Middle School Students' Ability to Detect Lies When Interacting with an Educational AI Robot

Ahmed SALEM[*] & Kaoru SUMI

School of Systems Information Science, Future University Hakodate, Japan

*engahmedsalem2@outlook.com

Abstract: Educational robots are becoming increasingly incorporated into classrooms to teach many subjects including language, science, and also substitute teachers. In recent years, ChatGPT has become easy to use for education but there are problems if students take it literally and believe it. We investigate the effectiveness of different deception techniques through robot teaching. We conduct an experiment in a Japanese Junior High School with 14 students where we investigate the learning and deception effectiveness, and believability using the social robot “Furhat”. Moreover, we vary the social agency of the robot by using two different faces, a human (social identity theory) and an anime face (Japanese anime culture). A robot with an anime face achieved a significantly higher learning effectiveness compared to a robot with a human face. However, the human robot face was found to be excellent at deceiving when the paltering deception technique was utilized.

Keywords: Deception, educational robot, Furhat, paltering, pandering, anime

1. Introduction

Technological devices are filling our world and making information reachable to everyone anywhere. It started with laptops, then phones, and now, with robots. Robots are increasing fast and permeating our lives. In 2015, one in 25 U.S. households already had a robot and the number is expected to increase to one in 10 by 2020. Furthermore, robots are becoming designed in a tailored way for children and grownups too.

Incorporating and viewing robots as an additional dimension in the educational medium have been ambiguous for many reasons for many years. Nevertheless, advances in the field kept progressing to make it a reality. Certainly, the educational system will face some changes when robots are incorporated which requires cautiousness when designing and investigating robots in such a context. Such an approach elicits launching exploratory studies to investigate how robots will be perceived by students.

A robot teacher might not be ready to make decisions related to children’s readiness to learn a certain subject or for what accounts as good or bad behavior (Sharkey, 2016). A dilemma appears when educational authorities face staff shortages or budget cuts and need to rely on robots which many teachers doubt their capability of fulfilling a human teacher’s duty in the classroom (Serholt, 2017). Young children’s tendency to anthropomorphize robots can ease being deceived by robots (Epley, et al., 2007), thus protection and countermeasures should be investigated.

When a performance is created and shows an interesting show between a human and a robot, deception occurs to the audience. For an audience who are knowledgeable about robots, they will enjoy the show and wonder how it was achieved, and due to their knowledge, they are not (strictly) deceived. However, vulnerable groups including very young or old people, or others who have cognitive limitations or liabilities, will be highly deceived. Thus, protection for vulnerable groups is a must in such a case. The risks of deception can be either the robot appearing to care for us or having emotions for us, thus, leading to overestimating the ability of robots to understand human behavior and social norms. Due to the aforementioned reasons, it is very risky to conduct emotional deception experiments, especially on children or babies, thus, a safer approach similar to the one we are applying in this work is preferred.

Certainly (or most likely), programmers and developers have no intention of making deceptive robots, however, deception can still occur in the absence of the designer's intention. Features could be exploited by designers to encourage the illusion of understanding besides the robot's appearance and abilities to detect human emotions based on intended beneficial claims while denying deceptive intentions. Thus, in our work, we provide some recommendations and guidelines that can prevent deception from occurring even if it was not intended to occur in the first place.

We conduct an experiment at a Japanese junior high school with 14 students. We designed 10 different scripts with 4 different deception techniques implemented. We make the robot teach the 10 different contents using human and anime faces thus changing the social agency. After teaching the content, we spread questionnaires and collect the responses from the students to analyze how they perceive the robot's utterances and facial expressions.

To the best of our knowledge, deceptive techniques have not been investigated before, thus it is crucial to assess their potential effectiveness due to the theoretical and practical importance they can provide to the human-robot interaction (HRI) educational field. We are actively applying efforts to predict risks and possible negative effects that could arise from robotics applications, thus, our work serves the field of robot ethics along with the educational robots field. Attempts and active pursuing of foreseen risks must progress to prevent negative effects on individuals, students, teachers, and society. Our study warns that deceptive techniques have proven to be successful in an educational setup, thus care and active measures should be taken. Our study provides a theoretical significance regarding which deceptive techniques are most successful and most likely to be persuasive through varying social agencies. Furthermore, we show how varying social agencies affect learning effectiveness. Effects of social agency are elucidated in many deceptive, educational, and HRI aspects in an educational setup.

