Affect-Based Trust Estimation in Human-Robot Collaboration

Establishing a correlation between physiological response and human trust in robots.

Medialogy Master Thesis

MTA191042

Aalborg University
10th Semester Medialogy
Rendsburggade 14
9000 Aalborg, Denmark

AALBORG UNIVERSITY
STUDENT REPORT

School of Information and Communication Technology
Aalborg University
http://www.sict.aau.dk

Title:
Affect-Based Trust Estimation in Human-Robot Collaboration

Theme:
Human-Robot Interaction

Project Period:
Spring Semester 2019

Project Group:
MTA191042

Author(s):
Anders Skaarup Johansen
Jesper Wædeled Henriksen

Supervisor(s):
Matthias Rehm

Page Numbers: 39

Date of Completion:
May 27, 2019

Abstract:

This project investigates whether there is a correlation between physiological response, trust, and the impact speed has on human-robot collaborative tasks. Using galvanic skin response, heartbeat-rate, pupil-dilation and body language for affect detection and comparing the physiological responses of those modalities with a subjective evaluation of the robots performance and behavior. While there wasn't a statistically significant correlation between trust and physiological response, it was discovered that speed plays a significant role when directly approaching or retracting from a person.

Anders Skaarup Johansen
<ajoha14@student.aau.dk>

Jesper Wædeled Henriksen
<jhenri13@student.aau.dk>

Contents

1 Introduction **1**

2 Theory **2**
2.1 Human Robot Interaction 2
2.2 Trust In Robots 3
2.3 Quantifying Emotions 4
2.3.1 Cardiac Response 5
2.3.2 Electrodermal Activity 5
2.3.3 Non-Verbal Communication 6
2.4 Multimodal Approaches 6

3 Proposed Solution **8**

4 Implementation of Solution **9**
4.1 Physiological Measurements 9
4.1.1 Signal Processing 10
4.1.2 Robot Programming 13

5 Evaluation **15**
5.1 Setup of Evaluation 15
5.2 Evaluation Design 16
5.3 Subjective Assessment 16

6 Results and Analysis **18**
6.1 Trust 18
6.1.1 Physiological Response 21
6.2 Posture 23
6.2.1 Additional Observations 26
6.2.2 Pupillary response 27

7 Conclusion **28**

A Code Snippets **33**

B Trust Questionnaire **34**

C Complete Trust Scores **37**

D Scoring Schemes **39**

Chapter 1

Introduction

In the world of robotics there are a wide range of target areas e.g. production, construction, cleaning. In most cases the robot is working alone. It might vacuum your house or cut your grass or be part of a production line where it receives some product that it process and pass along. In these production lines the interaction between the robots are not important as long as the production line is working (e.g. receiving the correct x and outputting the right y). However, in some areas a collaboration between humans and robots would increase the quality of the product, while also reducing the chance of stress related injuries. One of these areas is the butchery, Danish Crown. In order for man and machine to work together the safety of the humans have to be assured, but also the perceived safety is important. The workers who would have to start working with these robots might not be familiar with robots at all and could be very reluctant when they saw a robot swinging a knife that cuts pigs in half in under a second. Even though the robot was perfectly safe and had unlimited safety measures, the worker might still be hesitant when working with the robot. Therefor it is very important to reduce the stress and increase the trust that people have in the collaborative robots, as it has been shown that there is a correlation between trust in robots and the effectiveness of the collaboration between the robot and the human.

To create such a system the robot would need the ability to assess each human and be able to change behavior based on said assessment. These changes could be alter the speed, move away from the human, emergency stop, pause, wait for input, etc. Using an assessment system one could ease the human into the new shared work space and give them time to get familiar with the robot at a pace that would fit the individual. Such a system would require the robot to be able to sense when the human was uncomfortable. Research into human emotions has proved that emotions can be estimated using physiological measurements such as heart rate or galvanic skin response (Mehrabian 1996, Bütepage & Kragic 2017). Pupil dilation can help estimate cognitive load (Xu et al. 2011), and emotional arousal (Bradley et al. 2008).

This project aims to investigate the correlation between the emotional state of a human performing a collaborative task together with a robot and the amount of trust that the person has in the robot. This is achieved by conducting a series of tests where a human has their physiological responses measured while performing varying tasks in collaboration with a robot, at various speeds. The project also aims to investigate the impact that a robot's movement-speed has on human trust in the robot.

Chapter 2

Theory

2.1 Human Robot Interaction

The advancement in robotics in recent years has resulted in a drastic increase of humans and robots sharing work-spaces, and as such human-robot interaction (HRI) has become an increasingly important field to insuring that HRI and human-robot collaboration (HRC) tasks are safe and intuitive for the human.

When a human and an autonomous robot perform a task in close proximity, the robot needs to be aware of the human's actions and it needs to be able to adapt to those actions. Traditionally to avoid this issue, HRI has been limited to a master-slave principle, where the robot is used as a tool to perform an action instead of working in parallel with humans. When conducting collaborative work with a robot, it is important for the human to feel safe and trust the actions of the robot. By giving the robot the ability to sense how the human is feeling, and adapt to human social cues, it is possible to dynamically alter the robots actions to provide a more trustworthy and pleasant experience for the human, and reduce risk of serious injury.

According to Bütepage & Kragic (2017), HRI can be separated into three categories of interactions;

1. **Instruction**
   Where the robot performs a sequence of actions which are governed exclusively by the humans decision making. In this case the task follows the master-slave principle, and that prevents any advancements in social interactions or a parallel workflow. As is shown in Figure 2.1, the robot is used as a tool to work towards the humans goal.

2. **Cooperation**
   Where the robot and human are working towards a shared goal by completing subtasks separately. While the goal remains shared the subtasks are independent from each other and do not contribute towards the partners progress in any way. As shown in Figure 2.1 the work flow is entirely separated and parallel.

3. **Collaboration**
   Much like cooperation, in collaboration the human and the robot are working towards a shared goal. However, the subtasks performed by either partner has an interaction with the other partner's subtask to some extent. In Figure 2.1 it can be seen that this is very similar to cooperation with the difference being the workflows are interacting with each other.

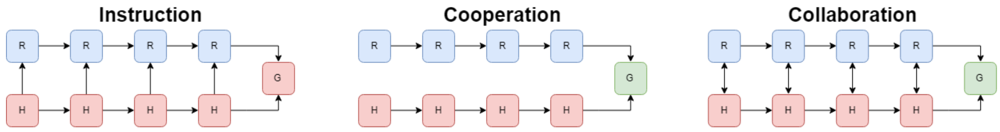


Figure 2.1: This figure illustrates the direction of interactions during an arbitrary HRI task, and shows the subtle but significant differences between the different categories of HRI

While collaboration and cooperation are very similar, they are conceptually different. An example that illustrates this would be a car manufacturing task. Cooperation would mean that the robot holds the door and then screw in the bolts by itself while the human would be doing the same to the other door on the other side of the vehicle. The collaborative version of that would be for instance the robot holding the door in the right position as the human screws in the bolts to keep the door in place. For this to be an efficient solution the robot needs to be accepted socially and emotionally by it's human partner.

2.2 Trust In Robots

It has been observed that during human-human interactions, as the humans interact they will both send and receive emotional signals which in turn adjusts their affective state, depending on the signals (Papousek et al. 2008). As shown by Papousek et al. (2008) this plays a very important role in how humans feel about the interaction. If the robot were to successfully assume a position in a social workspace, where it is trusted by it's human partners, it should be able to receive and respond to human social. For this to happen successfully, the robot must be able to detect and interpret the behavior and affective state of the human partner. For this project trust is defined as: *System automation trust is defined as having confidence in and entrusting the system automation to do the appropriate action*. This definition was borrowed from Biros et al. (2004).

According to Gaudiello et al. (2016) there are six features that define robot acceptance: representational, physical, behavioral, functional, social, and cultural. However, functional and social acceptances play a key-role for effective human-robot interaction and the others are less important and will not be considered (Gaudiello et al. 2016). Functional acceptance is defined as the robot's perceived ease of use, usefulness, accuracy, and innovativeness (Gaudiello et al. 2016). Social acceptance is defined as the robot's social presence, perceived sociability, and social influence (Gaudiello et al. 2016). The representational and cultural acceptance continuously vary depending on the human's familiarity with robots in general. Humans are likely to accept a rough physical appearance (Turkle 2011) or accept inappropriate behavior like humans do with children (Young et al. 2009). However, humans are much less likely to accept uselessness or deception from the robot (Kaplan 2005, Shaw-Garlock 2009, Heerink et al. 2010).

According to Nass & Moon (2000), Nass et al. (1996, 1995) when humans engage in collaborative tasks with machines, humans tend to unconsciously accept computers as social entities. Additionally Young et al. (2009) proposed that there is a correlation between robot efficiency and human trust in the robot. According to Yagoda & Gillan (2012) trust is a valid indicator of robot function and social acceptance. Weiss et al. (2009), Heerink et al. (2010) found that a set of social acceptance indicators, such as performance expectancy, attitude towards technology, and anxiety. In our experiment we will attempt to estimate the subjects attitude towards the robot using their emotional state. In order to measure the emotional state, a way of quantifying emotions was needed.

2.3 Quantifying Emotions

Starting with Darwin (1872) and later expanded on by Ekman et al. (1969), human emotion, regardless of cultural influences, can be categorised into 6 basic emotions; fear, anger, disgust, sadness, joy, and surprise. The strength of Ekmans model was that it was able to distinguish the different emotions clearly. According to Posner et al. (2005) the downfall of the model is that it is not designed to fit any emotions that are not included in it's defined categories, and thus have difficulty modeling some components of emotion.

Russell (1980) proposed a circumplex model where the horizontal axis represents valence while the vertical axis represents arousal. The exact center of the model is a neutral state. With this approach it is possible to map a humans affective state based on the those two dimensions.

Mehrabian (1996) introduced a three state model called the Pleasure-Arousal-Dominance (PAD) model. This model is used for measuring emotional states (e.g., anger, depression, elation, fear, relaxation). According to Mehrabian (1996) the PAD model showed considerable independence between the scales and the combination of the scales showed high reliability. Furthermore, distinctions between certain clusters of affectional states (e.g., fear, anger, sadness) required more than two dimensions to be clearly separated. Here are 8 labels of traits Mehrabian (1996) found to describe the temperament space:

•	Exuberant	(+P+A+D)	vs.	Bored	(-P-A-D)
•	Dependent	(+P+A-D)	vs.	Disdainful	(-P-A+D)
•	Relaxed	(+P-A+D)	vs.	Anxious	(-P+A-D)
•	Docile	(+P-A-D)	vs.	Hostile	(-P+A+D)

Table 2.1: 8 labels of traits to describe the temperament space by Mehrabian (1996)

However, Mehrabian (1996) used questionnaires to determine the PAD values. In order to give an indication of the affective state of a human in real-time other measurement tools had to be used.

