# Embodied Conversational Agent Behavior and its Impact on Trust in Other Agents

**Reza Moradinezhad**
Drexel University
Philadelphia, PA, USA
rm976@drexel.edu

**Erin T. Solovey**
Worcester Polytenchnic Institute
Worcester, MA, USA
esolovey@wpi.edu

Proceedings of the CHI 2020 Workshop on Mapping Grand Challenges for the Conversational User Interface Community, April 25, 2020, Honolulu, Hawaii, USA.


## Abstract
Although it takes longer to build trust toward embodied conversational agents (ECAs), once built, this trust is more resilient to errors than conventional (e.g. WIMP) user interfaces [5]. In our work, we are exploring factors that influence the process of building trust in an ECA through interaction, as well as how the behavior of one ECA can influence perceptions of trust in other ECAs.

## Author Keywords
Human Computer Interaction; Trust; Virtual Agents; Embodied Conversational Agents.

## ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: User Interfaces

## Introduction
The application of Embodied Conversational Agents (ECA) is spreading through a variety of fields such as medicine [17, 6, 11], therapy [15, 18], pedagogy [22], games [10], e-commerce [3, 2], and social robots [24, 10]. Building a trustworthy interaction between these agents and humans is critical for the success of the system. Specifically, it is important to consider how interaction with one agent affects the perception of a previous agent and expectations from an agent encountered in the future.


**Figure 1:** Participants familiarize themselves with both agents in the introduction, before beginning the experiment.

Interpersonal trust is defined as "the willingness of a party to be vulnerable to the outcomes of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" [20]. This type of trust has been widely studied as a measure of trust between humans and automation [16, 12]. Jian et. al. [14] propose a trust survey based on the values associated with interpersonal trust to measure the subjective trust between humans and automation. As for objective trust, Pak et. al. [21] introduce the "conformity with the agent" as a measure for behavioral or objective trust.

Previous studies have shown that user interfaces containing ECAs have been more favored by users and helped users to feel more comfortable sharing sensitive personal information in medical and therapeutic settings [17, 6, 18, 11]. Designers of ECAs have studied different facial expressions alongside other non-verbal behaviors to understand non-verbal indicators of trust in ECAs [8, 10, 4, 13]. Studies on inconsistent assistive agents suggest that not only the level of reliability of the agent has an impact on trust, but also the order of exposure [9], stage of the task where the agent makes a mistake (early vs late) [7], and the difficulty of the task [19] significantly affect the trust. In addition, Yuksel et. al. [23] suggest that attractiveness of the agent is as important as its reliability. Therefore, how users feel about the agent before working with them also plays a role in how trustworthy that agent is perceived.

## Experiment

We conducted a study to gain insight into understanding the process of trust building and repair based on previous and future interactions. 35 participants answered two sets of 50 general knowledge questions, while receiving assistance from two different agents. The agents could be either Coop-

erative (accurate 80% of the time) or Uncooperative (accurate 20% percent of the time), meaning each participant did one of the four possible conditions: 1) Cooperative in both sets (CC) 2) Cooperative first, Uncooperative second (CU) 3) Uncooperative first, Cooperative second (UC) 4) Uncooperative in both sets (UU). Figure 1 shows the interface and the two agents. Participants answered a trust survey (a modified version of Jian et. al. [14]) before each set (baseline, after set 1, after set 2). The results have implications for trustworthy interaction design for user interfaces that contain ECAs.

## Results

In the following we provide a summary of the results that we have obtained so far.

As expected, the performance of users was significantly higher when they were working with the cooperative agent compared to the performance with the uncooperative agent. Further analysis showed that the number of wrong answers even in the first 10 "easy" questions was significantly higher with uncooperative agent than cooperative one. This can be interpreted as over-trust negatively affecting the normal performance of the users, meaning even in cases where the correct answer was obvious to the users, they sometimes chose to comply with the agent and choose the wrong answer.

Both self-reported perceived trust measured by Jian et. al. [14] survey, and the behavioral trust, which was measured by the number of times the users complied with the agent's suggested correct answer, showed significantly higher trust for cooperative agents than uncooperative agents.

As expected, we found significant difference in overall trust in both set 1 and set 2 for all pairs of conditions with opposing agent behavior (C vs U). However, interestingly, we found significant differences between the first sets of CU and CC and second sets of UC and CC where the agent had the same cooperative behavior. This suggests that users find the cooperative agent less trustworthy if they don't have the experience of working with the uncooperative agent.

## Conclusion

The increase in use of ECAs in medical and therapeutic settings as well as e-commerce websites, and the introduction and popularity of home assistance systems such as Amazon Alexa and Google Home suggest that the future of HCI will be more like human-human interaction. This raises the question that how interaction with one ECA affects the perception of trustworthiness of another ECA. Our study shows that the experience of interaction with a less reliable agent will result in reporting higher trust scores for the more reliable agent. However, if the users only interacted with a highly reliable agent, the trust scores would be lower. This suggests that users will be more tolerant toward occasional errors of a highly reliable agent if they have the experience of interacting with a less reliable agent. Also, the results suggest that users find agents less trustworthy if they make mistakes on easy tasks, which is in line with "easy-errors hypothesis" introduced by Madhavan et al. [19].

