In AI We Trust: Investigating the Relationship between Biosignals, Trust and Cognitive Load in VR

Kunal Gupta
kgup421@aucklanduni.ac.nz
The University of Auckland
Auckland, New Zealand

Ryo Hajika
ryo.hajika@auckland.ac.nz
The University of Auckland
Auckland, New Zealand

Yun Suen Pai
yspai1412@gmail.com
The University of Auckland
Auckland, New Zealand

Andreas Duenser
Andreas.Duenser@data61.csiro.au
CSIRO
Hobart, Australia

Martin Lochner
mlochner@uwaterloo.ca
CSIRO
Hobart, Australia

Mark Billinghurst
mark.billinghurst@auckland.ac.nz
The University of Auckland
Auckland, New Zealand

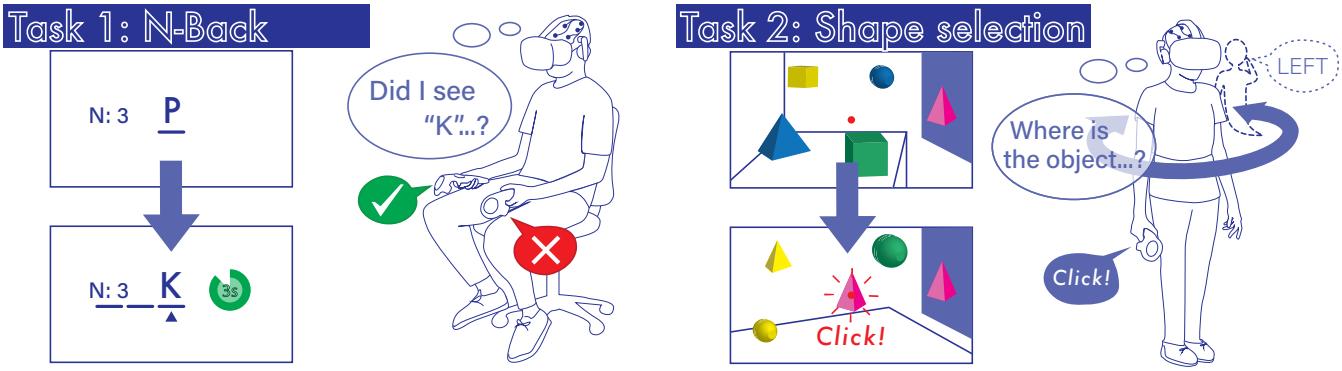

Figure 1: Our experiment to evaluate trust and cognitive load using the n-back task with an agent-assisted shape selection task

ABSTRACT

Human trust is a psycho-physiological state that is difficult to measure, yet is becoming increasingly important for the design of human-computer interactions. This paper explores if human trust can be measured using physiological measures when interacting with a computer interface, and how it correlates with cognitive load. In this work, we present a pilot study in Virtual Reality (VR) that uses a multi-sensory approach of Electroencephalography (EEG), galvanic skin response (GSR), and Heart Rate Variability (HRV) to measure trust with a virtual agent and explore the correlation between trust and cognitive load. The goal of this study is twofold; 1) to determine the relationship between biosignals, or physiological signals with trust and cognitive load, and 2) to introduce a pilot study in VR based on cognitive load level to evaluate trust. Even though we could not report any significant main effect or interaction of cognitive load and trust from the physiological signal, we found that in low cognitive load tasks, EEG alpha band power reflects trustworthiness on the agent. Moreover, cognitive load of the user decreases when the agent is accurate regardless of task's cognitive load. This could be possible because of small sample size, tasks not stressful enough to induce high cognitive load due to lab study and comfortable environment or timestamp synchronisation error due to fusing data from various physiological sensors with different sample rate.

VRST '19, November 12–15, 2019, Parramatta, NSW, Australia

ACM ISBN 978-1-4503-7001-1/19/11…$15.00
https://doi.org/10.1145/3359996.3364276

CCS CONCEPTS

• **Human-Centered Computing** → **Empirical Studies in HCI**.

KEYWORDS

Trust, Cognitive Load, Physiological signals, Virtual Assistant, Virtual Reality

ACM Reference Format:
Kunal Gupta, Ryo Hajika, Yun Suen Pai, Andreas Duenser, Martin Lochner, and Mark Billinghurst. 2019. In AI We Trust: Investigating the Relationship between Biosignals, Trust and Cognitive Load in VR. In *25th ACM Symposium on Virtual Reality Software and Technology (VRST '19), November 12–15, 2019, Parramatta, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3359996.3364276

1 INTRODUCTION

As data-driven artificial intelligence agents are becoming more common (e.g. Google Assistant, Amazon Alexa, and Tesla self-driving

Table 1: List of Previous Work Regarding Sensing of Cognitive Load and Trust

Author	Sensing	Platform	Cognitive Load	Trust
Samson et al. [Samson and Kostyszyn 2015]	-	Trust game	Raven P Matrix Test	EIS Trust Scale
Hu et al. [Hu et al. 2016]	EEG, GSR	Desktop car driving game	-	Sensing result
Dey et al. [Dey et al. 2019]	EEG	VR shape selector	N-back	-
Daniel et al. [McDuff et al. 2016]	HR, HRV	Desktop ball control, Berg card sorting	Dundee Stress	-
Akash et al. [Akash et al. 2018]	EEG, GSR	Desktop car driving game	-	Sensing result
Dong et al. [Dong et al. 2015]	EEG	Desktop matrix game with virtual agent	-	Sensing result
Zhang et al. [Zhang et al. 2017]	EEG, Eye gaze, ECG, EMG, SKT, RESP	VR driving	Sensing result	-
Khawaji et al. [Khawaji et al. 2015]	GSR	Desktop text chat	Sensing result	Sensing result
Hale et al. [Hale and Antonia 2016]	Motion	Virtual avatar mimicry	-	Sensing result
Salanitri et al. [Salanitri et al. 2016]	-	VR	-	Technology Trust Measure
Gerry et al. [Gerry et al. 2018]	EEG	VR shape selector	N-back	-
Our method	EEG, GSR, HRV	VR shape selector	N-back	System Trust Scale

cars), trust is becoming an increasingly important factor in human-computer interaction (HCI). For example, if users are going to rely on Google Assistant for weather forecasts, they need to be able to trust the information that is given to them. Similarly, users of self-driving cars need to be able to trust that the car will be able to take them to their destination without incident. Similar issues occur in virtual environments where user interaction with virtual agents, objects, and even other human participants projected as a virtual avatar rely heavily on our trust in both virtual and living entities within the virtual space.

