary for initiation. Especially, since the Affectiva-trigger generally decreased faith in the system and perceived competence of the system.

3.3 The DIY-Assistant Experiment

3.3.1 Overview. In this experiment [27], we incorporated proactive dialog behavior into a virtual DIY assistant with advanced planning, reasoning, and dialog capabilities. The AI agent was used to assist users with a complex DIY task that involved the use of electric tools. An initial version of the agent was developed in collaboration with Robert Bosch GmbH for a nationally-funded project [4]. Here, also an initial study was conducted to test its applicability for DIY tasks. The results showed that the AI agent was more trustworthy than a baseline version without any interactive features, i.e. voice commands and question answering were not supported. The baseline provided only static, pre-defined instructions in the form of text and images. The instructions were presented as a slide mimicking the state-of-the-art of current online guides for DIY [5]. Further, the AI agent reduced the setup time for tool deployment.

In a follow-up experiment, we investigated the impact of proactive dialog on trust in HAITs by applying *Notification* proactive dialog actions and using a user activity tracking method as trigger mechanism. The trigger mechanism is based on a classification algorithm trained on movement data from a connected electric drill to predict the user's actions

and identify intervention points during the plan-based dialog. The primary goal of the investigation was to study the impact of proactive dialog on trust in HAITs depending on the user's activity. For comparison, we used the proactivity of the AI agent as a between-subject variable and compared it with a reactive agent variant. The system is described in Fig. 3.

3.3.2 Findings. The study demonstrated that incorporating proactive dialog into the interaction between the user and AI agent resulted in a significantly higher level of trust compared to the non-proactive baseline when comparing the initial trust measurements before the experiments and the overall measures after the experiment. This finding fostered our observations from our previous studies and validated the application of developed proactive actions in realistic use cases. The specific DIY use case provided quite intuitive results, as we expected that an actively engaging system was perceived as a more trustworthy assistant. Particularly by novices who seem to be in need of more natural and social assistance. Although no significant differences were observed between the proactive and baseline conditions regarding overall measurements of trust and its components, the results showed that timely proactivity tended to enhance trust toward assistants. The proactive condition was associated with higher overall trust and cognitive-based trust scores, while the baseline condition was associated with higher affect-based trust scores. This difference may be due to users' familiarity with reactive systems. Additionally, the study found gender-dependent effects of proactive dialog, with females tending to accept the proactive assistant more, while males preferred to work with a non-proactive assistant due to their experience with DIY tools and speech assistants. Overall, the study highlights the benefits of a medium-level of proactive system behavior, particularly when interacting with novice users who may feel more comfortable interacting with an actively engaging system.

4 PREDICTING TRUST FOR PROACTIVE DIALOG

The experiments conducted on proactive dialog strategies indicated that determining when and to what extent to be proactive largely depended on the user and context. The decision on proactive behavior could either benefit or harm the HCT relationship, depending on whether it was implemented appropriately or inappropriately. Therefore, both usability-based measures and the HCT relationship should be considered when deciding on appropriate proactive AI behavior. Previous research on trust has identified various human, machine, and context-related factors that influence the trust relationship [18, 35, 43, 45], indicating that an extensive set of information needs to be considered when modeling trust in proactive mixed-initiative interaction. While there are several data corpora available for conventional dialog modeling, such as DSCT [59] and MultiWoz [6], none of them are adequate for modeling proactive dialog. This is because proactive behavior is either absent or underrepresented in these corpora [2], and trust-related features are not sufficiently annotated. To address this gap, a new data corpus was created for this purpose, which involved developing an AI agent prototype for personal advising with a proactive dialog model. The agent collected personal and dialog data in a serious gaming scenario, resulting in a trust-annotated data corpus containing interactions with the proactive assistance system. This corpus was then used to develop a user model for predicting the perceived trustworthiness of a proactive system. The goal was to accurately model and predict trust using user-, system-, and context-related features during mixed-initiative dialog. The following sections provide details on the data collection method, the corpus, and the methods used to predict the HCT relationship.

4.1 Data Collection

As described in Kraus et al. [31], we gathered a dataset of 308 dialogs, with a total of 3696 exchanges between users and a proactive dialog agent. The data was collected through an online game via the clickworker [1] framework, where users had to make strategic decisions to manage a company with the agent's help. Each exchange was annotated with the user's self-reported measures of trust in the system, including competence, predictability, and reliability. The data also included objective features such as task complexity, exchange duration, user actions, and static user information such as age, gender, personality, and domain expertise. The game was structured as a turn-based planning task, where the user made decisions based on the options presented by the agent. The agent used natural language to provide suggestions and information to help the user make the best choice. We designed the agent to provide suggestions generating the most points dependent on the task step to avoid any unintended negative effects on the system's trustworthiness. The agent used a combination of rule-based and randomized strategies for selecting different proactive dialog action types to create a diverse dataset.

