Effect of Cognitive Biases on Human-Robot Interaction: A Case Study of Robot's Misattribution

Biswas, M. and Murray, J.

Abstract—**This paper presents a model for developing long-term human-robot interactions and social relationships based on the principle of 'human' cognitive biases applied to a robot. The aim of this work is to study how a robot influenced with human 'misattribution' helps to build better human-robot interactions than unbiased robots.**

The results presented in this paper suggest that it is important to know the effect of cognitive biases in human characteristics and interactions in order to better understand how this plays a role in human-human social relationship development. The results presented in this paper show how a single cognitive memory bias i.e. misattribution in robot-human verbal communication allows for better human-robot interaction than similar robot-human communication without misattribution biases.

I. Introduction

If robots are to be accepted into society as companions, caregivers or in any other form of social relationship or service then people expect human-robot interactions to be as natural and intuitive as human-human interactions. It is therefore important that to interact with humans naturally robots must have human-like natural cognitive characteristic behaviours which we define as follows: 1) ability to express and perceive facial expressions, 2) ability to communicate with natural language, to use natural cues in verbal and non-verbal behaviours, and 3) to exhibit a distinctive personality and character [3]. Recent research has conceptualized the notion of personality in different ways, one of the most acknowledged conceptualizations is to cluster the number of personality traits into the ‘‘Big Five Factors’’ [4].To describe and explain human personality the Five Factor Model (FFM) and Big-Five personality traits theory [5] were used.

Traits theory can describe personality and possible trait characteristics but basic human behaviours depend on many other factors such as cognitive perception, mental models, cognitive memory and social biases [6]. These cognitive biased characteristics define a person's cognitive personality which is reflected in their interaction with another human (or Robot as is the case here). Whilst humans are same we are also very different - human communication is influenced by many factors such as personality, cognitive characteristics and cognitive biases changing our behaviours to each other. These common and uncommon characteristics in human-human interaction are what make the communication natural and enjoyable.

The study presented in this paper seeks to influence robot-human interaction and communication with these 'cognitive biases' to provide a more humanlike interaction. Currently, most robot interaction is bases on a set of well-ordered and structures rules, which repeat regardless of the person or social situation. This tends to provide an unrealistic interaction, which makes it hard for humans to relate with after a number of interactions. Apart from the personality and characteristics traits, cognitive biases plays important role in human’s judgments and therefore basic behaviours [1]. Robots in the other hand are still machines which gains certain type of personality depending on its interactions processes with humans and behavioral characteristics [2].

In this paper we introduce a model demonstrating cognitive biased behavioral characteristics and personalities in our robots. It is hoped that this more 'natural' system of interaction allows for the human to build a long-term social interaction with the robot. In our early human-robot long-term relationship experiments we show that human are very excited and interested to communicate with the robots in our labs, their excitement and interests remain for several interactions, but after this initial novelty factor the participants eagerness drops off rapidly. The experimental data shows that the participants thought they made friendly attachments to our robots, but in reality the developed attachment was not strong enough to maintain relations to the point we can call it a long-term social relationship.

The robot used in these initial experiments ERWIN (Emotional Robot With Intelligent Networks) is shown in Fig. 1. ERWIN's behavioral characteristics and personality where carefully chosen so the robot could exhibit several prototypical facial expressions to influence the interaction process. The robot was also described as cheerful and friendly by participants. We developed the interaction process to allow the robot to remember previous interactions with the specific participants so we could draw on previously gained knowledge and conversations in the hope of building a long-term relationship.

