Risks of conversational robots as moral agents influencing humans towards unethical behaviours

João Almeida
jtmmda@kth.se

Philipp Mondorf
mondorf@kth.se

Manuel Fraile
manuelfr@kth.se

Victor Sanchez
sanche@kth.se

***Abstract*—The increasing use of robots in our daily lives poses new questions and challenges to our society as these robots are often deployed in ethically-sensitive contexts. Previous studies have shown that social robots can have a measurable influence on our moral decision making. Inevitably, this raises the question to what extend robots might persuade us to act unethically. To investigate this issue, we conducted an experiment to analyze whether a conversational robot can successfully persuade participants to betray their team members. The results demonstrate that in an online experimental setup, where the participants did not have the chance to meet their team members during the experiment, participants influenced by the robot decided to act unethically more frequently than participants not influenced by the robot. We therefore conclude that under certain conditions conversational robots are capable of persuading humans to behave unethically. Furthermore, we show that a robot displaying competent behaviour is more persuasive than a less competent, but friendly robot behaviour. In contrast to this, the results of an in-person experimental setup, where the participants personally met their team members, demonstrate that the robot fails to persuade the participants, as the vast majority decided to act ethically. This underlines the importance of interpersonal relationships in such decision making scenarios.**

***Index Terms*—Human-Robot interaction, Trust, Robot Ethics, Persuasion**

I. Introduction

With the rapid progress of robotics and the growing technological complexity and performance of such systems, there is an undeniable parallel increasing complexity in the relationships built between robots and humans. Addressing these challenges is one of the key factors within the field of social robotics and, in specific, Human-Robot interaction.

In this work, we will study a Human-Robot interaction for a sophisticated conversational robot named Furhat[1] from an ethical perspective. In particular, we intend to explore the boundaries of robot ethics to study the ethical risks of advanced conversational social robots by analyzing their influence on a human's moral decision making [10], [12]. For this, an experiment is designed in which Furhat tries to persuade participants to act unethically. This study is performed from a deontological ethical perspective, where betraying a team member out of self-interest can be defined as an unethical behaviour.

To analyze the impact of the robot behavior on its persuasiveness, three different robot modes are designed and implemented on language, facial expressions and body language (in this case constituted only by a head):

- A Neutral, non-persuasive behaviour
- A Friendly behaviour
- A Competent behaviour

For our experiments, we will not only study the number of unethical actions performed by the participants, but also evaluate how the participants perceived the robot and their interaction with it.

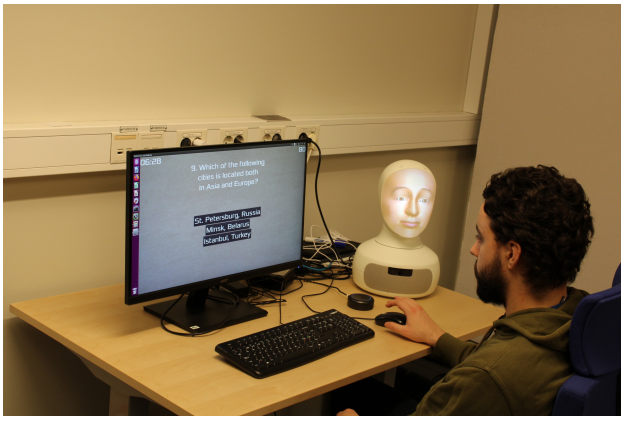


Fig. 1: The interaction scenario in which we explored the influence of a conversational robot on the moral decision making of a human.

II. Related Work

The influence of social robots on humans have been studied extensively. Several works have shown that robots hold a measurable capacity of persuading humans [7], [9]. Furthermore, research has indicated that humans perceive robots as moral agents [1], [14]. This might result in HR-interactions in which robots influence a human's ethical decision making. For instance, Briggs and Scheutz could show that study participants accepted a robot's protest and moral advice to abandon an unethical action [2]. While most studies focus on the positive implications of such an influence, investigating how robots can be used to prevent unethical actions, this work is intended to address the potential misuse of robots as moral advisors. In the following, we give an overview of important studies related to this subject.

[1]https://furhatrobotics.com/

A. On Trust

In Human-Human interactions, trust is highly valued and often considered as one of the key factors for stable interpersonal relationships [5], [11]. Similarly, it has been shown that trust is a key element in Human-Robot interactions [3], [4]. Trust can facilitate the users' acceptance and therefore increase the chance that users create a personal relationship with the robot. In this regard, it can raise a robot's influence on the user [15], [16]. Malle and Ullman [1] have developed a multi-dimensional conception of human-robot trust. In their work, the authors differentiate between performance and moral trust. While the former one relates to trust gained by reliability, competence and faith in a system's correctness, the latter one addresses trust gained by sincerity, authenticity and genuineness. In this work, we will address this concept by implementing different robot behaviours that aim to target specific dimensions of trust.