Although important, there is a lot of research that needs to be conducted in trust and interactive technology. For starters, defining trust is rather tricky, even in the field of psychology [Hernandez-Ortega 2011]. Understanding factors that influence trust is also important. Trust is affected by our personal experience interacting with an entity as we evaluate its reliability over time. However, other factors like how an agent sounds or looks can influence our trust in that particular technology [Davis et al. 2009; McDonnell and Breidt 2010; Qiu and Benbasat 2005]. Voice, intonation, appearance, and motion also play an important role in influencing trust.

Researchers have been studying trust in human-computer interfaces since the late 80s [Muir 1987], and previous research has shown a strong correlation between cognitive load and trust [Samson and Kostyszyn 2015]. Trust is often measured by using a set of subjective surveys (e.g. the System Trust Scale (STS) [Jian et al. 2000]). However in recent years, researchers have begun to explore psycho-physiological measures such as EEG or GSR and they were able to develop a general trust sensor model with a mean accuracy of 72% and a classifier based model with an accuracy of 79% [Akash et al. 2018]. Compared to survey-based methods which are usually only issued at the end of a task, physiological signal sensing can be used as a tool for continuous and real-time evaluation, allowing for a myriad of integrated solutions with technological devices and real-time sensing.

In our research we are interested in the correlation between trust and cognitive load, and are using VR to explore this. The hypothesis is that agents can assist people performing a task and so reduce the person's cognitive load, but only if the agent is trusted. So we are interested in how we can reliably measure trust, and if we can use Virtual Reality to manipulate trust and cognitive load. We approach the sensing of trust by measuring it using a multi-sensory approach of EEG, GSR and HRV signals. By recording several physiological signals for the experiment, we can establish which signals show the best correlation with cognitive load and trust. This opens the door for more potential future works, especially for use cases like developing a machine learning model or different VR input modalities that can be integrated into wearables like the HMD, haptic glove and so on. We also perform an N-back Test to understand the cognitive load level of each participant so that we can specifically design a study environment in VR based on that.

We use a VR shape-selection task which has previously been shown to increase cognitive load [Dey et al. 2019]. For our implementation, we added a virtual agent giving instructions to the participant to assist them in the task. We also added a countdown timer to differentiate between the low and high cognitive load task. Depending on the condition, the reliability of the agent is changed, while the system measures the aforementioned physiological signals. From this, we then determine the correlation between each of the signal types with cognitive load and trust. With this information, we suggest several application scenarios in VR that can benefit from this.

Compared to previous work, our research work makes the following novel contributions:

(1) we explore the relationship between physiological signals with trust and cognitive load,

(2) we develop an experimental design method in VR to evaluate the relationship between trust and cognitive load based on individual's cognitive load levels.

2 RELATED WORK

In this section, we summarize previous related work on cognitive load and trust sensing methods, and how this is beneficial to VR as a platform. We summarize our findings in Table 1.

2.1 Physiological Signal Sensing

Physiological signals, bio-potentials, or bio-signals are electrical potential differences that exist in the human body that can be measured as electrical signals. These exists in the form of EEG, GSR, HRV, heart rate (HR), electromyography (EMG), skin temperature (SKT), respiration (RESP), etc. These signals often correspond to some kind of physiological [Abouelenien et al. 2017], cognitive [Augereau et al. 2018; Grimes et al. 2008; Rozado and Dunser 2015], or emotional state [Szwoch 2015]. In human-computer interaction (HCI), these signals are also exploited as a form of input and interaction, such as using EEG to compare different devices or interaction techniques in controlled virtual environment[Frey et al. 2016], measure cognitive workload experiences by the users interacting with computers using functional near infrared spectroscopy (fNIRS) [Hirshfield et al. 2009], using EMG sensing by muscle activation as a trigger [Pai 2016] or electrooculography (EOG) sensing for subtle inputs [Lee et al. 2017]. Our work employs a multimodal approach, using multiple physiological signals to measure cognitive load and trust, as well as the relationship between them.

2.2 Trust

Trust is a psycho-physiological state that involves a firm belief about another's intention and one's willingness to act by following their words, expressions, decisions, or actions [Susan and Holmes 1991]. In both face-to-face and virtual (remote) human-human interaction, trust is considered as an important factor to achieve successful outcomes because of how it influences information exchange among people, coordination, assistance and collaboration among individuals [Jarvenpaa et al. 1998]. The more we trust a colleague, the better the collaboration outcome. The definition of trust has varied throughout various literature, with some claiming that it can be categorized into persistence, technical competence and fiduciary responsibility [Barber 1983], and others claiming that it can be divided into dispositional, situational and learned [Hoff and Bashir 2015]. Another important consideration is the factors that influence trust. Apart from experience, trust can be influenced by not just the physicality of the entity, but also aspects like culture and gender which are demographic factors [Akash et al. 2017]. In robotics, trust towards robots is a combination of human-related (expertise, competency, experience, demographic, comfort, etc.), robot-related (reliability, predictability, failure rates, personality, etc.), and environmental (membership, culture, communication, complexity, etc.) factors [Hancock et al. 2011]. These factors can be physiologically sensed, such as shown by Akash et al. [Akash et al. 2018] who used EEG and GSR to measure trust using a desktop-based driving game.

2.3 Relationship between Cognitive Load and Trust

Cognitive load is defined as the amount of working memory required for a task and has been well explored among researchers. For example, Zhang et al. [Zhang et al. 2017] measured cognitive load using a combination of physiological signals to assist autistic individuals in driving. The author fused eye gaze, EEG, peripheral physiology modality (ECG, EMG, RSP, SKT, PPG and GSR) and performance modality into a feature set for classification and achieved an 83% accuracy. Recent work has also used EEG to measure cognitive load in VR environments. For example, Gerry et al. [Gerry et al. 2018] and Dey et al. [Dey et al. 2019] measured cognitive load using EEG to create a VR adaptive training system. Gerry et al. showed clear activity in the alpha band (8 - 13Hz) whereas Dey et al. found an increase in alpha synchronicity when the presented task is harder.

Trust however, as mentioned previously, is a psycho-physiological state. This means that it carries cognitive components because it is the calculation of subjective probability given a specific situation, which explains its correlation with cognitive load [Rempel et al. 1985; Samson and Kostyszyn 2015]. Among related works that aimed to correlate cognitive load with trust are Samson et al. [Samson and Kostyszyn 2015] and Khawaji et al. [Khawaji et al. 2015] who used a trust game and GSR sensing in a text chat environment respectively. However, neither work addressed this in a VR environment nor how such work can be used to contribute towards VR interface design.

In summary, researchers have explored the use of physiological cues for measuring cognitive load and trust using various means. We have summarized these findings on Table 1. To our knowledge, there have been no previous experiments in VR to measure trust with physiological signals. There has also been little previous work that explores the relationship between cognitive load and trust. However, the shape selector task in VR presents a useful baseline for us to iterate from, while EEG and GSR has been used to measure trust outside of VR. Our main research question is the following: What kind of physiological signal features can provide a continuous objective measure of trust and cognitive load in VR? Our research will help address this question, especially for experiments conducted in a VR environment. Finally, we would like to suggest key applications and scenarios where both cognitive load and trust can be leveraged to improve current VR experiences.