4.2 Trust Modelling and Estimation

We developed a new model for predicting trust in interactions with proactive AI agents based on the collected data set [30]. The model represents trust on a discrete, ordinal scale and formulates the prediction problem as a multi-class classification task, with the target classes being the distinct trust values ranging from 1 to 5. At each step of the interaction, user-specific and task-related features, along with the proactive dialog action type, are quantified and concatenated into a feature vector that is then fed to a classifier. We compared the performance of different classification algorithms - Support Vector Machine (SVM), Gated Recurrent Unit (GRU) Networks, and Extreme Gradient Boosting - and found that typically for smaller data sets, the SVM approach outperformed all other approaches. The SVM achieved an F_1-score of 0.533, Cohen's κ of 0.363, a Spearman's ρ of 0.426, and an extended accuracy [2] eA [48] of 0.895 by evaluating the different algorithms using cross-validation. We believe that even though the scores may appear low, the proposed model is useful for including trust as a metric for training a trust-adaptive AI agent, given the complexity of trust measurement.

5 LEVERAGING TRUST FOR PROACTIVE DIALOG

Based on the data corpus, we created a simulated environment for training and evaluating the RL-based proactive agent for effective HAITs. For training the agent, we implemented a corpus-based user simulator that interacted with the implemented RL-based agent. To be able to adapt the dialog to the estimated user trust in the agent, we used the supervised learning-based trust state model presented in the previous section which allows for estimating the simulated user's current trust in the agent's action. The estimate was then included in the proactive agent's state and its reward model. In the following, we present the implementation, training, and evaluation of the trust-adaptive proactive dialog agent. A depiction of the architecture is presented in Fig. 4.

5.1 Trust Adaptation using Reinforcement Learning

5.1.1 User Simulation. The aim of user simulation was to create realistic user behavior and characteristics in order to train and test proactive dialog policies. To achieve this, personal and dialog data were collected from a previous corpus

[1] www.clickworker.de
[2] describes the accuracy of the predictor given an ordinal set of classes, where the distance between the classes is also important for the quality of performance.

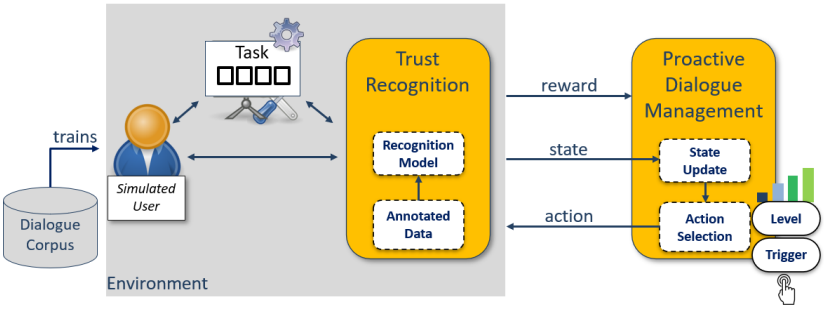

Fig. 4. Architecture of the proposed RL framework for implementing trust-adaptive proactive dialog agents. We formulate the collaboration process as an MDP and train an RL-based proactive dialog agent on interactions with a simulated user based on our data collection. For integrating user trust estimations in the state and the reward function of the agent, we utilize our trust estimation module for predicting user trust in the agent's actions in real-time.

collection and used to create distinct user types based on variables such as age, gender, technical affinity, the propensity to trust, domain expertise, and personality traits. Task-related behavior was simulated by generating values for game scores, help requests, task step duration, and perceived difficulty, with the probabilities of specific user behavior based on structured data distributions. This allowed for the reproduction of user behavior in response to proactive system actions. The quality of the user simulator was evaluated using the Kullback-Leiber distance [32], which measures the similarity between distributions of behavior generated by the simulator and those of real users. The user simulator achieved a score of 0.172, indicating realistic performance. The simulated user-specific and task-related features, as well as proactive system actions, were then used as input to the trust state model which then could estimate the simulated user's perceived trust in the system in real-time.

