Human-Robot Trust Assessment From Physical Apprehension Signals

Hald, Kasper

DOI (link to publication from Publisher):
10.54337/aau429766207

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

HUMAN-ROBOT TRUST ASSESSMENT FROM PHYSICAL APPREHENSION SIGNALS

BY
KASPER HALD

DISSERTATION SUBMITTED 2021

AALBORG UNIVERSITY
DENMARK

Human-Robot Trust Assessment From Physical Apprehension Signals

Ph.D. Dissertation

Kasper Hald

Dissertation submitted February, 2021

Dissertation submitted: February, 2021

PhD supervisor: Prof. MSO Matthias Rehm
Aalborg University

Assistant PhD supervisor: Prof. Thomas B. Moeslund
Aalborg University

PhD committee: Associate Professor Claus Brøndgaard Madsen (chair)
Aalborg University

Professor with Special Responsibilities Kerstin Fischer
University of Southern Denmark

Associate Professor Antonio Sgorbissa
University of Genoa

PhD Series: Technical Faculty of IT and Design, Aalborg University

Department: Department of Architecture, Design and Media Technology

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-895-7

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2021

Curriculum Vitae

Kasper Hald

2014 - Graduated M.Sc. in Medialogy at School of Information & Communication Technology, Aalborg University

Specialized in Interaction.

Master thesis title: System Control, Navigation & Selection Using Off-the-Shelf Hardware for a Classroom Immersive Virtual Environment.

2014-2017 - Research Assistant at The Department of Architecture, Design & Media Technology, Aalborg University

Developing and evaluation tablet games for assessment and rehabilitation of acquired brain damage.

2018-Submission - Ph.D. Fellow at The Department of Architecture, Design & Media Technology, Aalborg University

Abstract

In my research I have worked towards enabling real-time human-robot trust assessment by inferring decreases human trust from signs of physical apprehension from the robot. The goal is to help enable robot-augmented production, where robots assist production staff to relieve them of repetitive and strenuous tasks. To ensure safe and productive human-robot collaboration we have to ensure an appropriate level of trust in the robot, as too much trust can lead to dangerous situations, whereas too little trust can lead to loss in productivity. My main hypothesis is that if the user experiences a decrease in trust, they will increase their distance from the robot by stepping or leaning away from it.

I designed and developed an augmented reality enabled human-robot collaboration cell, using projection to display task critical information within the shared work space. Using a Rethink Robotics Sawyer robot I performed a series of experiments where participants performed repeated tasks with the robot. In the middle of the experiments I would have the robot disrupt the participants' expectations in order to decrease their trust towards it. Using an infrared camera for body tracking I assessed changes in their movement to correlate it with decreases in trust.

Through my experiments I found that sudden increases in robot movement speed would decrease trust in the robot, while decreases in speed or having the robot change movement trajectory during collaboration had no effect. Also, if the robot would perform an action that went counter to the shared objective, trust decreased significantly. Despite provoking decreases in trust, I was not able to find consistent correlations with apprehensive movements. I also experimented with repairing trust using explanations of why the robot made a mistake, but they had no effect.

Resumé

I min forskning har jeg arbejdet på at muliggøre måling af menneske-robottillid ved at læse fald i tillid ud fra tegn på fysisk tilbageholdenhed i forhold til robotten. Målet er at muliggøre robot-udvidet produktion, hvor robotten assisterer de ansatte for at hjælpe med gentagende og besværlige opgaver. For at sikre produktiv og sikker menneske-robotsamarbejde er vi nødt til at opretholde et passende niveau a tillid til robotten, da for meget tillid kan føre til farlige situationer, hvorimod for lidt tillig kan føre til fald i produktivitet. Min hypotese er, at hvis brugeren oplever et fald i tillid, vil de øge deres afstand til robotten ved enten at træde eller læne sig væk fra den.

Jeg designede og udviklede en menneske-robotsamarbejdsstation udstyret med augmented reality, som kan vise opgavekritisk information i arbejdsområdet ved hjælp af projektioner. Med en Rethink Robotics Sawyer robot udførte jeg en serie af eksperimenter, hvor deltagerne udførte gentagende opgaver i samarbejde med robotten. Midtvejs gennem eksperimenterne fik jeg robotten til at bryde deltagerens forventninger for a nedsætte deres tillid til den. Hjælp af et infrarødt kamera til sporing af brugeren målte jeg ændringer i deres bevægelser for at korrelere dem med fald i tillid.

Gennem mine eksperimenter fandt jeg ud af, at pludselige forhøjelser af robottens bevægelseshastighed ville sænke tillid til robotten, hvorimod sænkelse i hastighed eller ændringer i robottens bevægelsesmønster under samarbejdet ikke havde nogen effekt. Derudover, hvis robotten udførte en handling, som gik imod det delte samarbejdsmål, ville tillid også falde. På trods af at jeg kunne provokere fald i tillid, var jeg ikke i stand til at finde konsistente korrelationer med tilbageholdende bevægelser. Jeg testede også reparation af tillid ved hjælp af forklaring om, hvorfor robotten lavede en fejl, men de havde ingen effekt.