Kreibig (2010) mapped the response of various physiological signals and the corresponding emotion. According to Kreibig (2010) the anger, anxiety, surprise and fear have an increase in heartrate (HR) and electrodermal activity (EDA). Sadness can have both increases and reductions in HR and EDA reactions. Joy have no change in EDA response but an increase in HR. This means that a third dimension was needed to separate the positive and negative feelings from each other, as many of them would get the same HR and EDA responses. Pupil dilation has been proven as a reliable way of detecting emotional arousal (Henderson et al. 2018). Similarly Xu et al. (2011) found that pupil dilation could be a feasible way of detecting cognitive load which in most cases would imply some sort of emotional arousal. Kawai et al. (2013) and Babiker et al. (2013) found that negatively valenced stimuli trigger larger pupil dilation than positively valenced stimuli. According to Bradley et al. (2008) the response time of pupil dilation is around 0.2 seconds which is very fast for a physiological response. This means that the pupil response can be used early to indicate when the subject is having an emotional response and in conjunction with HR and EDA could give an indication which emotion it is.

2.3.1 Cardiac Response

Based on the Implicit-Affect-Primes-Affect (IAPE) model, proposed by Gendolla (2012, 2015), it is possible to measure a effort-related cardiac pre-ejection period (PEP) response induced by emotional stimuli, namely by subjecting the participants to flashes of images that were designed to provoke certain positive and negative emotions. Chatelain & Gendolla (2015) showed that it was possible to measure changes in the effort-related PEP response when subjected to said stimuli. Framorando & Gendolla (2018) further showed that there is differences in reaction to the stimuli between genders when the duration of exposure changes. Briefly (25ms) flashed images provoked a stronger reaction with sadness-primed images as opposed to anger, meanwhile the opposite was observed when the duration was longer (780ms). This pattern was exclusively observed in males. Additionally Freydefont et al. (2012), Goshvarpour et al. (2017) showed that it is possible to make similar estimation of sadness and anger using only cardiac activity. In this report we want to investigate if HR can be used to help detect when the robot cause an negative emotional response in humans. Goshvarpour et al. (2017) proposed a method using exclusively cardiac-activity measurements to classify emotions that are otherwise indistinguishable with arousal measurements. Their study aimed to classify the emotional state into the following classes; Happiness, Peacefulness, Sadness and fear. Using electrocardiography (ECG) to measure heart-rate variability (HRV) and finger pulse activity for pulse-rate variability (PRV) with a lagged pointcare plot, and further quantifying these measures using fitted ellipses as shape indices. From these plots the short term variability (SD_1), the overall variability (SD_2), a ratio of the two ($\frac{SD_1}{SD_2}$), and the area of the fitted ellipse ($Area$), can be calculated. SD_1, SD_2, $\frac{SD_1}{SD_2}$ and $Area$ from both HRV & PRV plots are then fed to a support vector machine (SVM) for classification. In the paper, Goshvarpour et al. (2017) proposes two ways of feeding the variables to the SVM, either by combining the features before feeding them to the SVM or by classifying them with separate SVMs and then passing those classifications into a combining classifier and using the result of that classifier in conjunction with the initial seperate classifications to recognize the emotion.

2.3.2 Electrodermal Activity

As proven by Boucsein (2012), Critchley (2002) the intensity of arousal that is experienced has a proven correlation with the activity in the eccrine sweat glands. As the intensity of arousal increases so does the activity in the sweat glands. While it is possible to infer the level of arousal it is not a representation of the specific emotion as both positive and negative stimuli cause an increase in arousal. The activity of the eccrine sweat glands have an important role in human thermo-regulation and sensory discrimination. They also showed a strong response to emotional stimuli as the skin become increasingly conductive, and those changes in electrical activity can be measured. EDA is the resistance of the skin with different levels of activity in the sweat glands. The most common solution uses a fixed voltage system, also known as an exosomatic method, where a constant voltage (commonly 0.5V) is applied to the skin between two electrodes and is then calculated, using Ohm's law, as such:

$$C = \frac{V}{R} \tag{2.1}$$

where V is the fixed voltage, R is the resistance, and C is skin conductance which is gained by measuring the current flowing between the electrodes. Skin conductance have to ability to react fairly quickly. According to LabScribe2 (2013) the response time of EDA is approximately 1.35 seconds from the beginning of stimuli to the reaction in EDA which is less ideal for real-time affect estimation. LabScribe2 (2013) used questions as stimuli. We expect the EDA response to visual stimuli is quicker as there is less comprehension to be done by the participants.

2.3.3 Non-Verbal Communication

It has been showed that non-verbal communication such as eye contact, facial-expression, posture, body movements and inter-personal distance are ways where humans communicate emotion both conciously and subconsiously. They provide valuable information about the emotions or attitude that a person is feeling in a given situation. Non-verbal communication play an important role in interpersonal interaction. For HRI this information can be very valuable to adjust the robots behavior. In the field of social robotics, a lot of studies have been done trying to identify and classify body-language so robots can respond appropriately or in case of human-like robots learn how to mimic human emotions through body language (McColl & Nejat 2012, Beck et al. 2010). The assumption is that people will use the same innate body language when communicating with a robot as they would when communicating with another human. The experiments conducted by Bull (2016) thoroughly investigate the importance of posture in interpersonal communication, as opposed to bodily movement. These studies are aimed at investigating postures that are held for a duration longer than 1 second. The investigations confirmed the observations done by Mehrabian & Friar (1969) that people tended to lean forward when expressing a positive emotion and backwards when the expressing a negative emotion. Additionally Mehrabian & Friar (1969) shows that people felt more comfortable with distances greater than the mean distance when having a negative attitude towards what was being communicated, and distances smaller than the mean distance when they had a positive attitude towards what was being communicated. Additionally the orientation of people during conversation was less conclusive as presented in Bull (2016), Mehrabian (1967), the orientation of people varies greatly dependant on the sender and receiver of the interpersonal interaction.

2.4 Multimodal Approaches

In recent years, EDA has been a common physiological measurement used to infer cognitive load (Shi et al. 2007, Nourbakhsh et al. 2012, Haapalainen et al. 2010) and stress (Panigrahy et al. 2017, Villarejo et al. 2012), in addition many recent studies show great promise when using EDA in conjunction with other modalities as shown by Goshvarpour et al. (2017), Haapalainen et al. (2010), Gombos (2006), Xu et al. (2011), Healey & Picard (2000).

Rani et al. (2004) proposed a novel affect-based framework to detect the level of anxiety that the user is feeling via biofeedback sensors. The sensors measured HR, EDA, and facial electromyography (EMG) (of the jaw and eyebrow), treating these signals as a multivariate problem, they were able to create an anxiety index which would detail the level of anxiety that the user was experiencing. Reliable correlations were found between the self-reported anxiety level and the physiological activity showing that is possible to utilize a multiple modalities to reliably infer a human-user′ affective state. Using such a framework the robot can then make the appropriate modifications to its behavior.

Attempting to do emotion recognition through facial motion has been a common approach, either by classifying the facial expression into Ekmans 6 basic emotions(Ekman et al. (1969)) or by detecting the action units (AU) of the Facial Action Coding System (FACS). However, Stienen & de Gelder (2011) propose a method that would address a problem that is often neglected in affect recognition that rely on a computer-vision approach, namely by measuring in real-time. Given that it was designed as a human-computer interaction (HCI) focused solution where the user's gaze is normally focused on the monitor, resulting in the subject facing the machine for the majority of the process. A very common problem when tracking finer details of the face using computer vision is that it relies on tracking the position of the face between frames. Not only does this introduce a lot of noise for the signal processing, it becomes hard to track points that are not visible to the camera due to the subject turning away or getting occluded by the robots appendages. With a collaborative HRI task, the user cannot be expected to be looking at the robot for prolonged amount of time, as they should be looking at their part of the collaborative task as opposed to the actions of the robot.

Chapter 3

Proposed Solution

In our experiment we try to estimate an emotional state using HR, EDA, pupillary responses, and posture. We are most interested in the negative emotions such as: anger, disgust, fear, sadness. Anger, disgust, and fear all have a high arousal rating which can be measured with EDA and HR. However, sadness is hard to detect as it has a high variance in the measured response. Sadness is not that relevant for our test. If the robot is making the participant sad it is unlikely to be connected to the robots actions and therefore beyond our scope. We compare the physiological data with a self reported trust score in order to see whether or not there is a correlation between the participants emotional state and their trust in the robot. How the trust score is collected is described in Chapter 5. The following hypotheses will be tested:

Hyp 1 *There is correlation between the trust score and the physiological measurements.*
It could be interesting to see if there is a way to monitor when humans are uncomfortable with the actions of a robot before the human move away from the robot. This project explores the possibility of using physiological measurements to see if they can be correlated with situations where the human feels uncomfortable.

Hyp 2 *Humans are more afraid the faster the robot moves.*
To have an effective relationship between the human and machine, the human needs to trust that the machine will not hurt them. To ensure this trust the movements of the robot have to communicate that the human is safe. We investigate if speed has an effect of when the human no longer feels safe in the work-space shared between man and machine.

Hyp 3 *Humans will get more comfortable with the robot over the course of the exercises*
As mentioned trust is very important to the relationship between robot and human. But it would be interesting to see how fast trust can be build between the two. Therefor we investigate if there are any change in the trust during our experiments.

Hyp 4 *Humans will have less trust to the robot when the robot is operating behind them.*
The most frequent occurrence of robots are in production lines and/or industrial production. In some environments the robots work very closely with the humans and in these scenarios it would be best if the robot's work space could be positioned anywhere while the human still feels safe. Therefor we investigate if the position of the robot's work space would have an effect on humans' trust.

