Explain yourself!
Effects of Explanations in Human-Robot Interaction

Jakob Ambsdorf*, Alina Munir, Yiyao Wei, Klaas Degkwitz, Harm Matthias Harms, Susanne Stannek, Kyra Ahrens, Dennis Becker, Erik Strahl, Tom Weber, Stefan Wermter

Abstract— Recent developments in explainable artificial intelligence promise the potential to transform human-robot interaction: Explanations of robot decisions could affect user perceptions, justify their reliability, and increase trust. However, the effects on human perceptions of robots that explain their decisions have not been studied thoroughly. To analyze the effect of explainable robots, we conduct a study in which two simulated robots play a competitive board game. While one robot explains its moves, the other robot only announces them. Providing explanations for its actions was not sufficient to change the perceived competence, intelligence, likeability or safety ratings of the robot. However, the results show that the robot that explains its moves is perceived as more lively and human-like. This study demonstrates the need for and potential of explainable human-robot interaction and the wider assessment of its effects as a novel research direction.

Fig. 1. The real NICO robot on the left and its simulated counterpart. This study was conducted using two simulated robots.

I. INTRODUCTION

Explainable Artificial Intelligence (XAI) [1] promises to provide humanly understandable explanations about the actions, recommendations, and underlying causes of Artificial Intelligence (AI) techniques. Explanations of an algorithm's decisions can increase transparency and build trust in users of artificial intelligence and robotic systems [2, 3]. In Human-Robot Interaction (HRI), robots that explain their actions can increase users' confidence, trust, and reduce safety risks in human-robot cooperative tasks [4, 5], whereas miscommunication can create confusion and mistrust [6].

While the field of XAI has the potential to enhance human-robot interaction, there is still a lack of research on the effects and perception of robots that explain their actions to users. Specifically, due to the currently limited functionality and challenges to generate humanly understandable explanations from deep neural networks, the research on the interaction of humans and explainable robots is in its early stages of development [7]. Despite the increasing attention on the application of XAI and its benefits, only a small amount of user studies have been conducted [7, 8]. A considerable part of research in the field focuses on the design of theoretical frameworks for human-robot interaction studies that leverage XAI [9, 10].

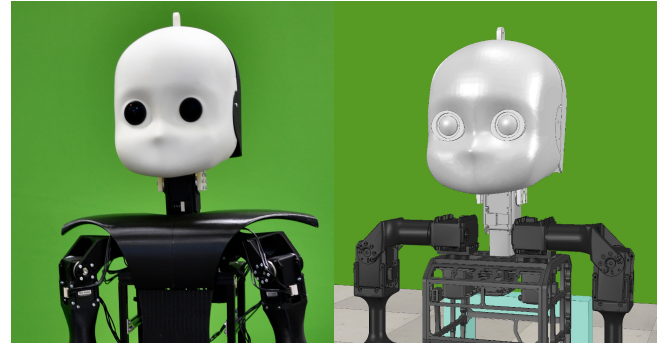To measure the effect of XAI in human-robot interaction, we conduct a study in which two simulated robots compete in a board game. While one of the robots explains the reasoning behind each of its game-play decisions, the other is constrained to only announcing each move without any explanations. For the game scenario, the game Ultimate Tic-tac-toe was selected. The game is similar in rules to regular Tic-tac-toe but is less well-known, more challenging to play, and game states are difficult to analyze [11]. For measuring the effect of a robot explaining its actions, the Godspeed series [12], perceived competence, and participants' predictions about which robot will win are assessed. While the Godspeed questionnaire measures the general robot perception, perceived competence measures an underlying dimension of forming trust.

The technical components of the study are implemented in robot simulation software and can be transferred to real robots (see Figure 1), enabling the study to be conducted in a lab experiment in the future. Due to the COVID-19 pandemic, the study was conducted entirely online.

II. RELATED WORK

The research on explainable artificial intelligence reaches back to rule-based expert systems in the 1970's [13, 14]. With the introduction of increasingly performant models, the importance of providing explanations and justifications for their actions was stressed early on [15, 16]. The recent success of deep learning [17] and the ubiquitous use of such artificial intelligence systems, which are largely considered to be black-boxes, has intensified the discussion and need for explainable AI considerably. Thereupon, explainable artificial systems have been declared an economic priority, which causes governmental regulations and technical requirements [18, 19]. For example, citizens of the

European Union, who are affected by automated decision-making systems, have the right to an explanation of that decision [20]. Such regulations enforce that AI systems are transparent, explainable, and accountable for safety-critical applications [21, 22, 23].

Consequently, there has been a surge in research of explainable systems under various names such as comprehensible, understandable, and explainable AI [24]. Various attempts at taxonomizing the methods for explaining neural networks and other machine learning algorithms have been proposed recently. The properties of these methods, such as their dependence on the employed model, the degree of faithfulness to the prediction process and their means of communication, vary considerably among them. We direct the reader to corresponding surveys and taxonomies for a comprehensive overview [1, 24, 25, 26]. The majority of the methods that have been proposed generate explanations that are understandable only for experts, rather then potential users [27]. On the other hand, in the field of autonomous robotics [28, 29] and in human-robot interaction [7, 30], explaining actions to a user rather than an expert is particularly important. This is especially pronounced in human-robot cooperative tasks [31]. In this context, explanations provide new information to the user, thereby assisting in understanding the reasoning behind actions, as well as providing an intuition about the robot's capability [10, 32].

