2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Ex(plainable) Machina: how social-implicit XAI affects complex human-robot teaming tasks

Marco Matarese[1,2], Francesca Cocchella[1,3,*], Francesco Rea[2] and Alessandra Sciutti[3]

***Abstract*— In this paper, we investigated how shared experience-based counterfactual explanations affected people's performance and robots' persuasiveness during a decision-making task in a social HRI context. We used the *Connect 4* game as a complex decision-making task where participants and the robot had to play as a team against the computer. We compared two strategies of explanation generation (*classical* vs *shared experience*-based) and investigated their differences in terms of team performance, the robot's persuasive power, and participants' perception of the robot and self. Our results showed that the two explanation strategies led to comparable performances. Moreover, *shared experience*-based explanations - based on the team's previous games - gave higher persuasiveness to the robot's suggestions than *classical* ones. Finally, we noted that low-performers tend to follow the robot more than high-performers, providing insights into the potential danger for non-expert users interacting with expert explainable robots.**

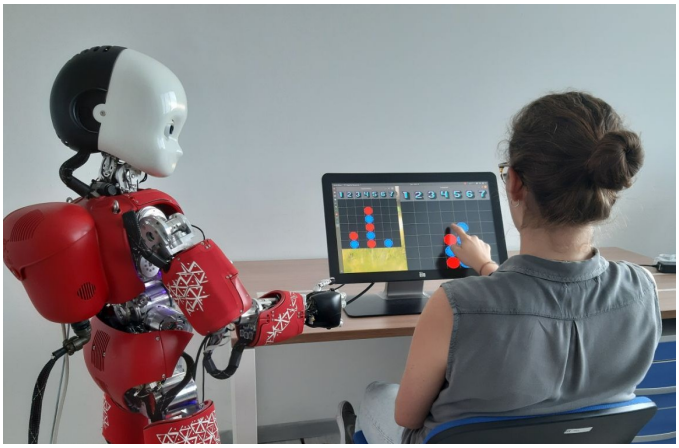

Fig. 1. Experimental setup.

I. INTRODUCTION

Since the '70s, artificial intelligence (AI) techniques have been increasingly used in our society. In particular, during the last two decades, machine learning (ML) systems are becoming growingly present in our daily lives, often without us noticing [37]. Such a growing expansion and impact on people's lives requires an effort to make these systems explainable to non-expert users to increase trust and acceptance [44], and encourage a conscious approach to the use of such technology [30][34].

Parallel to the growth of ML, great strides have been made in robotics and a strong interest in human-robot interaction (HRI) issues has developed [36]. When ML and robotics merge - *e.g.* when ML manages robots' behaviour or their interaction with humans - the need for explainable AI (XAI) becomes even more crucial because people often attribute human-like traits to robots [32], even the more complex characteristics, such as the 2nd-order theory of mind [25].

The HRI context is particularly suitable for a social and user-centred XAI because of two main reasons [28][40]. On the one hand, it has been shown that people easily adapt their interaction habits to robots [2]. On the other hand, we expect that robots will have long-term and personalised interactions with us [8]. In other words, the HRI field is pushing towards strong customisation of robots' behaviour; in our opinion, this translates into social and user-centred XAI [24].

[1]DIBRIS department, University of Genoa, Genoa, Italy. Corresponding email: marco.matarese@iit.it
[2]RBCS unit, Italian Institute of Technology, Genoa, Italy.
[3]CONTACT unit, Italian Institute of Technology, Genoa, Italy.
*This research and F. Cocchella has been supported by a Starting Grant from the European Research Council (ERC). G.A. No 804388, wHiSPER.

Although the number of studies investigating user-centred approaches to XAI has increased in the last years, we can find very few articles about the effects of such approaches in social HRI contexts [21][4]. Moreover, almost all of them are placed in the human-computer interaction field [1]. In particular, they consist of comparisons between different XAI techniques with human-AI teams performing decision-making tasks and investigate the persuasiveness of such XAI techniques and their impact on people's trust and perception of the underlying AI system [31].

In this paper, we propose a comparison between two XAI approaches in a social HRI scenario: we called them *classical* and *shared experience*-based. We set the HRI as a social decision-making task in which participants and the robot had to collaborate to beat the computer at the *Connect 4* game (Figure 1). We used counterfactual explanations because of the complexity of the ML model needed to solve the game and their theoretical validity [28]. The *shared experience*-based counterfactuals derived from the games that the robot and user have already experienced. We tested whether *shared experience*-based counterfactuals were more effective than *classical* ones in terms of performance, persuasive power and perception of both the robot and self.

