Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-32 (2016) 048–053

Real-Time Sensing of Trust in Human-Machine Interactions ⋆

Wan-Lin Hu * **Kumar Akash** * **Neera Jain** * **Tahira Reid** *

* *School of Mechanical Engineering, Purdue University, West Lafayette, IN 47906 USA (e-mail: hu188@purdue.edu, kakash@purdue.edu, neerajain@purdue.edu, tahira@purdue.edu).*

Abstract: Human trust in automation plays an important role in successful interactions between humans and machines. To design intelligent machines that can respond to changes in human trust, real-time sensing of trust level is needed. In this paper, we describe an empirical trust sensor model that maps psychophysiological measurements to human trust level. The use of psychophysiological measurements is motivated by their ability to capture a human's response in real time. An exhaustive feature set is considered, and a rigorous statistical approach is used to determine a reduced set of ten features. Multiple classification methods are considered for mapping the reduced feature set to the categorical trust level. The results show that psychophysiological measurements can be used to sense trust in real-time. Moreover, a mean accuracy of 71.57% is achieved using a combination of classifiers to model trust level in each human subject. Future work will consider the effect of human demographics on feature selection and modeling.

Keywords: human-machine interface, modeling, real-time, categorical data, classifiers, discriminant analysis, human brain, intelligent machines, physiological models

1. INTRODUCTION

Motivation and Problem Definition: Advances in sensing, communication, and control systems have spurred the development of a number of *smart* systems and services. Increasing levels of automation have resulted in humans being displaced as the primary decision-maker in roles such as power plant operators and aircraft pilots (Jian et al., 2000). Additionally, in what are broadly being called Human-Agent Collectives, we expect to see a growing need for cooperation between humans and machines in a variety of situations, including disaster relief (Jennings et al., 2014; Sadrfaridpour et al., 2016). It is well established that human trust in automation is central to successful interactions between humans and machines (Yagoda and Gillan, 2012; Lee and See, 2004; Sheridan and Parasuraman, 2005). Here, *machine* refers broadly to any automated system, such as an autonomous robot or a process control system in a power plant. Therefore, we are interested in using feedback control principles to design machines that are capable of *responding to changes in human trust level in real-time*. However, in order to do this, we require a sensor for measuring human trust level *online*.

Trust itself can be classified into three categories: dispositional, situational, and learned (Hoff and Bashir, 2015). Dispositional trust refers to the component dependent on demographics (e.g. gender, culture) whereas situational and learned trust depends on time-varying factors such as task difficulty, self-confidence, and experience. Therefore, situational and learned trust factors influence *real-time human decision-making during interactions with automated systems*. Researchers have attempted to predict human trust using dynamic models that rely on the experience and/or self-reported behavior of humans (Lee and Moray, 1992; Jonker and Treur, 1999). However, it is not practical to use human self-reported behavior as a feedback control variable. An alternative is the use of psychophysiological signals to *sense* trust level (Riedl and Javor, 2012). While these measurements have been correlated to human trust level, they have not been studied in the context of real-time trust sensing.

Background on Psychophysiological Measurements and Trust: There are several psychophysiological measurements that have been studied in the context of human trust. We focus here on electroencephalography (EEG) and galvanic skin response (GSR). EEG is an electrophysiological measurement technique that captures the cortical activity of the brain (Handy, 2005), and a powerful technique to observe brain activity in response to a specific event is through an event-related potential (ERP). An ERP is determined by averaging repeated responses over many trials to eliminate random brain activity (Handy, 2005). GSR is a classical psychophysiological signal that captures arousal based upon the conductivity of the surface of the skin. It has been used in polygraph tests for many decades (Grubin and Madsen, 2005).

Some researchers have studied trust via EEG, especially with ERPs. Boudreau et al. (2008) found a difference in peak amplitudes of ERP components in human subjects

⋆ This material is based upon work supported by the National Science Foundation under Award No. 1548616. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

10.1016/j.ifacol.2016.12.188

while they participated in a coin toss experiment that stimulated trust and distrust. Long et al. (2012) further studied ERP waveforms with feedback stimuli based on a modified form of the coin toss experiment conducted by Boudreau et al. (2008). The decision-making in the 'trust game' (Ma et al., 2015) has also been used to examine human-human trust level. Finally, researchers have examined GSR in correlation with human trust level. Khawaji et al. (2015) found that the average of GSR values, and the average of peaks of GSR values, are significantly affected by both trust and cognitive load in the text-chat environment.

