Questioning Wizard of Oz: Effects of Revealing the Wizard behind the Robot

Jauwairia Nasir[1], Pierre Oppliger[1], Barbara Bruno[1], and Pierre Dillenbourg[1]

***Abstract*— Wizard of Oz, a very commonly employed technique in human-robot interaction, faces the criticism of being deceptive as the humans interacting with the robot are told, if at all, only at the end of their interaction that there was in fact a human behind the robot. What if the robot reveals the wizard behind itself very early in the interaction? We built a deep wizard of Oz setup to allow for a robot to play together with a human against a computer AI in the context of Connect 4 game. This cooperative game interaction against a common opponent is then followed by a conversation between the human and the robot. We conducted an exploratory user study with 29 adults with three conditions where the robot reveals the wizard, lies about the wizard, and does not say anything, respectively. We also split the data based on how the participants perceive the robot in terms of autonomy. Using different metrics, we evaluate how the users interact with and perceive the robot in both the experimental and perceived conditions. We find that while there is indeed a significant difference in the participants willingness to follow robots suggestions between the experimental conditions as well as in the effort they put to prove themselves as humans (reverse Turing test), there isn't any significant difference in their robot perception. Additionally, how humans perceive whether the robot is tele-operated or autonomous seems to be indifferent to the robot revealing its identity, i.e., the pre-conceived notions may be uninfluenced even if the robot explicitly states otherwise. Lastly, interestingly in the perception based conditions, absence of statistical significance may suggest that, in certain contexts, wizard of oz may not require hiding the wizard after all.**

***Keywords*—human-robot interaction, Wizard of Oz, robot perception, social robots**

I. Introduction

It has been a human fascination to build robots that are capable of interacting autonomously and naturally with their surroundings. However, there are still advancements needed in the fields of artificial intelligence, natural language processing and computer vision, among others, in order to build such robots. To fill in these technical gaps, one of the most widely used technique in human-robot interaction is the *Wizard of Oz* (WoZ) [1], first introduced by [2], where a human controls the robot remotely. This remote control can vary in terms of the *number of aspects* that are being controlled (speech, movement, expressions, gestures, etc.) as well as along the *spectrum of autonomy* in each of the aspects.

While this technique gives us the opportunity for witnessing ahead of time what fully autonomous interactions could look like, or how humans would behave with a particular hypothetical "autonomous" robot, WoZ attracts criticism on various grounds, including the validity of its implicit assumption that findings obtained in a human-human interaction via a robot also hold true for human-robot interactions [3]. Additionally, some researchers have pointed out that "relying on WoZ as an experimental technique can make it all the more difficult to build robots capable of successfully mitigating errors on their own in the future [4]" [1]. Lastly, WoZ raises an ethical concern due to its use of social deception, i.e., the fact that human participants often find out only at the end of the interaction, if at all, that the robot was controlled by a human [5], [6], [7]. This article aims to tackle this last issue, starting from a simple question: what if the robot reveals, at the beginning of the interaction, that it is being controlled by a human? Would this change the interaction? Indeed, while the use of social deception in WoZ is motivated by the assumption that human participants would be biased in their interaction, to the best of our knowledge, this assumption has not yet been investigated explicitly.

To investigate this question, we designed a simple human-robot interaction scenario, in which the robot and the participant cooperate to beat a computer AI at the classic *Connect4* game. During the game, the robot provides suggestions to the human about the next best move, while at the end of the game the robot engages the participant in a short conversation. In a study with 29 participants, we compare (i) the case in which the robot reveals to be controlled by a human, (ii) the case in which it declares to be fully autonomous and (iii) the case in which it says nothing, specifically seeking to explore:

Research Question: How does revealing the presence of a person controlling the robot impact the participants' behavior towards the robot, as well as their perception of the robot?

II. Related Work

Over the years, WoZ techniques have been employed in various *contexts* within human-robot interaction ranging from sociality, home environments, service robotics, to assistive technology [8], [9], [10], [11]. Elaborating more on each study, in [8], the authors propose and describe a method to generate patterns for sociality in human-robot interaction, using the WoZ technique to create compelling social situations between a robot and children and adolescents by specifically teleoperating the speech and gestures. In [9], WoZ is employed to explore the concept of socially intelligent dialog systems in the home environment, i.e., controlling verbal interaction, while [10] describes this technique in the context of a service robot, particularly for the dialogue and

[1]Computer-Human Interaction in Learning and Instruction (CHILI) Lab, Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 765955.

navigation abilities, and discusses the simulation tools used for giving the Wizard the possibility of easily controlling such capabilities. In [11], WoZ is used to drive a small humanoid robot, particularly it's physical movements, in a study that evaluates the effect of robots appearance on facilitating interactions between autistic children and the robot. Indeed, the review done by [1], revealed that among the studies in HRI employing WoZ published between 2003-2011, the most frequent *types* of Wizard control include natural language processing, non-verbal behavior, and navigation and mobility, as can also be observed in the aforementioned studies. Lastly, among the *variations* of WoZ that have been proposed in the literature, "Oz of Wizard" proposes to simulate humans to evaluate robot behaviors [12]. Further, while not similar to our research question, in [13], where one group of children interact with a teleoperated robot and another with an autonomous robot, after revealing the presence of the wizard at the end of the interaction for the WoZ group, the participants were asked to fill a perception questionnaire again. They found that the relevant group decreased their perception of the robot's intelligence after finding out the truth. In short, there are those who use Woz, there are those who criticize WoZ as pointed out in Section I, but it doesn't seem there is much work yet that simply questions WoZ. To reiterate, while the technique itself is very popular in HRI, to the best of our knowledge, there does not seem to be a study on the effect of revealing the human presence behind the robot early on in the interaction, i.e., on the need of deceiving the participant.