Trust has been identified as the foundation of human-robot interaction for collaborative and cooperative tasks [33, 34]. Explanations can be considered a human-like attribute, and research has shown that humans expect explanations in conjunction with a robot's actions [35]. An accurate understanding of the robot's intent, which is derived by explanations of the robot's decision-making, can consequently increase trust [5]. Additionally, research suggests that the robot's display of human-like attributes can lead to fostering trust as well [36]. However, trust can be lost quickly in the event of inconsistent actions [37] or robot failure [38].

One dimension of human-robot trust is the robot's competence to achieve the user's desired goal [39]. The perception of a robot as competent is a requirement for the development of trust [40] and changes the social human-robot relationship [41]. The robot's competence has been suggested as one of the main factors determining the preference of one robot over another in HRI scenarios [42]. Concurrently, users tend to over- or underestimate a robot's competence based on their perception and observed robot behavior [43]. Despite the significance of perceived competence for successful human-robot interaction and its link to trust, there has been, to our knowledge, little work directly assessing how explanations can impact perceived competence.

To assess the perception of robots, remote [44, 45, 46] and video-based studies [47, 48, 49] have been successfully conducted. Prior work found that video-based HRI studies generally achieve comparable results with their live counterparts [50]. However, it has been shown that certain aspects, such as empathy for a robot and willingness to cooperate, can be greater in physical interaction with a robot as opposed

to interacting with the robot remotely [51, 52]. Due to the COVID-19 pandemic, researchers further explored the possibilities of remote HRI studies and noticed that this type of experiment can increase the participants' effort and their frustration [53]. In contrast to live HRI studies, video-based studies have the advantage of reaching more participants and consistent experiment conditions among participants [54].

III. METHODS

A. Participants

Data for the study were collected from February 19th, 2021, to March 3rd, 2021. For the study evaluation, only the data of participants that left a remark about the study's objective and required at least 20 minutes to complete the study are considered. This left a total of 92 participants for evaluation. Detailed demographic information is shown in Table I.

TABLE I
DEMOGRAPHIC INFORMATION ABOUT THE PARTICIPANTS

Gender	Frequency	%
Male	42	45.66
Female	43	46.74
Other	7	7.60
Age	**Frequency**	**%**
< 20	4	4.35
20 − 25	40	43.48
26 − 30	26	28.26
31 − 35	9	9.78
36 − 40	2	2.17
> 40	11	11.96
Nationality	**Frequency**	**%**
Asia	5	5.44
Germany	69	75.00
Rest of Europe	7	7.61
Rest of the world	2	2.17
No information	9	9.78
Occupation	**Frequency**	**%**
Full or part-time employed	30	32.61
Not employed	2	2.17
Student	54	58.70
Retired	2	2.17
Other	4	4.35
Profession	**Frequency**	**%**
Computer science relation	48	52.17
Other	44	47.83
UTTT Experience	**Frequency**	**%**
Experience with UTTT	8	8.70
No experience with UTTT	84	91.30

B. Scenario

The selected study design aims to provide a plausible scenario, in which two robots interact and one explains its actions. The setting is intended to be natural and simple to understand, while simultaneously difficult to analyze to obscure the objective of the experiment. To this end, we chose a game scenario, in which two robots play a board

game and one explains its moves. The goal of the participant is to bet on the winner of each game, after observing a game-play snippet. Prior work showed that the use of a game scenario can increase participant engangement [55].

According to the aforementioned requirements, the two-player board game Ultimate Tic-Tac-Toe (UTTT) was selected. UTTT is a relatively unknown game with game rules that are quick to understand, but the game is challenging to play due to its large state space [56, 57]. The game board consists of nine nested Tic-tac-toes that compose a larger 3x3 board. Each smaller board is called a local board and the player has to win three local boards to form a line on the global board, just like in regular Tic-tac-toe. Thus, it is non-trivial for the participants to analyze the robots' game-play.

To examine the influence of XAI, we use two humanoid robots with different behaviors. Both robots have the same capability to win, as they use the same reinforcement learning algorithm (see Section III-C.1) to play UTTT. Each turn, the respective robot provides a verbal cue in addition to its game-play. Both robots can play UTTT against each other autonomously, and simulation software was used to record videos of complete games. Figure 2 shows both robots in the simulation environment, sitting at a table across from each other with the UTTT board game in front of them.

In the study, participants are shown short sequences of pre-recorded games without knowledge about the robots' capabilities and the outcome of the game. Each short game-play video shows a sequence of three moves by each robot, in which both robots comment on their moves via speech output. While one robot merely announces its moves, the robot using XAI provides an elaboration for the selected move. To actively involve the participants, they have to bet on the robot they believe to win the game. The video snippets were kept short to increase the difficulty of judging the winner from the presented game-play alone, under the assumption that participants would rely more on the robots' utterances to judge their capability.