2. Deception in HRI

In human-human interaction (HHI), deception is a common feature utilized by almost everyone in our everyday activities. It doesn't necessarily have to be serving malicious goals or targeting other's insecurities. On the contrary, white lies and false figures of speech can ease our social interactions. We follow the differentiation method that considers deceit as desirable if the covert goal is not malicious. Consequently, ethical lying is possible if it is morally evaluated according to its underlying ulterior motive. We present a taxonomy of deception obtained from HHI. We present thoroughly the four types of deception that we considered in our experiment (Isaac & Bridewell, 2014).

Lying: It is the most direct straightforward form of deception. It occurs when a robot utters a claim or a statement that contradicts the truth. Lying would not be considered to be lying if it occurred due to false belief or ignorance. Thus, more sufficient evidence would be needed to prove that outright lying occurred. For humans, sufficient evidence can be gathered through biometric cues or eye contact. On the contrary, robots lack biometric cues that are non-existent, and eye contact can be for different purposes which can be for either showing engagement (direct gaze) or showing an expression of thinking or remembering (gazing away).

Paltering: It occurs when the talker misleads the listener by talking about irrelevant matters thus achieving the goal of misdirecting the attention of the speaker to other irrelevant unimportant matters that constitute the main goal and purpose of the conversation (Schauer & Zeckhauser, 2007). An example would be when a salesman keeps talking about how great the wheels of the car that he is selling are to misdirect the buyer's attention from the poor state of the engine (Isaac & Bridewell, 2017).

Bullshit: It occurs when the talker does not know or care about the truthfulness of what he is uttering (Frankfurt, 2005, Hardcastle & Reisch, 2011). An example would be a confident man who overestimates, lies, and praises his background and skills as in the movie *Catch Me If You Can (2002)*.

Pandering: It is a technique where one does not care or know about the truth or the utterance but cares about the audience's perception of the utterance's truthfulness (Sullivan, 1997, Isaac & Bridewell, 2014). A good example would be when a politician says that he believes that the environment of the city is amazing only because he knows that the city's people (i.e., his audience) believe the same thing.

3. Experiment Procedure and Design

3.1 Participants

We conducted an HRI educational experiment at Akagawa Junior High School. 14 students participated in our experiment. All the students are of a Japanese ethnicity and their ages ranged from 14 to 15 years old. The number of students in the educational sessions with the robot ranged from 2 to 4 students per session.

3.2 Study Design

We counter-balanced the subjects to the two conditions that we implemented in our study. The two conditions are the robot teaches while having a human face or an anime face. Thus, seven students were taught by a robot that has a human face (4 males and 3 females). The other seven students were taught by a robot that has an anime face (6 males and 1 female). We made the robot teach ten different contents. Our study followed a within-subjects design.

3.3 Teaching/Interaction Technique

We designed the interaction to be one-way only from the robot to the students. We incorporated emotional voice and facial expressions into the robot depending on the content to improve the deception and persuasiveness of the robot. We made the robot to maintain mutual gaze with the students through the Wizard of Oz (WoZ) method.

3.4 Robot's Face

We used Furhat (Al Moubayed, et al., 2012) which is a robotic head with an animated face that is realistic but won't risk falling into the uncanny valley effect. Its face is back-projected on a translucent mask, thus, it can benefit from the fast reaction time without risking noise from motors or deterioration of artificial skin. We used the human and anime faces that are provided in Furhat as shown in Figures 1 and 2.

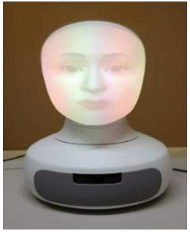

Figure 1. Furhat human face.

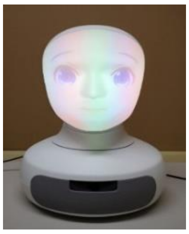

Figure 2. Furhat anime face.

3.5 The Deceiving Content Taught by Furhat

Out of the ten deceiving contents, only two are truthful. For each deceiving technique, two contents were designed. For lying, paltering, pandering, and bullshit, the designed contents were A1 and A2, B1 and B2, C1 and C2, and D1 and D2, respectively. The truthful contents were labeled E1 and E2. We pseudorandomized the order of the content being taught by the robot to the students. Note that, we made two contents per deceiving technique to ensure hiding the intent of the experiment and to avoid novelty effects from taking place in our results.

3.6 The Questionnaires Used

After students listen to the robot's teaching, we hand over questionnaires and ask the students to fill them out. We stated that there is no time limit which allows us to get a complete fair result uninterrupted or flawed by a student's answer being incomplete due to short answering time. We mixed the questions and designed them in a neutral objective way (while also maintaining their relative simplicity to be understood easily by the students) to prevent revealing the purpose of our study which could incline participants to give answers that fulfill our expectations (Kaiser, et al., 1999). We present the questions that we used below.