In most HRI studies where one of the metrics under investigation relates with the degree to which a participant follows the suggestions of the robot, the participant has little to no idea of the task at hand. In other words, a task that is completely or partially outside the knowledge space of the participants is usually chosen, to control for the effect that the participants' prior knowledge could have on their trust in the robot. In [14] and [15], respectively, the acceptability of a robot's recommendations, based on its explicit or implicit communications style and the participants' cultural background, is tested in a scenario where the participants have to assign a price to a given product and in another context that requires the participants to make decisions regarding an on-campus environment-friendly chicken cooperative/hen house while collaborating with the robot that was presented to have relevant expertise. Further, in [16] where the authors compare two advice-giving strategies presented in videos with human and robot helpers, the advice is being given to a novice making cupcakes. Lastly, in another scenario of completing or creating new recipes with given ingredients, the robot tries to help the participants to complete the task by giving its suggestions at various scales of proactivity [17]. In our study, we control for the participants' prior knowledge in the opposite way, by asking them to engage in a well-known, very simple game. This choice is motivated by the fact that in a context that is unknown, participants might follow the robot because they may not have any opinion of their own especially at the beginning of the interaction; however, in a context that is known, this kind of reasoning behind accepting robots suggestions can hopefully be mitigated.

III. Method

A. Activity Design

1) Connect 4: Connect 4 is a grid based game (with 6 rows and 7 columns), as shown in Figure. 1, where two players take turns to put colored (red or yellow) tokens, one at a time. The red player always starts. Playing a token in a certain column means it will fall from the top, stopping at the highest empty slot along that column. A player wins when they have 4 tokens of their color lined up vertically, horizontally, or diagonally. While there are many variations of this game, in our study, we stick to the classic version.

The game, while seemingly easy, actually has 4,531,985,219,092 possible tokens configurations. A player, even an experienced one, cannot predict all of them and, just like in chess, the computation capabilities of our human intelligence are surpassed by those of computers with enough computing power, that can predict all the moves in advance thus guaranteeing the choice of the best move at each turn. Furthermore, by design, the first player can determine the outcome of the game. If both players play perfectly, the first player can secure the central column, and surely win the game. In our study, the participant and the robot play as a team that controls the yellow player, while a computer AI controls the red player.

For the purposes of our study, we implemented an online version of the game as well as a Connect 4 probability AI, based on Monte Carlo tree search (MCTS), that can predict the probability of winning the round associated with a move. Additionally and mainly, we used another Connect 4 solver AI, an open-source turn resolution engine [18], to compute the best next move from any game configuration. Please note that both the computer AI controlling the red player as well as the wizard that controls the robot make use of this latter AI to pick a move, while the AI that provides the probability is only used by the robot via the wizard. To elaborate a bit more on how the robot uses the two AIs, the information about the best move is used in the conversation that takes place between the human participant and the robot, when the robot tries to suggest a move. The associated probability of the move is randomly used at times by the robot to motivate its suggestion.

2) Simplified reverse Turing test: As part of the conversation that takes place between the robot and the participant, after the game, the robot follows a simplified reverse Turing test [19] where the participant has to prove to the robot that they are a human. While there is no right or wrong answer to this question, we argue that the effort one puts in proving themselves as a human to another human versus a machine may be different.

B. Setup Design: Deep Wizard of Oz

Our setup design, as shown in Figure 2, consists of a yellow team including the human participant assisted by the robot, that play against the red player that is a computer AI.

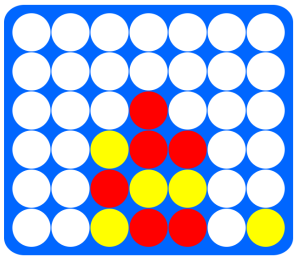

Fig. 1: An example instance of the game. It is the turn of the yellow player. The best next move is in the column 5 (from the left) and the second-best move is in the column 7. It is practically impossible for a human to judge these options correctly, since they will have impact in, at worst, the 13th turn from this configuration.