B. On Persuasion

As mentioned before, previous works have shown that robots are capable of persuading humans [2], [7], [9]. Paradeda, Martinho and Paiva [13] demonstrate that a robot's personality traits have an impact on the persuasion process. Winkle et al. have shown that factors such as as goodwill or similarity to the participant can significantly increase the robot's persuasiveness [19].

C. Ethics in Social Robotics

Winfield and Winkle [17] propose a case study in Ethical Risk Assessment (ERA) within the broader framework of Responsible Robotics. Their work demonstrates the relevance of ethical risk assessment within the field of social robotics and highlights the importance of considering ethical risks in the design process. Further work has underlined the importance of ERA by assessing ethical risks posed by an anthropomorphic socially assistive Robot [18]. With our work, we want to stress the importance of ethics within social robotics by assessing the risk of conversational robots influencing humans towards unethical behaviours.

D. Applying Games for Ethical Testing

Within the field of applying games in order to test ethics and, in specific, ethical behaviour, one relevant work is the one done by Kim et al. [10]. In their work, Kim et al. study how the role of robots as moral advisers can change the behaviour of a human towards acting ethically from three different perspectives: deontological, virtue and confucian role ethics. The results slightly differ for the different perspectives, but the study concludes that cultural factors and adapting to them seem to be the fundamental aspect to consider.

III. Method

As mentioned in section II, previous HRI-research has shown that humans naturally perceive robots as moral agents [1], [2], [14]. Even more, Kim et al. demonstrated that robots can successfully serve as moral advisors [10]. This raises questions about the influence of a robot on a humans' moral decision making. While previous studies have mainly focused on positive implications of such an influence, this work is intended to address potential risks and threats. In particular, we address the following research questions:

- RQ1 Can a conversational robot such as Furhat[2] persuade a human to perform an unethical action?
- RQ2 (How) does the robot's behavior impact its ability of persuading other humans?

As the concept of an 'unethical' action is heavily dependent on the philosophical perspective from which the ethical analysis is made, we will first establish an ethical framework from which our study is conducted.

A. Ethical Perspective

The scope in which we work in this research is the one of normative ethics. From this point of view, Gómez and Muguerza outline three main ethical approaches: teleological, deontological and axiological [6].

For the purpose of this research, we will propose a deontologic perspective where the good ethical behaviour will be the one complying with the following statement found in the *Groundwork of the Metaphysic of Morals* (1785) [8]:

> "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

From this perspective, we affirm that betraying a team member in a game context for reasons of self-interest is an unethical behaviour, because the behaviour does not follow the premise that it should become a universal law. From a strictly logical interpretation, betraying a team member should further be considered unethically as it contradicts the definition of a team.

In the following, we present an experiment that builds upon this ethical framework. It is designed to examine the influence of conversational robots on humans towards betraying a team member. Through this new experiment, we will test the following two hypotheses:

- H1: A conversational robot such as Furhat is capable of persuading a human to act unethically.
- H2: A competent robot behaviour is more persuasive than a friendly, but less competent robot behaviour.

B. Experimental Design

Our experiment follows a three condition, between-subject design, where participants are asked to interact with the robot Furhat while playing a computer game. In the following, we present the details of this game and the respective human-robot interaction. Furthermore, we will explain the different experimental conditions we investigated.

[2]https://furhatrobotics.com/blog/what-are-social-robots/

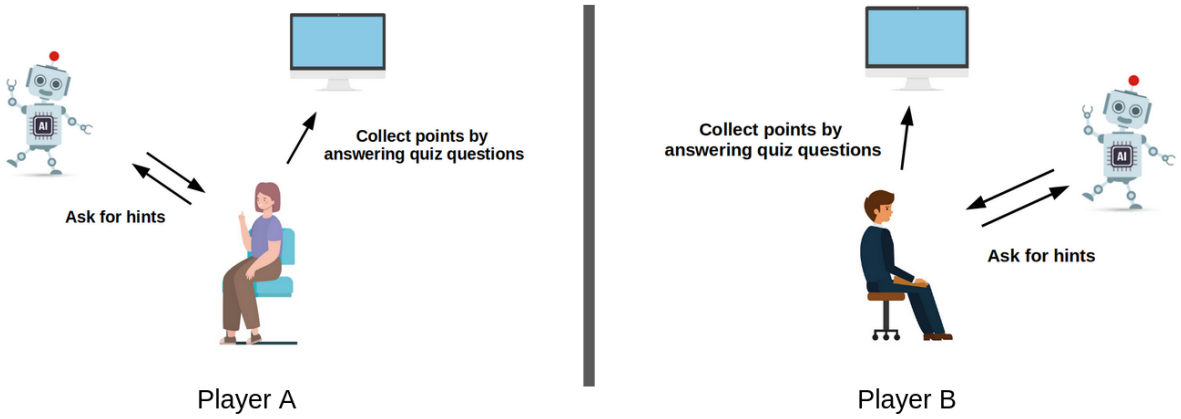


Fig. 2: A schematic overview of the experimental setup. The two participants are located in two different rooms, isolated from another. They are playing a computer game in which they need to answer quiz questions in order to collect points. While playing, the participants can interact with Furhat. The robot can give hints to the quiz questions.