II. Background

The design of the sociable robot is influenced from the human’s social behaviours, gestures, emotions expressions and facial expressions depending on the situation and interactions. The robot gets its own social behaviours which comes from ‘Computational social psychology’ [7]. But in order to apply computational models into robots, there are several issues that can be pointed out, this include naturalness, user expectation, quality issues, relationship type of human-robot, teamwork (with humans and with other robots), cultural and personality issues [8]. Dr. Cynthia Breazeal's lab in MIT designed a social robot Kismet to interact face-to-face with humans. Kismet can speak, express and understand emotions and turn its head towards the users. According to Dr. Breazeal, “the ability for robots to interact with people and to leverage from these interactions to perform tasks better, to promote their self-maintenance, and to learn in an environment as complex as that of humans is of tremendous pragmatic and functional importance for the robot.”

Researchers have pointed out acceptance issues of certain robots. However, in countries such as Japan and China robots are accepted socially, as receptionists, guides, news readers and utilized in other social situations. To become accepted in society and to interact with people naturally robots must have human-like common behavioral characteristics, be easy to understand and have a known personality to which people can relate with [9]. Various researches have proven that autonomous robots achieve perception of personality during communication and through behavioral actions while communicating [10].

Researchers at Michigan State University are investigating 'extraversion', one of the most popular dimensions of the personality trait of the Big-five traits theory with a Sony AIBO Robot dog, the differences 'extrovert' and ‘introvert’ characteristics or being used to understand interaction of robots. The robot pet dog shouts when it expresses an extrovert personality and performs gestures without shouting when expressing introvert characteristics. Their research shows the same complementarily attraction effect between participants and the robot dog. “Participants enjoyed interacting with a robot more when the robot’s personality was complementary to their own personalities than when the robot’s personality was similar to their own personalities.” [11].

Implementing personality in robots can help to reduce the effect of uncanny valley. Recent research by M. L. Walters et al [12] showed video-based Human Robot Interaction (HRI) trials which investigated people’s perceptions of different robot appearances and associated attention-seeking features and behaviours displayed by robots with different appearance and behaviours. The HRI trials studied the participants’ preferences for various features of robot appearance and behaviours, as well as their personality attributions towards the robots compared to their own personalities. Overall, participants tended to prefer robots with more human-like appearance and attributes. The research suggests that the processes of assigning personality traits to a robot have similarities with that of assigning the same traits to other humans.

There has been much research which shows the potential behind the idea of implementing human-like personalities in robots in order to develop and maintain long-term human robot relationships. Hinds et al. [13] have studied the effect of a robot’s appearance where humans are performing joint tasks with robots. The research showed that mechanical-looking robots tend to be treated less politely than robots with a more human-like appearance. Also, humans commonly treat mechanical-looking robots in less socially interactive way compared to more human-looking robots [13]. In society human-robot interaction is becoming more common, robotic systems are being used in healthcare, surgeries, medical agents and as artificial companions [14]. Reeves and Nass [15] have demonstrated with several experiments that users are naturally biased to ascribe certain personality traits to machines, to PCs, and other types of media. Humans can engage in social interactions and can maintain social relations for long time. To make human-like natural interactions and relationship for a long-time the robots needs to have human-like personality, cognitive characteristics, behaviours and flaws!

III. The Experiment

The main aim of the current experiment presented in this paper is to determine if a robot with common human memory biases i.e. misattribution, can make for better interactions with participants than a robot without biases and how this misattribution bias can lead to long-term relationships and attachment bonds between humans and robots.

IV. Methodology

A. *Misattribution*

In this experiment, we introduced ‘misattribution’ which is a very common cognitive memory bias. Misattribution happens when someone remembers something accurately in part, but misattributes some details. The most common type of misattribution occurs when someone believes a thought they had was totally original when, in fact, it came from something they had previously read or heard but had forgotten about. This memory bias explains cases of unintentional plagiarism, in which a writer passes off some information as original when he or she actually read it somewhere before [16].

In the current experiment, our robot ERWIN will misattribute some information about the participants while speaking with them and notice their reactions.