3 EXPERIMENTAL EVALUATION

For this study, we investigated the relationship between human trust of a virtual assistant and the cognitive workload while performing an agent-assisted task in VR. We gathered participants' physiological information, specifically EEG, GSR and HRV as objective measures to establish the trust relationship along with their behavioural data and traditional self-reporting questionnaires for behavioural and subjective measures respectively.

3.1 Prototype Design

In this section, we describe the training/task system and its components along with a virtual voice agent used to assist the player to complete the task. The experimental task chosen was target object detection in an immersive VR environment. The VR application included a voice agent that informs the user in which direction (relative to the user) a target object is located. The voice-assisted VR system has six components:

- OpenBCI EEG Electrode cap with Cyton-daisy module for 16-channels support at 125Hz sampling rate [OpenBCI [n.d.]],
- OpenBCI GUI v4.1.5 for EEG data acquisition and streaming to Unity [OpenBCI [n.d.]],
- Shimmer3 GSR+ Sensing device for sensing GSR and HRV signals at 128Hz sampling rate,
- Java application for Shimmer data acquisition,
- Unity 3D game engine 2019.2.4 [Unity [n.d.]] for the n-back and shape selector task, and
- First generation HTC Vive VR HMD [Vive [n.d.]] for the VR environment display and to enable interactions.

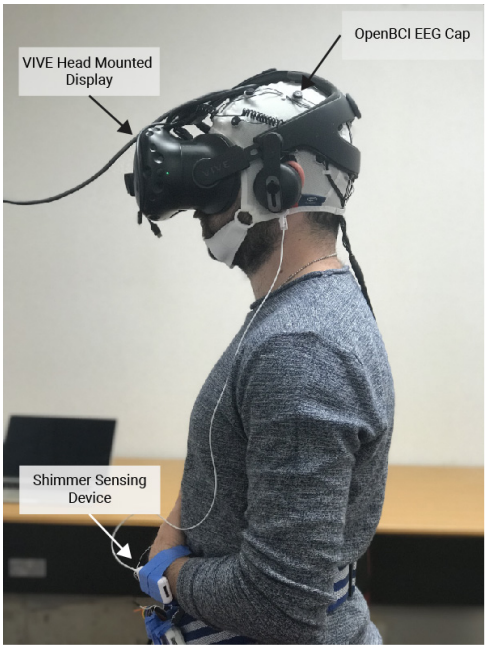

Figure 2: Participant with OpenBCI EEG Cap, Shimmer GSR+ Sensing Device, and HTC Vive HMD setup

3.1.1 Virtual Reality System. We used the HTC Vive virtual reality Head Mounted Display(HMD) [Vive [n.d.]] to enable the participate to perform tasks in virtual environment using Unity Virtual Reality (VR) [Unity [n.d.]] application.

As a baseline cognitive workload test, we designed a delayed digital recall task (n-back) for n = 1, 2, and 3 where incremental values of n indicate increasing difficulty level. The n-back task is a standard test which asks people to recall nth-number or character n before the currently displayed one [Kirchner 1958]. We implemented this in VR where a character was displayed for only a period of 0.5 seconds. Then for the next 3 seconds, the participant has to choose the matching or non-matching character by pressing right controller's trigger or left controller's trigger signifying true or false respectively to make their decision, before the next character appears. We implemented a circular progress bar to inform the participants about the time remaining to press the trigger. If they didn't make any choice before time was up, it was considered as an incorrect response. We recorded all the events including new round, trigger hit, times up, correct and incorrect choices, round completion time, etc. (see figure 3). Previous research has shown that EEG can be used to measure increasing cognitive load as the n-back task become more difficult [Dey et al. 2019]. In this case for each of n = 1, 2, and 3, we measured the EEG alpha power level (8 - 13Hz), GSR and HRV to establish a baseline measure for each user. This allows us to check if the main task reflects the correct cognitive load level based on the physiological signals.

For the main task, we used the Shape Selector Task of Dey et al. [Dey et al. 2019] and modified as per our system requirements. In this task, blocks of different shapes (cube, sphere and pyramid) and different colors (red, yellow, blue, green) were displayed in virtual reality with a "target" block (displayed on a pink platform for higher visibility). The objective for the user was to search for the "target" block among the displayed blocks and press the trigger of the controller when the participant had positioned a head pointer over the target (see figure 4). Participants were told that they should complete the task as fast as possible and that they had a limited amount of time. They were advanced to the next level after selecting the correct target shape object. At each level the number of distracting objects and their movement increased, so the task became more difficult.

We also introduced a virtual assistant agent who verbally informs the player about the direction (left or right depending on the angle between the players' head orientation and relative target position) of the target in order to help the player complete the task quickly and efficiently. The virtual assistant is not visually represented, but uses an audio cue of "Left" or "Right" to guide the user. In different conditions the accuracy of the assistant is either 100 percent or 50 percent [Akash et al. 2017].

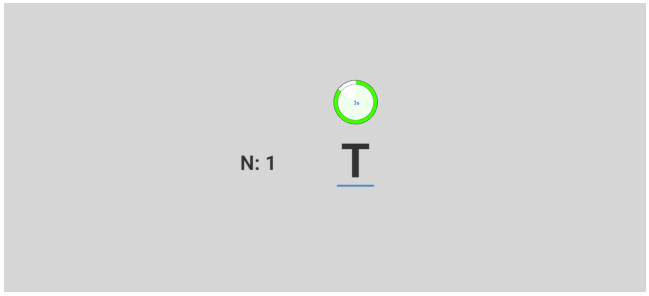

Figure 3: Virtual Environment Tasks: n-Back task

3.1.2 Physiological Sensing Devices. We measured participants' EEG physiological response using a 16-channel OpenBCI EEG Cap [OpenBCI [n.d.]] with gel-based electrodes. GSR and HRV physiological signals were measured using the Shimmer Sensing Device

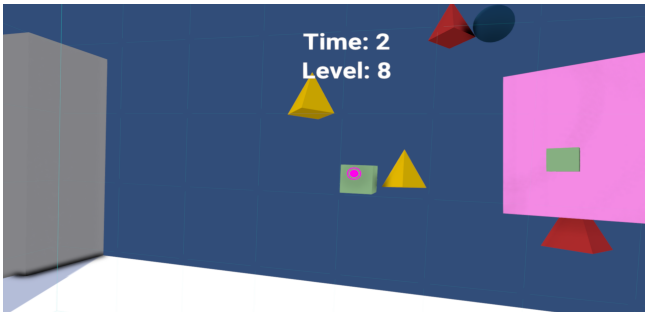

Figure 4: Virtual Environment Tasks: Shape Selector Task

[Shimmer [n.d.]] while performing the n-back test and shape selection tasks. For EEG, we used the data from electrodes placed near the pre-Frontal lobe responsible for decision making, and ability to concentrate (FP1 and FP2), and the electrodes at the parietal and occipital lobe [Dey et al. 2019] for measuring cognitive load, i.e. P3, Pz, P4, O1 and O2. We placed the Shimmer Device GSR electrode on the index and middle finger of the participants' non-dominant hand and HRV sensor electrode on the earlobes at the non-dominant side.