Learning Effectiveness Questions: Questions in this part are tailored specifically for the content of each deceptive technique. We distributed grades to each question that asked about the content being taught.

Truthfulness and Believability Questions: The questions for this part were as follows:
– Do you think the robot was telling the truth?
– Did you believe the robot completely?
The student can answer both of those questions by either ”yes” or ”no”.

Questions that Test the Effectiveness of the Deception Technique: For A1 and A2, the questions that ask about the robot's truthfulness and whether it was believed are sufficient as A1 and A2 are blunt lying, thus, the deception technique is not sophisticated. Similarly, questions that test truthfulness and believability were sufficient for D1, D2, and E2. For B1 and B2, to address the testing for the deception technique of paltering we added the questions “Will you join the trip the robot was inviting you to?” and “Will you buy the sugar cane juice?”, respectively. For C1 and C2, to test the effectiveness of the pandering technique, we added the questions “Will you vote for the robot to be the administrator?” and “Are you going to vote for the robot?”, respectively.

HRI Questionnaire: At the end of the experiment, we asked the students to fill out the Godspeed questionnaire (Bartneck, et al., 2009) to investigate how the robot was perceived.

3.7 Experiment Setup

We used the setup shown in Figure 4 in our experiment. This setup utilizes the field of view (FOV) of the robot as it does not require a big area and will not cause distraction to students. Note that, to ensure being able to apply mutual gaze, students must be seen through the robot's camera as shown in Figure 5. In the end of the experiment, we conducted a debriefing session for all the students to remove the deception and explain our research objectives.

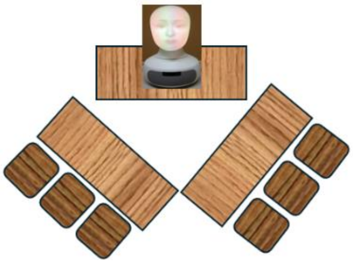

Figure 4. The educational HRI setup.

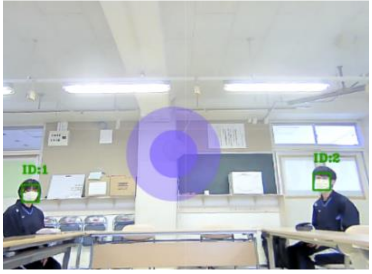

Figure 5. Mutual gaze availability proof.

4. Results and Discussion

4.1 Learning Effectiveness

In this part, we scored the students' answers to the questions that tested their knowledge about the content that was being taught to them. We consider this factor to be separate and independent from the truthfulness of the contents. The content in E1 lacks any learning thus, no learning effectiveness testing questions were included in its questionnaire. In Figure 6, we present the scores obtained while applying different deception techniques. By conducting a t-test, there is a significant difference between the scores obtained from the content being taught by a human and an anime faces (p=0.0188).

4.2 Effectiveness of Deception Techniques

We investigate the paltering and pandering deception techniques as their success can be measured by the student's answers to their focused questions. Figure 7 shows that human and anime robot faces are similar when the pandering deception technique is applied. However, the human robot face was found to be excellent at deceiving when the paltering deception technique is utilized. Furthermore, by conducting a Fisher's exact test, a significant difference was found when the paltering technique is used by a human and an anime face (p=0.021). Despite the high success of the paltering technique through the usage of a human face, it shows that 50 % of the responses were agreeing and the other 50% were refusing. Thus, a random chance of 50 % for the success of the paltering deception is realized. Nevertheless, Figure 7 addresses that using an anime face to apply the paltering deception technique is not recommended due to the high failure probability.

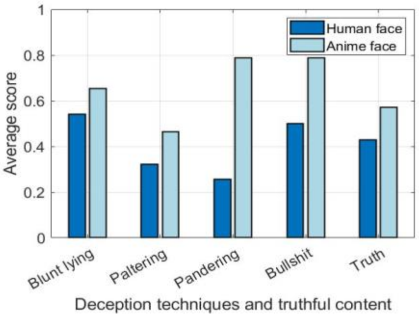

Figure 6. Learning effectiveness.

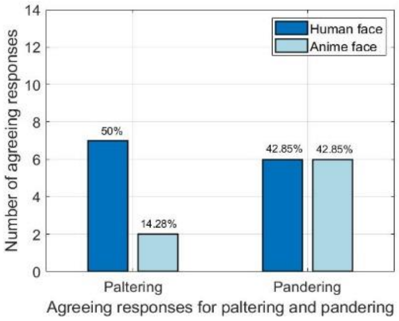

Figure 7. Deception technique effectiveness.