During the pilot study, participants reported difficulties distinguishing the robots in the questionnaire assessment. As
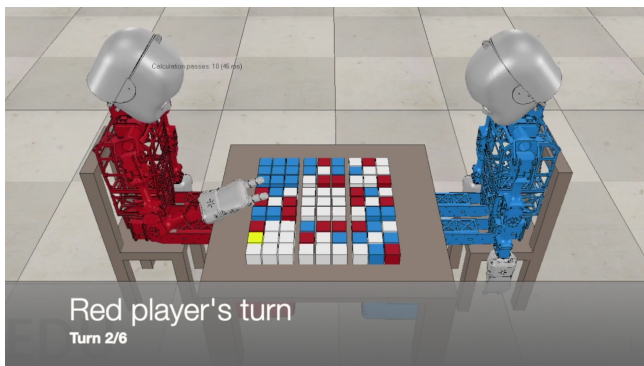

Fig. 2. Experiment setup: Two humanoid robots play a game of UTTT against each other in CoppeliaSim. In addition to voice output and robot movement, a video overlay indicates the robot's turn.

a result, the robots were colored differently and were given clearly distinguishable voices.

C. Materials, Software & Implementation

To enable two humanoid robots to play against each other and conduct the study, a variety of components are required. The center of the experiment are two Neuro-Inspired Companion (NICO) [58], which are child-sized humanoid robots. The modules that allow these two robots to play UTTT in the simulation software are implemented in the Robot Operating System (ROS) [59]. For simulation and game-play recording, the robot simulation software CoppeliaSim [60] is used. The voices for both robots are created using Google Cloud Text-To-Speech[1]. The experiment was conducted using LimeSurvey [61]. It was used to provide the participants with the game-play videos and assess the questionnaires and bets on the robots.

1) Reinforcement Learning: To present the study participants with non-trivial and unscripted game-play, a reinforcement learning agent was trained to play the game of UTTT. The reinforcement learning agent's neural network architecture is based on the model presented by Mnih et al. [62], where it is originally used to play Atari games. In contrast to Atari games where each pixel has to be processed, in UTTT only the current board state, estimated by a computer vision model, serves as the agent's input, and the agent returns a probability distribution over the board for the best move.

The agent was trained for 80.000 games against different types of simple strategy agents and itself. Initially, the agent was bootstrapped with 20.000 games against a strategy that selects a random move unless it can win a local board. Afterward, the agent was trained with 40.000 games against a strategy that does a random move unless it can win a local board or block the opponent from winning a local board. Finally, the agent was trained for 20.000 games against a random mix of the previous strategy (30%) and self-play (70%).

2) XAI-Algorithm: For generating explanations of the reinforcement learning agent's moves, a post-hoc approach is utilized. The XAI-algorithm processes the board and the agent's next move using a set of predefined rules to justify the move. The derived explanations are human justifications of the robot's actions. Overall, 12 rules were implemented in the XAI-algorithm, which are applied to generate the explanation. As an example of a generated explanation, consider the following scenario: The agent should not play in fields such that his opponent will get a chance to win within his next move. The agent's immediate goal is to force the opponent to play on a different local board. In this case, the explanation will be: "I did not play in row x and column y because my opponent could win the game if I send him to the corresponding local board." A second situation can be that the agent blocks the opponent in a local board, where he already has placed two marks in a row. Given this event,

the generated explanation is "I have blocked my opponent here, so he won't be able to win this local board".

D. Measurements

For the evaluation of the empirical study, the Godspeed questionnaire, perceived competence ratings, and bets on the winning robot are collected.

The Godspeed questionnaire measures the participant's overall perception of the robot. It measures the dimensions of anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety [12].

The Perceived Competence Scale (PCS) [63, 64] is applied to measure the robot's competence as perceived by the participant. It is a 4-item questionnaire that was initially used in the Self-Determination Theory [65]. The questions are adjusted to the subject of the study by converting the perspective of the questions from "me" to "robot". The first dimension considers the participant's belief that the robot will "make the right move", whereas the second dimension assesses their belief of the robot to "win a game".

As a direct measure of the robots' perceived competence, the participants are requested to place a bet on the robot they believe to win the game. These bets are assessed after each game-play video.

E. Study Design and Procedure

The study is constructed as a between- and within-subjects design to investigate how the robot's explanation for its game-play impacts the participant's perception of the robot. Specifically, two test groups and one control group are utilized. In the test group condition, either the red or blue robot explains its moves during the game. This is in contrast to the control group, where neither robot utilizes XAI and instead only announces its moves. The experiment conditions can be summarized as follows:

- **XAI red:** The red robot (on the left) explains the moves, while the blue robot only announces its moves.
- **XAI blue:** The blue robot (on the right) explains the moves, while the red robot only announces its moves.
- **Control:** Both robots only announce their moves.

For recruitment, participants are provided with a link to the survey and are randomly assigned via LimeSurvey to one of the three conditions (XAI red, XAI blue, control). The study begins with a general introduction to the experimental procedure. Afterward, the participants receive detailed introduction videos for the rules of UTTT and the experiment setup. Then, they are provided with an example game-play snippet and an introduction to the betting procedure.