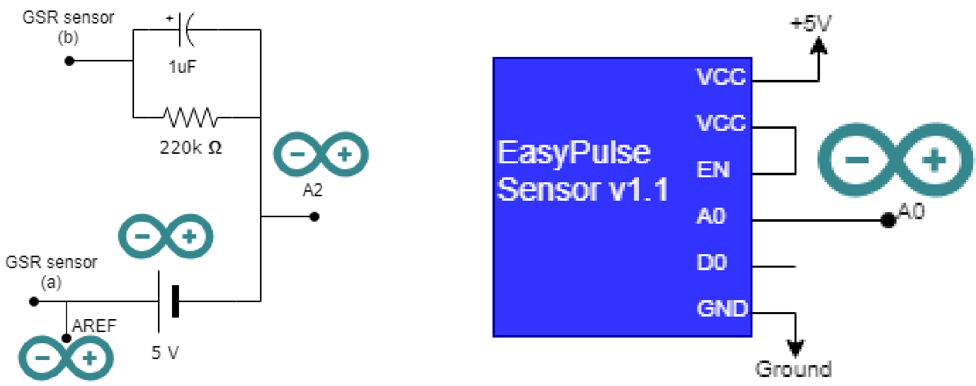
Chapter 4

Implementation of Solution

To measure the physiological data, sensors were required. How these sensors were created and how the signals were processed are detailed in this chapter.

4.1 Physiological Measurements

We measured physiological data in order to infer a emotional state. We measured HR, EDA, and pupil dilation. We used an Arduino Uno to send data from the EDA and HR sensors to a computer with a sampling rate of 20 Hz (see Figure 4.1).



(a) Diagram of the EDA measuring circuit. **(b)** Diagram of the HR measuring circuit.

Figure 4.1: Diagram of the electrical circuits measuring HR and EDA. The Arduino symbols on the diagrams signify that it is Arduino connections.

We used an Easy Pulse v1.1[1] to measure HR. Easy Pulse is a plug and play photoplethysmography (PPG) tool. Because PPG measures the volume of the blood it is possible to measure HR from the signal, as the volume increases with each ventricular contraction. EDA was measured using custom made copper-wired sensors. We used a hardware low pass filter to filter some of the noise from the EDA signal (see Figure 4.1a). A picture of the implemented circuits can be seen in Figure 4.2.

According to Mehrabian (1996) some of the feelings we are looking for have a high positive arousal change (panic, anxious, hostile). Therefore we need to look for increases in the EDA signal in order to try to track these emotions.

[1]http://embedded-lab.com/blog/easy-pulse-version-1-1-sensor-overview-part-1/

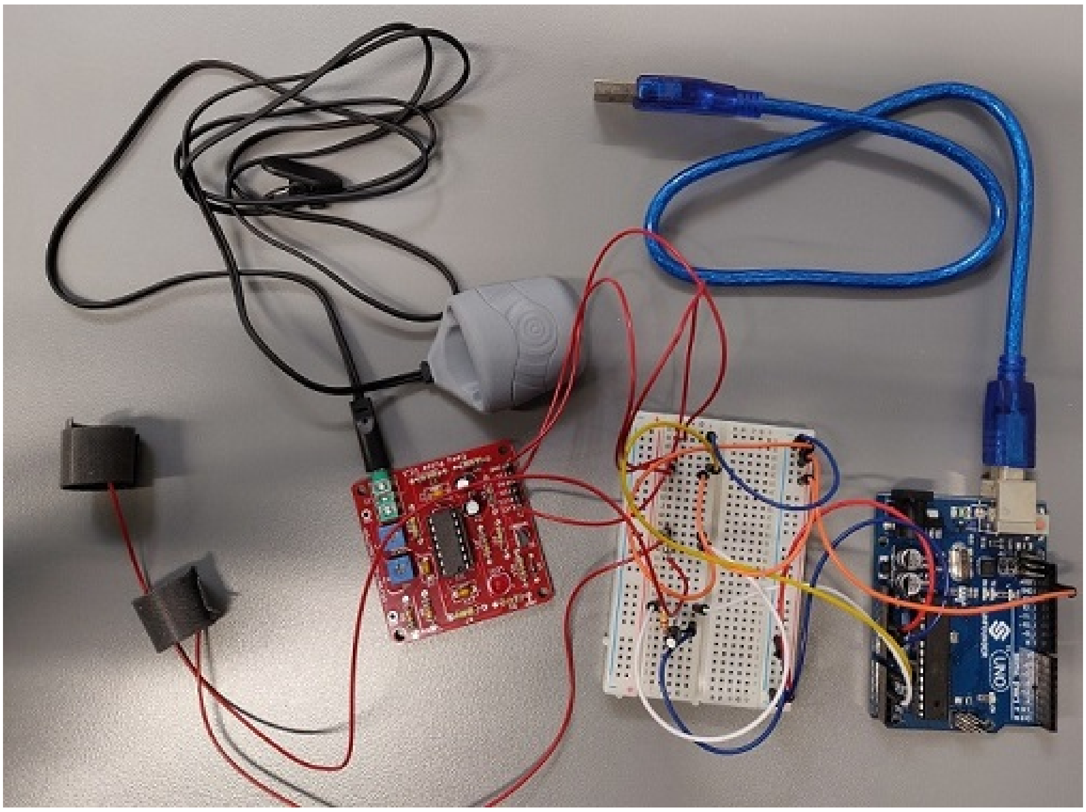


Figure 4.2: The HR and EDA sensors and their connections.

4.1.1 Signal Processing

Cardiac Response

Each heartbeat contains 2 big peaks that should be easily detected using PPG. They are known as the greater systolic-peak and smaller diastolic-peak respectively. This allows for easy calculation of inter-beat interval (IBI), HR, and systolic-diastolic peak-to-peak time (PPT). However, due to movement induced signal noise and artefacts, the signal is smoothed using a moving average filter, to the point where the diastolic peak is undetectable which is fine as we do not need the diastolic peak for HR peak detection. However, this make the detection of the systolic peak much more robust and significantly reduced the amount of false positive (FP) (see Figure 4.3).

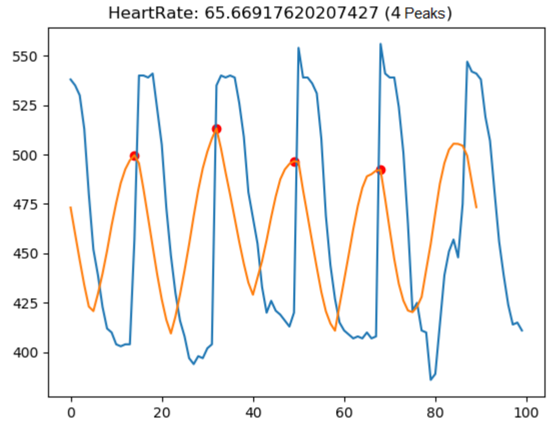


Figure 4.3: Before and after smoothing of the cardiac response signal

The PPG measurement produces a signal that is very susceptible to movement artefacts and thus prone to a lot of noise. Therefor the signal is smoothed using a moving average filter with a kernel of 10 samples (see Figure 4.3). To robustly measure the HR, a method proposed by Scholkmann et al. (2012) was used. The algorithm provided a flexible automatic peak-detection method for noisy periodic and quasi-periodic signals.

The signal is processed with a window that consists of the 100 latest samples, which results in the HR estimation being done over a 5 second period. Within this window the average time per peak is calculated as $HR = \frac{t_{latest} - t_{first}}{n_{peaks} - 1}$ where t_{first} is the time of the first detected peak in the window, t_{last} is the time of the latest detected peak within the window and n_{peaks} is the total number of peaks detected within the window. During testing the sensors were always turned on at least 5 seconds before the start of the test. The chosen window size of 100 samples could be decreased to reduce the delay, however that would decrease the amount of peaks within the window resulting in a less stable HR estimation. During pilot tests, the window size were tested starting at 60 samples, and increasing with 10 sample increments, and 100 samples was chosen as the best balance between delay and stability.

Pupillary response

To detect the puppilary response of the participants the position of the eyes for each participant needed to be reliably located and tracked. This is done using facial landmark functionality of the Dlib[2] library (see Figure 4.4). The Dlib library uses and implementation of the method described in Kazemi & Sullivan (2014), which utilizes an ensemble of regression trees to achieve very fast facial landmarks recognition.

[2]https://www.dlib.net

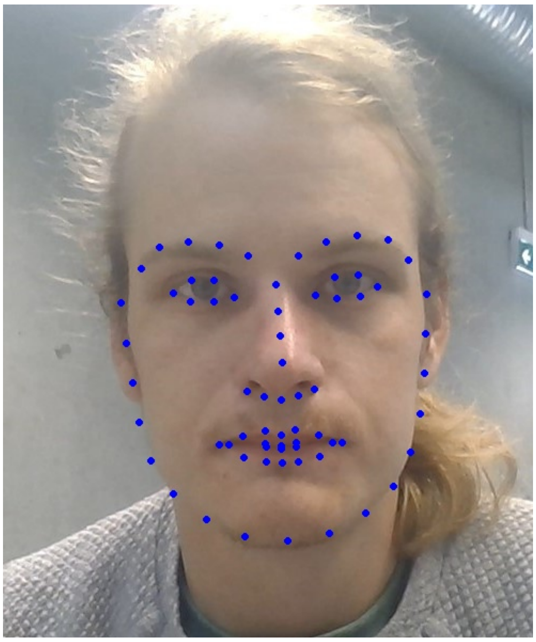
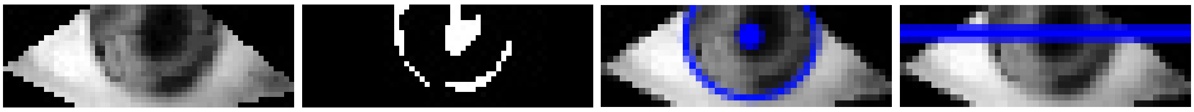
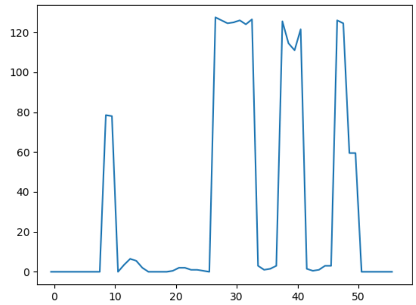


Figure 4.4: Example showing the 68 facial landmarks using the Dlib library

It is then possible to isolate and exclude the area around the eye and then the eyes are cut into two new images. We are looking for circular objects with very high contrast to their surroundings, in terms of color as perceived by an RGB camera. To increase the contrast a histogram equalization is performed on the image of each eye, resulting in a greater distance in terms of values from the "black" pupil, the colored iris, and the white sclera. Once the eye is properly segmented edge detection is used to identify the areas with the biggest changes in color, which naturally occurs at the transition between the sclera, the iris and the pupil. The resulting mask shown in Figure 4.5b can be used to perform a Circular Hough Transform (see Figure 4.5c). From this transform the circle with the highest accumulated score is chosen as the correct prediction of either iris or pupil. The the position of the circle center is then used to extract a single row of the image which crosses through the center of the pupil. The gradient of the signal is then used to determine where the biggest contrast between neighbouring values is located, which results in a signal with 4 big peaks representing the position of each transition between the sclera, iris, and pupil as seen in Figure 4.5e. From those the positions (P_1, P_2, P_3, P_4) the width of the iris can be calculated as $w_{iris} = P_4 - P_1$ and the width of the pupil can be calculated as $w_{pupil} = P_3 - P_2$ resulting in a pupil ratio of $PupilRatio = \frac{w_{pupil}}{w_{iris}}$.