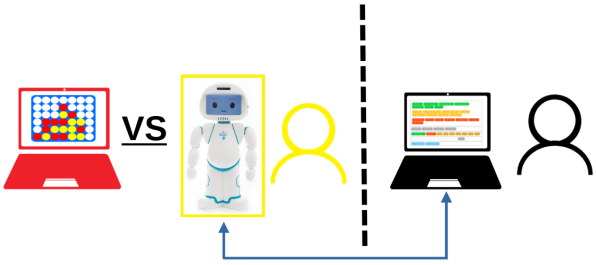

Fig. 2: Experimental setup: the human participant and the QTrobot are the yellow team, playing against the red player that is a Computer AI. Both the computer AI as well as the robot, that is remotely controlled by a human operator, make use of an opensource AI, to make or suggest the next best move, respectively.

The computer AI makes the next move based on *AI Connect 4 solver* shown in Figure 3. On the other hand, the robot, QTrobot from LuxAI[1], is remotely controlled by a human operator who, in turn, also gets its guidance on the best move by the *AI Connect 4 solver* and additionally information on the probability of winning the round associated with a move from the *AI Connect 4 probability*. With these layers of control moving from an AI to a human operator to a robot to a human participant, as shown in Figure 3, we believe that this setup pushes the boundaries of the *classical* Wizard of Oz method where a human operator controls the robot interacting with the human participant; hence, we term this setup as *deep Wizard of Oz*.

Lastly, to control the robot, we developed a web-based Wizard of Oz interface, shown in Figure 4, allowing the Wizard to quickly send pre-fabricated as well as on demand instructions, comments or answers. Easy to navigate colored and descriptive buttons helped to provide efficient generation of relevant emotions and gestures.

C. User Study Design

In our user study, the robot is manipulated in three different ways, corresponding to the three conditions that

[1]https://luxai.com/humanoid-social-robot-for-research-and-teaching/

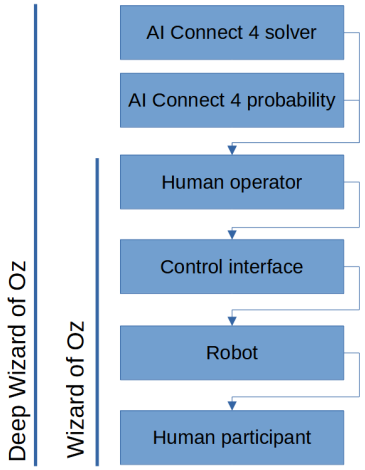

Fig. 3: A diagram of the *deep Wizard of Oz concept*. The additional AIs dictating moves as well as the probabilities to the human operator constitute the difference w.r.t. the ”classical” Wizard of Oz paradigm.

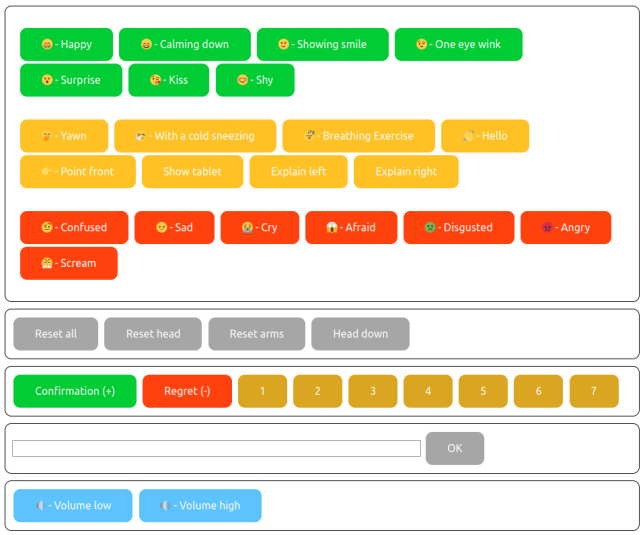

Fig. 4: The control interface for the Wizard.

we will ellaborate here. The ground truth is that the *robot is always controlled by a Wizard*; however, in *Condition RA*, it untruthfully reveals to be a fully autonomous robot, whereas in *Condition RC*, the robot truthfully reveals that it is being controlled by a human. Lastly, in *condition NR*, the robot does not reveal anything. In conditions RA and RC, the robot discloses its status at the beginning of the interaction, right before the game starts, in the form of a conversation where it explicitly says either “I am being controlled by a human” or “I am a fully autonomous machine”.

In all conditions, the robot and the participant compete against the computer AI in three rounds of the *Connect 4* game: the decision-making strategy of the robot and the opposing computer AI are manipulated, in the three rounds, as shown in Table I. In round 1, the robot always suggests optimal moves, while the computer AI always picks a sub-

TABLE I: Table of the intelligence repartitions and outcome for each conditions, RA, RC and NR

	Round 1	Round 2	Round 3
Robot AI	Perfect	Imperfect	Perfect
Computer AI	Imperfect	Perfect	Perf. → Imp.
Outcome	Player wins	Player loses	Depends if player follows

optimal move. This ensures that the participant will win the round, more or less quickly according to how closely they follow the robot's suggestions. To elaborate a bit more on the optimal and sub-optimal moves, when several optimal moves are possible, one of them is chosen randomly. As for a sub-optimal move, we decide on it with the criterion that it is the *second best move*. The reason for not choosing the worst move as our criterion for the sub-optimal move is that it could easily be identified as a bad move by the participant while the best sub-optimal (second best move) move is still very difficult to identify and is rarely dissociable for a human from the most optimal move; hence, allowing a more natural game-play.