5.1.2 Training the RL-based Proactive AI Agent. RL allows an agent to learn strategies for solving complex problems by maximizing a reward, e.g. see Sutton and Barto [55] for more information. To apply RL to adapt proactive dialog behavior, the interaction between the simulated user and the agent needed to be modeled as a Markov Decision Process (MDP), with defined dialog states, actions, and rewards. Dialog states were modeled using both static and dynamic knowledge of the task, while the action space was defined using the four defined proactive dialog acts. The reward function was designed to promote trustworthy, successful, and efficient behavior, with higher rewards given for trustworthy behavior. A Deep-Q-Network (DQN) [41] approach with a stacked Multi-Layer Perceptron (MLP) was used for function approximation. The DQN was trained using the RMSProp algorithm with ADAM optimization [22] and a mini-batch size of 64. The trained RL-based strategy was evaluated against rule-based and static proactive dialog strategies.

5.2 Findings

Our experiments found that the user's perceived trust in the proactive dialog agent is linked to its level of proactivity. A more reactive agent is perceived as more trustworthy because the user has more control over the final decisions. On the other hand, a more proactive agent is better at achieving task success because we designed it to provide the suggestion that results in receiving the most points for the given task. Therefore, a proactive dialog agent needs to balance both reactive and proactive behavior to enhance HAITs. The study found that including a trust measurement in the dialog policy is beneficial for making decisions. The RL-based agent provided the best compromise between task efficiency and

trustworthy behavior for improving collaboration. Observing the proactive dialog act types the RL-based agent selected, the predominant use of the *Notification*-action becomes evident (38 % of all actions), while the extreme levels *None* and *Intervention* were selected well-balanced (23 % and 25 %). In previous studies, we found that notifying behavior was perceived as most appropriate with regard to a proactive system's trustworthiness. Therefore, the perception of this proactive act type may be more invariant to the respective situation and user, while purely reactive and autonomous system behavior needs to be adapted to the user and the situation.

Finally, it may be useful to consider the RL-based strategy proactive dialog act type selection dependent on the system's lastly obtained trust value or game score. Under consideration of the user's last perceived trust in the system, the proactive dialog act type *Notification* was primarily used by the agent when trust is low to medium, while the *Suggestion* action and highly proactive behavior were more extensively used for higher trust levels. At the highest trust levels, a reactive approach may be more useful to avoid harming the HCT relationship.

We describe how the system's proactive dialog act selection was affected by the user's performance in the game. The system primarily used medium proactivity when the user was unsuccessful, and a *Notification* action when the user had low task success. It used an *Intervention* action for high trust values and high task success, while a *None* action was used after observing successful user behavior but the user trust in the agent's action was low. Overall, it could be observed that the AI agent gradually gave more control to the user after detecting successful user behavior.

In this experiment, we assumed that assume that the proactive agent is always right and therefore maximizing trust was ought to be beneficial as there was no risk in following the system's action from a task-centered perspective. In more realistic scenarios, where the agent may express faulty or non-optimal performance a calibrated trust level needs to be achieved. Trust calibration is defined as the process in which a user sets an appropriate trust level corresponding to the machine's trustworthiness and uses it in accordance with its abilities and limits [42].

6 CONCLUSION

This work presents a summarization of our recent work on integrating the concept of trust in the development and implementation of proactive agents for collaboration in HAITs. We describe our method on how to measure the impact of proactive dialog strategies on the HCT using subjective, psychologically validated questionnaires. Here, we let users rate their trust and sub-components, e.g. reliability, competence, etc., at different times during the experiments. For creating a baseline, we also apply a baseline measurement of trust before the experiments.

We conducted experiments to examine the relationship between proactive dialog act types and the HCT relationship, which depends on specific user characteristics, task properties, and cognitive-affective user states. Our findings showed that proactive dialog mainly affects the user's cognitive-based trust, including perceived competence, reliability, and understandability. We also found that a medium-level of proactive dialog had positive results for inexperienced users with low technical affinity when more context-related features were included as a trigger mechanism for proactive decision-making. To improve collaboration, we suggested incorporating various user states and contributed a novel user model that predicts the user's trust level during an ongoing dialog, which could be used as dialog adaptation criteria. By implementing a trust-adaptive proactive AI agent, we were able to enable trusted and task-effective proactive behavior. Finally, we proposed a novel approach that includes trust and task metrics in a reward function for RL-based decision-making, which proved to be beneficial for improving the HAIT. Overall, we believe that technical systems that can recognize and measure social aspects such as trust during interaction with humans are essential for building effective HAITs.