Contents

Preface

The Ph.D. fellow position is sponsored in part by Innovation Fund Denmark to be part of the project, Augmented Cellular Meat Production. Collaborators in the project and the research include Danish Crown, Danish Technological Institute, Egatec AS, Khora as well as the University of Southern Denmark and Augsburg University.

I would like to thank Danish Meat Research Institute, part of Danish Technological Institute, for input on framing human-robot interaction experiments for relevance to the context of industrial meat production as well as the Lab for Human-Centered Artificial Intelligence at Augsburg University for the research collaboration. I would also like to thank my supervisors, the Human-Machine Interaction group as well as my other colleagues at Aalborg University for their help and feedback during my research.

Kasper Hald
Aalborg University, February 9, 2021

Part I

Introduction

Chapter 1

Scope & Contribution

In this chapter I present the initial motivation for the research into robot-augmented production, after which I outline the main frameworks and terminology surrounding the research objectives. I present the main frameworks informing our approach to human-robot collaboration (HRC) and human-robot trust (HRT). Lastly, I present the objective and scope of the research going into this thesis.

1.1 Robot-Augmented Production

In recent years the long-term effects of working in industrial production lines have been observed, showing that performing strenuous repeated tasks over a career of many years highly increase the risk of acquiring quality-of-live-affecting musculoskeletal deceases in later life [7]. In Denmark there is an especially high risk in working in industrial meat production [1]. Because the work is set up in a production line where each staff member performs a few tasks repeatably on every cut of meat at a conveyor belt, it is common for the staff to be doing the same motions repeatably for long periods of time.

This thesis is in part motivated by these issues with the goal of advancing the field of human-robot interaction (HRI) in the context of close-proximity HRC. Introducing collaborative robots in industrial meat production can potentially help relieve the human worker, from here referred to as the operator, of the strenuous tasks, as it can assist in tasks such as positions or flipping over the cuts meat, or even assisting in cutting. HRC is a sub-category within HRI. Whereas HRI can be used to describe a brought variety of contexts, such as talking to a socially enabled robot, HRC is specifically about working together with the robot to achieve a common objective. While HRC also encompasses remotely controlling a robot, such as for search-and-rescue operations, I focus on close-proximity collaboration, which introduces some unique challenges, as the scenario does not allow for safety barriers.

While the field of close-proximity HRC is not new, the body of research have increased throughout the 2000s [3]. The focus of the majority of the research has been on the practical elements of the collaboration, such as operator tracking as well as task and motion planning. Due to the nature for meat processing where the cuts come in many varieties, the system sensors and planning software have to be very advanced to adapt and perform the tasks correctly. In addition, this means that the robot will have to move differently for each task, making the motions less predictable to the operator. Therefore an appropriate level of operator trust is necessary to prevent accidents, as sudden arm movements sparked by uncertainty is hazardous in close proximity to a robot equipped with powerful or sharp tools. It is therefore the focus of this thesis to develop methods of real-time trust assessment in the context of close-proximity HRC.

A necessary part of HRT is communication between the system running the robot and the operator, as system transparency and the properties of Explainable Artificial Intelligence improves operator trust. Because of this I also evaluate non-obstructive display modalities to facilitate system communication to aid in the collaboration, such as for displaying relevant task information and planned collaboration steps. With fast-paced production work in mind where the staff stand at a work surface, the initial design is based on the ability to overlay information on the work surface and the task subjects, those being the cuts of meat. I therefore start by investigating the capabilities of augmented reality (AR) technology. An illustration of the HRC production scenario is shown in Figure 1.1.

1.2 Human-Robot Collaboration

While HRI and HRC are not new fields of research, as experiment on humans interacting with social robots and collaborative manipulators have been done for decades, the main body of empirical research into close-proximity HRC has been done after the turn of the millennium. The topics of optimizing collaborative manufacturing and operator safety seeing particular growth after 2010 and HRT being tested less frequently [3]. The increase in publications of these topics involving close-proximity HRC experiments is shown in Figure 1.2.

When referring to social robots versus manipulators, social robots are designed with mimicking social interaction as their primary function, often having anthropomorphic features such as arms and a face for expression, while manipulators are designed for moving objects, primarily. There are seldom overlaps within research where manipulators are tested for their utility in

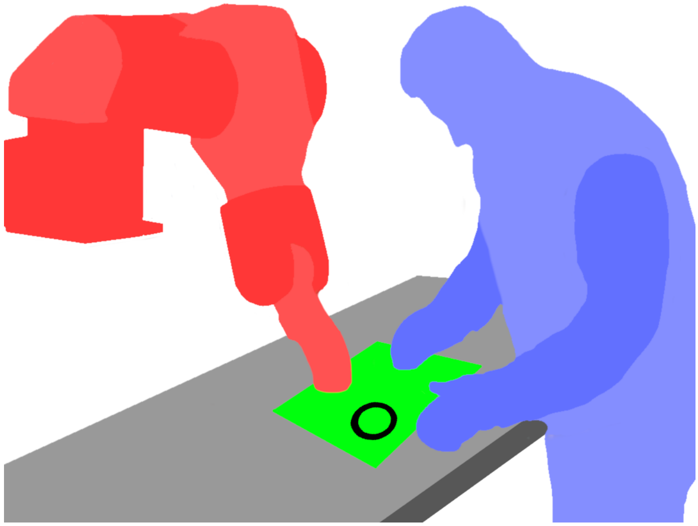


Fig. 1.1: A robot and an operator working in close proximity on a shared subject (green) with a circle projected onto it using AR to notify the operator of an area of interest.

social interaction, but the type of robot and whether it is anthropomorphic has been shown to affect operator perceptions of and trust in the robot [4].