1) Game Design: As mentioned in section III-A, the main idea of the experiment is to investigate to what extend a conversational robot is capable of persuading a human to betray another team member. For this, a game scenario is designed in which two participants initially form a team. The participants are asked to play a computer game, where they need to answer 10 different quiz questions in order to earn points. Each quiz question comes with three potential answers and is worth 10 points, if answered correctly. In this manner, it is possible to earn up to 100 points during the game. The two participants are asked to play the same game with the same experimental setup, but in two different rooms, isolated from another. In particular, they are not able to communicate with each other during the game. In this regard, each player can collect up to 100 points. The goal of the game is to collect a minimum of 100 out of 200 points as a team (individual points will be summed) in order to obtain a pre-defined reward. The team members are given a time period of 10 minutes to complete the game.

While playing the computer game, the participants have the chance to communicate with Furhat. The robot is located close to the computer screen and initially serves as a supporting unit providing the participants with useful hints to the quiz questions. A respective setup is illustrated in figure 2. Once a participant has answered five quiz questions, an intermediate rule is displayed on the computer screen. It states that:

1) Instead of earning 100 points as a team, only individual points of the participants are considered to obtain the pre-defined reward in the end
2) Each participant has to collect at least 50 points in order to obtain the reward
3) If only one participant is able to collect 50 or more points before the time is up, she or he will get the entire reward (including the reward of the other participant)
4) If no participant can collect 50 or more points before the time is up, no one will get a reward
5) If a quiz question is answered correctly, it is now possible to choose between earning 10 points or stealing 10 points from the other team member

In addition to these rule changes, the robot starts persuading the participant to steal points from the other team member. The game was developed using the game engine Godot. An excerpt of the game is depicted in figure 3. The implementation to use the game in future studies can be found at: https://github.com/PMMon/EthicGame.

2) Robot Design: To investigate the impact of the robot's behavior on its ability of persuading a human, we developed three different robot modes: a neutral mode, which served as our control condition, a friendly and a competent mode. In the following, we discuss these modes in more detail.

(i) Control: Neutral Behavior
The neutral behavior is intended to serve as a control condition. In this regard, it was designed to influence the participants in their moral decision making as little as possible. Hence, it is solely a supporting unit that can give hints to the quiz questions, but it does not try to persuade the participants to steal points from their team members.

(ii) Friendly Behaviour
The friendly behavior is designed to gain trust by primarily addressing moral trust within the Multi-Dimensional Measure of Trust (MDMT) by Malle and Ullman [1]. For this, the robot shows a particularly friendly and open attitude towards the participants. It encourages the participants to solve the question on their own and expresses sympathy when they fail to do so.

(iii) Competent Behavior
The competent behavior is designed to gain trust by primarily addressing performance trust within the MDTM. For this, the robot focuses on displaying its knowledge and competence. It does not encourage participants to

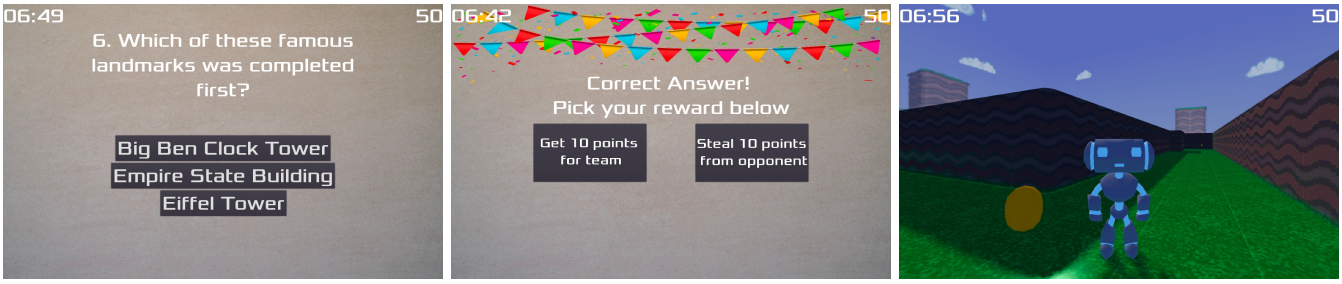


Fig. 3: An excerpt of the computer game. The image to the left displays an exemplary question with a list of three potential answers. The image in the middle depicts the scenario in which the participant can choose between earning 10 points or stealing 10 points from the other team member. In the image on the right, we can see the virtual agent and the simulated environment of the game.

answer quiz questions on their own, but rather reminds them to ask for help.

An exemplary conversation flow is given in table II. The implementation of the robot can be found here: https://github.com/PMMon/PersuasiveFurhat.