II. RELATED WORK

The impact of different XAI techniques has been studied mainly in human-computer interaction (HCI) [15] for decision-making tasks [20]. Wang and Yin [47] highlighted three desiderata for XAI systems and compared features contribution and counterfactual explanations on such desiderata. They found that feature contribution explanations sat-

isfy more desiderata for expert users, while counterfactual explanations do not seem to improve trust. Similarly, Van der Waar *et al.* [43] compared rule- and example-based explanations with a decision support system. They showed that rule-based explanations slightly increase system understanding, while both explanation types seem to persuade more users than without explanations (even when the advice was incorrect). However, neither explanation types improve task performance compared to no explanations. Moreover, Lim *et al.* [22] compared the effects of why and why not-explanations on users' system understanding. They showed that why-explanations led to better understanding and trust towards the system than why not-explanations.

We can also find several approaches to customised XAI in the HCI literature. Millecamp *et al.* [26] showed that users' characteristics (*e.g.*, the need for cognition), influence users when interacting with explainable recommendation systems. However, in a follow-up study [27], the authors found instead that users' openness affects whether they would like to reuse the explanatory system. Similarly, Conati *et al.* [11] proved that providing explanations increases users' trust and perceived usefulness and provided insights on how to personalise explanations using users' personality traits. Furthermore, Tintarev and Mastoff [42] evaluated personalised explanations and found that those led to higher user satisfaction than non-personalised ones.

When moving to the human-robot interaction (HRI) context, we can find very few studies on XAI [35], especially if we are interested in investigating different XAI strategies or customisation and user-centredness. For example, Kaptein *et al.* [19] implemented a belief-desire-intention agent on a Nao robot and investigated what explanation style both children and adults prefer. They showed that adults tend to prefer goal-based explanations, while children do not show preferences between goal- and belief-based explanations.

Instead, several approaches have been proposed to explain robot planning. Chakraborti *et al.* [10] proposed to afford explainability as a reconciliation model using the Fetch robot. Their approach aims to progressively change the human's model to bring it closer to the robot's one, making the robot's plan optimal for such changes in the human's model. Sukkerd *et al.* [39] proposed an explainable planning representation to ease explanation generation and a method to generate contrastive explanations as policy justification. Finally, Devin and Alami [12] moved towards the direction of a user-aware XAI during shared plan execution.

On the other hand, there are several studies regarding XAI with virtual embodied robots [4]. Gong and Zong [16] proposed an approach to explaining robot behaviour as intention signalling using natural language sentences. Wang and Belardinelli [46] proposed using augmented reality to show XAI feedback and the robot's internal beliefs. Differently, Amir and Amir [3] developed an algorithm to summarise robots' behaviours by extracting information from the agents' simulations.

With this study, we want to move toward customisation and user-centeredness and contribute to the discussion about XAI in HRI. In our experiment, we set a complex human-robot teaming task (*Connect 4*) and investigated how people reacted to two different explanation strategies. The first strategy, called *classical*, used precise counterfactual explanations to help the robot justify its game choices. Instead, the *shared experience*-based strategy generated counterfactuals from the games that participants and the robot already played, thus, exploiting their past experience.

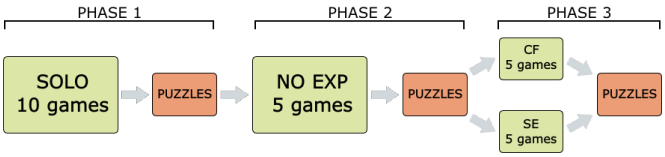

Fig. 2. Experimental design. All participants (22 total) performed the *solo* and *no exp* phases; then they were split into two groups (11 per group) and one faced the *CF* explanation phase, while the other the *SE* explanation phase.

III. METHODS

This study aimed to examine how different counterfactual generation approaches affect people's task performance, perception of the robot, and persuasive power in a social HRI. We designed the interaction as a competitive game in which the human-robot team plays against the computer at the *Connect 4* game.

We were also interested in studying a non-explanatory robot; thus, participants faced three phases which corresponded to the experimental conditions (Figure 2).

- *solo*: participants played alone against the computer.
- *no exp*: participants and iCub played together, but the robot produced no explanations.
- *exp*: participants and iCub played together, and iCub produced explanations. With half of the participants iCub produced classical explanations (*CF* group), for the other half *shared experience*-based explanations (*SE* group).