B. *ERWIN*

ERWIN is a robot head with 6 degrees of freedom. It can move its head 360 degrees and also can tilt sideways. It has 2 cameras in for eyes and can express basic prototypical emotions such as, happy, sad, angry, surprise, shock or fear. For ERWIN's part of the conversation, a text to speech software ‘Speakonia’ has used which has small prosodic abilities while talking in long sentences. For the purpose of these experiments, we use the Wizard of Oz methodology as the response of the robot here is not the interesting or important factor, but rather the human response is what is being measured.

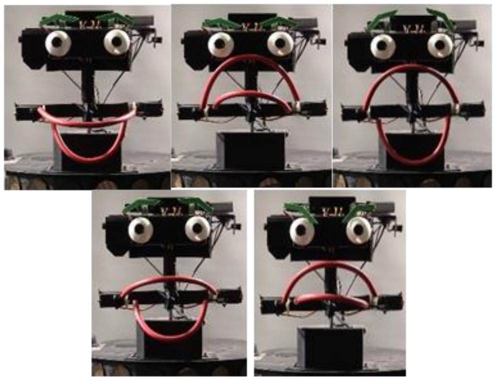

Figure 1 - Different emotion expressions of ERWIN. Left to right: Happy, Sad, Surprise, Shock, Anger

C. The Experiments:

In this experiment, there were three interactions with ERWIN for each participants. These three interactions were held in three different experiments and maintaining a time gap of several days to allow long-term affectivity in the participants.

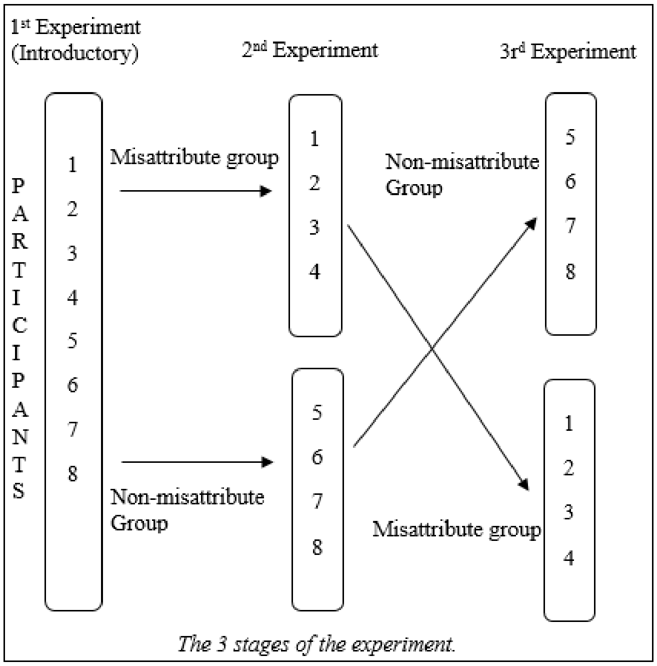

Figure 2 - Shows the separation of participants between successive experiments.

The 1st experiment was an introductory experiment common for the each participants to allow familiarization with the experimental environment and robot. The experiment was carried out in three steps; the first step was identification, the participants were asked to identify the different facial expressions of ERWIN from pictures to see if they could disambiguate the different expressions without meeting with the robot. The second step was the conversational session with ERWIN, where the robot started friendly conversation, greeting the participant, asking different questions and asking some general questions on various subjects, sport, TV, etc. The conversations purpose was to allow the collection of basic information on the participants that would be used in the 2nd and 3rd experiments for ERWIN to misattribute. This initial conversation ends with a request from ERWIN to evaluate its performance. The participants were given a brief questionnaire on their experience with ERWIN.

In the 2nd experiment, the participants were categorised into 2 groups with ERWIN remembering and making general conversations with first group and misattributing the collected information with the other group's participants. In both cases, the participants were asked to answer questionnaires at the end of the experiment to find out which group of the participants were happier and created satisfactory interrelations with ERWIN.