3.2 Participants and Design

We conducted a 2x2 within-subjects pilot study with two dependent variables - Cognitive Workload (CL) and Virtual Agent Accuracy (Acc) as factors. A total of 13 participants (6 Female and 7 Male; Age M: 26.61 , SD: 3.93) completed four conditions (LCL-LA, HCL-LA, LCL-HA and HCL-HA) according to Table 2.

Table 2: Experimental Conditions

	Cognitive Load: Low	Cognitive Load: High
Accuracy: Low	A: LCL-LA	B: HCL-LA
Accuracy: High	C: LCL-HA	D: HCL-HA

The order of condition for each participant was arranged in a Latin Square to eliminate potential ordering effects. To customize the Shape Selector Task as per the experiment conditions, for low cognitive load or easy tasks conditions (A and C) participants had 10 seconds to complete each level, whereas for high cognitive load or difficult tasks conditions (B and D), participants had only 5 seconds to complete each level. For low accuracy tasks (A and B), the agent's accuracy was 50 percent whereas for high accuracy tasks (C and D), the agent's accuracy was 100 percent. There were 20 levels to complete for each condition, making the total number of trials per participant 4 x 20 = 80 trials. From initial pilot testing, we found that the given time of 5 seconds and 10 seconds for the easy and difficult tasks respectively was sufficient yet challenging enough to locate the correct shape. Moreover we only look at 100% and 50% accuracy for the agent because it represents the opposite ends of a reliable and unreliable agent [Akash et al. 2018]. 100% accuracy indicates that the agent never lies, whereas 50% accuracy indicates that it is simply a matter of luck; above 50% would increase the chances of it being more accurate, and any lesser would make its unreliability more predictable.

All of the participants were above 18 years of age, native English speakers or fluent in English, familiar with computers and smartphones, and had some experience with virtual environments. They also all had some experience with using virtual assistants like Google Assistant, Apple Siri, Bixby, Amazon Alexa, etc. for tasks such setting an alarm, searching for a nearby cafe, and setting up a destination for car navigation.

3.2.1 Cognitive Load. Subjective cognitive load was measured using six questions based on the NASA Task Load Index questionnaire [Hart 1986] considering mental demand, physical demand, temporal demand, performance, efforts and frustration while doing the task. The average Task Load Index was calculated based on the NASA TLX score calculation technique that suggests a higher cognitive load for a higher TLX score. For physiological measures, we use EEG, GSR and HRV physiological cues to measure cognitive load. These sensors are worn by the participants throughout the experiment.

3.2.2 Trust. We used twelve questions based on the System Trust Scale (STS) developed by [Jian et al. 2000] to measure trust. These questions considered system deception, underhanded behaviour, trust, dependability, reliability, etc. factors. They are answered on a 5-point Likert-type scale (1= strongly disagree, 5= strongly agree). Additionally, we use a behavioural measurement of trust. To achieve this, we constantly log the direction of the head movement of the participant relative to the target (left or right), along with the direction informed by the agent (left or right as well) within the same timestamp throughout the experiment.

3.3 Procedure

We arranged a room with minimum radio frequency interference as there was a risk of extra noise in the physiological signals due to such interference. After welcoming the participants, they were first given a copy of the Consent Form (CF) and Participant Information Sheet (PIS) to fill at the start of the session with an opportunity to ask any questions about the study. Once they signed the CF, they were asked to complete the pre-task questionnaire including questions regarding demography, previous VR and Virtual assistant experience. After that we asked them to wash and dry their hands, then put the GSR and HRV sensors on their non-dominant hand. We then setup the OpenBCI EEG cap with gel in the electrodes followed by the Vive VR HMD, as shown in Figure 2. The entire setup had to be carefully completed as the EEG cap electrodes could be displaced from their position because of the HMD straps, resulting in faulty EEG data.

After resting for 10 minutes, we explained the n-back task and ran a practice session for participants to get familiarized with the interface and the task. Next, participants were asked to complete a delayed digital recall task (n-back) for n = 1, 2 and 3 for at least 15 correct answers. During the n-back test, EEG, GSR and HRV sensor data was recorded. After the n-back test, participants were asked to fill out the NASA TLX self-assessment questionnaire [Hart 1986].

Table 3: System Trust Scale questions

Q#	Statement
Q1	The system is deceptive
Q2	The system behaves in an underhanded manner
Q3	I am suspicious of the system's intent, action, or outputs
Q4	I am wary of the system
Q5	The system's actions will have a harmful or injurious outcome
Q6	I am confident in the system
Q7	The system provides security
Q8	The system has integrity
Q9	The system is dependable
Q10	The system is reliable
Q11	I can trust the system
Q12	I am familiar with the system

Post n-back task, we explained the Shape Selector task to the participant followed by a practice round of the task. Next, participants completed the shape selector tasks (4 tasks as in Table 2) in Latin-square order. In these tasks, the virtual assistant provided additional audio information about the direction (left or right relative to their head orientation) of the target object. As soon as the user selected the correct block, they advanced to the next level where the level number indicates the difficulty level and is decided based on the number of displayed objects and task complexity. There were a total of 20 levels per condition. Each condition took 10 minutes and there was 5 minutes of rest between each condition.

After each task, subjects completed the NASA TLX self-assessment questionnaire and the System Trust Scale self-assessment questionnaire [Jian et al. 2000]. At the end of the experiment, we conducted non-structured, open-ended interviews with the participants to understand their perspective and experiences while performing the tasks. Each session took approximately 60 minutes.

4 RESULTS

This section reports the experiment's analysis and results. First we report on System Trust Scale questionnaire, and then participants' head movement direction as a behavioral measure of trust recorded via head tracking. Next, we report on NASA TLX questionnaire for all the conditions followed by EEG physiological analysis with NASA TLX.

For GSR and PPG data collected through the Shimmer device, we converted PPG into HRV using the algorithm provided in Shimmer API. Later, we smoothed the GSR and HRV data using the moving average filter with a window of 3 seconds and then normalized it. For further GSR analysis, we used the Ledalab toolkit [Benedek and Kaernbach 2010] and performed continuous decomposition analysis to separate the tonic and phasic components. 'We later used Maximum Phasic component and Net phasic component as features to be used from GSR. As for HRV, we used Kubios [Tarvainen et al. 2014] to get mean Heart noRate, Low Frequency (LF), High Frequency (HF) and LF/ HF ratio for each condition and these features were used for further tests. Due to some technical problem with the Shimmer Device, we have lost participant 2 and 10 data, so excluded the results from them.

