AFRL-AFOSR-JP-TR-2018-0006

Impact of Human like Cues on Human Trust in Machines: Brain Imaging and Modeling Studies for Human-Machine Interactions

Soo-Young Lee
Korea Advanced Institute of Science and Technology

01/05/2018
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ IOA
Arlington, Virginia 22203
Air Force Materiel Command

FORM SF 298

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
09-01-2018	Final	30 Sep 2014 to 29 Sep 2017

4. TITLE AND SUBTITLE
Impact of Human like Cues on Human Trust in Machines: Brain Imaging and Modeling Studies for Human-Machine Interactions

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA2386-14-1-4035

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Soo-Young Lee

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Korea Advanced Institute of Science and Technology
291 Daehak-ro, Yuseong-gu
Taejon, 305701 KR

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD
UNIT 45002
APO AP 96338-5002

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR IOA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-JP-TR-2018-0006

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
When a human and an intelligent machine work together as a team, human trust can influence performance. Yet, an electrophysiological signature of trust has not been isolated. In order to isolate such a signature, the research team recorded fMRI or event-related potentials while subjects were playing two cognitive games. At the first experiment, human subjects played a theory-of-mind bilateral game with two types of computerized agents: with or without humanlike cues. At the second experiment, human subjects played a unilateral game in which the human subjects played the role of the Coach (or supervisor) while a computer agent played as the Player. Electrophysiological activities in brain regions belonging to the theory-of-mind network correlated with perceived capability, especially when a machine opponent had some human-likeness. In particular, the research shows that activity in the left parietal region correlating with a human players future behavior can be identified as the neural signature of capability-based trust. These results reveal that brain signals underlying trust as influenced by perceived capability and human-likeness might be useful for performance optimization of human-machine systems.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			ROBERTSON, SCOTT
Unclassified	Unclassified	Unclassified	SAR	5	19b. TELEPHONE NUMBER *(Include area code)* +81-042-511-7008

Final Report for AOARD Grant FA2386-14-1-4035

"**Impact of Human-like Cues on Human Trust in Machines: Brain Imaging and Modeling Studies for Human-Machine Interactions**"

December 28, 2017

Name of Principal Investigators (PI and Co-PIs): Soo-Young Lee
- e-mail address : sy-lee@kaist.ac.kr
- Institution : Korea Advanced Institute of Science and Technology
- Mailing Address : 291 Daehak-ro Yuseong-gu, Daejeon 34141, Republic of Korea
- Phone : +82-42-350-3431 (O), +82-10-2682-2064 (M)
- Fax : +82-42-350-8490

Period of Performance: 09/30/2014 – 09/29/2017

Abstract:
When a human and an intelligent machine work together as a team, human trust can influence performance. Yet, an electrophysiological signature of trust has not been isolated. In order to isolate such a signature, we recorded fMRI or event-related potentials while subjects are playing two cognitive games. At the first experiment, human subjects played a theory-of-mind bilateral game with two types of computerized agents: with or without humanlike cues. At the second experiment, human subjects played a unilateral game in which the human subjects played the role of the Coach (or supervisor) while a computer agent played as the Player. Electrophysiological activities in brain regions belonging to the theory-of-mind network correlated with perceived capability, especially when a machine opponent has some human-likeness. In particular, our research shows that activity in the left parietal region correlating with a human player's future behavior can be identified as the neural signature of capability-based trust. These results reveal that brain signals underlying trust as influenced by perceived capability and human-likeness might be useful for performance optimization of human-machine systems.

Introduction:

In the age of intelligent machine the human trust in machines becomes much more important for efficient operation of autonomous systems (AS) and decision support systems (DSS) where human and machine work as a team. Several studies showed that the human trust, distrust, and mistrust in machines significantly affected overall system performance. Auto-pilot, air traffic control, and anti-aircraft warfare systems may be examples of those systems. It was found that the human decision process is affected by the trust in the team-mates, i.e., partners or counterparts. For the efficient operation as a team, the human operator should neither distrust nor mistrust the machine team-mates. The trust is dynamic, i.e., previous history may change the current level of trust. Also, it was found that the human trust in machine team-mates has different characteristics to that in human team-mates. Therefore, it is important to understand the impact of humanlike cues on sequential trust development, maintenance, and degradation in machine team-mates for the better design of AS and DSS.

It is understood that trust has multidimensional characters. Barber had identified 3 axes for the trust space, i.e., Persistence (Reliability), Technical competence (Ability), and Fiduciary responsibility (Moral and social obligation). Also, the trust dynamics are greatly affected by humanlike cues such as facial expression, speech, and personality.