(a) **(b)** **(c)** **(d)**



(e)

Figure 4.5: This figure shows an illustration of the pupil-ratio calculation method, where (a) is the segmented eye, (b) is the edge-detection mask, (c) is the Circular Hough Transform, (d) shows the line at which the pupil gradient signal is calculated and (e) is the resulting graph.

Any signal that has an erroneous pupil ratio due to the missing peak detection or impossible ratio such as when the algorithm detects a pupil larger than the iris, the algorithm will return a 0 pupil ratio. This means there will be holes in the signal and thus the general trend-line, excluding zeroes should be considered the correct pupil-ratio.

4.1.2 Robot Programming

We used the Sawyer robot[3] for our experiments. This robot can be programmed with RoS[4] or the embedded Intera system. We used the Intera system as it had a lot of functionality already implemented, such as movement, landmark tracking, and vision tasks. This made it very fast and easy to implement our tasks. Our tasks were made from landmark tracking, movement, and pressure measurements. The robot can be seen in Figure 4.6.

[3]https://www.rethinkrobotics.com/sawyer/
[4]https://www.ros.org/

Figure 4.6: The Sawyer robot.

Chapter 5

Evaluation

This chapter documents the experiments and the methods used for our evaluation.

5.1 Setup of Evaluation

The setup of the evaluation is as shown in Figures 5.1 and 5.2. In the water task the participant would be seated facing the robot at a distance of roughly 2 meters from the base frame of the robot. In the LEGO task had the participant seated with their right side facing the robot, at a distance of roughly 50cm.

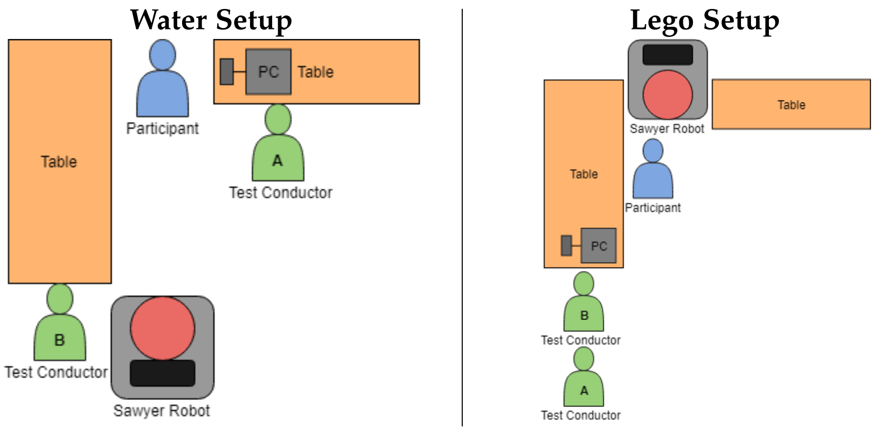


Figure 5.1: Illustration of the evaluation setup.

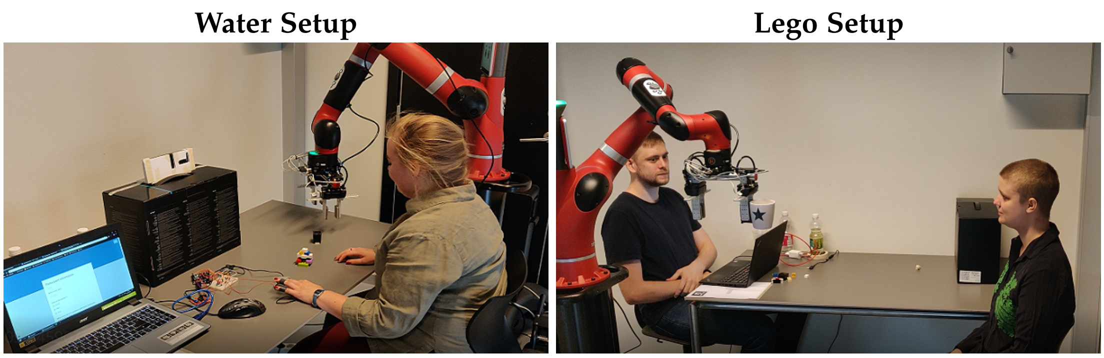


Figure 5.2: Picture of the real-life evaluation setup.

5.2 Evaluation Design

For evaluation we used a set of tasks and questionnaires. The tasks was solved one at a time followed by a questionnaire to determine trust score. The tasks was designed to elicit an emotional response by the participant, mainly stress / anxiety / general uncomfortableness. There was 4 tasks in total.

1. Received glass of water from robot.
2. Gave glass of water to robot
3. The participant build LEGO and place the finished LEGO into the claw of the robot. The robot took the LEGO and put the LEGO in another location close to participant. Robot was positioned to the side of the participant operating within their field of view.
4. This task was similar to the previous task but the robot would put the finished LEGO behind participant. The participant was told not to turn around.

We used a latin square in order to balance the test and the task was executed at 70% and 100% speed. During the different tasks we recorded HR, EDA, and pupil dilation in order to establish the emotional state of the participant. We also recorded video of the participant completing the tasks. Before each evaluation baseline levels of the participant's HR and EDA was established. According to Mehrabian (1996) participants return to baseline emotional baselines at different speeds. Therefore, the participant was asked to complete a survey in between each task. This served 2 purposes: we needed a trust evaluation of each tasks and as a break to get them as close as possible to their emotional baseline before the next task.

5.3 Subjective Assessment

We measured the participant's subjective assessment of the robot and the tasks they performed using a 14 question questionnaire. The questionnaire were developed by Schaefer (2013) to evaluate the participants trust level after each task (see Table 5.1). All the questions are marked in a Likert scale from 1-7, where 1 is "Strongly Disagree" and 7 is "Strongly Agree". The questionnaire are not randomized.

Function successfully	Act consistently
Reliable	Predictable
Dependable	Follow directions
Meet the needs of the mission	Perform exactly as instructed
Have errors (Reverse Coded)	Provide appropriate information
Unresponsive (Reverse Coded)	Malfunctioning (Reverse Coded)
Communicate with people	Provide feedback

Table 5.1: List of trust evaluation parameters developed by Schaefer (2013).

Reverse coded means that the result needed to be inverted for evaluation purposes (e.g. Strongly disagree -> strongly agree).

Posture

Additionally to pupillary response the videos gathered of each participant during the tests will also be used to estimate emotion based on upper-body posture. Based on the scoring scheme developed by Bull (2016), a new scoring scheme was developed to fit an HRI task. The scheme described by Bull (2016) was developed from a set of varying interpersonal talks, where the subject would be be face-to-face with another human. However, the investigations presented by Bull (2016) has stationary human subjects, whereas this experiment presents an active collaborative task between a human and a robot. The assumption of this scoring scheme is that the participant will convey the same inter-personal information with their body language when interacting with the robot as they would when interacting with another human. However, during the task the claw of the robot, which they are interacting with, performs a series of movements. Therefor some of the measurements have to be adapted to the new context while others have to be entirely excluded.

The list of considered actions are put into 3 groups; Leaning, Orientation and Position can be seen in Table 5.2:

Leaning	**Orientation**	**Position**
Lean towards robot	Turn towards robot	Lower torso
Reduces lean towards robot	Reduces turn towards robot	Raise torso
Lean away from robot	Turn away from robot	Move chair towards robot
Reduces lean away from robot	Reduces turn away from robot	Move chair away from robot

Table 5.2: List of the considered actions for each category

The description for each observation is further explained in Appendix 6.7. The criteria for a pose to be considered is that it lasts for longer than 0.5 seconds and is a still pose rather than part of a fluid motion, this is done to exclude FP and false negative (FN) due to unexpected movements.

Chapter 6

Results and Analysis

This chapter contain the results and an analysis of each of the hypotheses.

6.1 Trust

First we are looking at the hypothesis "*Humans will get more comfortable with the robot over the course of the exercises*". The trust scores during the water exercises went up, while the trust score during the LEGO exercise remained similar to each other (see Tables 6.1, 6.2).

	Test 1	**Test 2**	**Test 3**	**Test 4**
Total	**102,357**	**105,357**	**110,571**	**111,214**
Mean	**5,118**	**5,268**	**5,529**	**5,561**
Variance	**0,391**	**0,469**	**0,167**	**0,643**

Table 6.1: Water trust scores change during the tests.

	Test 1	**Test 2**	**Test 3**	**Test 4**
Total	**106,071**	**95,071**	**100,143**	**98,857**
Mean	**5,3036**	**4,7536**	**5,0071**	**4,9429**
Variance	**0,504**	**0,474**	**0,963**	**0,988**

Table 6.2: LEGO trust scores change during the tests.

Even with a relative small number of exercises the water experiment has an increase in trust. Why there is a difference in water and not in LEGO could be due to the fact that the participants had a more difficult time handing the LEGO to the robot. In the water experiment the robot was holding the claw in a sideways manor and the claw was pushed backwards while in the LEGO experiment it was holding the claw in the same way but the participant had to push the claw sideways. The robot would close the claw when a certain pressure was applied. However, many participants started small and building up the pressure instead of giving a swift motion. This would require a lot more power from the participant than doing one swift motion. This could impact the way the trust scores are evaluated as e.g. "functions correctly" and "act consistently" could be impacted by the way the user interacted with the robot. The change in score would only be affected if the participants handed the robot the LEGO using the slow-increase-in-pressure technique at different tests.