The rest of the data were manually inspected and cleaned (short windows with high peaks and flat data were removed). We choose to only take the EEG signals from the last 5 rounds of the shape-selector task as it should better reflect the participant's cognitive load or trust level for that particular condition. We took the average alpha band power (8 - 13Hz) from the selected seven electrodes and used a median filter to further remove noise.

To calculate the final trust score, we first reversed the rating for the negative valence questions i.e. Q1-5. For example, if someone rated 2 for Q1, we changed it to 4 and so on. Then we calculated a score from the System Trust Scale questionnaire by averaging the ratings for all the questions. We performed the repeated measure two-way ANOVA (α = 0.05) on the STS scores for all conditions to determine participant's trust perception on the Virtual agent system. The test revealed that there was a significant main effect of accuracy on participants ($F(1,12) = 24.870, p < 0.001$). There was no significant main effect of cognitive load ($F\ (1,\ 12) = 2.004,\ p = 0.185$) and no significant interaction between CL and Accuracy ($F\ (1,\ 12) = 0.276,\ p = 0.610$). As the STS scores for all of the conditions were normally distributed, we ran the Pearson Correlation test to determine the relationship between the STS responses for tasks conditions A, B, C, and D. There was a statistically significant positive correlation between A - C ($r_s(12) = 0.580,\ p=0.048$) and a stronger and positive correlation between B - C ($r_s(12) = 0.737,\ p=0.006$).

To understand user behaviour, we use head tracking to record participants' head movement and directional assistance provided by virtual agent after every new round for all the tasks. We followed the already discussed belief of trust [Susan and Holmes 1991] that if the participant is willing to follow agent's assistance i.e. directional information, this can indicate that the participant is trusting the agent. For the behavioral analysis, head movement in the direction suggested by the agent was considered as trust and assigned 1 , whereas head movements opposite to the suggested direction was no-trust and assigned as 0 for every frame. At the completion of each condition, we averaged the assigned head movement data that gave us a value between the scale of 0 to 1 and used it as a behavioral trust parameter. A repeated measure two-way ANOVA (α = 0.05) revealed that there was a significant main effect of Accuracy ($F(1,12) = 20.769,\ p = 0.001$) on the averaged head movement data to search for target objects to complete the tasks in these conditions. No significant difference was found for Cognitive Load conditions ($F(1,12) = 0.322,\ p = 0.582$) and interaction effect between Accuracy and Cognitive Load ($F(1,12) = 2.156,\ p = 0.170$).

We calculated the unweighted average NASA TLX Score [Hart 1986] and performed a repeated measure two-way ANOVA test (α = 0.05), finding that there was a significant main effect of cognitive load conditions ($F(1,12) = 8.091,\ p = 0.016$) and interaction effect of cognitive load and accuracy ($F(1,12) = 10.832,\ p = 0.007$) on participants for perceived task load. There was no significant effect of accuracy on the task load.

As the NASA TLX Score was reported normally distributed through Shapiro Wilk test, for further investigation to determine the relationship between task conditions A, B, C, and D for NASA TLX score, we performed the Pearson product-moment correlation. Results revealed that there is a strong, positive statistically

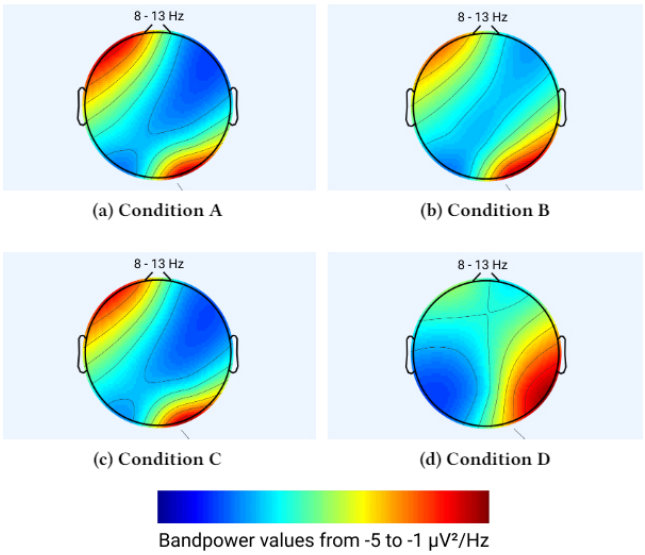

Figure 5: Heatmap of the average EEG bandpower across the 7 selected electrodes while performing the tasks (Red to blue color shows high to low active region)

significant correlation between A - B ($r_p(12) = 0.856, p < 0.001$), A - C ($r_p(12) = 0.918, p < 0.001$), A - D ($r_p(12) = 0.891, p < 0.001$), B - C ($r_p(12) = 0.831, p=0.001$), B - D ($r_p(12) = 0.919, p < 0.001$), and C - D ($r_p(12) = 0.936, p < 0.001$) for perceived Task Load including cognitive load.

Next, we analyzed the relationship of EEG alpha waves (8 - 13Hz) on task conditions A, B, C, and D using Spearman's rank-order correlation. There was a strong, positive statistically significant correlation between A - C ($r_s(12) = 0.706, p = 0.010$). No significant correlation was found between A - B ($r_s(12) = 0.266, p = 0.403$), A - D ($r_s(12) = 0.217, p = 0.499$), B - C ($r_s(12) = 0.558, p = 0.059$), B - D ($r_s(12) = 0.448, p = 0.145$), and C - D ($r_s(12) = 0.343, p = 0.276$) conditions. This is also seen in Figure 5 where Condition A and C were closest in terms of brain region of activation.

Through HRV analysis, we extracted LF, HF and LF/ HF ratio as features. We used the LF/ HF ratio feature for our results and performed the repeated two-way ANOVA test. The test reported that there was no significant main effect of CL ($F(1,11) = 0.669, p = 0.433$) or Accuracy ($F(1,11) = 0.061, p = 0.810$) and no significant interaction effect of CL and Accuracy ($F(1,11) = 3.176, p = 0.105$). Then we conducted the Spearman's rank-order correlation test on it and found a strong, negative correlation between B and C ($r_s(11) = -0.682, p = 0.021$). No significant correlation was found between A - B ($r_s(11) = -0.536, p = 0.089$), A - C ($r_s(11) = 0.318, p = 0.340$), A - D ($r_s(11) = 0.309, p = 0.355$), and B - D ($r_s(11) = 0.309, p = 0.355$), and C - D ($r_s(11) = 0.045, p = 0.894$).