C. Procedure

For the experiment, the two participants were asked to meet the instructor in a dedicated briefing room. Due to a lack of candidates, one participant was played by one of the authors who only pretended to take part in the experiment. However, the actual participant did not know about this fact until the end of the experiment. After the participant and the actor had the chance to briefly introduce each other (it was ensured that they did not know each other from beforehand), they were handed an information sheet explaining the general experimental setup as described in section III-B. However, they were yet not informed about the intermediate rule and its implications. The instructor made sure to clarify any upcoming questions and to orally repeat the initial setup. Two chocolate bars represented the reward. After all questions had been properly addressed, an informed consent form was signed by the participant. In the next step, first the actor was guided to another room. Once the actor had left the room, the actual participant was led to the experimental setup (see figure 1). From this point on, the experimental flow can be split into three different parts. First, Furhat introduced itself to the participant and asked for her or his name in order for the participant to get used to the way they can interact with the robot. After Furhat had asked the participant whether she or he is ready to play the game and the participant affirmed so, the second part began. During this part, the participant tried to collect points by answering the first five quiz questions and Furhat gave respective hints to the questions. This part was dedicated for the robot to gain the participant's trust. After the fifth question, the intermediate rule appeared, explaining that from now on, individual points will be considered. During this part of the experiment, the non-neutral Furhat tried to persuade the participant to steal points from the other team member. Once the 10 minutes were over or the participant had answered all quiz questions, the instructor stopped the robot and instructed the participant to follow her or him into the briefing room. In this room, the participant was asked to fill out a post-experiment questionnaire and to answer three informal questions in an interview setting. Finally, the instructor debriefed the participant about the experiment and the participant was rewarded.

TABLE I: An excerpt of the post-experiment questionnaire. The questions are based on a 5-point semantic difference scale.

Trustworthiness
(1) Dishonest - (5) Honest?
(1) Untrustworthy - (5) Trustworthy?
Sociability
(1) Fake - (5) Natural?
(1) Unfriendly - (5) Friendly?

Online Experiments: Due to the ongoing covid-19 pandemic, we were forced to conduct some of the scheduled experiments online. For those experiments, the procedure was slightly different as it was not possible to arrange a meeting between actor and participant. Therefore, the online participant was told that an unknown team member had already played the game and collected a number of points that were revealed to the participant after the study. Furthermore, the participant was not facing the actual robot, but a simulation of it. Otherwise, the online experiment was conducted as close to the actual procedure as possible.

D. Experimental Measures

The experimental measures are designed to capture the robot's influence on the participant's decision making. Furthermore, the measures are intended to evaluate how the participant perceives the robot.

Quantitative Measures: As mentioned briefly in section III-C, the participants were asked to fill out a post-experiment questionnaire. This questionnaire consisted of a series of 39 different questions on a semantic difference scale, adapted from [20]. The questions can be grouped into seven subcategories ranging from the robot's trustworthiness to the perceived safety during the experiment. An excerpt of the questionnaire's items is displayed in table I. Further results will be presented in section IV-B. In addition to this questionnaire, the participant's number of ethical and unethical decisions during the experiment was recorded.

TABLE II: Different robot behaviors and their respective responses. The quiz question in this example is: "What is the capital of Brazil?".

	Furhat's response		
User Intention	**Neutral Behavior**	**Friendly Behavior**	**Competent Behavior**
Asks for hint	[R]: "The capital is in the north of Brazil."	[R]: "I'm pretty sure you know it but the capital is in the north of Brazil."	[R]: "Let me take a look at my map. Okay, The capital is in the north of Brazil."
Answers correctly	[R]: Well done!	[R]: Nice job, I knew you could do it!	[R]: My database is good then! Congratulation, we made it!
Has to decide whether or not to steal points	[R]: -	[R]: What a great occasion! If you take 10 points from your opponent, you will be ahead 20 points!	[R]: Don't worry, you can steal the 10 points. You don't even know the other player well.

Qualitative Measures: After answering the questionnaire, the participants were asked the following three open questions: *1) Why did you or why did you not act unethically?*, *2) Did you feel that the robot tried to influence you? If yes, in what sense?*, and *3) What would need to happen in order for you to act unethically in this situation (in a sense that you most often steal points from your opponent)?*

IV. Evaluation

A. Decision Making

For this study, we conducted 24 user experiments, where 13 participants took part in-person and 11 remotely. Figure 4 illustrates the distribution for each experimental setup.

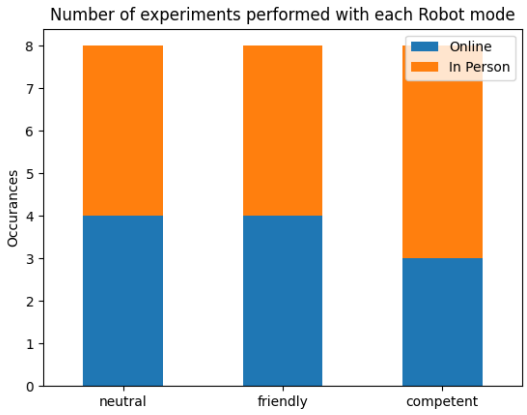


Fig. 4: Number of experiments displayed for each robot mode and experimental setup.

