Trust of Virtual Agent in Multi Actor Interactions

Halimahtun M. Khalid
Damai Sciences, A-31-3 Suasana Sentral, Jalan Stesen Sentral 5,
*504700 Kuala Lumpur, Malaysia**

Liew Wei Shiung
Faculty of Computer Science and Information Technology, University of Malaya,
50603 Kuala Lumpur, Malaysia

Voong Bin Sheng
Damai Sciences, Suite R26-11 Dua Sentral, No. 8 Jalan Tun Sambanthan,
50470 Kuala Lumpur, Malaysia

Martin G. Helander
Damai Sciences, A-31-3 Suasana Sentral, Jalan Stesen Sentral 5,
50470 Kuala Lumpur, Malaysia
E-mail: mahtunkhalid@gmail.com, liew.wei.shiung@gmail.com, sh3ng527@gmail.com, mahelander@gmail.com

Abstract

Trust is crucial when integrating virtual agents in human teams. Our study investigated the combined use of psychological and physiological measures in predicting human trust of agents undertaking social tasks. The psychological measures comprised trust scores on ability, benevolence and integrity. The physiological measures included facial expressions, voice, heart rate and gestural postures. Subjects interacted with two avatars. A neuro-fuzzy algorithm extracted rules from the psychophysiological data to predict trust. Results revealed that trust can be predicted with 88% accuracy.

Keywords: Trust, virtual agent, measurement, human-agent interaction.

1. Introduction

Trust is a key element in the development of effective human-agent-human relationships, as trust affects system effectiveness of safety, performance, and usability. With the development and integration of virtual agents in human teams, the issue of predicting trust has become a focal concern[1]. One of the gaps in research is the lack of a reliable measure of human-robotic trust. Past studies have emphasized subjective measurements only[2]. We present a method where subjective (general trust, psychological) measures and objective (physiological) measures were mapped to predict human trust of humanoid agents in performing social tasks in a multi-actor and bilingual contexts. In this instance, the role of the humanoid agent or avatar was to mediate between a teleoperator and clients using communication dialog in either English or Bahasa Melayu (Malay language).

Therefore, the main objective of our study was to estimate human trust of the avatar in a business mediation context. We hypothesized that trustor's (evaluator) trust of the trustee (agent) is influenced by factors of gender, ethnicity and trust components.

2. Method

In this section, we describe the study as follow.

2.1. *Subjects and location*

Forty-eight subjects (25 males and 23 females) aged between 21 to 46 years (mean age 26 ±5 years) participated in the study. Twenty-two of the subjects were Malays and 26 were Chinese. To achieve ecological validity the experiment was conducted in an office setting of a R&D studio.

2.2. *Avatars and equipment*

Two avatars were designed for the experiment. One represented a male Chinese and the other a female Malay using humanoid characteristics derived from an earlier study. Figure 1 illustrates the avatars.

Fig. 1 Male and female avatars

Microsoft Kinect V2 technology was used to record the audiovisual data, and a computerized scale was used to record subjective measures. Data was logged directly into a server via a gigabit router and Ethernet cables, which connected the server to the laptops used by the subjects. Headset microphones were used to record voices.

B. The experimenters and the server for recording data were positioned in an observation room.

2.3. *Measures and ROS integration*

We measured trust subjectively using criteria of: (a) General Trust and (b) Psychological Trust. General Trust is made up of three factors: Ability, Benevolence and Integrity (ABI)[3]. For each factor there were four attributes presented as semantic differential word pairs. Altogether there were 12 attributes in the General Trust scale as summarized in Table 1. These attributes were replicated from previous studies[4,5]

Table 1. General Trust as defined by three main categories, each consisting of four polar attributes

Trust Category	Positive	Negative
ABILITY	Competent	Incompetent
	Knowledgeable	Not Knowledgeable
	Qualified	Unqualified
	Skilled	Unskilled
BENEVOLENCE	Cheerful	Not Cheerful
	Friendly	Unfriendly
	Kind	Unkind
	Pleasant	Unpleasant
INTEGRITY	Dependable	Undependable
	Ethical	Unethical
	Honest	Dishonest
	Reliable	Unreliable

The above attributes in Table 1 were embedded in the interactive dialog.