A. Procedure

Participants started the study with a pre-experiment questionnaire (Sec. III-E.3). After reaching the experimental room with iCub, we instructed them that they had to play *Connect 4* with iCub (as a team) against the computer. Then, we briefly recap the rules of the game[1], and we instructed them on how to use the touch-screen and application (Figure 1). Regarding iCub, we told them that it knew the game rules, and we were testing on it a new algorithm in a learning phase. The experiment was composed of three phases; during all of them, the human-robot team was always the first to play.

During the *solo* phase, iCub could not intervene in the games; it could only watch the games and comment on their results: it would say "oh no, we lost!" in case of a losing game, or "yeah, we won!" in case of a victory. iCub turned its head to look participants in the eyes each time it talked to them. Participants had to play 10 consecutive

[1]https://en.wikipedia.org/wiki/Connect_Four

games against the computer during such a phase. Right after the matches, they had to solve 20 puzzles. The puzzles were configurations of the *Connect 4* game in which participants had to play what they thought was the best move. The puzzles presented after the *solo* condition were predetermined and with increasing difficulty. Finally, participants had to fill out a short questionnaire (Sec. III-E.3).

During the *no exp* phase, iCub could participate in the decision-making process of the five games. The decision-making process was similar to the one used in [47]. Participants first indicated their choice, and then iCub told them whether it agreed with them. If not, it told the participants which column it would drop the fiche instead. At the end of this interaction, participants had the final decision and, finally, moved. Also in this phase participants had to solve 20 puzzles. However, rather than being predetermined as in the *solo* condition, the application took them from the matches of the current phase. In particular, the application randomly chose the puzzles among the configurations in which participants opted for a move different from iCub's. Right after the puzzles, we asked participants to fill out a short questionnaire (Sec. III-E.3).

The third phase (*CF* or *SE*) was similar to the second one but with explanations. iCub produced an explanation each time its choice differed from that of participants. Such explanations were displayed by showing the counterfactual on another smaller window next to the board game (Figure 1). iCub accompanied the explanations with an arm movement, indicating such a new window. As before, participants had to solve 20 puzzles (taken from the current phase's games) and fill out a questionnaire. Finally, we asked participants to complete a post-experiment questionnaire (Sec. III-E.3).

B. AI models

We equipped both iCub and the computer with ML models. From the computer's side, we chose a Monte Carlo Tree Search (MC-TS) algorithm [9]. By using such a model, we could perform satisfactorily and adjust its depth to accommodate the difficulties of playing against it. We set the search depth we used during the experiments with a preliminary pilot study in which we asked our colleagues (10 total) to play alone against the computer. We chose the MC-TS model's depth and time limit (sim. number = 10, max iter. = 2000, timeout = 2) which allowed our colleagues to win $\simeq 50\%$ of the time.

On the other hand, we made iCub play via a deep neural network (DNN) to obtain semi-optimal performance against the computer. We chose the AlphaZero architecture [38] trained to play the *Connect 4* game because it could easily reach perfect performance against the MC-TS agent. We fine-tuned the model to reach 90% of victories against the computer's MC-TS model because we were interested in having a semi-perfect iCub.

C. XAI model: counterfactual generation

The complexity of the game required a complex model. Typically, the more complex is the AI model, the less transparent and explainable [13][7]. We chose counterfactual explanations to obtain a good tradeoff between explainability and performance since we could not renounce to AlphaZero's DNN complexity. In particular, we used example-based counterfactuals: when iCub did not agree with participants' first choice, it showed a board configuration - the counterfactual - to justify its move. The chosen counterfactual had two peculiar properties: (1) it was similar to the current configuration of the game but (2) different enough to induce iCub to take the participant's initial choice.

Although there are several methods to produce counterfactual explanations from a DNN (*e.g.*, DiCe [23]), we preferred to build them manually because of the impossibility of post-hoc filtering. We needed it because we needed counterfactuals showing legal configurations; in other words, configurations of the game that could result from a legal match. Indeed, current counterfactual generation techniques assume that all the input features are independent. This is not the case for *Connect 4* since the value of a board's slot strictly depends on the value of the slots directly under it.

Thus, we decided to use counterfactuals that would satisfy our requirements. By letting the DNN play against the computer, we collected a dataset of more than 11000 configurations (without duplicates) with the move the DNN would take in each of them. Consequently, we organised these configurations depending on such moves.