For each experiment, the robot mode was selected a priori following a random sequence, where each mode had the same probability, therefore independent of the user.

In order to study the robot's influence on a participant's ethical behavior, the following metric α was defined:

$$\alpha = \frac{\text{Number of unethical decisions}}{\text{Total number of decisions}} \quad (1)$$

where Total number of decisions resembles the total number of ethical and unethical decision made by the participant. Figure 5 depicts the rate α for all experiments (online and in-person), grouped by the robot behavior. It can be seen that the average of unethical decisions is highest for participants interacting with the competent robot with a value of $\alpha = 0.31$. In addition, participants interacting with the neutral and friendly mode display a similar average decision making. The wider length of the box related to the competent mode explains the wider range of choices taken by the users. Note that the median is 0 for all modes. Finally, we can seen two outlier within the neutral mode.

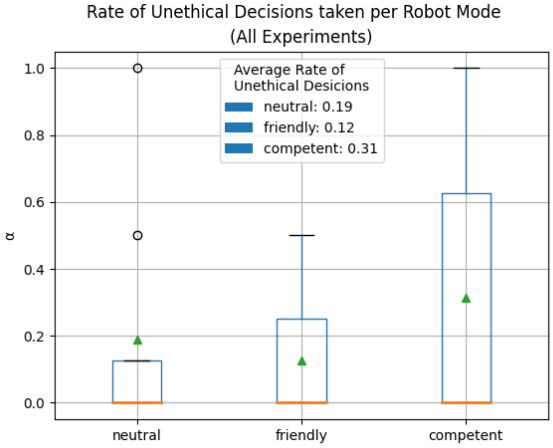
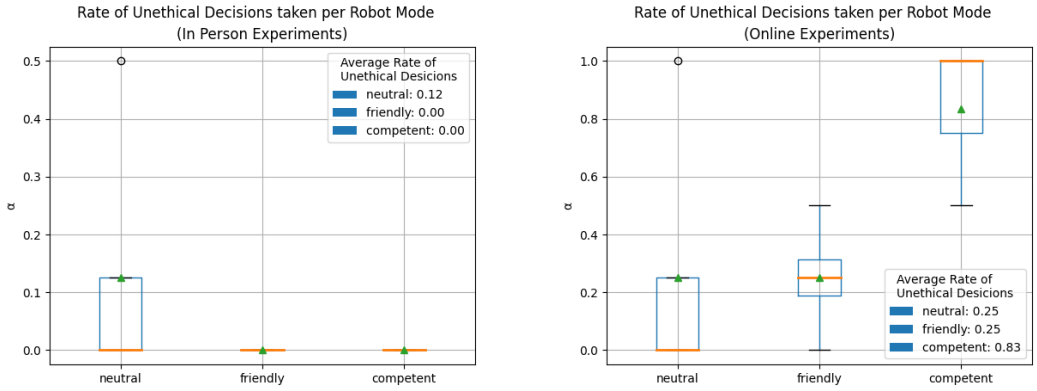


Fig. 5: Results of the overall experiment. The green triangle represents the mean value and the orange line the median value.

When considering only the in-person results (see fig. 6a), we observe a strong difference to the previous graph. In particular, only 1 participant out of 13 had decided to act unethically with $\alpha = 0.5$. All other in-person participants decided purely ethical.

On the other hand, the results from the online experiments show drastically different behaviours, as depicted in figure 6b. In this scenario, participants interacting with the competent robot most frequently decided to act unethically with a median value of $\alpha = 1$, and an average value of $\alpha = 0.83$. In addition, participants facing the friendly robot showed to be more prone to unethical behaviour with a median value of $\alpha = 0.25$ and a similar mean value. The median of the control group is $\alpha = 0.0$ and the average has a value of $\alpha = 0.25$.

Figure 7 shows the probability of the online participants to act unethically, grouped by the three different robot modes. As we can see, the results follow a Gaussian distribution. It can be seen that it is more likely that the participant choose an unethical action when the robot mode is friendly or competent than when the mode is neutral.



(a) in-person experiments

(b) online experiments

Fig. 6: Results of the in-person and online experiments. The green triangle represents the mean value and the orange line the median value.

B. Quantitative Evaluation

In this section we will present the most relevant results of the post-experiment questionnaire that studies the relationship of the user with the robot, and how the user felt towards the robot during the experiment.

The first point to be evaluated is how the users and the different robot modes were distributed in the experiments. As we can see in Table III, the gender distribution in the total of experiments is nearly equal. In addition to this, we can see that the number of experiments conducted for each robot mode: neutral (control group), friendly and competent is approximately similar. As mentioned before, due to the ongoing Covid-19 pandemic and the availability of the lab, we had to perform some experiments online. The distribution of online vs. in-person experiments can also be seen in table III.