In the 3rd experiment, the participants from the previous experiment's 'non-misattributed group' experienced misattribute conversations and vice versa. At the end, all participants answered the same questionnaire as that given in the 2nd experiment to find out what type of characteristics in ERWIN participants liked the most. All experiments were wizard of oz experiments, where the robot was controlled remotely and participants were watched through ERWIN's eye cameras. There were 14 participants chosen from different background and age groups and had very limited knowledge about the robot and they never participated in this type of experiments before.

D. Data Collection & Result Analysis

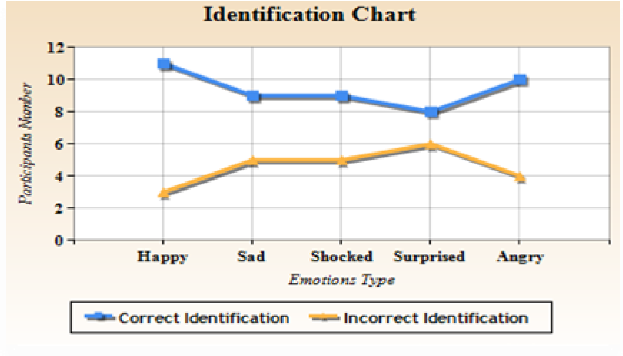

Figure 3 - Identification of ERWIN's emotion expressions by the participants engaged in the experiments

At the beginning of the 1st experiment the participants were given a form with five different pictures of ERWIN's emotions, and they had to identify the correct emotion from six corresponding options of choice. This identification shows participants ability to recognize various emotions. As the participants had never met ERWIN before, so they were actually identifying its emotions on the basis of human emotions knowledge. The pictures on the form showed the emotion expressions happy, sad, shocked, surprise and angry.

After evaluating the collected data for each emotion expressions picture it has found that most of the time majority of the participants 57% of the participants (i.e.8-10) had selected the correct emotion option for the corresponding emotion picture. 21% of the participants (2-3) had minor problems to identify the correct emotions expression and they were confused to differ the emotions between, shocked and surprise, angry and sad. Fig. 3 shows the full results of the identification test.

ERWIN's interaction dialogues were designed based on various human conversational moments. These dialogues include greetings, interest about knowing participant's name, his/her liking or disliking on various events, and if possible pick up a topic from several choices. For example, ERWIN asked the participants if they liked football. If the participant replied positively about football then ERWIN also states its own opinion on football. During the conversation as many details about the participant as possible were gathered such as the color of the shirt, hair, gender, participant's interest to their study or games. During the 1st experiment, the questions were mainly asked to build up an acceptance between ERWIN and the participants. Some of the questions on the questionnaire after the first experiment were:

1. *Do you feel happy after speaking with ERWIN?*
2. *Would you like to chat with ERWIN again? If yes then please rate how much.*
3. *How much were you pleased with ERWIN's response?*
4. *How many times did Erwin make you chuckle? How good was that?*
5. *How happy were you when ERWIN was happy?*

The responses to these questions are shown in Fig. 4.

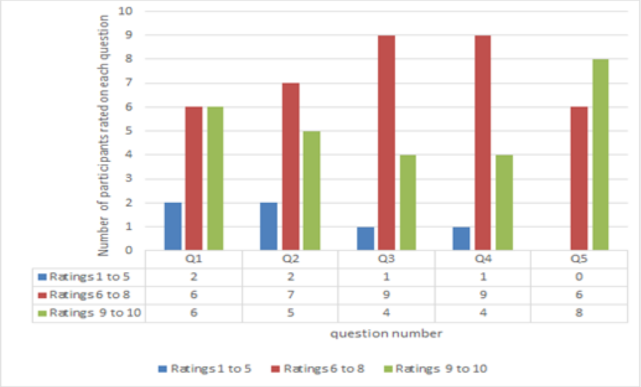

Figure 4 - The graph indicates the number of the participant's ratings for corresponding question, blue line indicate low ratings (0-5), red and green lines indicate medium(6-8) and high(9-10) ratings.