Another theory of why the LEGO building experiment had a lower test score could be because the task simply was more unpleasant than the water exercise. The robot moved in close proximity to the participant, it moved behind them, they were somewhat boxed in with the robot on one side and the facilitator on the other. Many reasons could help explain this. However, as previously mentioned, this is a small number of exercises and a pattern could be found if the experiment had more exercises. We could also have discovered this if we had used the same participants for both experiments. That would mean that we could directly compare the two experiments trust scores and having eliminated all the external factors (such as interaction method, or difference of opinion in how to score the robot).

To evaluate our hypotheses "*Humans are more afraid the faster the robot moves*" and "*Humans will have less trust to the robot when the robot is operating behind them*" an ANOVA test was performed to identify the impact of the different tasks, and the different speeds of the task. First we wanted to see if there was a significant difference between our tasks. Therefor we compare the fast and slow tasks with one another (see Table 6.3, 6.4).

RESUME						
Groups	Count	Sum	Average	Variance		
WaterFast	40	201,0714	5,0268	0,4178		
LegoFast	40	183,5	4,5875	0,6003		
ANOVA						
Source of Variation	SS	dof	MS	F	P-value	F crit
Between Groups	3,8594	1	3,8594	7,5817	0,0073	3,9635
Within Groups	39,7059	78	0,5090			
Total	43,5653	79				

Table 6.3: Comparison of the fast version of the two categories of tasks.

RESUME						
Groups	Count	Sum	Average	Variance		
WaterSlow	40	214	5,35	0,390		
LegoSlow	40	179,429	4,486	0,680		
ANOVA						
Source of Variation	SS	dof	MS	F	P-value	F crit
Between Groups	14,940	1	14,940	27,935	1,11E-06	3,963
Within Groups	41,714	78	0,535			
Total	56,654	79				

Table 6.4: Comparison of the slow version of the two categories of tasks.

The ANOVA show that the trust scores for the two tasks are distinctly different both in the fast and slow versions. There are a large drop in trust scores for the LEGO building exercise compared to the water exercise. This is also what we initially expected with the hypothesis "*Humans will have less trust to the robot when the robot is operating behind them*". This hypothesis will be further discussed later. The observed drop in the trust scores from the water scores to the LEGO building scores could be because the participants in general felt the LEGO building exercise was less safe than the water exercise.

Next we investigate the hypotheses "*Humans are more afraid the faster the robot moves*" and "*Humans will have less trust to the robot when the robot is operating behind them*". To test these hypotheses the results from an ANOVA where we compare the tasks's fast and slow versions are analyzed below (see Table 6.5, 6.6).

	RESUME	Water Give	Water Take	Total		
Fast	Count	20	20	40		
	Sum	97,286	103,786	201,071		
	Average	4,864	5,189	5,027		
	Variance	0,465	0,337	0,418		
Slow	Count	20	20	40		
	Sum	108,143	105,857	214,000		
	Average	5,407	5,293	5,350		
	Variance	0,248	0,545	0,390		
Total	Count	40	40			
	Sum	205,429	209,643			
	Average	5,136	5,241			
	Variance	0,423	0,432			
	ANOVA					
	Source of Variation	SS	dof	MS	F	P-value	F crit
	Speed Difference	2,089	1	2,089	5,241	0,024	3,966
	Task Difference	0,222	1	0,222	0,556	0,457	3,966
	Interaction	0,964	1	0,964	2,4203	0,123	3,966
	Within	30,297	76	0,398			
	Total	33,573	79				

Table 6.5: Comparison of the water tasks to see if there is a difference in the speed of the robot's movement.

	RESUME	LEGO In Front	LEGO Behind	Total			
Fast	Count	20	20	40			
	Sum	91,643	91,857	183,5			
	Average	4,582	4,593	4,588			
	Variance	0,646	0,587	0,600			
Slow	Count	20	20	40			
	Sum	88,571	90,857	179,429			
	Average	4,429	4,543	4,486			
	Variance	0,706	0,683	0,680			
Total	Count	40	40				
	Sum	180,214	182,714				
	Average	4,505	4,568				
	Variance	0,665	0,619				
	ANOVA						
	Source of Variation	SS	dof	MS	F	P-value	F crit
	Speed Difference	0,207	1	0,207	0,316	0,575	3,966
	Task Difference	0,078	1	0,0781	0,119	0,730	3,966
	Interaction	0,053	1	0,053	0,081	0,775	3,966
	Within	49,803	76	0,655			
	Total	50,143	79				

Table 6.6: Comparison of the LEGO building task to see if there is a difference when the robot places objects in front and behind the participant.

First we take a look at the hypothesis "*Humans are more afraid the faster the robot moves*". From the ANOVA we see that there was a significant difference in the water test in terms of speed. However, there are no difference in the LEGO experiment. Therefor we cannot reject our null hypothesis that the speed had no influence in terms of trust in the robot. However, there is a larger variance in the fast give-task so larger data samples could reveal a clearer pattern.

We test the hypothesis "*Humans will have less trust to the robot when the robot is operating behind them*" by comparing the trust scores for the participants during the LEGO building exercise. The null hypothesis was there are no significant difference between the tasks when completed in front and behind the participants. The ANOVA show that we cannot reject our null hypothesis. There was no significant differences in terms of speed or task for the LEGO building experiment. However as many of the participants disregarded the instructions of the experiment and followed the robot with their eyes and turned their bodies, which meant the robot was never truly behind them, we do not feel this experiment can be considered conclusive in this matter. Future redesign of the experiment could include two robots or a spot or point that the participant had to look at during the test. One could also introduce a second task which is solved while the robot is working. We will further investigate the posture and gestures of the participants during this task in future sections.

6.1.1 Physiological Response

The hypothesis *There is correlation between the trust score and the physiological measurements* was tested by comparing observations in the participants's physiological data, their behavior during the test, and their trust score. Figures 6.1a, 6.1b show what we classify as a EDA response to the robot. Figures 6.2a and 6.2b show what we classify as no EDA response to the robot.

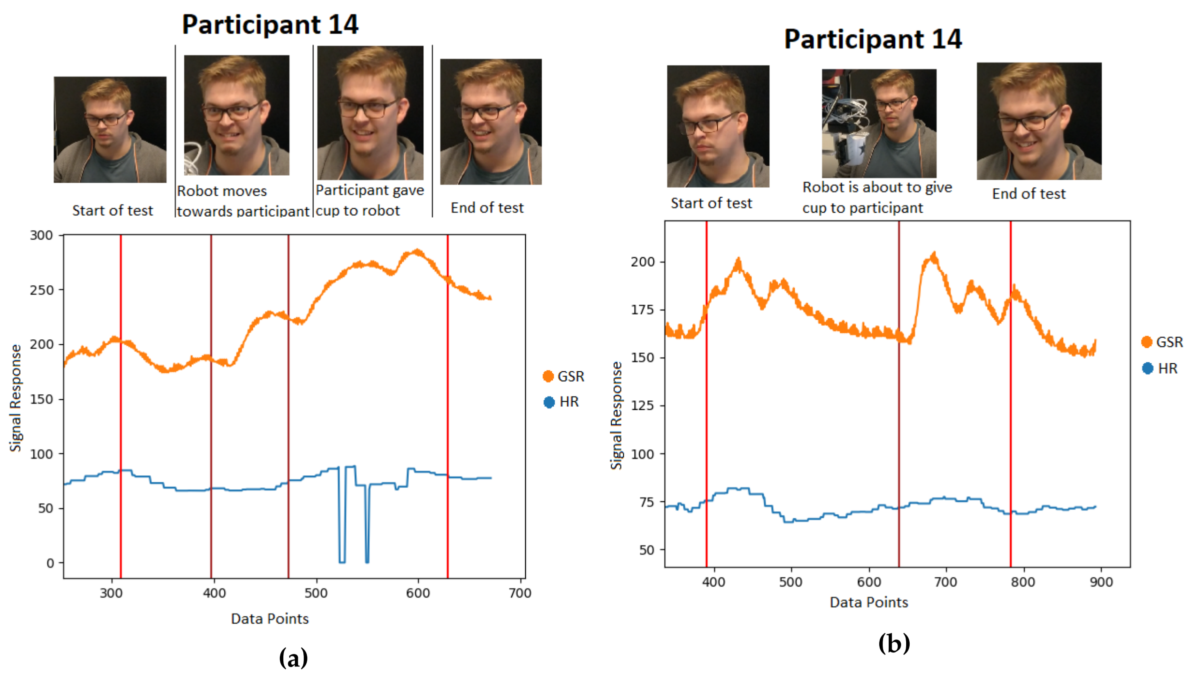
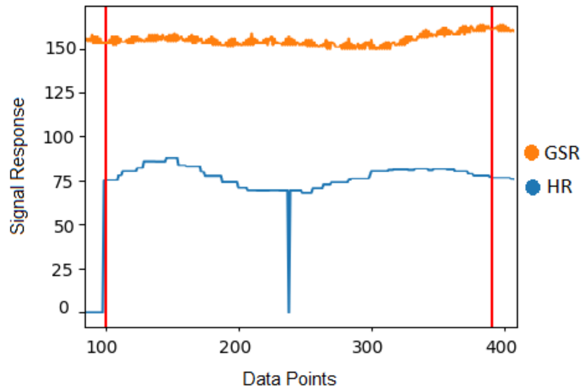
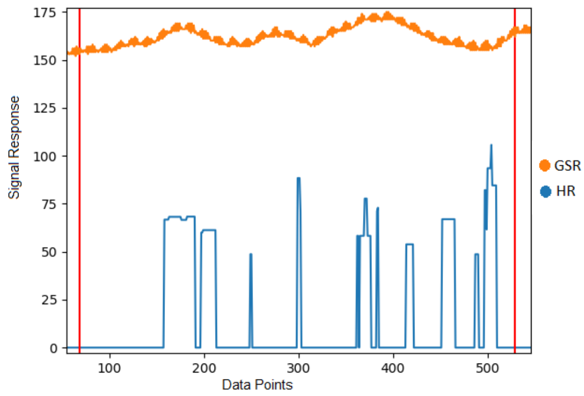


Figure 6.1: Signal that are classified as having a GSR response to the robot.



(a)



(b)

Figure 6.2: Examples of signals that were classified as no GSR response.