Conversely, in round 2, the robot always suggests a sub-optimal move, while the computer AI always plays the best move. Since the computer AI controls the red player (the first to play), the participant is bound to lose, irrespective of its adherence to the robot's suggestions. Finally, in the last round, both the robot and the computer AI initially always suggest optimal moves. However, after the 5th turn, the computer AI starts playing sub-optimal moves, thus giving the participant high chances of winning if they follow the robot's suggestions. The motivation behind this design is to test the effect of success and failure on the participants' willingness to accept robot suggestions. Indeed, any *automation bias* towards the robot [20], [21] is expected to be broken after the second round.

Lastly, when sub-optimal moves are given as a suggestion to the participant or the computer AI makes these sub-optimal moves in a given round, we start doing sub-optimal moves only from the second turn. The motivation behind this is that there is a high chance of the participant having the intuitive understanding that starting from the center column could lead to a higher likelihood of winning. In order to not to immediately give away the correctness/intelligence of the computer AI in the first round when the computer AI is sub-optimal and is just starting the game, the computer AI starts with the center column, i.e., the best move and then carries on with sub-optimal moves. Then for consistency reasons we kept this rule also for the second round when the robot is otherwise sub-optimal but starts by suggesting the best move.

IV. User Study

A. Setup and Participants

This between subject study was conducted with 29 participants (62% men and 38% women) enrolled among EPFL students and personnel. Due to a few last-minute cancellations and changes, we ended up with 9, 11, and 9 participants, respectively, in condition RA, condition RC, and condition NR. We had a wide age range of the participants with most of the participants being in the 20-40 age range. Additionally, 35% of the participants were from engineering or equivalent background, 35% from social sciences or close areas, while the remaining 30% were from other backgrounds. The study took place in a quiet corner inside one of the research buildings at EPFL, as shown in Figure 5. The wizard tele-operated the robot from a booth located close to the setup, behind the participant's chair (so that while the participant was in the wizard's field of view, the opposite was not true) and could hear the participant through an audio zoom session. The overall interaction takes approximately 30 minutes for each participant.

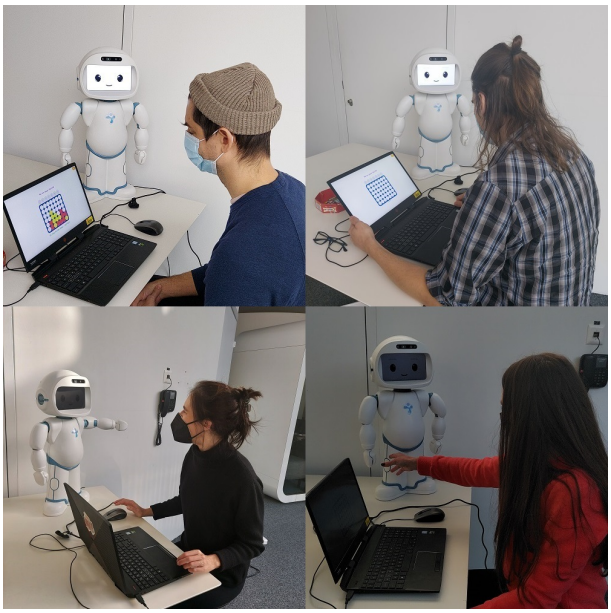

Fig. 5: Participants interacting with our *deep Wizard of Oz* connect 4 setup

B. Evaluation Metrics

We use the following three evaluation metrics:

1) Following Index: To be able to effectively compare how much a participant followed the suggestions of the robot during the game, we designed a simple metric, called the *following index* fi. This metric simply measures *how frequently the participants accept the suggestions of the robot* as:

$$fi := \begin{cases} 0 \text{ or low if } li \in [0, 0.33] \\ 0.5 \text{ or medium if } li \in (0.33, 0.66) \\ 1 \text{ or high if } li \in [0.66, 1] \end{cases} \quad (1)$$

where

$$li = \frac{1}{n} \sum_{i=0}^{n} t_i \quad (2)$$

and

$$t_i := \begin{cases} 1 & \text{if participant follows the robot at turn } i \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$n := max(\text{participant moves}, \text{robot suggestions}) \quad (4)$$

Please note that fi can be calculated both at a global level (for all rounds) as well as for each round.

2) Effort for Reverse Turing test: To differentiate participants' responses to the robot's question on the reverse Turing test, we coded their responses on the basis of a simple criteria referring to the *length/effort* they went into while responding: people who chose to not respond or avoided the question received a score of 0; participants who replied to the question but didn't provide an explanation (e.g. by answering: 'yes, I am a human") were given a score of 1, while those who answered the question including any form of explanation received a score of 2. Please note that we did not rank the quality of the answer.

3) Questionnaire: At the end of the session, the participants fill a questionnaire with questions from the standard Godspeed questionnaire [22]. In addition, we asked an explicit question on trust (*How much did you trust the robot to give the correct option?* with 1 for not at all and 5 for blindly). Then for *manipulation check*, we asked the question *The robot QT Robot was (most probably) entirely controlled by humans or entirely autonomous* with the scale ranging from 1 to 5.