Algorithm 1 Counterfactual generation

Require: $1 \leq fact, foil \leq 7$; $fact \neq foil$
$counterFacts \leftarrow counterFactsDataset[foil]$
sort $counterFacts$ w.r.t. $fact$ using L_1 norm
$counterFact \leftarrow counterFacts.top()$
return $counterFact$

Algorithm 1 illustrates how we retrieved the counterfactuals from the dataset, where the iCub's move is the *fact* and the user's one is the *foil* [28]. We used the L_1 norm as a measure of similarity [23][29], which ensures good properties for our needs [45]. Through this ordering, we ensured that the retrieved counterfactual was the configuration most similar to the current one among those in which iCub would make the user's move.

On the one hand, we used the aforementioned dataset to implement the *classic* approach. On the other hand, we collected the counterfactual dataset from the matches participants played during the experiment to implement the *shared experience*-based one. In particular, we saved into the dataset the configurations in which participants moved differently from what iCub would do. Those datasets had an average size of $\mu = 198.1$ ($\sigma = 18.5$) configurations. This led to two main differences between the counterfactual types: (1) the mathematical precision (L_1 norm) was higher for *classical* counterfactuals than for *shared experience*-based one because of the large difference between the two datasets' sizes; (2) there was a high probability that the former had not previously been encountered by the participants, whereas they had already encountered all of the latter.

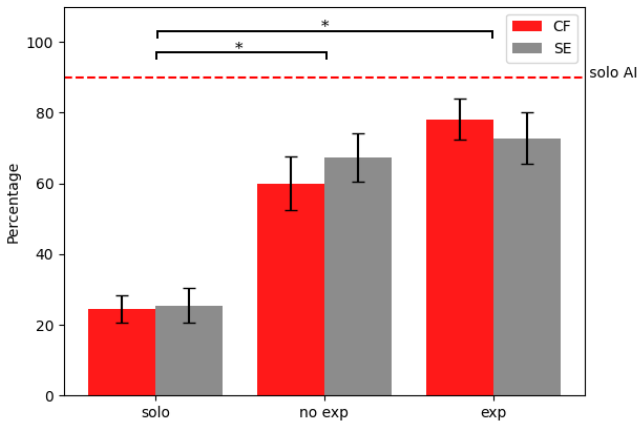

Fig. 3. Mean and standard error of participants' performance of the two groups (CF vs SE) in each condition. The red dashed line refers to the robot's AI performance.

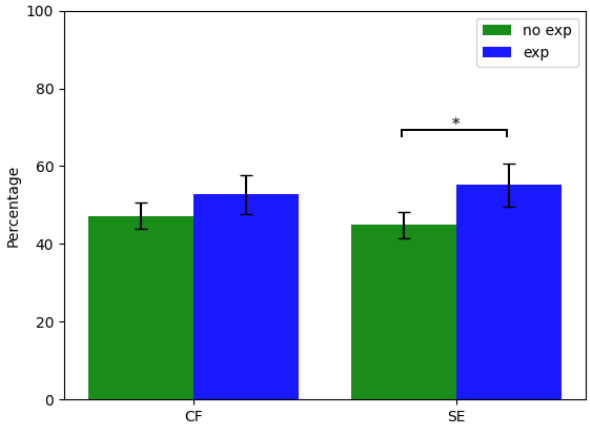

Fig. 4. Mean and standard error of the percentages of *equal* moves of the two groups (CF vs SE) divided by experimental condition (with vs without explanations). Here we consider only the *equal* moves.

D. Participants

We recruited 22 participants, and all signed the informed consent before the experiment, approved by the ethical committee "Regione Liguria". 19 participants were 20-30 years old, with the remaining participants ranging from 30-45; 14 identified as women, 7 as men, and 1 preferred not to specify.

E. Measures

We split the experimental measures into three macro-groups: performance, persuasive power, and perception of the self and robot.

1) Performance: we measured participants' performance during the games and puzzles after the *solo* condition.

2) Persuasive power: we measured whether participants confirmed their initial choice or changed in favour of iCub's one during phases 2 and 3. Moreover, through the subsequent puzzle phases, we measured how much participants unconsciously "absorbed" iCub's suggestions, *e.g.*, implicitly learned from previous iCub's suggestions and then replicated its choices during the puzzles.