From GSR analysis following Akash et al. [Akash et al. 2018] method, we used Net Phasic component for this research and performed the repeated two-way ANOVA test. The test reported that there was no significant main effect of CL (F(1,12) = 0.501, p = 0.495) or Accuracy ($F(1,12) = 3.406, p = 0.095$) and no significant interaction effect of CL and Accuracy. Then we conducted the Spearman's rank-order correlation test and could not find any correlation between A - B ($r_s(11) = 0.245, p = 0.467$), A - C ($r_s(11) = -0.304, p = 0.363$) , A - D ($r_s(11) = 0.548, p = 0.081$) , B - C ($r_s(11) = -0.012, p = 0.973$) , and B - D ($r_s(11) = 0.007, p = 0.983$) , and C - D ($r_s(11) = 0.327, p = 0.326$). At the end, we plotted a bar graph (Figure: 6) with the average net phasic GSR data of each conditions to observe the GSR pattern as used by Khawaji et al. [Khawaji et al. 2015].

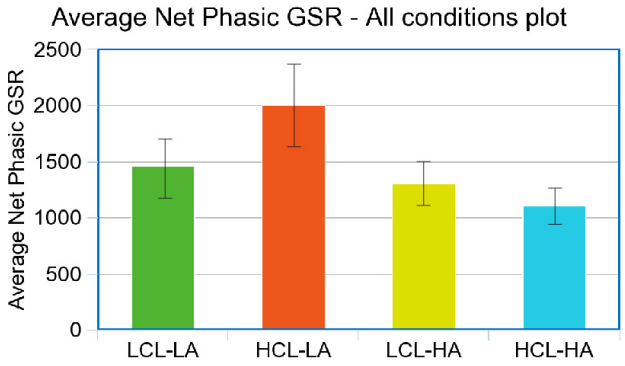

Figure 6: Plot of Average Net Phasic GSR for all conditions

5 DISCUSSION

From the results of the experiment, we found that participants had higher trust for high agent accuracy tasks as compared to the low agent accuracy tasks as indicated by one participant saying *"I can trust the system"*. The System Trust Scale questionnaire results indicate that the accuracy of the device had an effect on perceived trust, with the data showing a significant effect of accuracy of the agent on the answers to the questions. These results demonstrate that participants were able to identify whether the agent is accurate or inaccurate and whether they can trust the agent or not.

The accuracy of the device also had an effect on the participant's behaviour which aligns with what they perceived. Consequently, they followed the agent's direction when the agent was accurate and less so when the advice was less accurate. This was also For example, one participant said *"This guy is lying to me! I can't trust him anymore. He is like my husband!"* while performing the low accuracy task.

The significant main effect of perceived trust (System Trust Scale) on task conditions and behavioral trust (head movement relative to the direction told by the Agent) on conditions and a strong, positive correlation between conditions suggests that perceived trust has a correlation with behavioral trust through our Shape Selector Task. This means that a head movement task based on directions by a virtual agent can be used to evaluate how much the user trusts the agent. To allow users to experience the system's performance and calibrate trust, we only used the last 5 rounds of the shape selector task to evaluate the behaviour. We believe that at the initial stage, most users are primarily simply reacting to an order or request, as opposed to actually trusting it. This is similar to a reaction to an audio or visual stimuli because it simply grabs our attention, but not out of trust.

Regarding the n-Back test, we couldn't find any relationship between the baseline cognitive load and Shape Selector tasks cognitive load. We suspect this is simply due to the task not being difficult enough, or that participants simply guessed the answer, resulting in incorrect alpha values. However, we were clearly able to demonstrate through NASA TLX questionnaire analysis for all the conditions that the tasks conditions with low cognitive load (A, C) were significantly different than conditions with high cognitive load (B, D) along with main effect of cognitive load & Accuracy interaction. On further investigation, we found that for the low accuracy conditions, the task load was increased when the participant performed low cognitive load task and then a high cognitive load task (A and B) and vice versa. This was same for low to high cognitive load with participant performing in high accuracy tasks (C and D). This correlation suggests that our Shape Selector task system was able to increase the perceived task load when performing low and high cognitive load tasks. We believe this correlation can be improved by providing harder difficulty levels for the shape selector. We will explore it in our further studies.

In our experiment, we could not find any significant main effect or interaction effect of Cognitive Load and Accuracy on any of the physiological signals (EEG, HRV, and GSR). However, we found significant correlations between the conditions. As we already know that alpha band power from EEG is inversely proportional to cognitive load [Antonenko et al. 2010], our study suggested that during the low cognitive load tasks (A and C), the average alpha band power increased (indicating decrease in physiological cognitive load) when the participant performed with a trustworthy agent (high accuracy tasks, C and D). This means that with a task that requires low cognitive load, the alpha band power measurement reflects the trustworthiness of the agent. However, we have yet to prove this for a more difficult task with increased cognitive load.

From previous work [McDuff et al. 2016], we know that with the increase in cognitive load, the HRV's LF-HF ratio also increases. When we exposed the participants to a high cognitive load task when the agent was inaccurate (condition B), if their LF-HF ratio is high, their physiological cognitive load should be low for task with low cognitive load and high agent accuracy (condition C) as we have already determined that condition B is negatively correlated with condition C. This suggests that cognitive load should decrease when the agent is accurate regardless of the task's cognitive load. Whereas if the agent is inaccurate, task could result in high cognitive load despite of task being easy (low cognitive load).

As mentioned by Khawaji et al. [Khawaji et al. 2015], if the participant is asked to experience a low cognitive load task with an accurate agent (when the participant's trust is high), the average GSR should be at lowest level indicating a low physiological cognitive load. From our results (check Figure: 6), we observed that the average Net Phasic GSR was lowest even when the participant was performing a high cognitive load task with an accurate agent. However, we could not report its statistical significance. This could be possible because it was a lab study where the participants were in a comfortable environment that made them feel less stressed.

One assumption we can make from these pilot results is that, when the task load becomes higher, there is a possibility that the alpha band power would reflect the cognitive load more than trustworthiness of the agent. However, further studies with larger sample size would be needed to verify this.

5.1 Proposed VR Scenarios

Based on our findings, we believe that VR technology as a whole can benefit from increased understanding in trust and cognitive load. Virtual agents, as demonstrated in this study, can be optimized not just in terms of appearance, but in performance as well to better earn the trust of the user. Trust in technology should be carefully balanced; too much trust would result in negligence, whereas too little of it could render the technology useless. Besides virtual agents, understanding trust is useful for also designing the virtual environment itself. VR aided medical simulation have been researched and some are being used in practice [Oxford Medical Simulation [n.d.]; Stansfield et al. 2000; Willaert et al. 2012]. While most of these are focusing on cost reduction or realistic simulation, these do not cast a light on trainee's state. Observing trainee's physiological status and trust for the VR aid may realize adaptive and effective teaching method to convey them how to handle a situation. Advances in wireless communication technology and light-weight wearable AR system enable the development of novel remote collaboration systems. Although the technology is already used in some of the industries [Google Glass [n.d.]; Posada et al. 2015], more research is needed in order to improve the remote collaboration itself. Measuring workers trust and cognitive load for the system could help users manage remote collaboration tasks, for example in critical task environments such as factories, help in finding bottlenecks in workflow and improve overall productivity.