Although many researches had been reported on trust with behavioral data on the 3 axes, only a few brain imaging studies have been reported on trust. Neural correlates of 'static' trustworthiness had been studied for human faces and online web without feedback process. For the 'dynamic' development of trust with feedback process, sequential bilateral trust games between human and machine counterpart had been utilized for the fiduciary responsibility (moral), i.e., the willingness to cooperate.

Experiment:

In this research, we design two cognitive games between human subject and machine agent, and measure fMRI or EEG signals to identify human trust on machine. The bilateral (reciprocal) interaction in human-machine systems. The human operator and machine counterpart make decisions alternatingly for cooperation in the reciprocal interaction systems. The this systems had been studied for the fiduciary responsibility ("willingness to cooperate") only with perfect reliability and ability of machine counterparts.

We believe that many practical applications such as auto-pilot and anti-warfare systems are closer to the unilateral (nonreciprocal) interaction systems. When human and machine partners work together as a team, or when human supervises autonomous machine, human trust in the machine is mainly contributed by the technical ability. (The autonomous machine should have high reliability and also be designed to serve the responsibility.) Therefore, it is important to measure the human trust level from brain signals based on the technical ability of machine partners.

Human subjects will play a cooperative game with a computer agent (opponent) for maximizing subject own payoff as well as opponent's payoff. We believe that a cooperation will be made based on the TRUST to his opponent, and the trust will be generated if subject begins to believe his opponent has a "goodwill" (willingness to cooperate) and "technical ability" (opponent can make an optimal decision for both players' payoffs). Thus, we will design a computer agent (opponent to human subject) with varying traits of goodwill and technical ability. Human subjects will evaluate their trust level of each computer agent after completing games with self-report questionnaire (e.g., "Willingness to cooperate" describes the opponent with happy face?). We will investigate the relationship between self-reported trust level and task performance (total payoffs) as well as neural responses. (More details are described in 11. Outcome evaluation criteria and evaluation methods)

Results and Discussion:
For both bilateral and unilateral games, we found brain areas and time representing the human trust on machine.

List of Publications and Significant Collaborations that resulted from your AOARD supported project: In standard format showing authors, title, journal, issue, pages, and date, for each category list the following:

b) papers published in peer-reviewed conference proceedings

-Suh-Yeon Dong, Bo-Kyeong Kim, Kyeongho Lee, and Soo-Young Lee, "A Preliminary Study on Human Trust Measurements by EEG for Human-Machine Interactions," 3rd International Conference on Human-Agent Interaction (HAI2015), pp. 265-268, Daegu, Korea, 21-24, October, 2015.

d) conference presentations without papers

-Invited Talk, "Understanding Brain Internal States for Intelligent Conversational Agents", 17th China-Japan-Korea Joint Workshop on Neurobiology and Neuroinformatics (NBNI2017), RIKEN, Wako, Japan, December 19-20, 2017

-Keynote Talk, "Artificial Intelligent Systems with Human Internal State Understanding: I Know Who You Are and What You Think", Computational Intelligence in Information Systems 2016 (CIIS2016), Bandar Seri Begawan, Brunei, November 18-20, 2016.

-Invited Talk, "Investigating Neural Correlates of Human Trust in Machine Using a Theory-of-Mind Game", 4th International Workshop on Advances in Neuroinformatics (AINI2016), Wako, Japan, May 28-29, 2016

-Keynote Talk, "Understanding Human Internal States: I Know Who You Are and What You Think", 3rd International Conference on Human-Agent Interaction (HAI2015), Daegu, Korea, 21-24, October, 2015.

-Invited Talk, "I Know Who You Are and What You Think: An EEG Study", 10th AEARU Workshop on Computer Science and Web Technology (AEARU-CSWT2015), University of Tsukuba, Japan, February 25-27, 2015.

-Keynote Talk, “I Know What You Think: Understanding Human Internal States for Artificial Cognitives Systems”, The 21st International Conference on Neural Information Processing (ICONIP2014), Kuching, Malaysia, 3-6 November, 2014.

e) manuscripts submitted but not yet published,

-S.-Y. Dong, B.-K. Kim, and S.-Y. Lee, “Trust Me, I’ll Be Back: Measuring Trust in Machines with a Two-Player Theory-of-Mind Game”

- S.-Y. Dong, E.S. Jung, and S.-Y. Lee, “Human Trust in Machines with Humanlike Cues”

f) provide a list any interactions with industry or with Air Force Research Laboratory scientists or significant collaborations that resulted from this work.

- Emotional Conversational Agent: The 4th Korean Flagship AI Project, from December 2016 to December 2020, about US$14,000,000 (Principal Investigator of the consortium project team consisting of 3 universities and 2 industrial companies)

Attachments: Publication of b) and e) listed above.