After acknowledging to have understood the study procedure, the participants watch three game-play videos of their respective experiment condition in random order. Directly after each video, they place their bet on the robot they believe to win the game. After the experiment, the Godspeed questionnaire, perceived competence questionnaire, and demographics are assessed, where the order of which robot is evaluated first was randomized. An overview of the experiment procedure is illustrated in Figure 3.
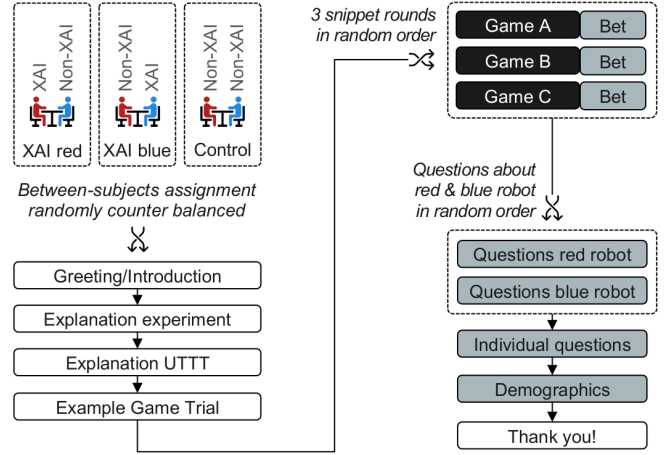

Fig. 3. Flow diagram of the experiment procedure

IV. RESULTS

For measuring the internal consistency of the Godspeed and perceived competence questionnaire, we calculate the Cronbach's Alpha [66] for each dimension. Each dimension of both questionnaires shows good internal consistency, except for the safety measure of the Godspeed questionnaire. The dimension of safety only consists of three items and the Cronbach's Alpha indicates an adequate consistency [67]. An overview of the estimated Cronbach's Alphas is shown in Table II.

TABLE II
CRONBACH'S ALPHAS OF THE QUESTIONNAIRES

Dimension	Godspeed Red Robot	Godspeed Blue Robot
Anthropomorphism	0.85	0.86
Animacy	0.90	0.89
Likeability	0.88	0.89
Intelligence	0.85	0.87
Safety	0.69	0.65
Dimension	**Perc. Competence Red Robot**	**Perc. Competence Blue Robot**
Right move	0.87	0.90
Ability to win	0.90	0.88

For the evaluation of the study data, the dimensions of the questionnaires and robots for each experiment condition (control (31 participants), XAI red (33 participants), XAI blue (28 participants)) are separated. This results in the evaluation of six different robots. An overview of the Godspeed dimensions with estimated means and standard errors is shown in Figure 4.

The individual dimensions of the Godspeed questionnaire are evaluated independently utilizing the Kruskal-Wallis test [68]. In case of significant differences among the robots, we utilize the pairwise Wilcoxon rank sum test as post-hoc analysis to estimate the differences among the robots [69].

The Kruskal-Wallis test suggests a difference in the anthropomorphism for the different robots ($p = 0.038$). However,
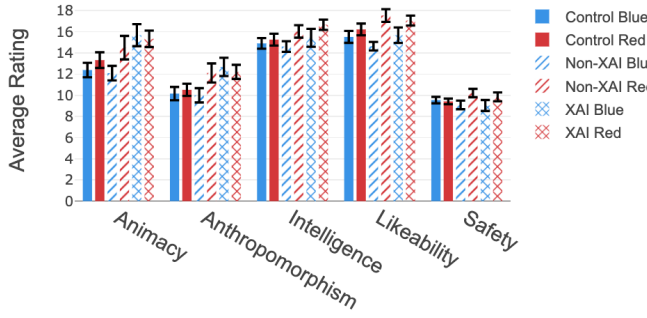
Fig. 4. Godspeed dimensions of the individual robots

The perceived competence questionnaire is evaluated identically to the previous procedure, and an illustration of the results of that questionnaire is shown in Figure 5.
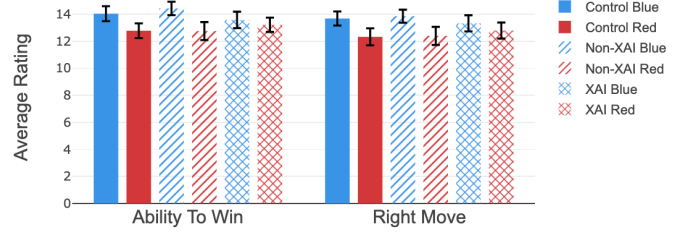

Fig. 5. Perceived competence dimensions of the individual robots

The Kruskal-Wallis test for the right move does not suggest a difference among the robots ($p = 0.225$). Correspondingly, there appears to be no difference in the robots' ability to win the game ($p = 0.205$).

The participants' betting behavior is a direct measure of the robots' ability to win the game. For the evaluation, the bets among the three games are aggregated for the different experiment conditions.

the pairwise Wilcoxon test does not reveal any significant difference. To infer if there is a difference in anthropomorphism between the XAI and non-XAI robot, the measures of anthropomorphism for both robots in the XAI experiment groups are combined. This leads to a sample of the non-XAI robots ($M = 10.97$, $SE = 0.56$) and XAI robots ($M = 12.42$, $SE = 0.53$). A Wilcoxon test reveals a significant difference between the XAI and non-XAI robot ($p = 0.043$).