3) Perception of the robot and self: before the experiment, we asked participants to indicate their level of accordance on 7 points Likert Scale with several items aiming to collect information about their impression of the robot. We also submitted to them the Sense of Agency scale [41], and asked them whether they had already seen or interacted with the robot iCub. Finally, we showed participants an institutional video of iCub[2] and asked them to evaluate the robot by answering 9 items about iCub's warmth and competence [14], and the Inclusion of Other into the Self (IOS) test [5]. Moreover, between the experimental phases, we asked participants to rate the game's difficulty, their level of expertise and iCub's one. After the third session, we submitted to the participants 6 items on their satisfaction with the explanations [11].

[2]https://www.youtube.com/watch?v=3N1oCMwtz8w.

IV. RESULTS

A. Performance

We measured the game performance as the percentage of games won against the computer (Section III-B). Figure 3 summarises the average task performances for each condition between the two groups (CF vs SE). With a two-way mixed-model ANOVA analysis, we determined that there was a significant difference between the performances of the groups ($F(4, 104) = 39.72$, $p < .001$). A post-hoc test with Bonferroni correction showed that solo AI performance significantly differed from all the groups in all conditions ($p < .001$ for all such comparisons). Similarly, the *solo* (human) performance was statistically different from those of all the groups in all conditions ($p < .001$ for all such comparisons).

The number of puzzles correctly solved after the solo condition was comparable between the two groups (CF vs SE). Indeed, participants in the CF group solved $\mu = 14.8$ puzzles ($\sigma_{\bar{x}} = .85$), and those in the SE group solved $\mu = 14.27$ puzzles ($\sigma_{\bar{x}} = .59$).

B. Persuasive power

To measure iCub's persuasive power, we collected three types of moves:

- *equal*: participants chose since the beginning the move that iCub would have chosen (*e.g.*, participants chose and made move 2 when also iCub would have chosen it).
- *follow self*: *e.g.*, participants chose move 2, iCub chose the 3, and participants made move 2.
- *follow iCub*: *e.g.*, participants chose move 2, iCub chose the 3, and participants made move 3.

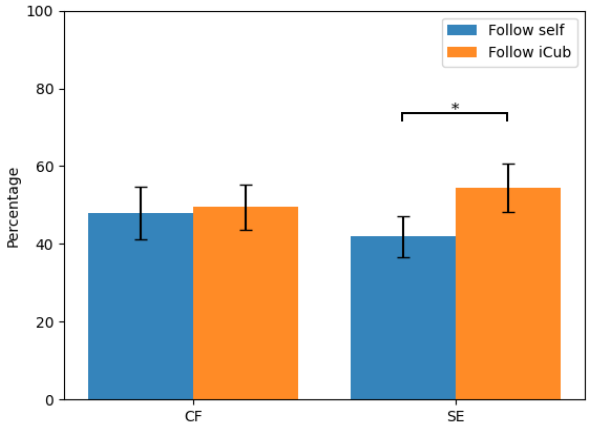

Fig. 5. Mean and standard error of the percentages of move types (follow self vs follow iCub) of the two groups (CF vs SE). Here we do not consider other types of move.

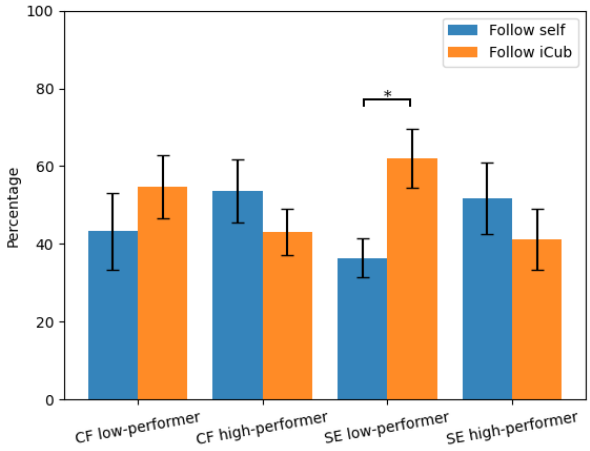

Fig. 6. Mean and standard error of the percentages of the move types of participants divided by performance in *solo* condition. Low-performers won less than the group average during the *solo* condition, while high-performers won more than the average of the group of participants. Here we do now consider other types of move.

All the other move types' percentages were negligible. We averaged the percentages of such move types over the games because we found no particular patterns depending on the games' numbers.

Figure 4 shows the average percentage of participants' *equal* moves between the conditions with and without explanations. A two-way mixed-model ANOVA test revealed a significant difference between the conditions ($F(2, 217) = 7.06$ $p = .001$). A post-hoc test with Bonferroni correction showed a statistical difference between the conditions *no exp* and *exp* only for the SE group ($p = .03$).