We see a EDA response to the robot's actions in 29 of the 39 participants at their first interactions with the robot. In 13 participants we observed a EDA response in the last test. The EDA signal changes approximately after a second after stimuli being applied. This is consistent with other research projects that analyze EDA response. 1 of the 40 participants did not have the EDA sensor attached properly which caused the EDA to be corrupted and the participant was excluded from the EDA results.

The HR measurement equipment would sometimes would stop working correctly and introduce artefacts into the signal making the signal useless during these periods. Examples of these are shown on Figures 6.3a, 6.3b, and 6.3c.

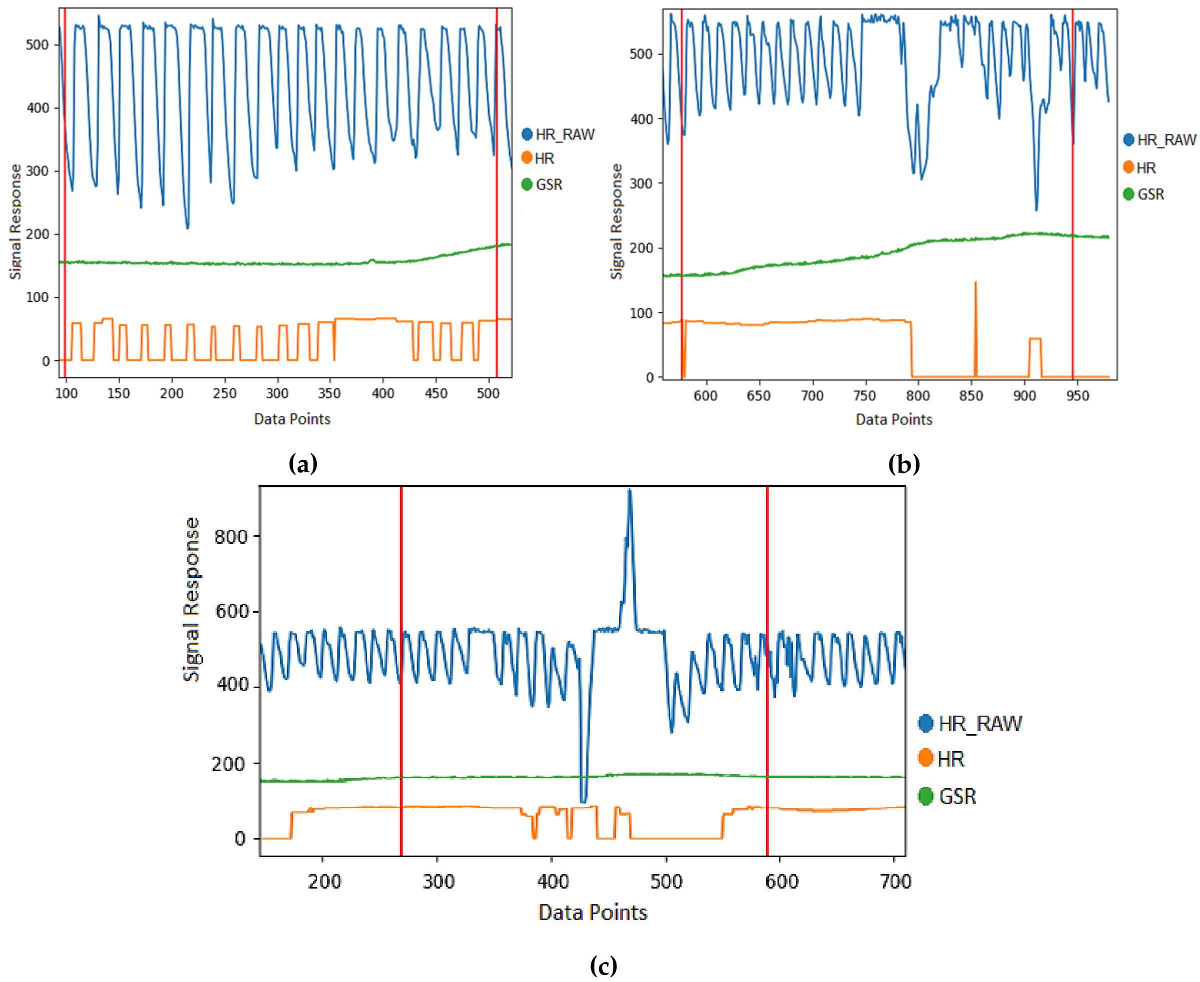


Figure 6.3: Examples of when the HR signal did not work.

This meant that about half of our HR data could not be used. From the usable data we observe that the HR most of the time would increase when the EDA would increase. As observed in some cases the cardiac response to stimuli was an increased HR and other times there was no significant change or a reduction in HR that was observed. Which leads us to believe that the emotions being are measured is a result of the participant experiencing different emotions. Further investigations would have to be conducted to establish what emotions in particular each set of responses has a correlation with. The participants with a increase in HR and EDA, had a 5,184 average trust score while the partcipants with no EDA response had an average trust score of 5,194. The same can be observed in participants which had a decrease in HR while having an increase EDA.These varied less than one standard deviation away from the mean, participants that had no EDA response but still had an increase in HR showed the same variation, within one standard deviation from the mean. However due to the low sample size of these observations, this cannot be proven to be the case with statistical significance.

6.2 Posture

In this section the results from the posture analysis are presented. The full description of each variable measured can be found in Appendix 6.7.

Torso posture scoring scheme

Action Category	Action description
LTR	Lean Towards Robot, torso is moved towards the robot.
RLTR	Reduces Lean Towards Robot, the torso approaches the vertical position.
LAR	Lean Away from Robot, torso is moved away from the robot.
RLAR	Reduces Lean Away from Robot, the torso approaches the vertical position.
TTR	Turn Towards Robot, torso is turned towards the robot.
RTTR	Reduces Turn Towards Robot, the torso approaches the vertical orientation from TTR.
TAR	Turn Away from Robot, torso is turned away from the robot.
RTAR	Reduces Turn Away from Robot, the torso approaches the vertical orientation from TAR.
LT	Lower Torso, the torso is lowered so the spine is collapsed from a straight sitting position.
RT	Raise Torso, the torso is raised so that the spine is straight from a collapsed position.
mCTR	move Chair Towards Robot, the chair is moved closer to the robot.
mCAR	move Chair Away from Robot, the chair is moved away from the robot.

Table 6.7: This table contains a list of actions that are considered for torso movement during the experiments, shown with an action identifier in the left column and an action description in the right column

Water Take

		LTR	RLTR	LAR	RLAR	TTR	RTTR	TAR	RTAR	LT	RT	mCTR	mCAR
Fast	Mean	0.058	0.1316	0.937	1.5	0.074	0.1316	0.289	0.08	0	0.15	0	0.084
	Variance	0.029	0.1537	2.56	5.8758	0.05	0.3116	1.245	0.06	0	0.19	0	0.061
	Frequency	0.11	0.105	0.47	0.474	0.11	0.053	0.11	0.1	0	0.1	0	0.11
Slow	Mean	0.289	0.5222	0.222	0.2944	0.072	0	0.322	0.07	0	0.09	0	0.028
	Variance	0.274	1.2151	0.284	0.5627	0.044	0	1.765	0.08	0	0.13	0	0.013
	Frequency	0.28	0.222	0.17	0.167	0.11	0	0.06	0.1	0	0.1	0	0.06

Table 6.8: This figure shows the results from the both the fast and slow version of the water task, where the robot had to receive the cup from the participant

As can be seen in Table 6.8 the participants spent much more time leaned towards the robot in the slow task as opposed to the fast task where more time was spent leaning away from the robot. However during both test great variance is observed as some participants were very active with their body language were others kept theirs very muted. With a few instances of people raising their backs from their normal seated position.

Water Give

		LTR	RLTR	LAR	RLAR	TTR	RTTR	TAR	RTAR	LT	RT	mCTR	mCAR
Fast	Mean	0.122	0.3444	0.794	0.5111	0	0.05	0.206	0.04	0	0.04	0	0
	Variance	0.124	0.8602	0.792	0.6599	0	0.0425	0.541	0.03	0	0.03	0	0
	Frequency	0.11	0.167	0.61	0.444	0	0.056	0.11	0.1	0	0.1	0	0
Slow	Mean	0.279	0.2263	0.163	0.0632	0	0	0.211	0.11	0	0	0.126	0.163
	Variance	0.693	0.2756	0.479	0.0718	0	0	0.798	0.22	0	0	0.287	0.479
	Frequency	0.11	0.158	0.05	0.053	0	0	0.05	0.1	0	0	0.05	0.05

Table 6.9: This figure shows the results from the both the fast and slow version of the water task, where the robot had to give the cup to the participant

As seen in Table 6.9 a similar reaction in terms of body movement to what is observed in the previous task, however at it also showed a tendency for participants to move the chair back and forth in the slow variant. This is an very unexpected observation and contradicts most of the other movement that the participants perform, as moving away from the robot would be considered a display of negative attitude it would normally be expected to occur with a much greater frequency and to a much larger degree in the faster version of the task.

Lego Front

		LTR	RLTR	LAR	RLAR	TTR	RTTR	TAR	RTAR	LT	RT	mCTR	mCAR
Fast	Mean	0.122	0.3444	0.794	0.5111	0	0.05	0.206	0.04	0	0.04	0	0
	Variance	0.124	0.8602	0.792	0.6599	0	0.0425	0.541	0.03	0	0.03	0	0
	Frequency	0.11	0.167	0.61	0.444	0	0.056	0.11	0.1	0	0.1	0	0
Slow	Mean	0.025	0.055	0	0	0	0	0	0	0	0	0	0.055
	Variance	0.012	0.0575	0	0	0	0	0	0	0	0	0	0.057
	Frequency	0.05	0.05	0	0	0	0	0	0	0	0	0	0.05

Table 6.10: This figure shows the results from the both the fast and slow version of lego building task where the robot would place the construction in front of the participant

As shown in Table 6.10, much more movement across the board can be observed in the fast task as opposed to the slow task, and again the pattern of creating distance between the robot and the participant is observed to a much greater extent in the fast variant as would be expected.