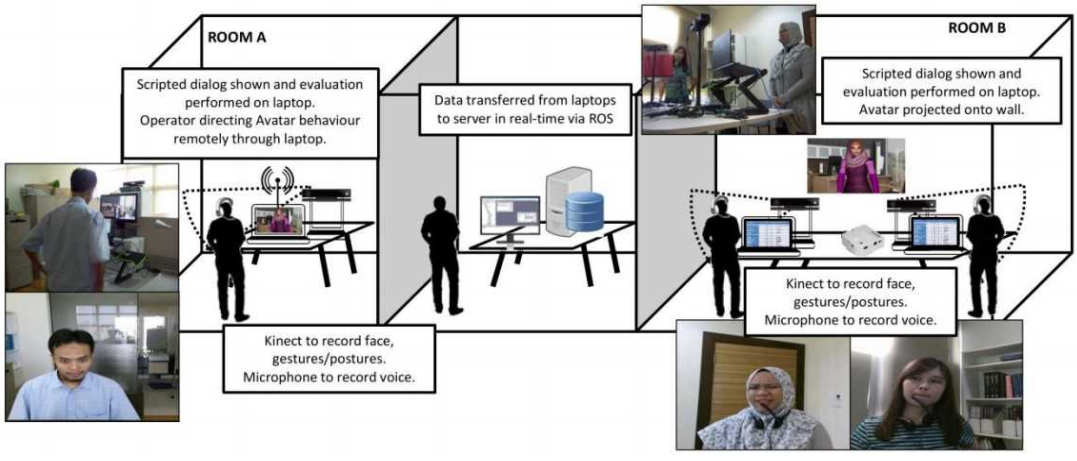

Fig. 2. Arrangement of experimental session

Figure 2 shows the layout of the experimental session, which comprised of 3 rooms. Trustor A (Teleoperator) was in Room A and Trustors B and C (Clients) in Room

Data collection and analysis were made possible via the Robot Operating System (ROS) framework, which integrated the agents, Graphical User Interface (GUI),

and hardware. The audiovisual and evaluation data were recorded on the server computer acting as ROS master. The avatar was controlled by a trigger signal sent by subjects through the GUI.

2.4. *Experimental design and dialog*

The experimental design was a within-subjects design with manipulations of gender, ethnicity, language and compliance to instructions by teleoperator, clients and avatar.

The dialog comprised a communication script of a Business scenario that concerns establishing a factory. This scenario was selected, as it was proven from the results of previous studies to be the most effective than a Disaster or Healthcare scenarios[1,2].

2.5. *Procedure*

First, subjects were briefed on the experiment and their tasks. Second, they signed an informed consent form and completed a subject profile sheet. Third, they were trained on the GUI to familiarize them with the online scoring method and dialog box.

Testing was done in a group of three subjects. The subjects were randomly assigned to the role of Operator or Clients based on a draw lot technique. Typically, the teleoperator in Room A initiated the interaction, followed by the robot confirming its task, and interacting with Trustor B and Trustor C in Room B, who then responded to the avatar, which in turn communicated back to the teleoperator. Subjects were instructed to act out the script in their respective roles. To test for systems operability and usability, a pilot study with two groups of subjects was carried out prior to the actual experiment.

Experimenters were positioned in a separate observation room and monitored the experiment via a screen and a headset. The computer server was placed in the observation room to record data in real time from the Kinects, headsets, and subject evaluation forms.

After completing each application scenario, subjects were given a 10-minutes break to complete a post-hoc questionnaire, which assessed their opinions and preferences regarding the avatar and its features.

2.6. *Data analysis*

We performed factor analysis, MANOVA and Pearson Correlations on the subjective data to test the hypotheses. The objective data, which was obtained from the Kinect and headsets, were analyzed separately. The video, audio, and evaluation recordings for each subject were segmented based on the ROS timestamps corresponding to the time when subjects evaluated each trust attribute. The segments were then processed as follow: Voice recordings were processed using PRAAT[6] and Audio Analysis Library[7]. Facial expressions were processed using OpenFace[8]. Heart rate signals were estimated from the subjects' faces using independent component analysis[9]. Gestural postures were estimated from Kinect's joint coordinates. All extracted objective features were then appended with contextual and subjective information such as the trust evaluation score, dialog condition, and the avatar type.

2.7. *Trust classification*

To predict trust, we mapped the objective features from the facial expressions, heart rate variability, voice, and gestures[10] of subjects to their subjective scores, to create a dataset. A neurofuzzy classifier ensemble method was then used to learn and predict subjects' trust of the virtual agents.

3. Results and Discussion

The results from MANOVA and correlation analyses revealed the following.

3.1. *Gender and ethnicity*

The gender and ethnicity of the trustee (virtual agent) had no significant effect on trust. But the interaction effects of gender and ethnicity of the trustors (human subjects) influenced their evaluations of the trustees, with $F(1, 2304) = 18.02$, $p<.001$. Gender-wise, female subjects were more trusting, giving higher trust scores than male subjects, $F(1,2304)=24.69$, $p<.001$. This finding confirmed our hypothesis on the influence of gender and ethnicity in trust evaluations.