6 LIMITATIONS

In this section, we describe some of the limitations faced during the study. First, the OpenBCI EEG hardware has limited spatial and temporal resolution. Other EEG sets may provide better signal to noise ratio. However, since we are collecting EEG with GSR and HRV simultaneously, we had to synchronise timestamps, which may lead to artificial upsampling of the data since each sensor has different sampling rates. The synchronizing was achieved by having the Java application (streaming the GSR and HRV signals) and OpenBCI GUI (streaming the EEG signals) both stream to Unity which outputs the final signal file containing all sensor values in the same timestamp, alongside other data such as head movement.

There were also some difficulties when using the HTV Vive HMD with the EEG cap, primarily because both devices are mounted on the head. When the HMD is mounted on the EEG cap, it clamps down on the prefrontal, parietal and occipital lobes (the tightening mechanism squeezes the front and back section with a knob). Regions that are not clamped down instead causes the electrodes to lose contact with the scalp, particularly around the temporal and frontal lobe. This results in poorer signals at that region, which needs to be considered if those channels are involved in future studies. Furthermore, the HMD needs to be tight enough, without causing discomfort to the participant, otherwise small movements will instead displace the HMD, inducing noise into the clamped regions as well. Finally, the lack of significant effects in the n-back task may be explained by some participants who claimed that they

were partly guessing the answer rather than trying to make the correct choice, leading to low cognitive load especially when n = 3.

7 CONCLUSION AND FUTURE WORKS

The primary objective of our evaluation study was to determine the correlation between physiological signals with trust and cognitive load while performing tasks with cognitive load and virtual agent accuracy as dependent variables. To explore this we modified a Shape Selector Task where a player searched for a target object in an immersive virtual environment. We also added a virtual agent who informed the user about the direction of the target object with respect to the user's head position with adjustable accuracy.

In terms of the trust, we found that most of the time there was a significant main effect of agent's accuracy on perceived trust with a significant correlation with behavioral trust. We also determined that there was a significant main effect of cognitive load on perceived task load across the task conditions. On further analysis, a significant positive correlation of cognitive load with accuracy of the agent was reported. These results demonstrate how the experimental design method in VR environment we have developed and presented here can be used to evaluate trust of a virtual agent.

In our physiological data investigation, we found that there was a strong and positive significant correlation of average alpha band power at the pre-frontal, parietal and occipital lobe of the brain with agent's accuracy in a low cognitive load task. Furthermore, there was a strong, negative significant correlation of LF-HF ratio with the agent's accuracy and cognitive load suggesting that if the agents is accurate, then the physiological cognitive load state can be decreased regardless of the task difficulty. The reason for this will need to be investigated in future research, but we suspect that this could simply be due to the alpha channels favoring cognitive load over trustworthiness (higher correlation with cognitive load than trust). Overall, we conclude that these results along with this pilot study is one of the first approach towards exploring the correlation between physiological signals with trust and cognitive load in VR.

For future works, we propose to either use another cognitive load task like the Stroop Test [Gwizdka 2010], or possibly tweaking the parameters of the n-back task and/or shape selector task by further increasing the difficulty to obtain more significant results in terms of cognitive load correlation. For example, we may set the time limit to per condition instead of per task and observe the differences. The conditions with low cognitive load can be assigned a time limit of 10 minutes to complete all the tasks, whereas the high cognitive load conditions require only 5 minutes to complete all the tasks. We may also experiment with a wider range of agent accuracy, such as being 75% accurate to establish the transition of trust.

For the EEG signals, we only select the time window of the last 5 rounds of the shape selector task. However, other time windows could possibly be experimented with, as well as signals from other frequency bands, such that the beta, gamma, theta and delta band [Jensen and Tesche 2002]. We would also like to combine eye-tracking with head movement to explore if it can provide a better and robust trust model instead of only head-movement.

We also plan to develop a machine learning model based on our currently available dataset to train a model capable of predicting both the trust and cognitive load level based on physiological signals in real time. Finally, we would like to use this model to optimize virtual entities, such as changing the appearance or voice of a virtual avatar based on a user's trust level.

ACKNOWLEDGMENTS

Part of the work was funded by the Data61, UTAS UniSA Automation, trust and workload CRP.

REFERENCES

Mohamed Abouelenien, Mihai Burzo, Rada Mihalcea, Kristen Rusinek, and David Van Alstine. 2017. Detecting Human Thermal Discomfort via Physiological Signals. In *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '17)*. ACM, New York, NY, USA, 146–149. https://doi.org/10.1145/3056540.3064957

Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A classification model for sensing human trust in machines using eeg and gsr. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 4 (2018), 27.

Kumar Akash, Wan-Lin Hu, Tahira Reid, and Neera Jain. 2017. Dynamic modeling of trust in human-machine interactions. In *2017 American Control Conference (ACC)*. IEEE, 1542–1548.

Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara Van Gog. 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review* 22, 4 (2010), 425–438.

Olivier Augereau, Benjamin Tag, and Koichi Kise. 2018. Mental State Analysis on Eyewear. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (UbiComp '18)*. ACM, New York, NY, USA, 968–973. https://doi.org/10.1145/3267305.3274119

Bernard Barber. 1983. *The logic and limits of trust*. Vol. 96. Rutgers University Press New Brunswick, NJ.

Mathias Benedek and Christian Kaernbach. 2010. A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods* 190, 1 (2010), 80–91.

Alanah Davis, John D Murphy, Dawn Owens, Deepak Khazanchi, and Ilze Zigurs. 2009. Avatars, people, and virtual worlds: Foundations for research in metaverses. *Journal of the Association for Information Systems* 10, 2 (2009), 90.

Arindam Dey, Alex Chatburn, and Mark Billinghurst. 2019. Exploration of an EEG-Based Cognitively Adaptive Training System in Virtual Reality. (2019).

Suh-Yeon Dong, Bo-Kyeong Kim, Kyeongho Lee, and Soo-Young Lee. 2015. A Preliminary Study on Human Trust Measurements by EEG for Human-Machine Interactions. In *Proceedings of the 3rd International Conference on Human-Agent Interaction (HAI '15)*. ACM, New York, NY, USA, 265–268. https://doi.org/10.1145/2814940.2814993

Jérémy Frey, Maxime Daniel, Julien Castet, Martin Hachet, and Fabien Lotte. 2016. Framework for electroencephalography-based evaluation of user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2283–2294.