## Future Work

Using our log data, we plan to build a machine learning model which is able to accurately predict whether a user is trusting an agent, meaning that they are going to choose the answer with HP feedback from the agent. This way, we can passively monitor the process of trust building and observe the fluctuations in the process. Most importantly, if an agent can know how likely a user is to choose the suggested answer, it can provide more helpful information and increase the performance. For example, in case of under-

trust, the agent can provide extra information and justification about why the user should trust the agent. Also, in case of over-trust, the agent can explain that there are limitations in its assistance and user should put more weight on their own knowledge when making the final decision. This type of real-time trust repair is more effective than traditional trust repair process in which the agent apologizes or provides explanations only after a mistake is made.

Aranyi et. al. [1] show that it is possible to build adaptive agents which can give feedback only based on brain activity measured by fNIRS. The reviewed studies support the idea that emotions and other feelings such as stress and self-awareness are associated with activation in the prefrontal cortex. This region is known to be involved in the supramodal coordination of perceptual and cognitive processes, and it is easily accessible by fNIRS sensors. This makes it possible to build systems which can use those emotions as input. An example of such a system would be an assistive ECA to serve as a math tutor or a recommender during creative thinking process which can adapt its assistance based on the user's mental state. Another example would be an ECA which help users in critical decision making. The ECA can adjust the extent of its suggestions and assistance based on the user's mental state. For instance, it would come up with fewer suggestions with longer delays between each of them when it detects the user is in normal state, but will increase the number of and frequency of suggestions if it detects the user is stressed or tired.

## Bios

*Reza Moradinezhad* is a fifth year PhD candidate in Computer Science at Drexel University College of Computing and Informatics. After graduating with his B.S. in IT Engineering from University of Mazandaran, Iran in 2013, and after two years of teaching and freelancing in Website and Windows application development, he started his PhD program at Drexel University in 2015. His research involved using physiological data, namely brain data acquired by functional Near Infrared Spectroscopy (fNIRS) as a means of feedback for Human Computer Interaction (HCI). His current project looks at factors of trust between Humans and Embodied Virtual Agents (EVAs). He looks at how an agent's inconsistent behavior can affect the human's trust. He also explores how one agent's behavior can cause a human to perceive another agent's trustworthiness differently.

*Erin T. Solovey, Ph.D.* is an Assistant Professor of Computer Science at Worcester Polytechnic Institute. Her research expertise is in human-computer interaction, with a focus on accessibility and emerging interaction techniques, such as brain-computer interfaces and human-agent interaction.

## References


[1] Gabor Aranyi, Florian Pecune, Fred Charles, Catherine Pelachaud, and Marc Cavazza. 2016. Affective interaction with a virtual character through an fNIRS brain-computer interface. *Frontiers in computational neuroscience* 10 (2016), 70.

[2] Timothy Bickmore and Justine Cassell. 2001. Relational agents: a model and implementation of building user trust. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 396–403.

[3] Justine Cassell and Timothy Bickmore. 2000. External manifestations of trustworthiness in the interface. *Commun. ACM* 43, 12 (2000), 50–56.

[4] Justine Cassell, Hannes Högni Vilhjálmsson, and Timothy Bickmore. 2004. Beat: the behavior expression animation toolkit. In *Life-Like Characters*. Springer,

163–185.

[5] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. Sage Publications Sage CA: Los Angeles, CA, 263–267.

[6] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, and others. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[7] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International journal of human-computer studies* 58, 6 (2003), 697–718.

[8] Aaron C Elkins and Douglas C Derrick. 2013. The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group decision and negotiation* 22, 5 (2013), 897–913.

[9] Jean E Fox and Deborah A Boehm-Davis. 1998. Effects of age and congestion information accuracy of advanced traveler information systems on user trust and compliance. *Transportation Research Record* 1621, 1 (1998), 43–49.

[10] Aimi Shazwani Ghazali, Jaap Ham, Emilia I Barakova, and Panos Markopoulos. 2018. Effects of robot facial characteristics and gender in persuasive human-robot interaction. *Frontiers in Robotics and AI* 5 (2018), 73.

[11] Matthew Gombolay, Xi Jessie Yang, Bradley Hayes, Nicole Seo, Zixi Liu, Samir Wadhwania, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. 2018. Robotic assistance in the coordination of patient care. *The International Journal of Robotics Research* 37, 10 (2018), 1300–1316.

[12] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2009. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 528–533.

[13] Jennifer Hyde, Elizabeth J Carter, Sara Kiesler, and Jessica K Hodgins. 2016. Evaluating animated characters: Facial motion magnitude influences personality perceptions. *ACM Transactions on Applied Perception (TAP)* 13, 2 (2016), 8.

[14] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.

[15] Sin-Hwa Kang and Jonathan Gratch. 2012. Socially anxious people reveal more personal information with virtual counselors that talk about themselves using intimate human back stories. *Annual Review of Cybertherapy and Telemedicine* 181 (2012), 202–207.

[16] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[17] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 19.

[18] Gale M Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.

[19] Poornima Madhavan, Douglas A Wiegmann, and Frank C Lacson. 2006. Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors* 48, 2 (2006), 241–256.

[20] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[21] Richard Pak, Nicole Fink, Margaux Price, Brock Bass, and Lindsay Sturre. 2012. Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 9 (2012), 1059–1072.

[22] Susanne Van Mulken, Elisabeth André, and Jochen Müller. 1999. An empirical study on the trustworthiness of life-like interface agents.. In *HCI (2)*. 152–156.

[23] Beste F Yuksel, Penny Collisson, and Mary Czerwinski. 2017. Brains or beauty: How to engender trust in user-agent interactions. *ACM Transactions on Internet Technology (TOIT)* 17, 1 (2017), 1–20.

[24] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *International conference on intelligent virtual agents*. Springer, 218–233.