For animacy, a significant difference is indicated by the Kruskal-Wallis test ($p = 0.008$). A pairwise Wilcoxon test suggests a significant difference for the blue XAI robot ($M = 15.68$, $SE = 1.04$) and red XAI robot ($M = 15.33$, $SE = 0.78$) to the blue control robot ($M = 12.39$, $SE = 0.68$) and blue non-XAI robot ($M = 12.09$, $SE = 0.69$) (Table III).

TABLE III

PAIRWISE WILCOXON TEST FOR ANIMACY

	Blue Robot Control	Blue Robot Non-XAI	Blue Robot XAI	Red Robot Control	Red Robot Non-XAI
Blue Non-XAI	0.687	—	—	—	—
Blue XAI	**0.027**	**0.027**	—	—	—
Red Control	0.415	0.415	0.191	—	—
Red Non-XAI	0.415	0.314	0.415	0.687	—
Red XAI	**0.027**	**0.027**	0.822	0.191	0.415

Likewise, for the dimension of likeability, the Kruskal-Wallis test shows a difference in the likeability among the robots ($p = 0.003$). The post-hoc pairwise Wilcoxon test (Table IV) reveals a significant difference among the likeability of the red non-XAI robot ($M = 17.54$, $SE = 0.60$) and red XAI robot ($M = 17.06$, $SE = 0.47$) to the blue non-XAI robot ($M = 14.64$, $SE = 0.40$). Neither the Kruskal-Wallis test for the intelligence ($p = 0.102$) nor the safety ($p = 0.243$) indicate a significant difference among the robots.

TABLE IV

PAIRWISE WILCOXON TEST FOR LIKEABILITY

	Blue Robot Control	Blue Robot Non-XAI	Blue Robot XAI	Red Robot Control	Red Robot Non-XAI
Blue Non-XAI	0.284	—	—	—	—
Blue XAI	0.799	0.276	—	—	—
Red Control	0.603	0.134	0.763	—	—
Red Non-XAI	0.086	**0.004**	0.156	0.203	—
Red XAI	0.134	**0.008**	0.216	0.324	0.777

TABLE V

BINOMIAL TEST FOR THE BETTING BEHAVIOR AMONG THE EXPERIMENT CONDITIONS

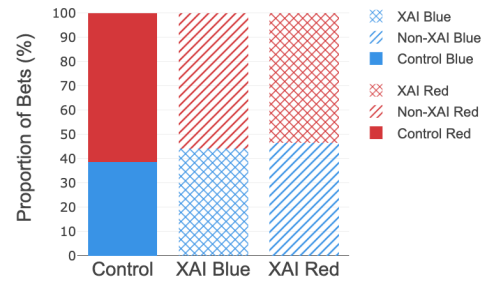| Condition	Bets Blue Robot	Bets Red Robot	p-value
Control	36	57	**0.038**
XAI Blue	37	47	0.326
XAI Red	46	53	0.547

Fig. 6. Betting behavior among the experiment conditions

For each experiment condition, a binomial test is used to estimate if either the red or blue robot was favored during the experiment. The findings are summarized in Table V and the betting behavior is illustrated in Figure 6. The results show that in the control group, the participants significantly preferred to bet on the red robot despite both robots behaving identically.

V. DISCUSSION

The evaluation of the Godspeed questionnaire shows that the XAI robots, which explain their moves to the participant, are perceived as more human-like and lifelike than

their non-explaining counterpart. Specifically, the behavior of explaining the reasoning behind a move in a board game was perceived as more human-like. Analogously, a significant increase of animacy ratings was observed. This is not surprising, since the XAI robot was communicating more than the non-XAI robot that only announced the moves during the experiment.

In contrast to the increase in these dimensions, the XAI robot was not perceived as more intelligent, likeable or safer than its counterpart. The explanation of the robot could have not been precise or convincing enough to affect the perception of the participants, in spite of the majority not having prior experience with the game UTTT. Furthermore, the provided game-play might have shown too few moves to convince the participants of the robot's capability. Instead of relying on the robot's utterances, the participants could have judged the game-play capability on the basis of single moves. In this case, participants would decide on the assumed better player quickly, without correcting their opinion after watching more game-play. The similarity in safety might arise due to both robots being identical except for one explaining the move, or due to the study being conducted online.

Surprisingly, the perceived competence questionnaire did not reveal a difference in competence between the XAI and non-XAI robots. Despite the XAI robot being perceived as more human-like and lively, the perception of the robot's competence was unaffected. This finding suggests that instead of a relationship among perceived competence and the anthropomorphism or animacy of the robot, there might be a link to perceived intelligence that was similarly unaffected by the XAI robot. A reason for not observing a difference in perceived competence could be that the effect of the manipulation by the robot explaining its move was too small. Alternatively, the participants might have analyzed only the board-state and game-play to estimate which of the robots will win the game to place their bet. This could have led to the suspicion that both robots have equal capabilities, which could result in a diminished effect of the manipulation.