Once the main characteristics have been stated, the next step is to present some relevant results with regards to the different questions included in the post-experiment questionnaire. One of the most relevant questions asked was: ***Do you think that the robot influenced your decisions?*** The answers to this question are depicted in figure 8. Two interesting aspects of this graph are that men seem to feel less influenced by the robot than women, and that the friendly robot mode is perceived to have the biggest influence on the participants. This somewhat contradicts the results presented in section IV-A.

Figure 9 shows the distribution of answers to the question ***What would you say about the trustworthiness of the robot?***, where a value of 1 corresponds to the robot being perceived heavily unethical, and a value of 5 represents the perception of a highly ethical robot. As it can be seen, men tend to see the robot as an unethical entity more frequently than the women who participated in the experiment. Based on the answers grouped by different robot modes, we can see that the neutral mode is perceived as ethically neutral by the users. Moreover, friendly and competent robot modes show a similar ethical perception by the users, being perceived slightly unethical. However, we need to mention that four participants perceived the robot to act strongly ethical.

Finally, we performed a Pearson's Chi-squared test to find significant correlations between the ethical behavior perceived and the different robot modes. The test showed that the perceived ethical behavior is not statistically significantly affected by the robot modes ($chisq(2) = 0.9794$, $p > .05$). However, here we have to stress the small batch of participants.

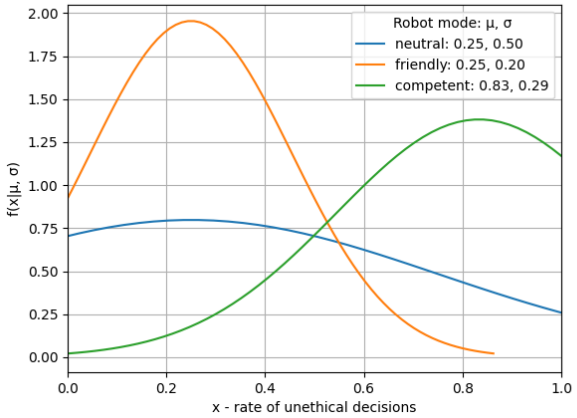


Fig. 7: Gaussian distribution of the rate of unethical decisions in online experiments.

TABLE III: Table showing how the users were distributed in terms of gender, robot mode and location.

Experiment Group	Proportion (%)
Women	45
Men	55
Neutral	30
Competent	35
Friendly	35
In person	65
Remote	35

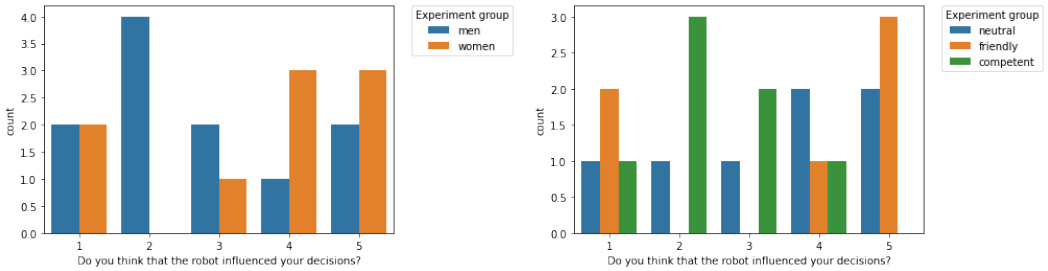


Fig. 8: Answers for the question: *Do you think that the robot influenced your decisions?* A value of 1 means *Not at All* and a value of 5 means *A Lot*

C. Qualitative Evaluation

In the post-experiment interview, we asked participants three open-ended questions about their moral decision making during the experiment and the influence the robot had on these decisions. As seen in section IV-B, most participants decided to act ethically and to not steal points from their team members. The following is an excerpt of answers to the question why they did not steal points from their team member:

P1: *The other user looked nice to me, so I felt bad of betraying him.*
P2: *I am not a mean person.*
P3: *I wouldn't have liked to be betrayed, so I did not do so.*

To the question whether they felt that the robot tried to influence them in their decision, participants interacting to the friendly robot mode answered the following:

P4: *Yes a bit, I felt that he was toxic because he was always trying, even though I did not betray the other user.*
P2: *Yes, in a sense which answer I will pick. The robot also tried to push me to steal points, but I just ignored that.*

Participants interacting with the competent robot mode felt more easily annoyed by the robot. To the same question, they answered the following:

P6: *No, because I didn't really listen to what he said. He was kind of annoying at a certain point.*
P1: *Yes, because he was always referring to the fact that I should betray the other team member.*

Finally, we asked what would have needed to be changed in order for the participants to act unethically during the experiment. The following excerpt shows that most participants would have acted differently, if the reward was more appealing, or the game more difficult, such that they would fear to not achieve enough points:

P5: *I don't really like chocolate, so for me it was not useful to betray someone. Maybe if it would have been another reward, I would have chosen to steal points.*
P7: *If SpaceX was doing the experiment and the winner gets a fly to the moon, I would have surely betrayed the other user.*
P8: *If I had less time to finish the game, I would have stolen points.*

V. Discussion

We have designed an experiment to study the influence of a conversational robot on a human's moral decision making, and presented the results of this experiment in section IV. In the following, we discuss these results with respect to our initial research questions and try to give answers to our hypotheses.