REFERENCES

[1] Ryan SJd Baker, Sidney K D'Mello, Ma Mercedes T Rodrigo, and Arthur C Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies* 68, 4 (2010), 223–241.

[2] Vevake Balaraman and Bernardo Magnini. 2020. Proactive Systems and Influenceable Users: Simulating Proactivity in Task-oriented Dialogues. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue-Full Papers, Virually at Brandeis, Waltham, New Jersey, July. SEMDIAL*.

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[4] Gregor Behnke, Marvin Schiller, Matthias Kraus, Pascal Bercher, Mario Schmautz, Michael Dorna, Michael Dambier, Wolfgang Minker, Birte Glimm, and Susanne Biundo. 2019. Alice in DIY wonderland or: Instructing novice users on how to use tools in DIY projects. *AI Communications* Preprint (2019), 1–27.

[5] Pascal Bercher, Gregor Behnke, Matthias Kraus, Marvin Schiller, Dietrich Manstetten, Michael Dambier, Michael Dorna, Wolfgang Minker, Birte Glimm, and Susanne Biundo. 2021. Do It Yourself, but Not Alone: Companion-Technology for Home Improvement—Bringing a Planning-Based Interactive DIY Assistant to Life. *KI-Künstliche Intelligenz* 35, 3 (2021), 367–375.

[6] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ–A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *arXiv preprint arXiv:1810.00278* (2018).

[7] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[8] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 815–824.

[9] J Michael Crant. 2000. Proactive behavior in organizations. *Journal of management* 26, 3 (2000), 435–462.

[10] Michael W Eysenck, Nazanin Derakshan, Rita Santos, and Manuel G Calvo. 2007. Anxiety and cognitive performance: attentional control theory. *Emotion* 7, 2 (2007), 336.

[11] Andrew Faulring, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. 2010. Agent-assisted task management that reduces email overload. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 61–70.

[12] Michael Frese and Doris Fay. 2001. 4. Personal initiative: An active performance concept for work in the 21st century. *Research in organizational behavior* 23 (2001), 133–187.

[13] Celina Friemel, Stefan Morana, Jella Pfeiffer, and Alexander Maedche. 2018. On the role of users' cognitive-affective states for user assistance invocation. In *Information Systems and Neuroscience*. Springer, 37–46.

[14] David Garlan and Bradley Schmerl. 2007. The RADAR architecture for personal cognitive assistance. *International Journal of Software Engineering and Knowledge Engineering* 17, 02 (2007), 171–190.

[15] Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2008. Simultaneous teleoperation of multiple social robots. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 311–318.

[16] Arthur Graesser, Sidney D'Mello, Patrick Chipman, Brandon King, and Bethany McDANIEL. 2007. Exploring relationships between affect and learning with AutoTutor. In *Proc Int Conf AIED*.

[17] Martin Thomas Hibbeln, Jeffrey L Jenkins, Christoph Schneider, Joseph Valacich, and Markus Weinmann. 2017. How is your user feeling? Inferring emotion through human-computer interaction devices. *Mis Quarterly* 41, 1 (2017), 1–21.

[18] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[19] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 159–166.

[20] Charles L Isbell and Jeffrey S Pierce. 2005. An IP continuum for adaptive interface design. In *Proc. of HCI International*.

[21] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv e-prints* (2014), arXiv–1412.

[23] Melina Klepsch, Florian Schmitz, and Tina Seufert. 2017. Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in psychology* 8 (2017), 1997.

[24] Johannes Maria Kraus. 2020. *Psychological processes in the formation and calibration of trust in automation*. Ph. D. Dissertation. Universität Ulm.

[25] Matthias Kraus, Diana Betancourt, and Wolfgang Minker. 2022. Does It Affect You? Social and Learning Implications of Using Cognitive-Affective State Recognition for Proactive Human-Robot Tutoring. https://doi.org/10.48550/ARXIV.2212.10346

[26] Matthias Kraus, Johannes Kraus, Martin Baumann, and Wolfgang Minker. 2018. Effects of gender stereotypes on trust and likability in spoken human-robot interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

[27] Matthias Kraus, Marvin Schiller, Gregor Behnke, Pascal Bercher, Michael Dorna, Michael Dambier, Birte Glimm, Susanne Biundo, and Wolfgang Minker. 2020. "Was That Successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant Using Multimodal Cues. In *Proceedings of the 2020*

International Conference on Multimodal Interaction (Virtual Event, Netherlands) *(ICMI '20)*. Association for Computing Machinery, New York, NY, USA, 585–594. https://doi.org/10.1145/3382507.3418818

[28] Matthias Kraus, Nicolas Wagner, Zoraida Callejas, and Wolfgang Minker. 2021. The Role of Trust in Proactive Conversational Assistants. *IEEE Access* 9 (2021), 112821–112836.