After each short game-play of the robots, the participants had to decide which robot would win the game. This betting behavior is a direct measure of confidence in the robot's ability. However, the results show that the XAI robot was not preferably selected as the winner. For the control group, the participants favored the red robot despite both exhibiting the same behavior. This might have had an effect on the study, since it appears that overall the participants favored the red robot over the blue robot.

Preferring the red robot could be a result of assuming the gender of the red robot as female and the blue robot as male. While both robots had a female voice according to Google's Text-To-Speech service, the blue robot's voice was slightly lower. In conjunction with the blue color of the robot, this might have affected the participants' perception of the robot. For example, it has been shown that gendering a robot's voice affects the perception of the robot in terms of stereotypes, preference, and trust [70, 71]. Other potential sources for this bias are a general preference of one colour over the other (regardless of gender biases), the type of moves carried out by the robot in the snippets, as well as the side that the robot is sitting on.

The preference of the red robot is also reflected in the assessment of likeability, where the red non-XAI robot was perceived as significantly more likeable than the blue non-XAI robot. Remarkably, providing explanations seems to cancel this effect, leading the blue XAI robot to not be perceived as significantly less likeable than any of the red robots.

VI. Limitations

The results demonstrate that explanations for actions influence the perception of the robot, however, there are several limitations that need to be addressed in future work.

Since UTTT does not allow symmetric game-play, both robots have to make different moves. Therefore, the selected game sequence might affect the participant's betting behavior and perceived competence. A scenario that allows identical moves might be preferred for the evaluation of the influence of XAI in human-robot interaction.

Further, the study was conducted online, which allows for little control over the engagement of the participants. Participants could have been distracted during the study, consciously evaluated the objective of the study, or might have been confused by the wording in the study without the possibility to ask for clarification. Finally, a higher number of participants for each experiment condition could permit a more accurate estimation of the difference in perception between the robots.

VII. Conclusion

In this study, a human-robot interaction scenario was designed that explores the effect of robots that utilize XAI to explain their actions. Two robots played the board game UTTT, in which one robot explained the reasoning behind its moves, while the other robots simply announced the next move. The study was conducted online by presenting three short sequences of game-play. After each sequence, the participants had to bet on the winning robot. The experiment concluded with the assessment of several traits each participant assigned to the robots.

In our findings, we could not show that a robot that explains the motivation behind its actions increases the perceived competence. On the one hand, humans might still have limited trust in a robot's ability to perform a specific task despite their provided explanation. On the other hand, to manifest the concept of perceived competence in the participants mind might require more time and evidence of the robot's capability than was provided in the short sequences of game-play.

However, a robot that explained its moves was perceived as more human-like and lively than a robot that only announced the moves. This demonstrates that robots that provide reasoning about their actions influence human perception. Such methods could be used to increase trust in robots, especially

since robots and artificial intelligence are often perceived as black boxes. The observed effect might be more pronounced in a human-robot cooperative task.

All the software components of this study are implemented in robot simulation software, which allows the NICO robot to play UTTT autonomously and react to a player's moves. In future work, a human could play against the robot and the effects on the perception for an XAI and non-XAI opponent could be evaluated. This study illustrates the potential and effects of XAI in HRI, and demonstrates that a robot explaining its behavior can be perceived as more lively and human-like.

REFERENCES

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] J. L. Herlocker, J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pp. 241–250, 2000.

[3] O. Biran and C. Cotton, "Explanation and Justification in Machine Learning: A Survey," *IJCAI-17 workshop on explainable AI (XAI)*, pp. 8–13, 2017.

[4] A. Weidemann and N. Rußwinkel, "The Role of Frustration in Human–Robot Interaction – What Is Needed for a Successful Collaboration?," *Frontiers in Psychology*, vol. 12, no. March, pp. 1–17, 2021.

[5] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 109–116, 2016.

[6] T. Nomura and K. Kawakami, "Relationships between robots' self-disclosures and humans' anxiety toward robots," in *Proceedings - 2011 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT 2011*, vol. 3, pp. 66–69, 2011.

[7] S. Anjomshoae, D. Calvaresi, A. Najjar, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, vol. 2, pp. 1078–1088, 2019.

[8] U. Ehsan, P. Tambwekar, L. Chan, B. Harrison, and M. O. Riedl, "Automated rationale generation: A technique for explainable AI and its effects on human perceptions," in *International Conference on Intelligent User Interfaces, Proceedings IUI*, vol. Part F1476, pp. 263–274, 2019.

[9] G. Ciatto, D. Calvaresi, M. I. Schumacher, and A. Omicini, "An abstract framework for agent-based explanations in AI," in *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, vol. 2020-May, pp. 1816–1818, 2020.

[10] L. Sanneman and J. A. Shah, "Trust considerations for explainable robots: A human factors perspective," *arXiv preprint arXiv:2005.05940*, 2020.

[11] W. George and J. E. Janoski, "Group actions on winning games of super tic-tac-toe," *arXiv preprint arXiv:1606.04779*, 2016.

[12] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[13] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, pp. 351–379, apr 1975.

[14] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable Artificial Intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, pp. 1–21, 2021.