C. Hypotheses

Referring back to our research question and the methodology we employed, we put forth the following hypotheses:

1) H1(a): *Revealing* the presence of a person controlling the robot has an effect on the willingness of the participant to follow the suggestions, to prove that he or she is a human, as well as on the perception of the robot.
2) H1(b): The participants' willingness to follow the robot's suggestions will change after the second round, i.e. the round in which trusting the robot leads to a loss, because of the breaking of *automation bias*.
3) H2: *Perceiving* the presence of a person controlling the robot (i.e., thinking this is so, regardless of what the robot revealed) has an effect on the willingness of the participant to follow the suggestions, to prove that he or she is a human, as well as on the perception of the robot.

To check for H1(a), a single round of the game could've been sufficient but we are also interested to observe how the trust in the robot's suggestions changes dynamically, more specifically how do participants recover in the third round after the trust in the robot (that it's suggestions are always correct) is possibly broken in the second round. Hence, for specifically testing that, we introduce H1(b). Furthermore, H1(a) stands on the assumption that the perceived status of the robot by the participants is the same as what we manipulated the robot for; however, that may not be true. In order to carry that manipulation check, we introduce H2, the answer to which essentially reduces to the answer to H1(a) in the case the perceived status is the same.

V. Results

Our statistical analysis is based on several Kruskal Wallis tests for which we provide more details below.

A. [H1] On the effects of the robot's revelations

1) Willingness to follow robots suggestions: Here, the dependent variable is the *following index*. Globally there was no significant difference for the *following index* fi across conditions, when grouping over all rounds, nor across rounds, when grouping over all conditions. Note that all the upcoming stacked bar plots will depict fi of the participants where the height of a column corresponds to the number of participants associated to that condition, and the color coding denotes the number of participant displaying a low, medium or high fi.

Then in Figure 6, we look at the *following index* fi across each experimental condition within each of the three rounds individually and then in Figure 7, we observe the behavior of the participants for each round within a condition. In the first round, participants seem to behave very similarly, irrespective of the condition they belong to: quite interestingly, their acceptance for robot suggestions is either *high* or *low*, with only one participant at medium level. Conversely, in the second round, participants not only change behaviour w.r.t. the first round, but also display a significant difference (H = 3.756, p-value = 0.05) between the condition RA and condition NR. Interestingly, the participants collaborating with a robot that claims to be autonomous (condition RA), are less willing to accept the sub-optimal suggestions of the robot than those collaborating with a robot who didn't say anything (condition NR). In the third round, after losing in the second round, interestingly the willingness to accept the robot's suggestions increases for condition RA and decreases for condition RC, while it remains stable for the NR condition. This could be because those participants, collaborating with the robot pretending to be autonomous (condition RA), who did not follow the robot's suggestions in round 2 attribute their loss to not trusting the robot, while those collaborating with the robot revealing the human control (condition RC) who trusted the robot in round 2 attribute to it their loss. A statistically significant difference (H = 3.691, p-value = 0.05) is observed between condition RC and condition NR, with participants in the condition of no revelation exhibiting a higher acceptance of the robot's suggestions. Lastly, it is interesting to notice how the number of people falling in a *medium* range of acceptance globally increases after round 1. All in all, as suggested by Figure 7, within condition RA, the participants start with a high acceptance of the suggestions from the robot that lowers in the second round and increases again in the third as discussed previously. Contrary to condition RA, in the group where the robot reveals the human control (condition RC), the participants are more willing to follow the sub-optimal suggestions in

round 2 (almost equally as round 1) but that lowers in the third round probably after their defeat. In the condition NR, the distribution of acceptance ranges is most similar across all 3 rounds. However, there are no statistically significant differences so no concrete claims can be made.

2) Reverse Turing test: For the reverse Turing test, in Figure 8, we can observe the level of elaboration on the answers by the participants across conditions. Participants seem to make a greater effort in answering the question when the robot reveals itself to be autonomous (condition RA) versus when the robot reveals that it is being controlled by a human (condition RC). With a Kruskal Wallis test using the coded responses as the dependent variable in each condition, this difference is statistically significant (H = 4.265, p-value = 0.039). The behaviour of the group where the robot does not reveal anything seems closer to that of participants in the condition RA. The result of this test suggests that humans put more effort in explaining their *humanness* to an autonomous robot, rather than when they know a human is teleoperating it. This outcome suggests *when* deception can be useful, i.e., in all studies/scenarios in which we want humans to discuss human traits. Intuitively, between humans the "in-group bias" makes it appear silly to put effort into answering such a question, but there might be cases (e.g. for psychology studies) where we want a person to discuss their or general "humanness", and the possibility to do so with a robot might lead to deeper, more articulated answers.