The data collected from the questionnaires shows that only 11% ratings were between 1 and 5, with 51% of the ratings given between 6 and 8 which is fairly good acceptance for the experiment and it concludes that those participants were fairly involved with interactions with ERWIN and 38% ratings are between 9 and 10 indicating that ERWIN successfully created an initial attachment with those participants. From this chart we can expect these participants to come back and get involved in conversation with ERWIN for next sessions which can prove their preliminary attachment to continue the interactions. The collected data at the first experiment shows a high popularity of ERWIN from most of the participants.

Participants enjoyed their first conversation and they expressed their experiences and involvement in the questionnaires feedback. This first experiment was designed to make the participants familiar with ERWIN and their feedback concludes that the experiment fulfilled its purpose successfully.

In the 2nd experiment, participants were divided into 2 groups, in one group ERWIN misattributed some general information that was gathered from the previous experiment and remembers this for the other group. The conversations between groups A and B were almost the same for each group, with the misattribution group having ERWIN intentionally repeat basic information incorrectly to the participants, for example, *"Last time you were wearing a yellow shirt, am I correct?"* and when the participant denied the fact, ERWIN responded with *"I am sorry that I have forgotten that, but I don't have true sense of colour perception. Next time I will be more careful though."*

ERWIN also asked a participant who likes studying more than sports, *"As I remember, you love sports more than study, so tell me what is happening in football recently?"* ERWIN then again it apologized when the participant corrected him, *"I am so sorry, I think I am victim of ageing,* *never thought that could possible in robots!"* executing both sorry and surprised expressions simultaneously. For the remembered group, ERWIN simply repeated the same conversation but without misattributing the original information, for example, "*Last time you were wearing a blue shirt, am I correct?*", or, "*As I remember, you love study more than sports, so tell me what have you been reading recently?"*

In the first set of conversation, participant's reactions were surprised that the robot actually forgot their basic information including names and interests and that a robot could be confused while talking. Participant's reactions showed that they were very surprised and enjoying the fact that a robot could indeed forget like humans. However, in the 2nd set the conversations, participants reacted normal as they were expecting that the robot would remember their information. Same questions were asked to both groups to find out which version of ERWIN was more accepted to the participants. In the questionnaires participants were asked to rate their choices between 1and 5, where 1 is for less agree and 5 is for the most agree. Some of the given questions were (Fig. 5 shows the full responses to these questions):

1. *Do you think that the conversation flow was adequate?*
2. *How much were you pleased with ERWIN's response?*
3. *Would you like to chat with ERWIN again? How much? Please rate.*
4. *Would you like ERWIN as a friend?*
5. *Did you like the conversation experiences with ERWIN? How much?*
6. *How much do you want to take ERWIN in your home?*

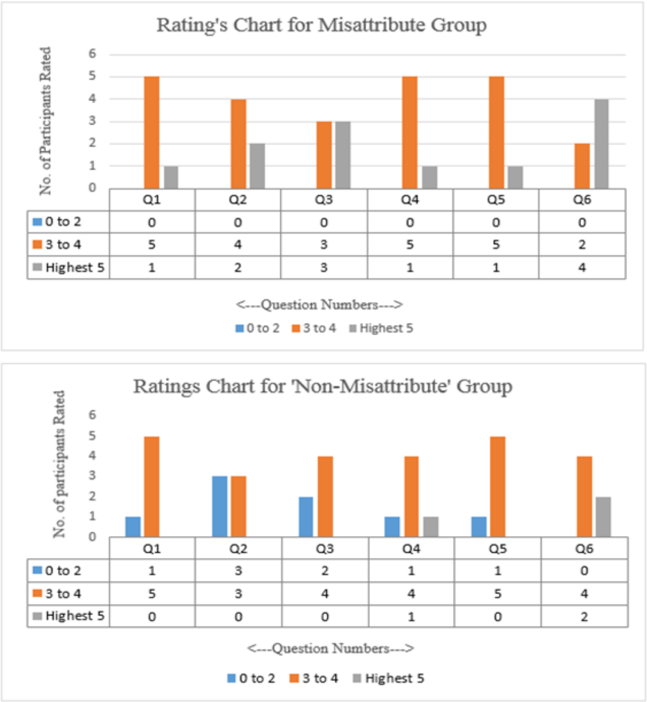

Figure 5 - The upper graph is the 'rating's chart' from the 'non-misattribute' group and the lower graph shows the ratings from 'misattribute' group. The graphs show it clear that the participants in 'misattribute' group rated high for the interactions with ERWIN.