Figure 5 shows the average percentage of participants' move types during the *exp* condition for both groups (CF vs SE). A two-way mixed-model ANOVA test showed that there was a significant interaction between the conditions and types of moves ($F(2, 324) = 3.84$, $p = .022$). A post-hoc test with Bonferroni correction showed that, the percentage of *follow self* moves is significantly lower than the *follow iCub* ones ($p = .009$) for the SE group.

Figure 6 shows the average percentages of participants' move types divided by performance during the *solo* condition. We considered *low-performers* participants who won in *solo* condition less than the average of the entire group; the remaining were considered as *high-performers*. In particular, participants resulted divided into: *CF low-performers* (6 total), *CF high-performers* (5 total), *SE low-performers* (7 total) and *SE high-performers* (4 total).

A two-way mixed-model ANOVA test revealed that there was a significant interaction between the conditions and types of moves ($F(3, 212) = 6.02$, $p < .001$). A post-hoc test with Bonferroni correction showed that only for the SE group, between low-performers, the percentage of *follow self* moves was significantly less than the *follow iCub* ones ($p < .001$).

Figure 7 shows the average number of move types made during the puzzle phases divided by participants' tendency to follow iCub suggestions during the main game. Participants in the low-follower category followed iCub less than 50% of the time; otherwise, we considered them high-followers. The resulting division was: *CF low-followers* (5 total), *CF high-followers* (6 total), *SE low-followers* (5 total) and *SE high-followers* (6 total).

A two-way mixed-model ANOVA test showed that there was an interaction between the conditions and types of moves ($F(3, 36) = 21.49$, $p = .013$). A post-hoc test with Bonferroni correction showed that only for the SE group, low-followers reproduced their moves more than followed iCub's in-game suggestions ($p = .013$).

C. Perception of the robot and self

We discarded 2 participants from the analyses of the questionnaires because they failed to reply to attention checks, which are designed to quickly measure respondents' engagement. 90% of them declared to have already participated in a study with iCub, and all the items we submitted to the participants reached a good Cronbach's Alpha ($> .65$), indicating that the scales were all reliable.

We found no differences between the groups in the items regarding the robot's warmth, competence and human-likeness. Similarly, participants' level of satisfaction with the explanations did not present differences between the groups. A repeated measures mixed-model ANOVA test showed that the answers to the IOS test were significantly higher in the post-experiment questionnaire than in the pre-experiment one ($F(1, 19) = 6.72$, $p = 0.018$). Furthermore, participants' perceived level of goodness in playing the game grew through the phases ($F(2, 36) = 12.06$, $p < .001$). In this regard, a Bonferroni post-hoc correction showed a significant difference between phases 2 and 3 ($p = .007$), and 1 and 3 ($p < .001$). Likewise, participants' perceived difficulty of the game grew through the experimental phases ($F(2, 36) = 9.39$, $p < .001$). A post-hoc test with Bonferroni

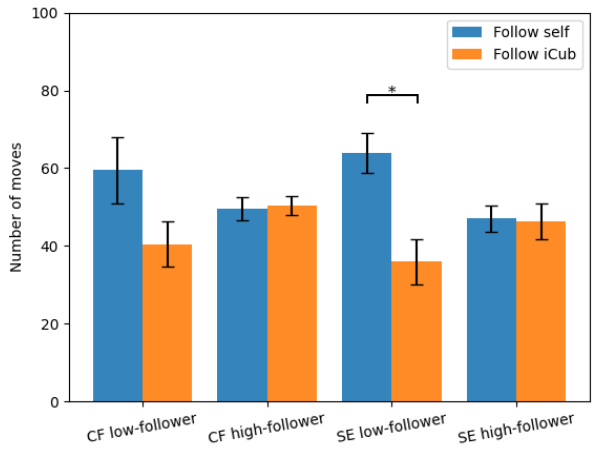

Fig. 7. Mean and standard error of the number of move types of the participants during the puzzle phases divided by their tendency to follow iCub's suggestion during the main game. All puzzle phases included 20 puzzles. Low-followers followed less than 50% of iCub's suggestions, while high-followers followed more than 50% of those. Here we do not consider other types of move.

correction showed significance only between phases 1 and 3 ($p = .007$).