LegoBack

		LTR	RLTR	LAR	RLAR	TTR	RTTR	TAR	RTAR	LT	RT	mCTR	mCAR
Fast	Mean	0	0	0.96	0.375	0.405	0.155	0.05	0.29	0.26	0.14	0.025	0.165
	Variance	0	0	3.164	0.4599	0.88	0.0865	0.023	1.28	0.22	0.1	0.012	0.182
	Frequency	0	0	0.55	0.4	0.3	0.25	0.1	0.1	0.3	0.2	0.05	0.15
Slow	Mean	0	0	0.645	0.17	0.535	0.28	0	0.16	0.35	0.08	0	0.165
	Variance	0	0	1.972	0.1341	0.925	0.2546	0	0.46	0.48	0.07	0	0.275
	Frequency	0	0	0.35	0.2	0.4	0.35	0	0.1	0.3	0.1	0	0.1

Table 6.11: This figure shows the results from the both the fast and slow version of lego building task where the robot would place the construction behind the particiapnt

As can be seen in Figure 6.11 there seem to be no indication of a positive attitude in the form of leaning forward in neither the fast or slow variant of the task. However, there is a very small difference in the general behavior for this task, as certain aspects that would be expected to be more prominent in the faster variant of the task that turn out to be more prominent in the slow task, such as LT and TTR and inversely variables such as mCTR and TAR are more prominent in the faster where the opposite was expected. The results from the posture indicate the same tendency as the trust scores in terms of the perceived danger of the different tasks, and the observed frequency, be it positive or negative, support the observations of the trust scores.

6.2.1 Additional Observations

Trust Scores

- During the water tests, there were 3 instances of the robot failing to grasp the cup. Observing the scores of the trust questionnaire for those tests, the participants showed no deviation from the norm in terms of trust within the task. A lower score would have been expected in the evaluation of *Function Successfully*, *Meet the needs of the mission*, *Have Errors*, *Act consistently* or *Malfunctioning*. However no significant deviation from the mean was observed.

- During the LEGO tests, 2 participants experienced that the robot failed to grasp the construction. Like the water test, the trust scores that were expected to drop, showed no significant deviation from the mean.

Participant reactions

- 2 participants voiced their discomfort with the robot holding the cup at eye level.

- 5 participants (4 male, 1 female) mentioned that the robot was "creepy" upon realizing it had eyes on it's display, whereas 3 participants (1 male, 2 female) thought the robot looked "cute" and "friendly".

- 3 Participants complimented the robot when the LEGO was placed in such a manor that it did not fall over when the robot released its grasp.

- 3 participants displayed a total of 5 instances of quickly retracting their arm when the pneumatic gripper pressurized or depressurized.

- 4 participants commented on the robots slow movement while it was getting into position for the next task, and described the sound and movement as scary and unpredictable.

- 6 out of 21 instances of visual nervousness occurs when the participant is not looking at the robot. Only a total 7 participants looked away from the robot during the tests and thus indicating that the participants are less nervous and are therefor less likely to get scared when they are not observing the robots movements.

- Out of 40 instances of the robot operating outside of the participants field of view, 26 instances had the participant turn around and look behind to further observe the robots movements.

6.2.2 Pupillary response

All though the pupillary response looked promising in the ideal setting where the test participant would not move their head, and predominately be looking straight ahead, the results showed to be very different during the actual tests, where the participants were significantly more prone to looking in a direction to such a degree that the view of the eye becomes to skewed for stable pupil-ratio prediction. Additionally the lighting varied slightly, which didn't seem to be a problem with people who had iris' of a lighter hue, making the iris-pupil color contrast naturally significant, but it dramatically impacted the accuracy for people with darker iris'.

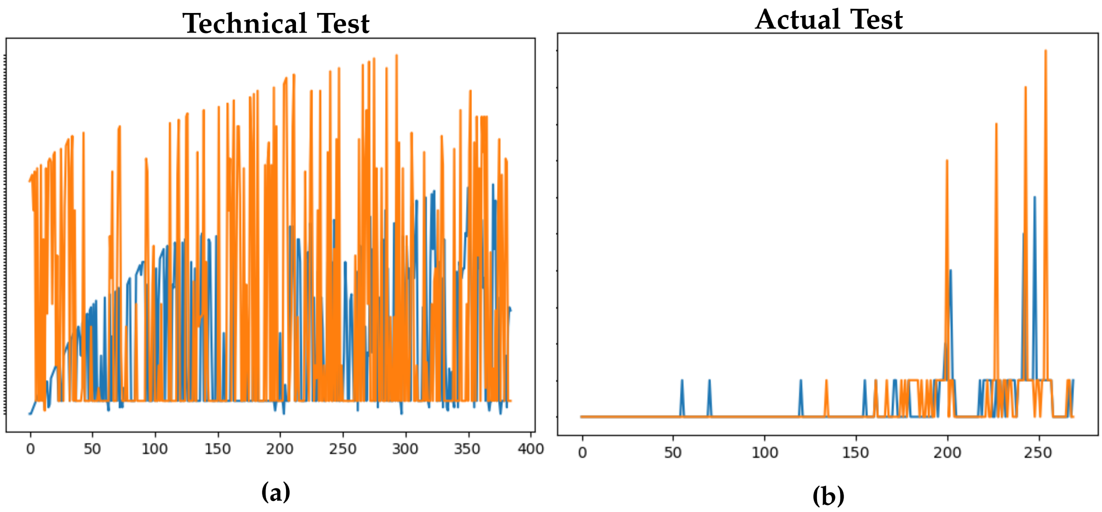


Figure 6.4: Graphs of the pupil-ratio algorithm working in ideal conditions (left) and test conditions (right)

As can be seen in Figure 6.4 the signals produced during the test were too unstable due to a combination of lighting changes, gaze direction and head orientation/movement. And therefore the results are unusable and render this aspect of the study inconclusive. Due to the unreliable results due to lighting and participant movements, changes to this version of the implementation would have to be made for it to be reliable. The algorithm works in ideal conditions and thus it would need some adjustments to facilitate test conditions closer to those that are ideal. For this a near-infrared camera can be used as it is less affected by lighting, and additionally the camera could be mounted on the participants head, so that the view of the eye is always static. This would still face problems however with gaze changes and blinking but could potentially facilitate conditions that would provide a reliable signal.

Chapter 7

Conclusion

There was no significant correlation between the trust scores and the physiological data. However, based on the physical responses the participants had during the test it seems likely a robot could evaluate the emotional state of a human during a collaboration task. It also seems likely that the emotional state could be estimated using physiological data as there were physiological responses after stimuli was applied and the response is clearly visible in the data.

Humans were more afraid the faster the robot moved in situations where the task had the robot arm moving directly towards or away from the participant to a statistically significant degree. The results regarding movement around the participant however show no statistical significant difference in terms of speed and can therefor not be verified. Humans did not have significantly less trust to the robot when the robot is operating behind them. However, the posture and gestures made by the majority of the participants during the tests indicate that they were less comfortable when the robot moved behind them. Our participants did get more comfortable with the robot over the course of the experiments. As expected the results confirm that the participants showed increase in trust scores as they gained more experience with the robot.

Bibliography

Babiker, A., Faye, I. & Malik, A. (2013), Pupillary behavior in positive and negative emotions, *in* '2013 IEEE International Conference on Signal and Image Processing Applications', pp. 379–383.

Beck, A., Cañamero, L. & Bard, K. A. (2010), Towards an affect space for robots to display emotional body language, *in* 'Towards an Affect Space for robots to display emotional body language', pp. 464–469.

Biros, D. P., Daly, M. & Gunsch, G. (2004), 'The influence of task load and automation trust on deception detection', *Group Decision and Negotiation* **13**(2), 173–189.
URL: *https://doi.org/10.1023/B:GRUP.0000021840.85686.57*

Boucsein, W. (2012), *Electrodermal activity*, Springer Science & Business Media.

Bradley, M. M., Miccoli, L., Escrig, M. A. & Lang, P. J. (2008), 'The pupil as a measure of emotional arousal and autonomic activation', *Psychophysiology* **45**(4), 602–607.

Bull, P. E. (2016), *Posture & gesture*, Vol. 16, Elsevier.

Bütepage, J. & Kragic, D. (2017), 'Human-robot collaboration: From psychology to social robotics', *arXiv preprint arXiv:1705.10146* .

Chatelain, M. & Gendolla, G. H. (2015), 'Implicit fear and effort-related cardiac response', *Biological psychology* **111**, 73–82.

Critchley, H. D. (2002), 'Electrodermal responses: what happens in the brain', *The Neuroscientist* **8**(2), 132–142.

Darwin, C. (1872), 'The expression of emotions in man and animals. new york: Philosophical library', *Original work published* .

Ekman, P., Sorenson, E. R. & Friesen, W. V. (1969), 'Pan-cultural elements in facial displays of emotion', *Science* **164**(3875), 86–88.

Framorando, D. & Gendolla, G. H. (2018), 'The effect of negative implicit affect, prime visibility, and gender on effort-related cardiac response', *Adaptive Human Behavior and Physiology* **4**(4), 354–363.

Freydefont, L., Gendolla, G. H. & Silvestrini, N. (2012), 'Beyond valence: The differential effect of masked anger and sadness stimuli on effort-related cardiac response', *Psychophysiology* **49**(5), 665–671.

Gaudiello, I., Zibetti, E., Lefort, S., Chetouani, M. & Ivaldi, S. (2016), 'Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers', *Computers in Human Behavior* **61**, 633–655.

Gendolla, G. H. (2012), 'Implicit affect primes effort: A theory and research on cardiovascular response', *International Journal of Psychophysiology* **86**(2), 123–135.

Gendolla, G. H. (2015), 'Implicit affect primes effort: Basic processes, moderators, and boundary conditions', *Social and Personality Psychology Compass* **9**(11), 606–619.

Gombos, V. A. (2006), 'The cognition of deception: The role of executive processes in producing lies', *Genetic, social, and general psychology monographs* **132**(3), 197–214.

Goshvarpour, A., Abbasi, A. & Goshvarpour, A. (2017), 'Fusion of heart rate variability and pulse rate variability for emotion recognition using lagged poincare plots', *Australasian physical & engineering sciences in medicine* **40**(3), 617–629.

Haapalainen, E., Kim, S., Forlizzi, J. F. & Dey, A. K. (2010), Psycho-physiological measures for assessing cognitive load, *in* 'Proceedings of the 12th ACM international conference on Ubiquitous computing', ACM, pp. 301–310.