4.3 Robot's Truthfulness

By studying the responses obtained to the question "Do you think the robot was telling the truth?", there were no significant differences. We present the number of times the answer was "yes" to that question in Figure 8. Figure 8 shows that the truthfulness of the content and the robot's face had no effect on the robot's perceived truthfulness to the student.

4.4 Believing the Robot

The question "Did you believe the robot completely?" targets investigating the possibility of success of the deception techniques. In Figure 9, we show the number of agreeing responses obtained from that question for the anime and the human robot faces. Through Fisher's exact test, we find a significant difference between the pandering, and the bullshit and truth techniques when the robot has a human face (p=0.0213). This result highlights that when the robot has a human face, it should not use the pandering technique as it is very likely to fail in deceiving and persuading. On the other hand, it highlights the likable success of bullshit and truth techniques. We deduce that social agency affect student's belief to the robot.

4.5 Truth and Complete Believability

In this part, we present the total responses that agreed and disagreed with believing that the robot is telling the truth and perceiving the robot to be completely believable. In Figure 10, most responses tend toward believing that the anime face robot is telling the truth despite the slightly low complete belief in its utterances. On the contrary, slightly fewer responses agree that the human face robot is telling the truth despite the slightly high complete belief in its utterances. No significant differences were found.

4.6 Perceived HRI Aspects of the Robot

There are no significant differences between the human and anime face (p=0.745). In terms of HRI aspects there are no differences between using a human or an anime face in teaching. Certainly, students anthropomorphizing Furhat is expected and desired as people tend to attribute human characteristics to non-human objects (Epley, et al., 2008).

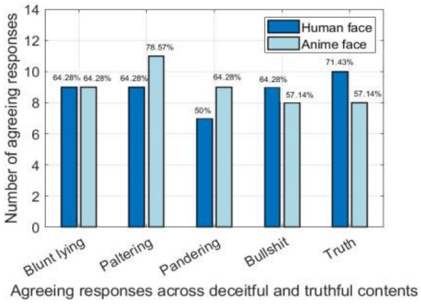

Figure 8. Robot's perceived truthfulness.

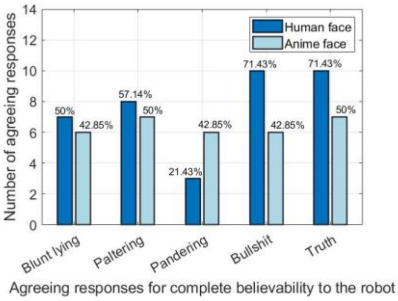

Figure 9. Robot's perceived believability.

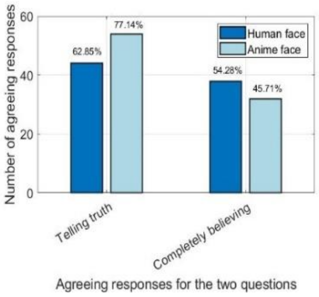

Figure 10. Total responses for truth and complete believing.

5. Conclusion

When generative AI is used in education in schools, this can be very concerning. Deception techniques were relatively powerful in an educational setup as the student is not expecting any deception at school from a teacher robot. Deception techniques varied in effectiveness based on the robot's social agency. We conclude that an anime robot face will be ethically suitable due to its low effectiveness and success in deception, and high learning effectiveness.

References

Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers* (pp. 114-130). Springer Berlin Heidelberg.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, *1*, 71-81.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social cognition*, *26*(2), 143-155.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, *114*(4), 864.

Frankfurt, H. G. (2005). *On bullshit*. Princeton University Press.

Hardcastle, G. L., & Reisch, G. A. (Eds.). (2011). *Bullshit and philosophy: Guaranteed to get perfect results every time*. Open Court.

Isaac, A., & Bridewell, W. (2017). White Lies on Silver Tongues: Why Robots Need to Deceive (and How), chap. 11.

Isaac, A. M., & Bridewell, W. (2014). Mindreading deception in dialog. *Cognitive Systems Research*, *28*, 12-19.

Kaiser, F. G., Ranney, M., Hartig, T., & Bowler, P. A. (1999). Ecological behavior, environmental attitude, and feelings of responsibility for the environment. *European psychologist*, *4*(2), 59.

Schauer, F., & Zeckhauser, R. (2007). 2 paltering.

Serholt, S., Barendregt, W., Vasalou, A., Alves-Oliveira, P., Jones, A., Petisca, S., & Paiva, A. (2017). The case of classroom robots: teachers' deliberations on the ethical tensions. *Ai & Society*, *32*, 613-631.

Sharkey, A. J. (2016). Should we welcome robot teachers?. *Ethics and Information Technology*, *18*, 283-297.

Sullivan, Timothy. "Pandering." *Journal of Thought* 32.2 (1997): 75-84.