HRC is a specific sub-topic HRI in which the robot and operator are working together to achieve an objective separate from the interaction itself. This often involves physical manipulation of one or several objects. This definition allows for a wide variety of scenarios, including the operator interacting with or controlling the robot remotely, or the operator and robot operating separately from one another, such as moving a collection of objects from one side of a room to the other. The common term for all types of collaborations between robots and humans is human-robot teams, but whether it is considered HRC is judged by the team composition and level of engagement between the human and robot [6].

Malik and Bilberg [6] proposed a model for referencing the types of HRC based on characteristics on three axes; interaction levels, team compositions and safety implications. Team composition describes the number human and robots involved in the collaboration. In this thesis I am focusing on teams composed of one human and one robot. Malik and Bilberg's levels of engagement are especially useful in illustrating the types of collaboration I am interested in. The levels named Isolation and Co-existence involve the operator and robot each performing their own tasks with no shared space, one with robot in a cage and one with them alongside one-another, respectively. In Synchronized Collaboration they do share a work space and goal, but

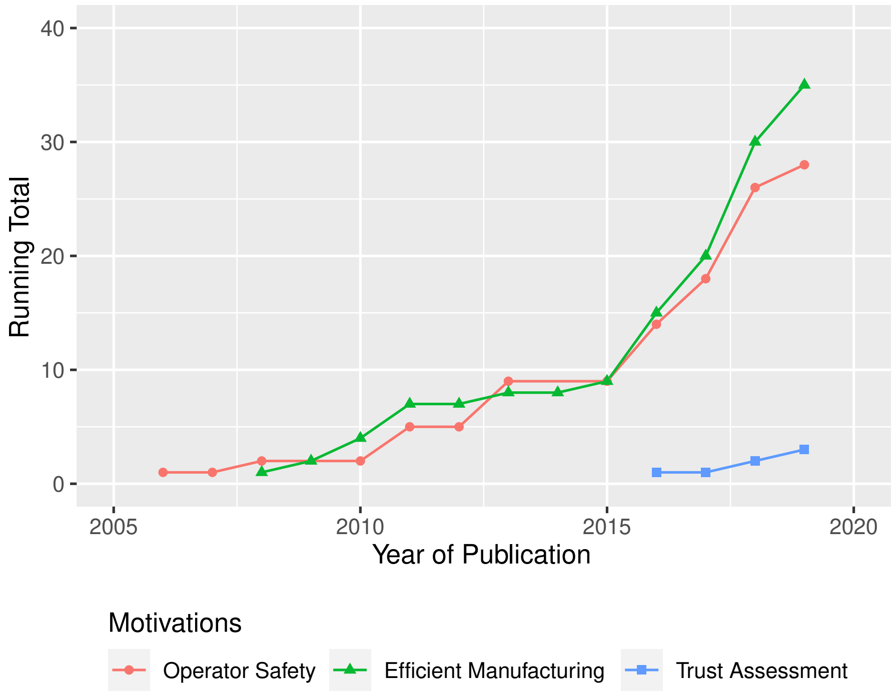


Fig. 1.2: Accumulative total of publications in empirical close-proximity HRC with stated motivations being optimizing collaborative manufacturing, operator safety and trust assessment. [3].

only one of them occupies the work space at any given time in what could be considered a turn-taking fashion. What separates the last two levels of engagement, Cooperation and Collaboration, is that while both involve a shared goal and work-space, which both occupy simultaneously, it is only when both operator and robot work on the same component at the same time, Malik and Bilberg label it as Collaboration. The levels of engagements is illustrated in Figure 1.3.

In this thesis, HRC is used in reference to mainly the Cooperation and Collaboration levels of engagement. When reviewing the literature of close-proximity HRC, our criteria was that publications had to document an experiment involving a human operator collaborating with a robot while within reach of the robot end-effector with no safety barriers. They also had to manipulate the same object or collection of objects during the task, but not necessitating that they had to touch the same object simultaneously. The literature review is documented in Paper A [3].

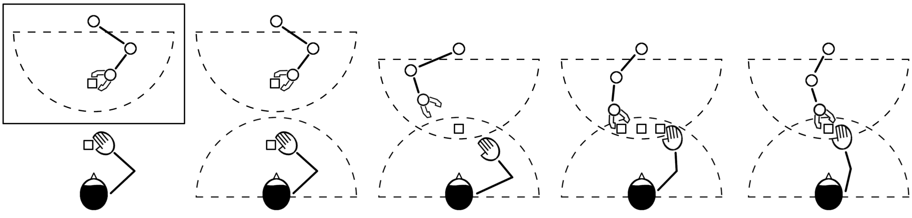


Fig. 1.3: The five levels of engagement in HRC, illustrating the overlapping work spaces of the operator and robot [6].