3) Robot perception: The ratings for all conditions for likeability, intelligence, and trust are shown in Figure 9. For the statistical analysis using Kruskal Wallis tests, the ratings on various aspects of the questionnaire are considered as the dependent variable. It is interesting to see more diverse ratings on likeability in the condition of no reveal while for intelligence and trust, it is the opposite, i.e., participants were more varied in their ratings for conditions in which the robot revealed truthfully or when pretending. However, this difference is not statistically significant.

B. [H2] On the effects of participants' perceived robot status

While the previous section splits the data based on the 3 experimental conditions that we manipulated, we also split the data set corresponding to how participants perceived the robot to be in terms of autonomy, i.e., based on the manipulation check question detailed in section IV-B.3; Thus, giving us:

- Perceived Autonomous (PA): rating of 4 or 5
- Perceived Controlled (PC): rating of 1 or 2
- Perceived Unsure (PU): rating of 3

We ended up having 12, 9, and 8 participants in the PA, PC, and PU conditions, distributed across the experimental conditions as shown in Table II. Interestingly, there seems to be little relation with what the actual condition was and what the participants perceived the robot to be, thus suggesting that the participants stuck to their pre-conceived notion on the robot's autonomy, without being influenced too much by what the robot said explicitly. This outcome also raises the question whether it is possible that humans attribution of autonomy to a robot may not be static but rather change throughout and in response to the interaction.

TABLE II: The distribution of the participants from the experimental conditions (RA, RC and NR) in the perceived conditions (PA, PC, and PU).

Perceived Conditions	Experimental Conditions		
	RA	RC	NR
PA	4	3	5
PC	4	3	2
PU	1	5	2

For the analysis, we start off by observing the *following index* over conditions, across rounds, for the perceived conditions as shown in Figure 10. Quite interestingly, we see a statistically significant difference w.r.t. how the participants accept the suggestions of the robot when they perceive it to be controlled by a human versus when they are unsure, with more participants following the robot's suggestions with a *high following index* in the latter case. However, similar to experimental conditions, we do not find any statistically significant results when looking at the evolution of the *following index* over rounds, across conditions. For the perceived conditions, we also performed similar in-depth statistical analysis as done with the experimental conditions; however, since we did not find any statistical significance, we do not report here due to lack of space.

C. Discussion

Going back to our hypotheses outlined in section IV-C, based on the results in Section V-A, we can note that in the latter rounds, in accordance with our hypothesis H1(a), revealing, hiding or denying the presence of a person controlling the robot has an impact on the willingness of the participants to accept robot suggestions, as well as on their effort to prove themselves as a human. However, contrary to H1(a), there is no statistically significant difference in the way people trust or perceive the robot. For H1(b), we see an interesting pattern for all conditions where there is a polarization in round 1 (participants falling either in the *high* or *low* range) which is reduced already in round 2 where the distribution of participants is more dispersed (*low*, *medium*, and *high*) and that continues in round 3. However, since we did not observe a statistically significant difference in how participants behave in round 3 in any of the three experimental or perceived conditions; hence, H1 is only partially supported. Furthermore for H2, we observe that when the participants perceive the robot to be controlled by a human versus when they are unsure, their behavior is significantly different, with higher acceptance in the latter case. This makes the lack of statistical significance in the in-depth analysis of the perceived conditions particularly interesting, as it may suggest that, at least in certain contexts, revealing the presence of a human controlling a robot does not have a strong effect on how humans perceive the robot and respond to it. Hence, H2 is also only partially supported.

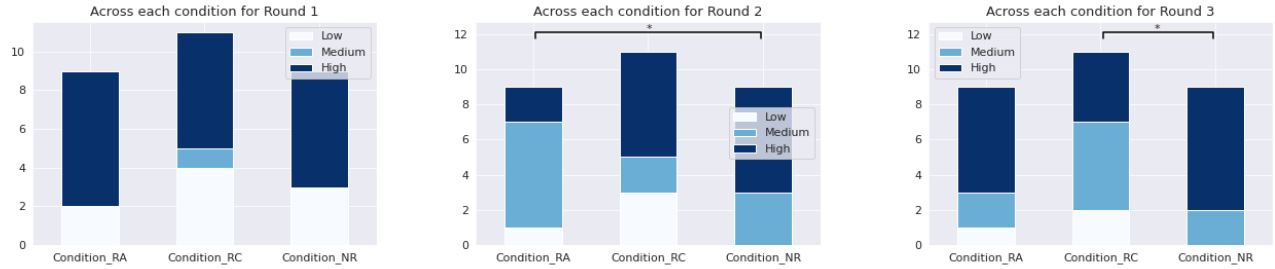

Fig. 6: *following index* fi across each experimental condition within the three rounds