However, ethnicity alone had no significant effect. Its interaction with trust components produced a significant effect, $F(2, 2304)=5.01$, $p=.007$. On average, Chinese subjects scored the agents highly on 'Ability' and 'Integrity' attributes, while Malay subjects scored the agents highly on the 'Ability' attributes only.

In addition, there were significant effects of intra-ethnicity on trust, with $F(2, 2304)=4.17$, $p=.016$. When

the groups consisted of just Chinese subjects, the evaluation of trust was higher than when the groups consisted of Malays only, or when it represented a mixed-ethnic composition. However, the gender composition of the group had no significant effect on trust evaluations. This finding suggests the importance of cultural and ethnic composition of teams in trust-related tasks.

3.2. *Subject preferences*

An analysis of the post-experiment questionnaire revealed that subjects' trust evaluation scores correlated significantly with their explicit Trust, $r=0.09$, $p<.001$, and Liking preferences for the agents, $r=0.06$, $p=.006$. Subjects who trusted one or more agents explicitly scored higher than subjects who explicitly trusted neither agent, $F(3,2304)=7.81$, $p<.001$.

In addition, subjects who explicitly liked the avatars scored highest; subjects who explicitly disliked the avatars scored second highest, and subjects who were undecided scored lowest, $F(2,2304)=3.681$, $p=.025$.

3.3. *Trust prediction*

Using the neurofuzzy classifier ensemble method[10], we were able to predict subject's trust in humanoid agents, with an accuracy of 88.3% $\pm$ 0.2% (F1-score 0.8401), when all physiological data was used. Greater accuracy at 90% was achieved when gestural posture data was excluded, whether for Low, Medium, or High Trust. Details of the analysis are described elsewhere[10].

4. Conclusion

This study has shown that virtual agents can be trusted to perform social tasks using communication dialogs in a multi-actor and bilingual contexts. The trust by humans varies with the gender and ethnicity of the evaluator. Also, trust can be easily evaluated for attributes of 'Ability' and 'Integrity' using dialogs than for 'Benevolence'. This implies that our design and modeling of the humanoid agents requires further anthropomorphism to appeal to humans, persona-wise.

Integration of the various data via ROS allows for real-time annotation of the data recordings. This enables segmentation of the relevant objective data to each trust score evaluation. As such, the accuracy of the neurofuzzy classifier was enhanced to predict trust.

Acknowledgment

We gratefully acknowledge the financial support by the US Aerospace Research and Development office (AOARD), Japan and the US Air Force Office of Scientific Research (AFOSR), Washington D.C. under Grant No. FA2386-14-1-0016.

References

1. P. Robinette, A. M. Howard, and A. R. Wagner, Effect of robot performance on human–robot trust in time-critical situations, in *IEEE Trans. Hum. Mach. Syst.* **47**(4) (2017) 425–436.
2. A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich, Common metrics for human-robot interaction, in *Proc 1st ACM SIGCHI/SIGART Conf. Hum Rob Interact* (Salt Lake City, Utah, USA, 2006), pp. 33–40.
3. R. C. Mayer, J. H. Davis, and F. D. Schoorman, An integrative model of organizational trust, in *Acad. Manage. Rev.* **20**(3) (1995) 709–734.
4. H. M. Khalid, W. S. Liew, P. Nooralishahi, Z. Rasool, C. K. Loo, and M. G. Helander, Predicting trust in social communication: Implications for human-robot interaction, in *Proc. HFES 2016 Int. Annu. Meeting* (Washington, USA, 2016), pp. 19–23.
5. H. M. Khalid, W. S. Liew, P. Nooralishahi, M. G. Helander, and Z. Rasool, Toward a theory of psychological trust for human-robot-human interaction: Effects of scenarios, gender, and ethnicity, in *Proc. SEANES Int. Conference* (Bandung, Indonesia, 2016), pp. 321–330.
6. P. Boersma, Praat: doing phonetics by computer, in *http://www. praat. org* (2006).
7. T. Giannakopoulos and A. Pikrakis, Introduction to Audio Analysis: A MATLAB Approach, in *Academic Press* (2014).
8. T. Baltrušaitis, P. Robinson and L. P. Morency, Openface: an open source facial behavior analysis toolkit, in *2016 IEEE Winter Conf Appl Comput Vis* (Lake Placid, New York, USA, 2016), pp. 1–10.
9. M. Lewandowska, J. Rumiński, T. Kocejko, and J. Nowak, Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity, in *2011 Federated Conference on Computer Science and Information Systems* (Szczecin, Poland, 2011) pp. 405–410.
10. H. M. Khalid, W. S. Liew, B. S. Voong, M. G. Helander and C. K. Loo, Technology for estimating trust in human-robot interaction, in *4th Asian Conf. Def. Tech.* (Tokyo, Japan, 2017).