1.3 Human-Robot Trust

Trust in a collaborative robot partner is very important, as it has an effect on the decisions of the operator [4]. In the research there is large overlap between trust in robots and trust in automation in general. Lee & See [5] defined trust in automation as the attitude that an agent will help achieve a goal in a situation involving uncertainty and vulnerability. This definition is with the caveat that it must consider the appropriateness of the trust as well as the context and the characteristics of the robot.

In describing trust calibration in human-robot teams, de Visser et al. [2] use a similar definition of trust, based on the willingness of a party to be vulnerable to the actions of another party based on the expectations towards that party. Both use definitions based on the willingness to be vulnerable as a results of the cognitive processes based on the goal, the context and the characteristics of the partner.

Both Lee & See [5] and Hancock et al. [4] present similar models on trust calibration for automation and robots, respectively. The goal of trust calibrations is to achieve the appropriate level of trust for the operator relative to the actual trustworthiness of the system, actual trustworthiness being an objective measure of the systems capabilities to achieve the goal. Cases of over trust can lead to misuse, which is potentially dangerous in HRC, while under trust can results in disuse, which could be considered non-optimal use. The process of trust calibration consists of trust-dampening actions in the case of over trust and trust-repairing actions in case of under trust. The concept of trust calibration is illustrated in Figure 1.4 [2].

In a meta-analysis Hancock et al. [4] categorized factors affecting trust development, grouping them between human-related, robot-related and environmental factors. The robot-related factors are further categorized between performance-based and attribute-based. The performance-based factors, among which are reliability, predictability and failure rates, are especially relevant in the context for robot-augmented production, as these factors

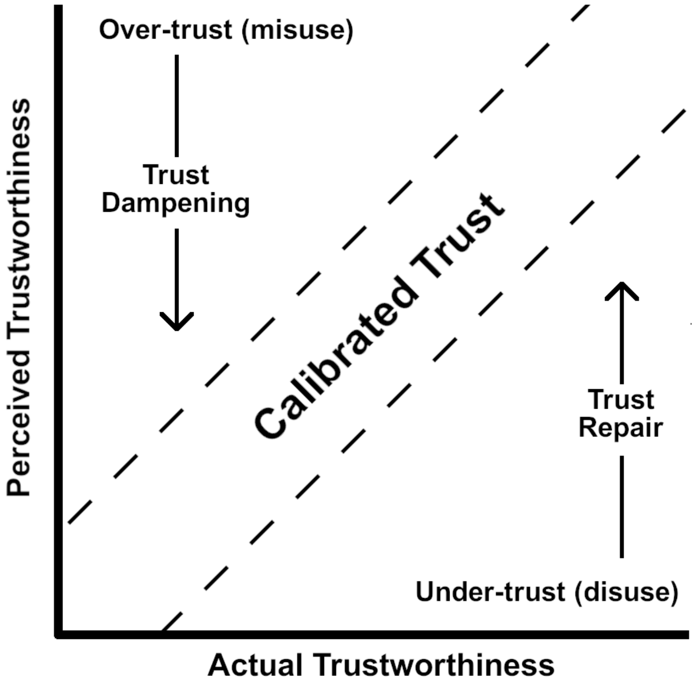


Fig. 1.4: Illustration of trust calibration, maintaining appropriate trust in the robot using trust repair and trust dampening [2].

will also have a large effect on operator safety.

1.4 Thesis Scope

The objective of this thesis is to contribute to safe robot-augmented production. As safe operations in HRC require appropriate levels of trust I look to enable trust calibration, which requires that one can infer the operator's current level of trust in the robot and the system. Operator trust toward a robot is often evaluated using questionnaires, which is not feasible in the context of robot-augmented production. Therefore I research the possibilities of non-obstructive sensors for real-time HRT assessment by correlating reported trust with sensor data in the context of close-proximity collaboration.

I approach the objective with the hypothesis that physical apprehension signals towards the robot can serve as an indicator of a decrease trust. As the definitions used for HRT [4, 5] involve the willingness to engage in a vulnerable situation, apprehension in the form of the operator increasing their distance to the robot should indicate trust dampening. Approaching this using computer vision-based motion tracking will also be useful for other applications, such as implicit human-robot communication. Rani et al. [8] described how the affective state of the operator can be used in implicit communication and inform the next actions of the robot in combination with the task objectives and the operator's actions. They showed that enabling an adaptive human-aware motion-planning system improved the operator's feeling of safety around the robot.

My approach is a series of HRC experiments featuring repeated collaborative tasks while the test subjects report their trust toward the robot. Throughout the experiments I introduce trust-dampening actions by disrupting the subject's expectations toward the robot in order to correlate the decrease in trust with the sensor data gathered. Due to the lock-down in 2020 I also investigated the options of trust assessment experiments using virtual reality (VR).