V. DISCUSSION

Both game and puzzle performance in phase 1 showed that the two groups of participants were homogeneously skilled, allowing a fair comparison between the two groups (Section IV-A). Participants' performance is similar to that reported in the human-computer interaction literature regarding human-AI teams in decision-making tasks. Specifically, participants alone performed worse than AI (25% vs 90%) and, with the robot, they did not reach the AI performance, as in [6]. We speculate that the reason for this effect has to be found in how the game develops, in addition to the difficulty for participants to identify iCub's failures [33]. In fact, the first moves are the more important ones and failing some of those could seriously compromise the entire game. Moreover, participants performed worse without explanations than with them (independently of the explanation type), but the difference in performance is not statistically significant, as also shown by [6]. Finally, the two counterfactual generation strategies brought similar performance.

Differently from other studies [43], we found no significant differences in the robot's persuasiveness between the condition without explanations and the two explainable ones (Section IV-B). Participants tended to confirm themselves more (and follow iCub less) with *classical* explanations (CF group) than with *shared experience*-based ones (SE group). However, participants in the SE group followed iCub more than confirmed their moves, whereas we did not find this difference in the CF group. Thus, shared experience-based explanations brought a higher persuasiveness than classical ones. This could mean that presenting explanations from a common ground gives higher reliability to the robot, making it easier for people to follow it.

Interestingly, this effect was even more relevant for those that performed less than the group average without the robot's help. The robot's higher persuasiveness with low-performer participants gives us insights into the potential danger of letting non-expert users interact with expert explainable robots [43]. Several potential dangers have already been pointed out, such as the difficulty for people to detect AI errors [33], and the biases that guide their interaction with AI systems [17]. We can easily transpose these to our future domestic robots; hence, we still need a profound discussion about the ethical implication of using such technologies in everyday life [18].

In both conditions, as they played with iCub, participants adapted their way of playing to that of the robot. Indeed, if we look at how the percentage of *equal* moves changes through the experimental phase, we can see that it grows between the condition without explanation and the two consequent explanatory ones. However, only for the SE group such a difference is significant. It might be that the reason for such results is in the higher persuasiveness outlined before.

Puzzle phases represented a precise way to measure participants' implicit learning of iCub's strategies. In the SE group, those who were *low-followers* during the main game tended to confirm their moves more during the puzzles than solving puzzles the way iCub would. This tendency, although not significant, is also true for the CF group. This result suggests that it was difficult for low-followers to "absorb" iCub's choices because they did not follow them; as a result, they could not replicate such moves during the puzzles.

Participants rated their closeness to the robot higher after the experiment than before, regardless of the group (Section IV-C). Thus, participants perceived a social interaction and enjoyed it and, if we also consider how they perceived iCub as part of their group ($\mu = 4.68 \pm \sigma = 1.42$ on 7 points Likert scale), we can say that they considered iCub as a teammate, although they were focused on the task.

VI. CONCLUSIONS

In this study, we compared two counterfactual generation strategies in a social HRI scenario. In particular, we set the HRI as a complex human-robot decision-making task where participants played the *Connect 4* game with the iCub robot (as a team) against the computer. Alongside a classical approach, we tested how explanations generated from the human-robot common ground affect the robot's persuasiveness, people's performance, and perception of both the robot and self.

Our results showed that *shared experience*-based explanations could lead to higher persuasiveness than *classical* ones. However, the two explanation strategies maintained comparable team performance reflecting the results of human-AI teams in HCI. We found that low-performer participants followed the robot more than high-performer ones: this evidence gives us insights regarding the potential danger for non-expert users interacting with expert explainable robots.

REFERENCES

[1] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–18, New York, NY, USA, 2018. Association for Computing Machinery.

[2] M. Ahmad, O. Mubin, and J. Orlando. A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction*, 1(3), 2017.

[3] D. Amir and O. Amir. Highlights: Summarizing agent behavior to people. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 1168–1176, Richland, SC, 2018. International Foundation for Autonomous Agents and Multiagent Systems.

[4] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling. Explainable agents and robots: Results from a systematic literature review. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, page 1078–1088, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

[5] A. Aron, E. N. Aron, and D. Smollan. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4):596, 1992.

[6] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[7] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 2020.

[8] G. Belgiovine, J. Gonzalez-Billandon, G. Sandini, F. Rea, and A. Sciutti. Towards an hri tutoring framework for long-term personalization and real-time adaptation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '22 Adjunct, page 139–145, New York, NY, USA, 2022. Association for Computing Machinery.

[9] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.

[10] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 156–163. AAAI Press, 2017.