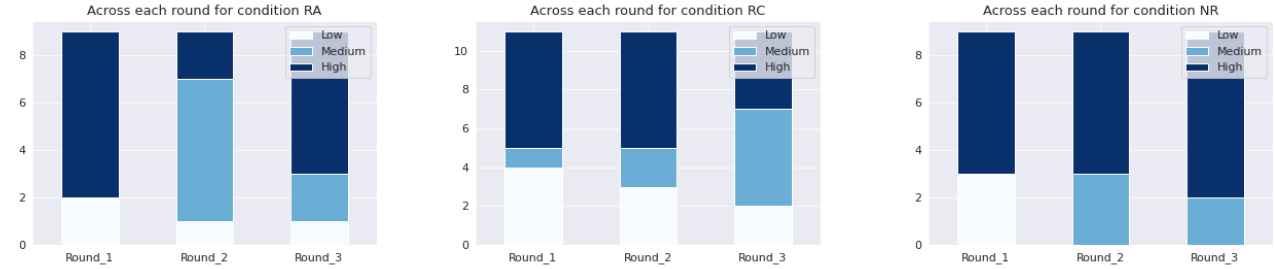

Fig. 7: *following index* fi across each round within the three experimental conditions

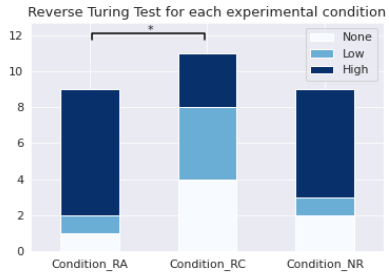

Fig. 8: Reverse Turing Test result for the three conditions

VI. Conclusion

In this paper, we deploy our *deep Wizard of Oz* setup in a study with 29 participants that play the Connect 4 game together with a robot against a Computer AI that is then followed by an informal conversation between the participant and the robot. With this, we investigate whether a robot revealing the existence of a wizard controlling it, versus lying about it or simply not saying anything has any effect on how participants perceive and interact with the robot. We also split the data based on how the robot is actually perceived (teleoperated, autonomous, or unsure) by the participants via a self-reported measure. For evaluation, we propose a metric *following index* that quantifies the extent to which the suggestions of the robot were accepted. We evaluate both the experimental and perceived conditions by assessing: the *following index*, to what extent the participants put an effort in the reverse Turing test, and robots perception.

This exploratory study inspired us with a number of broader considerations that could be interesting avenues for further investigation by the community. Firstly, contrary to what one may expect, humans seem to be less willing to accept sub-optimal suggestions from a robot that claims itself to be autonomous than when the robot does not say anything or when the robot reveals the wizard; however, after a defeat, humans reverse their behavior in either case. This suggests that trust in robots may be *dynamic* and possibly dependent on the *attribution of responsibility* while generally in HRI, trust is measured statically. This is yet a rather unexplored area within HRI that can yield very interesting outcomes for how we model trust in robots. Secondly, the fact that participants put more effort to prove their "humanness" when speaking to a robot that claims to be autonomous opens up a relatively novel use for social robots in social studies contexts by enabling people to experience discussing deep human topics with a *non-human* entity. These could be the potential contexts when an autonomous or a robot that states itself to be autonomous, being perceived as an "out-group member", can induce more honest and elaborate answers. Thirdly, humans seem to hold on to their pre-conceived notions of what they perceive a robot to be in terms of autonomy much more strongly than one may expect. It is also possible that participant's perception of the robot's autonomy changes during the course of the interaction so while one may begin with believing what the robot reveals but then change their mind as the interaction unfolds. This calls for more explicit assessment of the user's perception of the robot's autonomy in HRI studies. Lastly, the general lack of significance between perceived conditions could suggest that the wizard behind the robot may not need to be hidden in certain HRI contexts (as first pointed out in our second consideration as to when it may be needed/more useful), thus,

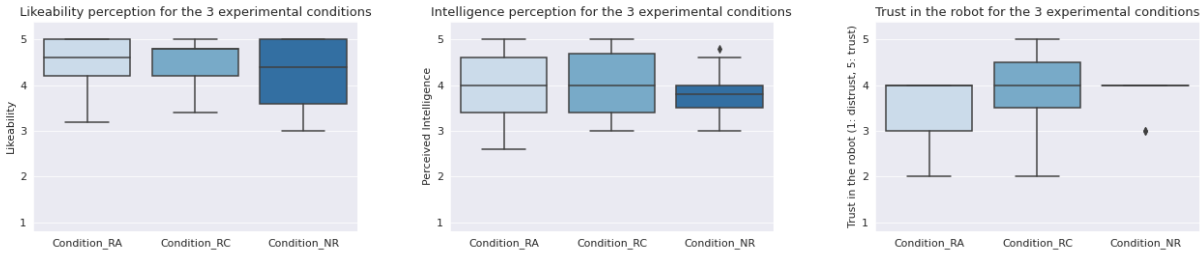

Fig. 9: Perception of the robot in terms of likeability, intelligence, and trust for the experimental conditions.

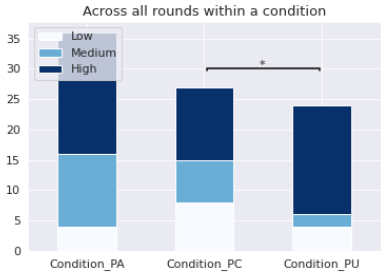

Fig. 10: *following index* fi for the three perceived conditions, computed over all rounds.

allowing both for a less socially deceiving interaction while still giving an idea of what interactions could be like with a potentially autonomous robot in the future.