A secondary objective is enabling HRC using non-obstructive displays for system communication based on AR. This allows us to evaluate the effects of robot transparency in trust-repairing actions. For this I did a test on trust repair through mistake explanation in collaboration with the Lab for Human-Centered Artificial Intelligence at Augsburg University. It also informs how to design the test scenarios for evaluating real-time trust assessment. This is why, despite being being a secondary objective, system communication is the first subject to be implemented and evaluated.

The topics and papers in the thesis and how topics derive from others are shown in Figure 1.5. Based on the research objective I aimed to address the following questions.

1. How can we enable communication between the operator and the system controlling the robot using AR?

2. How can we use system communication in trust-repairing actions to increase HRT?

3. How can we measure HRT throughout repeated close-proximity HRC tasks?

4. How can we lower the operator's trust toward the robot through trust-dampening actions?

5. How can we correlate body tracking as signals of physical apprehension with measured HRT throughout repeated close-proximity HRC tasks?

6. How can we perform HRT assessment experiments in VR?

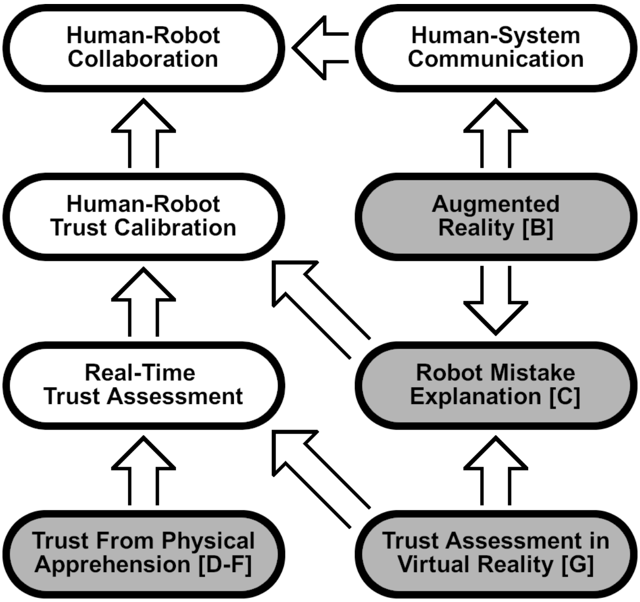


Fig. 1.5: The scope of the thesis, the main subjects in frames and with the arrows indicating which topics informed or enabled work on other topics. Subjects that I have performed experiments on are framed in grey and the paper indexes in this thesis are shown in square brackets.

1.5 Summary

In this chapter I described the motivations and challenges of implementing robot-augmented manufacturing technologies in industrial meat production, as well as the frameworks of HRC and HRT and trust calibration I am applying in the research. Lastly, I described the main objectives and scope of the thesis.

- People who work in manufacturing positions with a lot strenuous or repetitive movements have a high risk of developing musculoskeletal deceases. Introducing collaborative robots can potentially help reduce this risk.

- I focus on HRC at the levels of engagement where the robot and operator are either cooperating or collaborating, meaning they inhabit the same work area, either by taking turns or working on the subject simultaneously.

- Appropriate HRT is necessary for safe and efficient HRC, as too much trust may be dangerous and too little trust may lead to disuse of the robot. Trust can be calibrated to an appropriate level using trust-dampening and trust-repairing actions.

- HRT is a relatively novel field with most published experiments with close-proximity HRC having been done after 2010.
- System communication and transparency is important for maintaining trust. I am working with AR technology, as it is not obstructive and can be used to display task relevance information in the work area.
- As a step towards enable trust calibration I research real-time trust assessment using computer vision. I focus on signs of physical apprehension toward the robot through sudden motions or increased distance from the robot.
- In addition to testing trust assessment using trust dampening actions, I test system transparency and mistake explanations as trust-repairing actions.
- Throughout this thesis I will be addressing six research questions pertaining to system communication, AR, HRT assessment and VR.

References

[1] Arbejdstilsynet Denmark, "Anmeldte erhvervssygdomme 2012-2017," 2017. [Online]. Available: https://at.dk/media/4814/bilag-b-tabeller-med-anonymitet-final.pdf

[2] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[3] K. Hald, M. Rehm, and T. B. Moeslund, "Shared-Space Human-Robot Collaboration - Literature review on experiments in collaborative robots in close-proximity with human operators." Under review at the Autonomous Robots Journal, Springer, 2020.

[4] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[5] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, jan 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50{_}30392

[6] A. A. Malik and A. Bilberg, "Developing a reference model for human–robot interaction," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 13, no. 4, pp. 1541–1547, 2019.

[7] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of Electromyography and Kinesiology*, vol. 14, no. 1, pp. 13–23, feb 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641103001251?via{%}3Dihub

[8] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, jan 2004. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0263574703005319

Chapter 2

State of the Art

This chapter presents some of the existing research that informed the methods and procedures used throughout my studies. As human-robot trust (HRT) assessment is a main focus, I review previous empirical studies and assessment methods, as well as experiments into robot transparency and mistake explanations. To outline the technical requirements of my research I look at previous applications of augmented reality (AR) for system communication and body tracking utilized in human-robot collaboration (HRC). In addition, due to lock-down restrictions, I have researched the effects of performing HRC experiments using virtual reality (VR).