Healey, J. & Picard, R. (2000), Smartcar: detecting driver stress, *in* 'Proceedings 15th International Conference on Pattern Recognition. ICPR-2000', Vol. 4, IEEE, pp. 218–221.

Heerink, M., Kröse, B., Evers, V. & Wielinga, B. (2010), 'Relating conversational expressiveness to social presence and acceptance of an assistive social robot', *Virtual reality* **14**(1), 77–84.

Henderson, R. R., Bradley, M. M. & Lang, P. J. (2018), 'Emotional imagery and pupil diameter', *Psychophysiology* **55**(6), e13050.

Kaplan, F. (2005), Everyday robotics: robots as everyday objects, *in* 'Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies', ACM, pp. 59–64.

Kawai, S., Takano, H. & Nakamura, K. (2013), Pupil diameter variation in positive and negative emotions with visual stimulus, *in* 'Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics', SMC '13, IEEE Computer Society, Washington, DC, USA, pp. 4179–4183.
URL: *https://doi.org/10.1109/SMC.2013.712*

Kazemi, V. & Sullivan, J. (2014), One millisecond face alignment with an ensemble of regression trees, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1867–1874.

Kreibig, S. D. (2010), 'Autonomic nervous system activity in emotion: A review', *Biological psychology* **84**(3), 394–421.

LabScribe2 (2013), 'Galvanic skin response (gsr) and investigation into 'cheating''.

McColl, D. & Nejat, G. (2012), Affect detection from body language during social hri, *in* '2012 IEEE RO-MAN: the 21st IEEE international symposium on robot and human interactive communication', IEEE, pp. 1013–1018.

Mehrabian, A. (1967), 'Orientation behaviors and nonverbal attitude communication.', *Journal of communication* .

Mehrabian, A. (1996), 'Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament', *Current Psychology* **14**(4), 261–292.

Mehrabian, A. & Friar, J. T. (1969), 'Encoding of attitude by a seated communicator via posture and position cues.', *Journal of Consulting and Clinical Psychology* **33**(3), 330.

Nass, C., Fogg, B. & Moon, Y. (1996), 'Can computers be teammates?', *International Journal of Human-Computer Studies* **45**(6), 669–678.

Nass, C. & Moon, Y. (2000), 'Machines and mindlessness: Social responses to computers', *Journal of social issues* **56**(1), 81–103.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B. & Dryer, C. (1995), Can computer personalities be human personalities?, *in* 'Conference companion on Human factors in computing systems', ACM, pp. 228–229.

Nourbakhsh, N., Wang, Y., Chen, F. & Calvo, R. A. (2012), Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks, *in* 'Proceedings of the 24th Australian Computer-Human Interaction Conference', ACM, pp. 420–423.

Panigrahy, S. K., Jena, S. K. & Turuk, A. K. (2017), 'Study and analysis of human mental stress detection using galvanic skin response and heart rate sensors in wired and wireless environments', *Research Journal of Pharmacy and Technology* **10**(4), 1168.

Papousek, I., Freudenthaler, H. H. & Schulter, G. (2008), 'The interplay of perceiving and regulating emotions in becoming infected with positive and negative moods', *Personality and Individual Differences* **45**(6), 463–467.

Posner, J., Russell, J. A. & Peterson, B. S. (2005), 'The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology', *Development and psychopathology* **17**(3), 715–734.

Rani, P., Sarkar, N., Smith, C. A. & Kirby, L. D. (2004), 'Anxiety detecting robotic system–towards implicit human-robot collaboration', *Robotica* **22**(1), 85–95.

Russell, J. A. (1980), 'A circumplex model of affect.', *Journal of personality and social psychology* **39**(6), 1161.

Schaefer, K. (2013), 'The perception and measurement of human-robot trust'.

Scholkmann, F., Boss, J. & Wolf, M. (2012), 'An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals', *Algorithms* **5**(4), 588–603.

Shaw-Garlock, G. (2009), 'Looking forward to sociable robots', *International Journal of Social Robotics* **1**(3), 249–260.

Shi, Y., Ruiz, N., Taib, R., Choi, E. & Chen, F. (2007), Galvanic skin response (gsr) as an index of cognitive load, *in* 'CHI'07 extended abstracts on Human factors in computing systems', ACM, pp. 2651–2656.

Stienen, B. & de Gelder, B. (2011), 'Fear detection and visual awareness in perceiving bodily expressions.', *Emotion* **11**(5), 1182.

Turkle, S. (2011), 'Alone together'.

Villarejo, M. V., Zapirain, B. G. & Zorrilla, A. M. (2012), 'A stress sensor based on galvanic skin response (gsr) controlled by zigbee', *Sensors* **12**(5), 6075–6101.

Weiss, A., Bernhaupt, R., Lankes, M. & Tscheligi, M. (2009), The usus evaluation framework for human-robot interaction, *in* 'AISB2009: proceedings of the symposium on new frontiers in human-robot interaction', Vol. 4, pp. 11–26.

Xu, J., Wang, Y., Chen, F., Choi, H., Li, G., Chen, S. & Hussain, S. (2011), Pupillary response based cognitive workload index under luminance and emotional changes, *in* 'CHI'11 Extended Abstracts on Human Factors in Computing Systems', ACM, pp. 1627–1632.

Yagoda, R. E. & Gillan, D. J. (2012), 'You want me to trust a robot? the development of a human–robot interaction trust scale', *International Journal of Social Robotics* **4**(3), 235–248.

Young, J. E., Hawkins, R., Sharlin, E. & Igarashi, T. (2009), 'Toward acceptable domestic robots: Applying insights from social psychology', *International Journal of Social Robotics* **1**(1), 95.

Appendix A

Code Snippets

Arduino

The below is the code run on the Arduino:

```
void setup ( )
{
  pinMode(A0, INPUT); //HR
  pinMode(A2, INPUT); //GSR
  analogReference(DEFAULT);
  Serial.begin (9600);
}

float hr  = 0;
float GSR = 0;
void loop ()
{
  hr = analogRead (A0);
  GSR = analogRead(A2);
  delay(50);
  Serial.println(String(GSR)+',' +String(hr));
}
```

Appendix B

Trust Questionnaire

Participant Information

What is your age?

Your answer

What is your gender?

- ◯ Female
- ◯ Male
- ◯ Other: ______

State your occupation

Your answer

What is your general experience with robots? (Interaction, programming, usage, etc.)

	1	2	3	4	5	
No Experience	◯	◯	◯	◯	◯	Very Experience

Consent form

- ☐ I confirm that I have been informed about the purpose of the study and understood the information given to me. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.
- ☐ I understand that my participation is voluntary and that I am free to withdraw without giving any reason.
- ☐ I understand that the session will be video and audio recorded.
- ☐ I understand that direct quotes may be used anonymously from the recorded session in any write up, presentation or publication of this study.
- ☐ I understand that all personal information will be anonymised and treated confidentially.
- ☐ I agree to take part in the above study.

NEXT

Participant Information

* Required

Evaluation Form

On a scale on 1-7 please mark how you think the robot performed on each of the follow parameters

Function successfully *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Act consistently *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Reliable *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Predictable *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Dependable *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Follow directions *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Meet the needs of the mission *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Perform exactly as instructed *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Have errors *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Provide appropriate information *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Unresponsive *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Malfunctioning *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Communicate with people *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Provide feedback *

	1	2	3	4	5	6	7	
Strongly Disagree	○	○	○	○	○	○	○	Strongly Agree

Task Completed (filled by experiment conductors) *

Choose

BACK SUBMIT

Appendix C

Complete Trust Scores

Participant	Gender	Give Fast	Give Slow	Take Fast	Take Slow
1	Female	3,786	5,714	5,714	5,714
2	Male	4,786	4,929	4,714	4,714
3	Male	4,643	4,929	4,929	3,571
4	Male	4,500	4,571	5,286	5,214
5	Male	4,286	4,929	4,929	5,786
6	Male	3,714	4,214	4,571	4,929
7	Female	5,429	5,857	5,929	5,714
8	Male	5,857	5,857	6,000	5,786
9	Female	5,357	5,714	5,286	6,000
10	Female	5,429	4,929	4,286	5,214
11	Female	5,214	6,000	4,857	5,643
12	Female	3,857	5,714	5,714	5,786
13	Male	5,214	5,643	5,571	5,786
14	Male	4,071	5,929	4,000	5,214
15	Male	4,786	5,429	4,500	4,786
16	Male	5,786	5,786	5,857	6,143
17	Female	4,571	5,500	5,500	3,357
18	Male	5,500	5,214	5,286	5,500
19	Male	5,714	5,714	5,071	5,286
20	Male	4,786	5,571	5,786	5,714
Total	-	**97,285**	**108,142**	**103,785**	**105,857**
Mean	-	**4,864**	**5,407**	**5,189**	**5,292**
Variance	-	**0,465**	**0,248**	**0,336**	**0,544**

Table C.1: Show the trust score from the water cup experiments. Also show the mean and variance from each test.

Participant	**Gender**	**In front Fast**	**In front Slow**	**Behind Fast**	**Behind Slow**
21	Male	4,286	3,500	3,857	4,000
22	Female	4,643	6,000	5,286	5,643
23	Male	5,571	5,929	5,214	5,357
24	Male	4,714	4,571	4,929	4,714
25	Male	5,786	4,143	3,714	3,071
26	Female	4,143	5,071	4,286	4,643
27	Male	3,286	5,857	5,714	4,786
28	Male	5,357	5,429	5,214	5,500
29	Male	5,643	4,643	5,000	5,429
30	Male	5,286	4,929	4,643	5,714
31	Male	3,429	3,429	4,143	3,357
32	Female	4,214	5,214	4,143	4,571
33	Female	4,143	4,714	3,643	3,786
34	Male	3,500	3,929	3,357	3,786
35	Female	4,286	4,500	4,643	4,857
36	Male	5,071	5,214	5,500	5,714
37	Male	4,857	4,429	4,571	3,786
38	Male	5,357	5,500	5,357	5,429
39	Male	5,786	6,000	6,071	5,857
40	Male	5,571	4,500	5,071	5,071
Total	**-**	**94,929**	**97,500**	**94,357**	**95,071**
Mean	**-**	**4,746**	**4,875**	**4,717**	**4,753**
Variance	**-**	**0,619**	**0,555**	**0,737**	**0,650**

Table C.2: Show the trust score from the LEGO building experiments. Also show the mean and variance from each test.

Appendix D

Scoring Schemes