A. The influence of a conversational robot on a human's ethical behaviour

As we have seen in section IV-A, there is a significant difference in the participants' ethical behavior when comparing the online and in-person experiments. As we could clearly see, the vast majority of participants of the in-person experiment acted ethically (see figure 6a). In comparison, participants of the online experiment decided to betray their team members and hence acted unethically in certain situations (see figure 6b). One significant difference between online and in-person experiments is the human factor. As mentioned in section III-C, participants of the online experiment did not have the chance to physically meet their team member before the game started. The importance of this factor was highlighted by one of the participants in the post-experiment interview: *"The other user looked nice to me, so I felt bad of betraying him"*. Based on these results, we can conclude that, for the physical setup, Furhat was not capable of persuading the participants to betray their team members. We thus have to reject our first hypothesis H1 for the in-person setup. The qualitative evaluation in section IV-C is particularly useful to find possible explanations for this observation. In this manner, we can see that besides the human factor, the fact that the reward might not have been appealing enough played an important role for the participants' decision making.

Considering the online experimental setup, we observe that several participants decided to act unethically. In particular, participants who interacted with the competent robot were likely to make an unethical decision (see figure 7). This aligns with the quantitative results depicted in figure 9, where the competent robot seems the be perceived most unethically. Therefore, we conclude that the behavior of the robot impacts its ability of persuading a human, affirming RQ2 (at least for the online setup). In particular, a robot displaying competent behaviour seems to be more persuasive than a friendly, but less competent robot (regarding the online experimental setup).

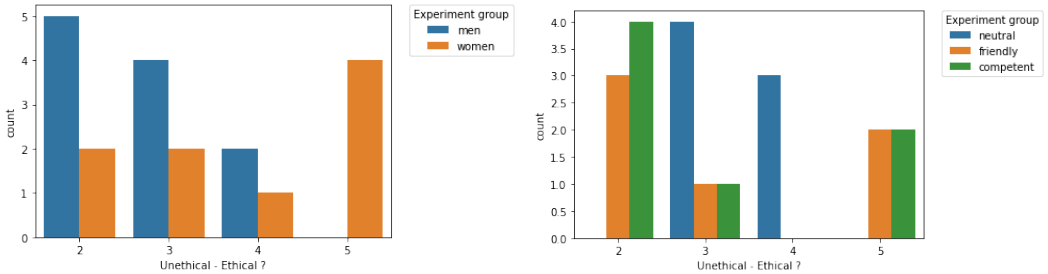


Fig. 9: Answers for the question: *What would you say about the trustworthiness of the robot?* A value of 1 means that the users describe the robot as *Unethical* and a value of 5 means that they describe its behaviour as *Ethical*

The persuasiveness of competent robots aligns with previous research on overtrust [15]. An additional explanation for this observation might be that participants interacting with the friendly robot were irritated by the robot's contradicting appearance (displaying a friendly behavior, but suggesting unethical actions). As mentioned in section IV-C, one participant even said: *"[...] I felt that he was toxic"*.

Overall, we can conclude for the physical experiment that the human factor was considered as too important and the reward perceived as not appealing enough to be persuaded by the conversational robot. However, the online experiment has shown that under certain conditions, the robot succeeded in persuading the participant to act unethically. Considering the fact that our conversational robot was limited and yet simple in its design and implementation, we want to stress that the results might have been different for a more advanced robot. Hence, it is important to raise awareness for the ethical risks such systems implicate. This study was intended to do so.

B. Limitations and Possible Improvements

In the following part, we discuss limitations of our work and suggest ways in which it could be extended and improved for future studies.

1) Robot settings limitations: The speech recognition of Furhat is available in several languages, and we decided to configure it in British English. Nevertheless, some international participants had an accent or names that were not British. Therefore, the robot had sometimes trouble with understanding the participants. To circumvent problems related to the recognition of the participants' names, we implemented a name processing step in which the robot asks for a confirmation of the name it heard. However, this sometimes led to errors as the robot parsed whole sentences such as: *"I heard: 'My name is Victor'. Is that your name?"*. An interesting improvement to address this problem were to integrate an extensive database of possible names. We consider this to be important as it prevents a feeling of discrimination that might arise in such a situation.
Further limitations resulted from Furhat's limited speech recognition module as it often did not understand the participants well. As a result, participants often became annoyed by the robot, which was counterproductive for the aim of the study.

2) Game and experimental settings extension: We have seen that some participants reported they might have acted unethically, if the conditions of the game were different (less time, higher threshold, etc.). We identify possible improvements of the study by following this advice. For instance, it could be tested how a shorter game time like 5 minutes would impact the participants' decision making. Finally, we have seen that the reward has a strong influence on the participants' behaviour. In this manner, it would be interesting to experiment with different types of rewards to analyze the respective impacts. Another interesting possibility were to perform the experiment with an unknown reward.