2.1 Human-Robot Trust Assessment

Schaefer [24] developed two scales with questions pertaining to the robot's characteristics, capability, dependability, predictability, errors and more. One scale with 40 items and a shorter version with 14 items. In both scales the participants rate each item in a scale of 0 to 100 in intervals of 10. The final trust score is the average of all item scores with negative qualities having inverted values. These scales were shown to be more sensitive and accurate than previous scales on trust in automation. Similarly, Charalambous et al. [4] developed a trust scale designed for industrial robots, specifically, with ten items pertaining to operator comfort, robot intimidation factor and reliability. Unlike the Schaefer scale, the items are statements graded according to agreement on a five-point scale. Kessler et al. [16] compared the Schaefer scale to a trust scale for automated systems by having participants rate the same robot with both scales. They found conflicting results, suggesting that they evaluate different constructs and are therefore not interchangeable.

Alternative measures of trust have previously been used in experiments. Rani et al. [21] utilized physiological measurements for affect recognition in the context of remote human-robot interaction (HRI). They measured inter-beat interval, relative pulse volume, electrodermal activity and facial electromyographic activity to potentially be used in an adaptive affect-based robot controller. Freedy et al. [12] inferred HRT from number of operator interventions as they were observing the operations of

Table 2.1: Summary of methods of assessing HRT in prior research.

Method	Features
Questionnaires [4, 14, 24]	Reported after interaction with the robot, but allows for high levels of detail.
Physiological measurements [21]	Automatic assessment with delay dependent on physiological reaction.
Number of operator interventions [10–12]	Based on operator behaviour and easily utilized in adaptive system.
Marker-based hand tracking [22]	Inferred from operator movements compared to previous movement patterns.

a collaborative unmanned ground vehicle based on perceived trustworthiness. Sadrfaridpour et al. [22] used marker-based hand tracking to infer trust using a neural network based on operator work speed compared to the robot in a HRC assembly task.

Researching previous trust experiments in with close-proximity robots, Dragan et al. [6] evaluated HRT dependent on robot motions patterns in collaborative task where the robot would hand the operator a cup of water. They used Hoffman's measure, which is a compound measure based on a series of questions, few of which reference trust and feeling of safety directly [14]. Comparing purely functional motions, based on motion planning for obstacle avoidance, to motions designed to be predictable, the predicable motions were more accepted by test participants. Bergman & Zandbeek [2] found that speed and stopping distance of an industrial robot at close proximity had significant effect trust based on questionnaires with five statements rated by agreement on a five-point scale. They conclude that speed and stopping distance should be considered as communicative cues.

In an example of application of trust, Floyd et al. [10, 11] implemented an adaptive behavioral framework for a robot based on inverse trust estimation, assessing operator trust based on the number of interruptions during the interaction, mainly adapting to under-trust. This was demonstrated in a simulated patrolling task with different types of simulated operators. Previously used methods of assessing HRT and their features are summarized in Table 2.1.

Computer vision-based body tracking in HRI is usually utilized for safety purposes. Both Morato et al. [19] and Tan & Arai [27] implemented skeleton tracking for safe HRC using a Kinect for standing work and a triple-camera setup for sedentary tasks, respectively. While I tested skeleton tracking during the thesis, I later adopted a top-down RGB-depth (RGB-D) camera setup. In a literature review on the possibilities and challenges of this type of tracking Liciotti et al. [18] outlined how this camera configuration has been successfully utilized in both people recognition and behaviour analysis. The uses include security and video analysis, intelligent retail environments and activities of daily living, but they only found one example involving HRI or robots in general.

2.2 Robot Transparency & Mistakes

Looking at trust and transparency in HRC, Boyce et al. [3] tested the effects of three transparency conditions on HRT in a human-in-the-loop robot simulations experiments with higher levels of transparency yielding higher trust scores using a modified trust in automation scale.

In a survey using a simulated robot testbed, where the human-robot team would perform reconnaissance missions to gather intelligence in a foreign town, Wang et al. [30] found that adding decision explanation capabilities increased transparency, trust and team performance. Based on their analyses, however, there was only significant differences when comparing explanation to no-explanation conditions for a simulated low-ability robot. A high-ability robot saw no gains in trust from enhanced decision explanations.

Kaniarasu & Steinfeld [15] tested the effects of the robot assigning blame after an error in a collaborative task. Comparing three robot personalities, one where the robot blamed itself, one blaming the user and one blaming the human-robot team, results showed that participants were annoyed at the robot blaming them, but they also showed less trust toward a robot that kept blaming itself.

Salem et at. [23] tested a home companion robot, comparing a robot that would make errors with one that did not. Results showed that the robot performing errors significantly affected perceived trustworthiness, while it did not have a significant effect on participants' willingness to follow the robot's instructions.

2.3 AR & VR in Human-Robot Interaction

Implementing AR in manufacturing is not a novel concept, and the umbrella term of AR-aided robot control, testing, assembly and transport has been coined as AR-aided manufacturing [20]. Using three case studies, Szafir [26] argued for the utility of virtual, augmented and mixed reality in HRI, demonstrating significant performance benefits.