Lynda Gerry, Barrett Ens, Adam Drogemuller, Bruce Thomas, and Mark Billinghurst. 2018. Levity: A Virtual Reality System That Responds to Cognitive Load. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, Article LBW610, 6 pages. https://doi.org/10.1145/3170427.3188479

Google Glass [n.d.]. *Google Glass - Glass Partners*. https://www.google.com/glass/partners/

David Grimes, Desney S Tan, Scott E Hudson, Pradeep Shenoy, and Rajesh PN Rao. 2008. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 835–844.

Jacek Gwizdka. 2010. Using Stroop task to assess cognitive load. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*. ACM, 219–222.

Joanna Hale and F De C Antonia. 2016. Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific reports* 6 (2016), 35295.

Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

Sandra G Hart. 1986. NASA Task load Index (TLX). Volume 1.0; Paper and pencil package. (1986).

Blanca Hernandez-Ortega. 2011. The role of post-use trust in the acceptance of a technology: Drivers and consequences. *Technovation* 31, 10-11 (2011), 523–538.

Leanne M Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert JK Jacob, Angelo Sassaroli, and Sergio Fantini. 2009. Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2185–2194.

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine* 49, 32 (2016), 48–53.

Sirkka L Jarvenpaa, Kathleen Knoll, and Dorothy E Leidner. 1998. Is anybody out there? Antecedents of trust in global virtual teams. *Journal of management information systems* 14, 4 (1998), 29–64.

Ole Jensen and Claudia D Tesche. 2002. Frontal theta activity in humans increases with memory load in a working memory task. *European journal of Neuroscience* 15, 8 (2002), 1395–1399.

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.

Ahmad Khawaji, Jianlong Zhou, Fang Chen, and Nadine Marcus. 2015. Using galvanic skin response (GSR) to measure trust and cognitive load in the text-chat environment. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 1989–1994.

Wayne K Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology* 55, 4 (1958), 352.

Juyoung Lee, Hui-Shyong Yeo, Murtaza Dhuliawala, Jedidiah Akano, Junichi Shimizu, Thad Starner, Aaron Quigley, Woontack Woo, and Kai Kunze. 2017. Itchy Nose: Discreet Gesture Interaction Using EOG Sensors in Smart Eyewear. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers (ISWC '17)*. ACM, New York, NY, USA, 94–97. https://doi.org/10.1145/3123021.3123060

Rachel McDonnell and Martin Breidt. 2010. Face reality: investigating the uncanny valley for virtual faces. In *ACM SIGGRAPH ASIA 2010 Sketches*. ACM, 41.

Daniel J. McDuff, Javier Hernandez, Sarah Gontarek, and Rosalind W. Picard. 2016. COGCAM: Contact-free Measurement of Cognitive Stress During Computer Tasks with a Digital Camera. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4000–4004. https://doi.org/10.1145/2858036.2858247

Bonnie M Muir. 1987. Trust between humans and machines, and the design of decision aids. *International journal of man-machine studies* 27, 5-6 (1987), 527–539.

OpenBCI [n.d.]. *OpenBCI - Open Source Brain Computer Interfaces*. https://shop.openbci.com/collections/frontpage/products/openbci-eeg-electrocap-kit

Oxford Medical Simulation [n.d.]. *Oxford Medical Simulation*. https://oxfordmedicalsimulation.com/

Yun Suen Pai. 2016. Physiological Signal-Driven Virtual Reality in Social Spaces. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology (UIST '16 Adjunct)*. ACM, New York, NY, USA, 25–28. https://doi.org/10.1145/2984751.2984787

J. Posada, C. Toro, I. Barandiaran, D. Oyarzun, D. Stricker, R. de Amicis, E. B. Pinto, P. Eisert, J. Döllner, and I. Vallarino. 2015. Visual Computing as a Key Enabling Technology for Industrie 4.0 and Industrial Internet. *IEEE Computer Graphics and Applications* 35, 2 (2015), 26–40.

Lingyun Qiu and Izak Benbasat. 2005. Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International journal of human-computer interaction* 19, 1 (2005), 75–94.

John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of personality and social psychology* 49, 1 (1985), 95.

David Rozado and Andreas Dunser. 2015. Combining EEG with pupillometry to improve cognitive workload detection. *Computer* 48, 10 (2015), 18–25.

Davide Salanitri, Glyn Lawson, and Brian Waterfield. 2016. The Relationship Between Presence and Trust in Virtual Reality. In *Proceedings of the European Conference on Cognitive Ergonomics (ECCE '16)*. ACM, New York, NY, USA, Article 16, 4 pages. https://doi.org/10.1145/2970930.2970947

Katarzyna Samson and Patrycjusz Kostyszyn. 2015. Effects of cognitive load on trusting behavior–an experiment using the trust game. *PloS one* 10, 5 (2015), e0127680.

Shimmer [n.d.]. *Shimmer GSR+ Sensor*. https://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor

Sharon Stansfield, Daniel Shawver, Annette Sobel, Monica Prasad, and Lydia Tapia. 2000. Design and Implementation of a Virtual Reality System and Its Application to Training Medical First Responders. *Presence: Teleoperators and Virtual Environments* 9, 6 (2000), 524–556.

D Susan and John G Holmes. 1991. The dynamics of interpersonal trust: Resolving uncertainty in the face of risk. *Cooperation and Prosocial Behavior; Cambridge University Press: New York, NY, USA* (1991), 190.

Wioleta Szwoch. 2015. Emotion Recognition Using Physiological Signals. In *Proceedings of the Mulitimedia, Interaction, Design and Innnovation (MIDI '15)*. ACM, New York, NY, USA, Article 15, 8 pages. https://doi.org/10.1145/2814464.2814479

Mika P Tarvainen, Juha-Pekka Niskanen, Jukka A Lipponen, Perttu O Ranta-Aho, and Pasi A Karjalainen. 2014. Kubios HRV–heart rate variability analysis software. *Computer methods and programs in biomedicine* 113, 1 (2014), 210–220.

Unity [n.d.]. *Unity*. https://unity3d.com/unity/whats-new/2019.2.4

Vive [n.d.]. *Vive*. https://www.vive.com/us/product/vive-virtual-reality-system/

Willem I. M. Willaert, Rajesh Aggarwal, Isabelle Van Herzeele, Nicholas J. Cheshire, and Frank E. Vermassen. 2012. Recent Advancements in Medical Simulation: Patient-Specific Virtual Reality Simulation. *World Journal of Surgery* 36, 7 (2012), 1703–1712.

Lian Zhang, Joshua Wade, Dayi Bian, Jing Fan, Amy Swanson, Amy Weitlauf, Zachary Warren, and Nilanjan Sarkar. 2017. Cognitive load measurement in a virtual reality-based driving system for autism intervention. *IEEE transactions on affective computing* 8, 2 (2017), 176–189.