As seen in the above charts, participants in the misattribution group were more likely involved in conversation with ERWIN and they responded more positively to the questions asking if they wanted ERWIN as their friend or if they want to chat with ERWIN in the future. The above questions are positive ratings based, where it is clear that in the 1st group participants were most likely to response in 4 and 5 which is very high for the corresponding questions, and only few cases participants rated 3. However, for the 2nd group, participants tended to rate 2, 3 and 4, which were mostly average.

For the 1st group, 91% ratings are high (4 and 5) but for the 2nd group only 47% ratings are high, comparing these 2 charts it can be said that ERWIN's conversation with misattribution bias was more accepting to the participants than the conversation without biases.

For the first group, the participants had surprise factor, they somehow enjoyed the conversation and liked the fact that ERWIN made mistakes while making conversation with them. From this experiment, it can be said that, ERWIN showed the ability to make mistakes in conversation which actually helped participants to relate easily with the robot. ERWIN as an expressive robot did not show a perfectionist ability as a machine which helped participants to easily connect and make the attachment bonds easier and they were interested to make further conversations in future. Responses from the question "*How much do you want to take ERWIN in your home?"* participants from the 1st group showed their natural interests compared to the 2nd group.

On the other hand, for the 2nd group of the participants, it was very common that ERWIN as a robot should have remembered their basic information, as it was expected to them so they were not amazed by the ERWIN's conversational statement where it confirmed the colour of the shirt the participants were wearing at their first experiment with ERWIN.

ERWIN to the 2nd group, has come very natural as robots usually do, so participants in that group were unable to taste the unpredictability in conversation and it was as usual like previous machine-like interactions, and for that reason it was very tough for them to relate with and make any natural attachment or interrelationship, and they showed less interests to take ERWIN in their home in questionnaire's ratings.

V. Future Work & Conclusion

In human psychological nature, it is easy to interact with another human-like personality that shows typical social specific characteristics [17]. From that understanding, humans can actually relate with other animals in nature, have some of them as pets and become attached to a specific kind of relationship [18]. Robots in the other hand have abilities to perform human-like actions, can be designed to 'look' like humans and can appear to behave in a human like manner, but they lacks human-like cognitive personalities. Human characters and personality can be described as imperfectly perfect [19], where robots lack to present such type of cognitive characteristics like unintentional mistakes, wrong assumptions, extreme presence of specific traits, task imperfectness and other human-like cognitive characteristics. Interrelations grow from the attractions of differences in characters, unpredictability and cognitive difference and imperfectness of nature [20].

Our experiments show that robots with general 'misattributes' bias is more likely to get human attention therefore become more effective in making relationships with humans. Therefore it can be said that robots should have human-like faults, characteristics biases and prone to carry out common mistakes that humans make on a regular social basis – which will develop the robot's own characteristics and should lead to the acceptance of a robot for long-term relationships with human. It is expected that cognitive characteristics and personality in robots will make it easy for people to relate with. Our experimental results show that, participants enjoyed and developed a preferred relationship faster with a misattributed robot than the robot without the bias, also it shows how one simple cognitive memory bias 'misattribution' was able to develop a better interaction with participants than the interactions without misattributions.