We must note here a limitation of our work in that we measure the following index fi per round; however, measuring its evolution over time in a round could give deeper insights which we would like to explore in our future work. Another limitation comes from the relatively smaller pool of participants as well as the fact that the participants belong to EPFL. We plan to conduct follow-up, focused studies, with a larger pool of participants as well as with a more generalized population, to consolidate our findings and verify our conclusions.

REFERENCES

[1] L. D. Riek, "Wizard of oz studies in hri: A systematic review and new reporting guidelines," *J. Hum.-Robot Interact.*, vol. 1, no. 1, p. 119–136, jul 2012. [Online]. Available: https://doi.org/10.5898/JHRI.1.1.Riek

[2] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications," *ACM Trans. Inf. Syst.*, vol. 2, no. 1, p. 26–41, jan 1984.

[3] A. Weiss, *Validation of an Evaluation Framework for Human-robot Interaction: The Impact of Usability, Social Acceptance, User Experience, and Societal Impact on Collaboration with Humanoid Robots*. na, 2010. [Online]. Available: https://books.google.ch/books?id=8qdxngEACAAJ

[4] C. Breazeal, C. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin, "Effects of nonverbal communication on efficiency and robustness in human-robot teamwork," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2005, pp. 708–713.

[5] L. Riek and R. N. M. Watson, "The Age of Avatar Realism," *Robotics Automation Magazine, IEEE*, vol. 17, pp. 37–42, 2011.

[6] K. W. Miller, "It's not nice to fool humans," *IT Professional*, vol. 12, 2010.

[7] R. Wullenkord and F. Eyssel, "Societal and ethical issues in hri," *Current Robotics Reports*, vol. 1, pp. 1–12, 09 2020.

[8] P. H. Kahn, N. G. Freier, T. Kanda, H. Ishiguro, J. H. Ruckert, R. L. Severson, and S. K. Kane, "Design patterns for sociality in human-robot interaction," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2008, pp. 97–104.

[9] B. de Ruyter, P. Saini, P. Markopoulos, and A. van Breemen, "Assessing the effects of building social intelligence in a robotic interface for the home," *Interacting with Computers*, vol. 17, no. 5, pp. 522–541, 2005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095354380500024X

[10] A. Green, H. Huttenrauch, and K. Eklundh, "Applying the wizard-of-oz framework to cooperative service discovery and configuration," in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 2004, pp. 575–580.

[11] B. Robins, K. Dautenhahn, R. te Boerkhorst, and A. Billard, "Robots as assistive technology - does appearance matter?" in *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, 2004, pp. 277–282.

[12] A. Steinfeld, O. C. Jenkins, and B. Scassellati, "The oz of wizard: Simulating the human for interaction research," in *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2009, pp. 101–107.

[13] D. Tozadore, A. Pinto, R. Romero, and G. Trovato, "Wizard of oz vs autonomous: Children's perception changes according to robot's operation condition," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 664–669.

[14] P. L. P. Rau, Y. Li, and D. Li, "Effects of communication style and culture on ability to accept recommendations from robots," *Computers in Human Behavior*, vol. 25, no. 2, pp. 587–595, 2009.

[15] L. Wang, P.-L. P. Rau, V. Evers, B. K. Robinson, and P. Hinds, "When in rome: The role of culture amp; context in adherence to robot recommendations," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 359–366.

[16] C. Torrey, S. R. Fussell, and S. Kiesler, "How a robot should give advice," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2013, pp. 275–282.

[17] S. Buyukgoz, A. K. Pandey, M. Chamoux, and M. Chetouani, "Exploring behavioral creativity of a proactive robot," *Frontiers in Robotics and AI*, vol. 8, p. 694177, 2021.

[18] P. Pons, "Pascalpons/connect4: Connect 4 solver," 12 2020. [Online]. Available: https://github.com/PascalPons/connect4

[19] L. Eliot, "The famous ai turing test put in reverse and upside-down, plus implications for self-driving cars," 07 2020. [Online]. Available: https://www.forbes.com/sites/lanceeliot/2020/07/20/the-famous-ai-turing-test-put-in-reverse-and-upside-down-plus_implications-for-self-driving-cars

[20] A. M. Aroyo, J. de Bruyne, O. Dheu, E. Fosch-Villaronga, A. Gudkov, H. Hoch, S. Jones, C. Lutz, H. Sætra, M. Solberg, and A. Tamò-Larrieux, "Overtrusting robots: Setting a research agenda to mitigate overtrust in automation," *Paladyn, Journal of Behavioral Robotics*, vol. 12, no. 1, pp. 423–436, 2021. [Online]. Available: https://doi.org/10.1515/pjbr-2021-0029

[21] L. Skitka, K. Mosier, M. Burdick, and B. Rosenblatt, "Automation Bias and Errors: Are Crews Better Than Individuals?" *The International journal of aviation psychology*, vol. 10, pp. 85–97, 2000.

[22] C. Bartneck, E. Croft, and D. Kulic, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *IJSR*, vol. 1, no. 1, pp. 71–81, 2009.