To determine the most appropriate AR hardware for the context, Elia et al. [8] proposed a 4-step process of evaluation depending on the specific manufacturing processes. The first step is pair-wise comparing the options using a multi-criteria analysis, ranking the options based on output modalities, reliability, responsiveness and agility. The second step is creating a judgement matrix from the comparisons. Step three is evaluation local weights and consistency of comparison, and step four is the final ranking of devices. Elia et al. also categorized the four main types of AR displays as head-mounted displays (HMD), handheld devices, projection-based displays and haptic force-feedback systems.

Kruijff et al. [17] categorized types of AR displays similarly to Elia et al., although excluding the haptic devices. They pointed out relevant issues of using see-though HMDs, as they have limited field of view (FOV), also pointed out by Szafir [26], and risk vergence-accommodation conflicts for virtual objects. For projection-based AR they pointed out the challenge presented by distortion of the images when projecting on uneven surfaces. Schwerdtfeger et al. [25] compared HMDs to laser projection-based displays, pointing out the low FOV and limited resolution of the HMD. In addition, the potential swimming effect of objects superimposed in see-through displays and having multiple planes of focus are significant challenges with high risk of eye fatigue. While they pointed out that projection-based AR addresses some of these problems, it also introduces other issues, mainly the requirement to pre-distort the image if it is to be projected onto an uneven surface and the issue of the operator occluding the projection if they reach into the work area.

While AR has been tested in the context of human-system collaboration, it is often implemented in sitting tasks, such as in a study by Green et al. [13], who utilized a HMD. Rather than using a see-through HMD, they used a stereo OLED display, which showed the feed from a camera mounted to the headset and angled downwards for more ergonomic posture during use. Charoenseang & Tonggoed [5] proposed an AR setup for sedentary collaborative assembly with a robot. Unlike the AR display types previously categorized, they used a top-down camera feed on a monitor in front of the operator, on which they displayed task-relevant overlays, yielding high usability scores. Similarly, Fang et al. [9] demonstrated a monitor-based AR interface for robot path planning with positive results. Lastly, Vogel et al. [28] proposed projections-based AR to display safe areas when working close to a robot.

During the thesis period the country went into lock-down, so I decided to explore evaluation of HRI using VR, allowing for experiments outside of the laboratory. Wainer et al. [29] tested the importance of robot presence on human-robot team performance and perception of the robot, comparing interaction with a co-located robot to a remote robot and a virtual robot, both presented on a screen. They found that the co-located robot was significantly favoured over the alternatives. Similarly, Bainbridge, et al. [1] compared a co-located humanoid robot to the same robot displayed on a screen with similar results.

Duguleana et al. [7] did a comparative study on the effect of immersive VR on HRI using a cave automation virtual environment. When compared the participants would give a real robot more space than the virtual one, but the subjects reported high engagement toward the virtual robot and a high degree of realism, 7.8 out of 10, relative to the real robot.

2.4 Summary

In this chapter I described state of the art of HRT with a focus on assessment methods and empirical studies as well as the effects of robots and transparency and the utility of AR, VR and body tracking in HRI. Below are the main takeaways.

- Schaefer [24] developed two HRT scales that have been proven to have higher validity than previous automation trust scales.
- Physiological measurements as well as marker tracking have previously been used in robot affect and trust assessment with promising results.
- While skeleton tracking in HRC is most often utilized for security purposes, top-down RGB-D cameras have successfully been used for people recognition and behaviour analysis.

- System transparency have been shown to increase HRT and human-robot team performance.
- A robot making mistake significantly decrease perceived trustworthiness.
- The main categories of AR displays are see-though HMDs, handheld devices, projection-based AR, haptic force-feedback systems and top-down camera feed with graphical overlays.
- While co-located robots are generally preferred over simulated or remote robots on a screen, a virtual robot may still be rated highly in terms of engagement and realism by users.

References

[1] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, “The effect of presence on human-robot interaction,” in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 701–706.

[2] M. Bergman and M. van Zandbeek, “Close Encounters of the Fifth Kind? Affective Impact of Speed and Distance of a Collaborative Industrial Robot on Humans,” *Human Friendly Robotics*, pp. 127–137, 2019.

[3] M. W. Boyce, J. Y. Chen, A. R. Selkowitz, and S. G. Lakhmani, “Effects of agent transparency on operator trust,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 179–180.

[4] G. Charalambous, S. Fletcher, and P. Webb, “The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration,” *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, 2016.

[5] S. Charoenseang and T. Tonggoed, “Human–robot collaboration with augmented reality,” in *International Conference on Human-Computer Interaction*. Springer, 2011, pp. 93–97.

[6] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, “Effects of Robot Motion on Human-Robot Collaboration,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI ’15*. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[7] M. Duguleana, F. G. Barbuceanu, and G. Mogan, “Evaluating human-robot interaction during a manipulation experiment conducted in immersive virtual reality,” in *International Conference on Virtual and Mixed Reality*. Springer, 2011, pp. 164–173.