[15] W. R. Swartout, "XPLAIN: a system for creating and explaining expert consulting programs," *Artificial Intelligence*, vol. 21, pp. 285–325, sep 1983.

[16] B. Chandrasekaran and W. Swartout, "Explanations in Knowledge Systems: The Role of Explicit Representation of Design Knowledge," *IEEE Expert-Intelligent Systems and their Applications*, vol. 6, no. 3, pp. 47–49, 1991.

[17] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] A. Bundy, "Preparing for the future of Artificial Intelligence," *AI & SOCIETY*, vol. 32, no. 2, pp. 285–287, 2017.

[19] L. Veiber, K. Allix, Y. Arslan, T. F. Bissyandé, and J. Klein, "Challenges towards production-ready explainable machine learning," in *OpML 2020 - 2020 USENIX Conference on Operational Machine Learning*, pp. 3–5, 2020.

[20] R. Hamon, H. Junklewitz, and I. Sanchez, *Robustness and Explainability of Artificial Intelligence - From technical to policy solutions*. 2020.

[21] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in explainable ai," *arXiv preprint arXiv:1810.00184*, 2018.

[22] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[23] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable AI for robotics," *Science Robotics*, vol. 2, no. 6, 2017.

[24] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.

[25] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215, IEEE, 2018.

[26] W. Samek and K.-R. Müller, "Towards explainable artificial intelligence," in *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, Springer, 2019.

[27] T. Miller, P. Howe, and L. Sonenberg, "Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[28] T. Sakai and T. Nagai, "Explainable autonomous robots: A survey and perspective," *arXiv preprint arXiv:2105.02658*, 2021.

[29] F. Hussain, R. Hussain, and E. Hossain, "Explainable Artificial Intelligence (XAI): An Engineering Perspective," *arXiv preprint arXiv:2101.03613*, 2021.

[30] R. Setchi, M. B. Dehkordi, and J. S. Khan, "Explainable robotics in human-robot interactions," in *Procedia Computer Science*, vol. 176, pp. 3057–3066, Elsevier B.V., 2020.

[31] 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, 2017.

[32] T. Lombrozo, "The structure and function of explanations," *Trends in cognitive sciences*, vol. 10, no. 10, pp. 464–470, 2006.

[33] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. Chen, E. J. De Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, 2011.

[34] Z. R. Khavas, S. R. Ahmadzadeh, and P. Robinette, "Modeling trust in human-robot interaction: A survey," in *International Conference on Social Robotics*, pp. 529–541, Springer, 2020.

[35] 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.

[36] M. Natarajan and M. Gombolay, "Effects of anthropomorphism and accountability on trust in human robot interaction," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 33–42, 2020.

[37] S. S. Sebo, P. Krishnamurthi, and B. Scassellati, "'I Don't Believe You': Investigating the Effects of Robot Trust Violation and Repair," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2019-March, pp. 57–65, 2019.

[38] M. Desai, P. Kaniarasu, M. Medvedev, A. Steinfeld, and H. Yanco, "Impact of robot failures and feedback on real-time trust," in *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 251–258, 2013.

[39] S. Park, "Multifaceted trust in tourism service robots," *Annals of Tourism Research*, vol. 81, p. 102888, 2020.

[40] L. Christoforakos, A. Gallucci, T. Surmava-Große, D. Ullrich, and S. Diefenbach, "Can Robots Earn Our Trust the Same Way Humans Do? A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI," *Frontiers in Robotics and AI*, vol. 8, no. April, 2021.

[41] L. Tian and S. Oviatt, "A taxonomy of social errors in human-robot interaction," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 10, no. 2, pp. 1–32, 2021.

[42] M. M. Scheunemann, R. H. Cuijpers, and C. Salge, "Warmth and competence to predict human preference of robot behavior in physical human-robot interaction," in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1340–1347, IEEE, 2020.

[43] E. Cha, A. D. Dragan, and S. S. Srinivasa, "Perceived robot capability," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, vol. 2015-Novem, pp. 541–548, 2015.

[44] C. Canning, T. J. Donahue, and M. Scheutz, "Investigating human perceptions of robot capabilities in remote human-robot team tasks based on first-person robot video feeds," in *IEEE International Conference on Intelligent Robots and Systems*, pp. 4354–4361, 2014.

[45] P. Schermerhorn and M. Scheutz, "Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task," in *ACHI 2011 - 4th International Conference on Advances in Computer-Human Interactions*, no. c, pp. 236–241, 2011.

[46] S. Kiesler, A. Powers, S. R. Fussell, and C. Torrey, "Anthropomorphic interactions with a robot and robot-like agent," *Social Cognition*, vol. 26, no. 2, pp. 169–181, 2008.

[47] M. L. Walters, M. Lohse, M. Hanheide, B. Wrede, D. S. Syrdal, K. L. Koay, A. Green, H. Hüttenrauch, K. Dautenhahn, G. Sagerer, and K. Severinson-Eklundh, "Evaluating the robot personality and verbal behavior of domestic robots using video-based studies," *Advanced Robotics*, vol. 25, no. 18, pp. 2233–2254, 2011.