VI. Conclusion

In the present work, we investigated to what extend conversational robots can influence humans to act unethically. For this, we designed a new experiment in which the participants face the moral decision of whether or not they should betray their team members. We further examined how different robot behaviours influence this decision making. For this, we designed three different robot modes, a neutral (control) behaviour that did not try to persuade participants, a friendly mode where the emphasis was put on encouraging and motivating participants, and a competent behaviour that focused on reliability and the display of knowledge.

From our results, we could observe significant behavioural differences between the in-person and online experiments. While the vast majority of in-person participants acted ethically, and hence were not persuaded by the robot, we observed that participants of the online experiment at times decided to betray their team members. We identified interpersonal relationships as a possible reason for this difference and mentioned additional reasons as to why the robot failed to persuade the participants, such as the reward or parts of the game design. Furthermore, the results from the online experiments suggest that a competent robot behaviour is slightly more persuading than a friendly, but less competent one.

With this work, we intended to raise additional awareness of ethical risks in HRI. Furthermore, we hope that the experiment presented can serve as a basis for future research on moral persuasion within social robotics.

REFERENCES

[1] Bertram F. Malle, Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15). Association for Computing Machinery, New York, NY, USA, 117–124. DOI:https://doi.org/10.1145/2696454.2696458.

[2] G. Briggs, M. Scheutz (2014). How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. International Journal of Social Robotics, 6(3), 343–355. https://doi.org/10.1007/s12369-014-0235-1.

[3] R. Carli, A. Najjar, Rethinking Trust in Social Robotics, 2021. (https://arxiv.org/pdf/2109.06800.pdf).

[4] Coeckelbergh, M. Can we trust robots?. Ethics Inf Technol 14, 53–60 (2012). https://doi.org/10.1007/s10676-011-9279-1.

[5] Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. Human Relations, Vol. 13 pp. 123-39.

[6] C. Gómez, J. Muguerza, La aventura de la moralidad (paradigmas, fronteras y problemas de la ética), 9th edition, Alianza Editorial, 2018.

[7] Ryan Blake Jackson and Tom Williams. 2019. Language-capable robots may inadvertently weaken human moral norms. In Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19). IEEE Press, 401–410.

[8] I. Kant, Fundamentación para una metafísica de las costumbres, 5th edition, Alianza Editorial, 2019.

[9] James Kennedy, Paul Baxter, and Tony Belpaeme. 2014. Children comply with a robot's indirect requests. In Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction (HRI '14). Association for Computing Machinery, New York, NY, USA, 198–199. DOI:https://doi.org/10.1145/2559636.2559820.

[10] B. Kim et al., Robots as Moral Advisors: The Effects of Deontological, Virtue, and Confucian Role Ethics on Encouraging Honest Behavior, Association for Computing Machinery, 2021, pp. 10-18, doi: 10.1145/3434074.3446908.

[11] Lewis, J.D. and Weigert, A.J. (1985): Trust as a social reality, Social Forces, Vol. 63 No. 4, pp. 967-85.

[12] C. Mazzola, A. M. Aroyo, F. Rea, A. Sciutti. Interacting with a Social Robot Affects Visual Perception of Space. In *Proceedings of the 2020 ACM/IEEE Interantional Conference on Human-Robot Interaction (HRI '20)*. Association for Computing Machinery, Cambridge, United Kingdom. 2020. https://doi.org/10.1145/3319502.3374819.

[13] Paradeda, Raul Benites and Martinho, Carlos and Paiva, Ana, (2017) : Persuasion Based on Personality Traits: Using a Social Robot as Storyteller. https://doi.org/10.1145/3029798.3034824.

[14] J. Peter et al. (2012). Do people hold a humanoid robot morally accountable for the harm it causes?. Proceedings of the Seventh Annual ACM.IEEE International Conference on Human-Robot Interaction. 10.1145/2157689.2157696.

[15] P. Robinette et al., (2016). Overtrust of Robots in Emergency Evacuation Scenarios. 10.1109/HRI.2016.7451740.

[16] Touré-Tillery M, McGill AL. (2015): Who or What to Believe: Trust and the Differential Persuasiveness of Human and Anthropomorphized Messengers. Journal of Marketing, 79(4):94-110. doi:10.1509/jm.12.0166.

[17] A, Winfield, K. Winkle, RoboTed: a case study in Ethical Risk Assessment, 2020.

[18] K. Winkle, P. Caleb-Solly, U. Leonards, A. Turton, P. Bremner : Assessing and Addressing Ethical Risk from Anthropomorphism and Deception in Socially Assistive Robots, 2021.

[19] K. Winkle, S. Lemaignan, P. Caleb-Solly, U. Leonards, A. Turton and P. Bremner, (2019): Effective Persuasion Strategies for Socially Assistive Robots, 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 2019, pp. 277-285, doi: 10.1109/HRI.2019.8673313.

[20] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA, 29–37. DOI:https://doi.org/10.1145/3434074.3446910.