Gaps in Literature: Although ERPs could show how the brain functionally responds to a stimulus, they are event triggered. It is difficult to identify triggers during the course of an actual human-machine interaction thereby rendering ERPs impractical for real-time trust level sensing. In addition, the use of GSR for measuring trust has not been explored. A fundamental gap remains in determining a static mathematical model that maps psychophysiological signals to human trust level and that is suitable for real-time sensing.

Contribution: In this paper we present a human trust sensor model based upon real-time psychophysiological measurements, primarily GSR and EEG. The model is based upon data collected through a human subject study and the use of classification algorithms to map continuous data to a categorical trust level. The proposed methodology for real-time sensing of human trust level will enable machine algorithm design aimed at improving interactions between humans and machines.

Outline: This paper is organized as follows. Section 2 introduces the experimental procedure and data acquisition. The methodology for data analysis is described in Section 3. The sensor modeling and classification results are presented and discussed in Section 4, followed by concluding statements in Section 5.

2. HUMAN SUBJECT STUDY

Prior investigation of human trust with respect to psychophysiological response has relied on experiments that do not mimic realistic human-machine interaction (HMI) scenarios (Boudreau et al., 2008; Long et al., 2012). We believe that the use of an experiment in a simple HMI context will result in trust models that are more broadly applicable. Thus we propose the following experiment that elicits human trust dynamics with respect to machines.

Participants: Thirty-one adults (20 males) from West Lafayette, Indiana (USA), aged 18-43 years participated in our study. All participants were healthy and one was left-handed. The group of participants were diverse with respect to their age, gender, major, and cultural background (i.e. nationality). The compensation was $15 per hour for their participation and each participant signed the informed consent form. The Institutional Review Board at Purdue University approved the study.

Stimuli and Procedures: When a participant came to the laboratory, we asked them to respond to a scenario in which they would be driving a car equipped with an image processing sensor. The algorithm used in the sensor

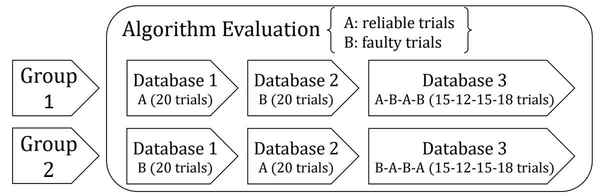

Fig. 1. Participants were randomly assigned to one of the two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty cases, were counterbalanced across groups.

would detect obstacles on the road in front of the car and the participant would need to repeatedly evaluate the algorithm report. We specifically informed the participant that the algorithm for image processing was in beta testing and that they would need to make their judgment of trust or distrust based on their experience with the algorithm.

There were two stimuli (*obstacle detected* and *clear road*). Both stimuli had a 50% probability of occurrence. Participants had the option to choose 'trust' or 'distrust' after which they received feedback of 'correct' or 'incorrect'. The trials were divided into two categories: reliable and faulty. In reliable trials, the algorithm correctly identified the road condition, which was in fact the stimuli. From the participant's perspective, this meant that choosing 'trust' would be marked as correct and choosing 'distrust' would be marked as incorrect. For the faulty trials, there was a 50% probability that the algorithm incorrectly identified the road condition.

Each participant completed 100 trials, along with four practice trials in the beginning of the study. The trials were divided into three phases, called databases in the study, as shown in Fig. 1. In database 3, the accuracy of the algorithm was switched between reliable and faculty according to a pseudo-random binary sequence (PRBS) in order to excite all possible dynamics of the participant's trust response. Figure 2 shows the sequence of events in a single trial. We validated the experimental design by collecting responses from 209 online participants (112 and 97 in groups 1 and 2, respectively) using Amazon Mechanical Turk (Amazon, 2005). The experiment elicited expected trust responses based on the aggregated data as shown in Fig. 3.

EEG Recording and Pre-processing: EEG, sampled at 256 Hz, was recorded from 9 scalp sites (Fz, Cz, POz, F3, F4, C3, C4, P3, and P4 based on the 10-20 system) using the B-Alert X10 EEG headset (Advanced Brain Monitoring, CA, USA) via iMotions (iMotions, Inc., MA, USA). All EEG channels were referenced to the mean of the left and right mastoids. The surface of the scalp and the mastoids were cleaned with 70% isopropyl alcohol wipes. Conductive electrode cream (Kustomer Kinetics, CA, USA) was then applied to each electrode including the reference. The contact impedance between electrodes and skin was kept to a value less than 40 kΩ.

Automatic decontaminated signals provided by the EEG system were used for model training and validation; that is to say, effects from electromyography, electrooculography, spikes, saturations, and excursions were minimized.