[48] H. Lehmann, A. Roncone, U. Pattacini, and G. Metta, "Physiologically inspired blinking behavior for a humanoid robot," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9979 LNAI, pp. 83–93, 2016.

[49] D. S. Syrdal, K. L. Koay, M. Gácsi, M. L. Walters, and K. Dautenhahn, "Video prototyping of dog-inspired non-verbal affective communication for an appearance constrained robot," in *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, no. May 2014, pp. 632–637, 2010.

[50] S. Woods, M. Walters, K. L. Koay, and K. Dautenhahn, "Comparing human robot interaction scenarios using live and video based methods: Towards a novel methodological approach," in *International Workshop on Advanced Motion Control, AMC*, vol. 2006, pp. 750–755, 2006.

[51] W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati, "The benefits of interactions with physically present robots over video-displayed agents," *International Journal of Social Robotics*, vol. 3, no. 1, pp. 41–52, 2011.

[52] S. H. Seo, D. Geiskkovitch, M. Nakane, C. King, and J. E. Young, "Poor Thing! Would You Feel Sorry for a Simulated Robot?: A comparison of empathy toward a physical and a simulated robot," in *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2015-March, pp. 125–132, 2015.

[53] C. Gittens, "Remote-hri: a pilot study to evaluate a methodology for performing hri research during the covid-19 pandemic," in *Proceedings of the 54th Hawaii International Conference on System Sciences*, p. 1878, 2021.

[54] M. Jung, M. J. S. Lazaro, and M. H. Yun, "Evaluation of methodologies and measures on the usability of social robots: A systematic review," *Applied Sciences*, vol. 11, no. 4, p. 1388, 2021.

[55] E. A. Boyle, T. M. Connolly, T. Hainey, and J. M. Boyle, "Engagement in digital entertainment games: A systematic review," *Computers in Human Behavior*, vol. 28, no. 3, pp. 771–780, 2012.

[56] G. Bertholon, R. Géraud-Stewart, A. Kugelmann, T. Lenoir, and D. Naccache, "At most 43 moves, at least 29: Optimal strategies and bounds for ultimate tic-tac-toe," *arXiv preprint arXiv:2006.02353*, 2020.

[57] I. Lakhmani and V. Punjabi, "A game based implementation of minimax algorithm using ai agents," *International Journal of Scientific Research and Engineering Trends*, 2020.

[58] M. Kerzel, E. Strahl, S. Magg, N. Navarro-Guerrero, S. Heinrich, and S. Wermter, "Nico—neuro-inspired companion: A developmental humanoid robot platform for multimodal interaction," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 113–120, IEEE, 2017.

[59] Stanford Artificial Intelligence Laboratory et al., "Robotic operating system."

[60] E. Rohmer, S. P. N. Singh, and M. Freese, "Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)*, 2013. www.coppeliarobotics.com.

[61] LimeSurvey Project Team / Carsten Schmitz, *LimeSurvey: An Open Source survey tool*. LimeSurvey Project, Hamburg, Germany, 2012.

[62] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[63] G. C. Williams, Z. R. Freedman, and E. L. Deci, "Supporting autonomy to motivate patients with diabetes for glucose control," *Diabetes care*, vol. 21, no. 10, pp. 1644–1651, 1998.

[64] G. C. Williams and E. L. Deci, "Internalization of biopsychosocial values by medical students: a test of self-determination theory.," *Journal of personality and social psychology*, vol. 70, no. 4, p. 767, 1996.

[65] R. M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being.," *American Psychologist*, vol. 55, no. 1, pp. 68–78, 2000.

[66] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[67] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Research in Science Education*, vol. 48, no. 6, pp. 1273–1296, 2018.

[68] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[69] G. D. Ruxton, "The unequal variance t-test is an underused alternative to student's t-test and the mann–whitney u test," *Behavioral Ecology*, vol. 17, no. 4, pp. 688–690, 2006.

[70] T. Nomura, "Robots and Gender," *Gender and the Genome*, vol. 1, no. 1, pp. 18–26, 2017.

[71] D. Gallimore, J. B. Lyons, T. Vo, S. Mahoney, and K. T. Wynne, "Trusting robocop: Gender-based effects on trust of an autonomous robot," *Frontiers in Psychology*, vol. 10, p. 482, 2019.

APPENDIX

A. Adjusted Perceived Competence Scale

1) "Make the right move":

- I feel confident in the $<blue|red>$ robot's ability to make the right move.
- The $<blue|red>$ robot is capable of making the right move. (The $<blue|red>$ robot has the possibility of doing this in the future.)
- The $<blue|red>$ robot is able to make the right move. (The $<blue|red>$ robot can currently do that.)
- I feel that the $<blue|red>$ robot is able to meet the challenge of choosing the right move.

2) "Win a game":

- I feel confident in the $<blue|red>$ robot's ability to win a game.
- The $<blue|red>$ robot is capable of winning a game. (The $<blue|red>$ robot has the possibility of doing this in the future.)
- The $<blue|red>$ robot is able to win a game. (The $<blue|red>$ robot can currently do that.)
- I feel that the $<blue|red>$ robot is able to meet the challenge of winning the game.