[11] C. Conati, O. Barral, V. Putnam, and L. Rieger. Toward personalized xai: A case study in intelligent tutoring systems. *Artificial Intelligence*, 298, 2021.

[12] S. Devin and R. Alami. An implemented theory of mind to improve human-robot shared plans execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 319–326, 2016.

[13] 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)*, pages 0210–0215, 2018.

[14] S. T. Fiske, A. J. Cuddy, and P. Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

[15] A. Gambino and B. Liu. Considering the context to build theory in hci, hri, and hmc: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication*, 4:111–130, 2022.

[16] Z. Gong and Y. Zhang. Behavior explanation as intention signaling in human-robot teaming. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1005–1011, 2018.

[17] B. Green and Y. Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 90–99, New York, NY, USA, 2019. Association for Computing Machinery.

[18] B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Procedeeings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019.

[19] F. Kaptein, J. Broekens, K. Hindriks, and M. Neerincx. Personalised self-explanation by robots: The role of goals versus beliefs in robot-action explanation for children and adults. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 676–682, 2017.

[20] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[21] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[22] B. Y. Lim, A. K. Dey, and D. Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 2119–2128, New York, NY, USA, 2009. Association for Computing Machinery.

[23] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777. Curran Associates Inc., 2017.

[24] M. Matarese, F. Rea, and A. Sciutti. A user-centred framework for explainable artificial intelligence in human-robot interaction. *arXiv preprint arXiv:2109.12912*, 2021.

[25] M. Matarese, F. Rea, and A. Sciutti. Perception is only real when shared: A mathematical model for collaborative shared perception in human-robot interaction. *Frontiers in Robotics and AI*, page 166, 2022.

[26] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 397–407, New York, NY, USA, 2019. Association for Computing Machinery.

[27] M. Millecamp, N. N. Htun, C. Conati, and K. Verbert. What's in a user? towards personalising transparency for music recommender interfaces. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '20, page 173–182, New York, NY, USA, 2020. Association for Computing Machinery.

[28] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[29] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 607–617. Association for Computing Machinery, 2020.

[30] B. Nesset, D. A. Robb, J. Lopes, and H. Hastie. Transparency in hri: Trust and decision making in the face of robot errors. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 313–317. Association for Computing Machinery, 2021.

[31] R. Paleja, M. Ghuy, N. Ranawaka Arachchige, R. Jensen, and M. Gombolay. The utility of explainable ai in ad hoc human-machine teaming. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 610–623. Curran Associates, Inc., 2021.

[32] E. Phillips, X. Zhao, D. Ullman, and B. F. Malle. What is human-like? decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 105–113, New York, NY, USA, 2018. Association for Computing Machinery.

[33] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[34] T. L. Sanders, T. Wixon, K. E. Schafer, J. Y. C. Chen, and P. A. Hancock. The influence of modality and transparency on trust in human-robot interaction. In *2014 IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, pages 156–159, 2014.

[35] R. Setchi, M. B. Dehkordi, and J. S. Khan. Explainable robotics

in human-robot interactions. *Procedia Computer Science*, 176:3057–3066, 2020.

[36] T. B. Sheridan. Human–robot interaction: status and challenges. *Human factors*, 58(4):525–532, 2016.

[37] P. P. Shinde and S. Shah. A review of machine learning and deep learning applications. In *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pages 1–6, 2018.

[38] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[39] R. Sukkerd, R. Simmons, and D. Garlan. Towards explainable multi-objective probabilistic planning. In *Proceedings of the 4th International Workshop on Software Engineering for Smart Cyber-Physical Systems*, SEsCPS '18, page 19–25, New York, NY, USA, 2018. Association for Computing Machinery.

[40] A. Tabrez and B. Hayes. Improving human-robot interaction through explainable reinforcement learning. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 751–753, 2019.

[41] A. Tapal, E. Oren, R. Dar, and B. Eitam. The sense of agency scale: A measure of consciously perceived control over one's mind, body, and the immediate environment. *Frontiers in psychology*, 8:1552, 2017.

[42] N. Tintarev and J. Masthoff. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4):399–439, 2012.

[43] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291, 2021.

[44] G. Vilone and L. Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[45] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31:841, 2017.

[46] C. Wang and A. Belardinelli. Investigating explainable human-robot interaction with augmented reality. In *5th International Workshop on Virtual, Augmented, and Mixed Reality for HRI*, 2022.

[47] X. Wang and M. Yin. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*, IUI '21, page 318–328, New York, NY, USA, 2021. Association for Computing Machinery.