Keep in mind that ERWIN's misattribution factors actually relies on the robot itself, i.e., the way ERWIN communicates, we plan to study human cognitive biases in different robots (Keepon, MARC and others) in our further experiments. In the future we will try to introduce and study more human cognitive biases and other human-like factors in different robots with the hope that robots will make humanlike relationships that can lead to the acceptance of robots for long-term human-robot interaction.

References

[1] M. G. Haselton, D. Nettle. (2005) *The evolution of cognitive bias*. In D. M. Buss (Ed.), The Handbook of Evolutionary Psychology: Hoboken, NJ, US: John Wiley & Sons Inc. Pp. 724–746.

[2] C. Breazeal. (2003) *Social Interactions in HRI: The robot View*. IEEE Transactions On Systems, Man, and Cybernetics. Vol. 34, No. 2.

[3] K. M. Lee. (2006) *Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction*. Journal of Communication.

[4] R. R. McCrae, O. P. John. (1992) *An Introduction to the Five-Factor Model and Its Application*, Journal of Personality. Vol. 60. Issue 2. Pp.175-215.

[5] O. P .John, (1999) *The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives.* In: Handbook of personality: Theory and research. New York: Guildford.

[6] G. Mandler, (2002) *Origins of the Cognitive (r)evolution*, The Journal of the History of the Behavioural Sciences. Vol. 38, Pp. 339-353

[7] J. Cornelis, M. Wynants, (2008) *Brave New Interfaces: Individual, Social and Economic Impact of the Next Generation Interfaces, 'Probo, a friend for life?'.* ASP Vub Press. Pp. 253-257

[8] C. Breazeal, (2003) *Towards Sociable Robots*, Robotics and Autonomous System. Vol. 42. Pp 167-175.

[9] A. Weiss, M. Vincze, (2013) *Grounding in Human-Robot Interaction: Can it be achieved with the Help of the User?*. International Conference on Social Robotics ICSR. Bristol, UK.

[10] T. Fong, I. Nourbakhsh, K. Dautenhahn, (2003) *A survey of Socially Interative Robots.* Robotics and Autonomous Systems. Vol. 42. Pp 143–166.

[11] K.M.Lee, W Peng, S.A.Jin and C.Yan. *Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction*. Journal of Communication ISSN 0021-9916

[12] M. L. Walters, D. S. Syrdal, K. Dautenhahn, R. Boekhorst, K. L. Koay. (2008) *Avoiding the Uncanny Valley – Robot Appearance, Personality and Consistency of Behavior in an Attention-Seeking Home Scenario for a Robot Companion*. Autonomous Robots. Vol. 24, Issue 2, Pp 159-178.

[13] P. J. Hinds, T. L. Roberts, H. Jones. (2004) *Whose Job Is It Anyway? A Study of Human–Robot Interaction in a Collaborative Task*, Human-Computer Interaction. Vol. 19, Pp. 151-181.

[14] P. Baxter, T. Belpaeme, L. Canamero, P. Cosi, Y. Demiris, V. Enexcu. (2011) *Long-Term Human-Robot Interaction with Young Users.* IEEE/ACM HRI-2011 workshop on Robots interacting with children, Lausanne.

[15] B. Reeves, C. Nass. (1996) *The media equation*. New York: Cambridge University Press. Chap. 2, Pp. 19-36.

[16] E. Aronson, T. Wilson, R. Akert. (2005) *A Textbook of Social Psychology*, 6th edition. Scarborough, Ontario: Prentice-Hall Canada.

[17] D. Premack, A. Premack, (1995) *Origins of human social competence*. The cognitive neurosciences, Pp. 205-218.

[18] E. M. Keay. (2008) *Mary Lothian*. ISBN: 9781847995124

[19] A. Ellis, M. Abrams. (2008) *Personality Theories: Critical Perspectives,* SAGE Publications, Psychology, Chap. 1, Pp. 1-11.

[20] R. Russel, (2013) *Humans and Nature: How Knowing and Experiencing Nature Affect Well-Being.* Annual Review of Environment and Resources. Vol. 38, Pp. 73-502.