[29] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2020. Effects of Proactive Dialogue Strategies on Human-Computer Trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3340631.3394840

[30] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2021. Modelling and Predicting Trust for Developing Proactive Dialogue Strategies in Mixed-Initiative Interaction. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 131–140.

[31] Matthias Kraus, Nicolas Wagner, and Wolfgang Minker. 2022. ProDial – An Annotated Proactive Dialogue Act Corpus for Conversational Assistants using Crowdsourcing. In *Proceedings of the 13th International Conference on Language Resources and Evaluation (LREC 2022)*.

[32] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[33] Theresa Law, Meia Chita-Tegmark, and Matthias Scheutz. 2020. The interplay between emotional intelligence, trust, and gender in human–robot interaction. *International Journal of Social Robotics* (2020), 1–13.

[34] John D Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International journal of human-computer studies* 40, 1 (1994), 153–184.

[35] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[36] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[37] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*. Elsevier, 3–25.

[38] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. 2013. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 881–888.

[39] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors* 55, 3 (2013), 520–534.

[40] Christian Meurisch, Cristina A Mihale-Wilson, Adrian Hawlitschek, Florian Giger, Florian Müller, Oliver Hinz, and Max Mühlhäuser. 2020. Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–22.

[41] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *nature* 518, 7540 (2015), 529–533.

[42] Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies* 27, 5-6 (1987), 527–539.

[43] Bonnie M Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922.

[44] Bonnie M Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996), 429–460.

[45] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39, 2 (1997), 230–253.

[46] Sharon K Parker, Helen M Williams, and Nick Turner. 2006. Modeling the antecedents of proactive behavior at work. *Journal of applied psychology* 91, 3 (2006), 636.

[47] Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2020. Improving worker engagement through conversational microtask crowdsourcing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[48] Niklas Rach, Wolfgang Minker, and Stefan Ultes. 2017. Interaction quality estimation using long short-term memories. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. 164–169.

[49] Pei-Luen Patrick Rau, Ye Li, and Jun Liu. 2013. Effects of a social robot's autonomy and group orientation on human decision-making. *Advances in Human-Computer Interaction* 2013 (2013), 11.

[50] Ruhi Sarikaya. 2017. The technology behind personal digital assistants: An overview of the system architecture and key components. *IEEE Signal Processing Magazine* 34, 1 (2017), 67–81.

[51] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[52] Thomas B Sheridan and William L Verplank. 1978. *Human and computer control of undersea teleoperators*. Technical Report. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

[53] Dominik Siemon. 2022. Elaborating team roles for artificial intelligence-based teammates in human-AI collaboration. *Group Decision and Negotiation* 31, 5 (2022), 871–912.

[54] Richard S Sutton. 1996. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in neural information processing systems*. 1038–1044.

[55] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[56] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[57] Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric P Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. *arXiv preprint arXiv:1905.11553* (2019).

[58] Cara Tannenbaum, Robert P Ellis, Friederike Eyssel, James Zou, and Londa Schiebinger. 2019. Sex and gender analysis improves science and engineering. *Nature* 575, 7781 (2019), 137–146.

[59] Jason D Williams, Matthew Henderson, Antoine Raux, Blaise Thomson, Alan Black, and Deepak Ramachandran. 2014. The dialog state tracking challenge series. *AI Magazine* 35, 4 (2014), 121–124.

[60] Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goals. *arXiv preprint arXiv:1906.05572* (2019).

[61] Jun Xu, Haifeng Wang, Zhengyu Niu, Hua Wu, and Wanxiang Che. 2020. Knowledge graph grounded goal planning for open-domain conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9338–9345.

[62] Neil Yorke-Smith, Shahin Saadati, Karen L Myers, and David N Morley. 2012. The design of a proactive personal agent for task management. *International Journal on Artificial Intelligence Tools* 21, 01 (2012), 1250004.

[63] Rui Zhang, Nathan J McNeese, Guo Freeman, and Geoff Musick. 2021. " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–25.

[64] Yutao Zhu, Jian-Yun Nie, Kun Zhou, Pan Du, Hao Jiang, and Zhicheng Dou. 2021. Proactive retrieval-based chatbots based on relevant knowledge and goals. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2000–2004.