[8] V. Elia, M. G. Gnoni, and A. Lanzilotto, “Evaluating the application of augmented reality devices in manufacturing from a process point of view: An AHP based model,” *Expert Systems with Applications*, vol. 63, pp. 187–197, nov 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0957417416303505

[9] H. Fang, S. Ong, and A. Nee, "A novel augmented reality-based interface for robot path planning," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 8, no. 1, pp. 33–42, 2014.

[10] M. W. Floyd, M. Drinkwater, and D. W. Aha, "Adapting autonomous behavior using an inverse trust estimation," in *International Conference on Computational Science and Its Applications*. Springer, 2014, pp. 728–742.

[11] ——, "How much do you trust me? learning a case-based model of inverse trust," in *International Conference on Case-Based Reasoning*. Springer, 2014, pp. 125–139.

[12] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *2007 International Symposium on Collaborative Technologies and Systems*. IEEE, may 2007, pp. 106–114. [Online]. Available: http://ieeexplore.ieee.org/document/4621745/

[13] S. A. Green, J. G. Chase, X. Chen, and M. Billinghurst, "Evaluating the Augmented Reality Human-Robot Collaboration System," in *2008 15th International Conference on Mechatronics and Machine Vision in Practice*. IEEE, dec 2008, pp. 521–526. [Online]. Available: http://ieeexplore.ieee.org/document/4749586/

[14] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," Tech. Rep., 2013. [Online]. Available: http://guyhoffman.com/publications/HoffmanRSS13Workshop.pdf

[15] P. Kaniarasu and A. M. Steinfeld, "Effects of blame on trust in human robot interaction," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 850–855.

[16] T. T. Kessler, C. Larios, T. Walker, V. Yerdon, and P. Hancock, "A comparison of trust measures in human–robot interaction scenarios," in *Advances in human factors in robots and unmanned systems*. Springer, 2017, pp. 353–364.

[17] E. Kruijff, J. E. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," in *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, oct 2010, pp. 3–12. [Online]. Available: http://ieeexplore.ieee.org/document/5643530/

[18] D. Liciotti, M. Paolanti, E. Frontoni, and P. Zingaretti, "People detection and tracking from an rgb-d camera in top-view configuration: review of challenges and applications," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 207–218.

[19] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-Time Human Tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, p. 011006, jan 2014. [Online]. Available: http://computingengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4025810

[20] J. Novak-Marcincin, J. Barna, M. Janak, and L. Novakova-Marcincinova, "Augmented Reality Aided Manufacturing," *Procedia Computer Science*, vol. 25, pp. 23–31, jan 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091301209X

[21] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, jan 2004. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0263574703005319

[22] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and control of trust in human-robot collaborative manufacturing," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 115–141.

[23] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 1–8.

[24] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[25] B. Schwerdtfeger, D. Pustka, A. Hofhauser, and G. Klinker, "Using laser projectors for augmented reality," in *Proceedings of the 2008 ACM symposium on Virtual reality software and technology - VRST '08*. New York, New York, USA: ACM Press, 2008, p. 134. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1450579.1450608

[26] D. Szafir, "Mediating human-robot interactions with virtual, augmented, and mixed reality," in *International Conference on Human-Computer Interaction*. Springer, 2019, pp. 124–149.

[27] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, may 2011, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/5942335/

[28] C. Vogel, M. Poggendorf, C. Walter, and N. Elkmann, "Towards safe physical human-robot collaboration: A projection-based safety system," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, sep 2011, pp. 3355–3360. [Online]. Available: http://ieeexplore.ieee.org/document/6094550/

[29] J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric, "The role of physical embodiment in human-robot interaction," in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2006, pp. 117–122.

[30] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

Chapter 3

System-Operator Communication

In this chapter I present my research into the system-operator communication modalities and how we can utilize them for trust-repair during close-proximity human-robot collaboration (HRC), as communication and system transparency are critical for maintaining operator trust toward the robot and the system. I present the design and development of the HRC cells used in the human-robot trust (HRT) assessment experiments described in Chapter 4. These include one real setup using projection-based augmented reality (AR) and a virtual reality (VR) setup based on the communication modalities of the real setup. This chapter includes the research and evaluation of appropriate AR display types as outlined in Section 2.3. Lastly, I present the experiment on the utilization of these communication modalities for trust-repairing actions through mistake explanation.

3.1 HRC Test Cells

The HRC test setup for my experiments into real-time HRT assessment is designed around projection-based AR as the main communication modality based on the results from the AR display experiment described in Section 3.1.1. For my experiments I used a Rethink Robotics Sawyer robot arm. The test cell is designed to let the operator be in close proximity to the robot for various collaboration scenarios with a table between them.

To allow for collaboration within the work space without occlusion of the projection by either the robot or the operator, I installed two projectors, mounted at either side above the work surface on an aluminium rig. By calibrating the projections such that they would overlap, both would have to be occluded at the same time for there to be a blocked area on the work surface. The full setup and demonstration of the projection are shown in Figure 3.1. In addition, this setup leaves room for a depth camera pointed down at the operator and work area from above. The setup was first

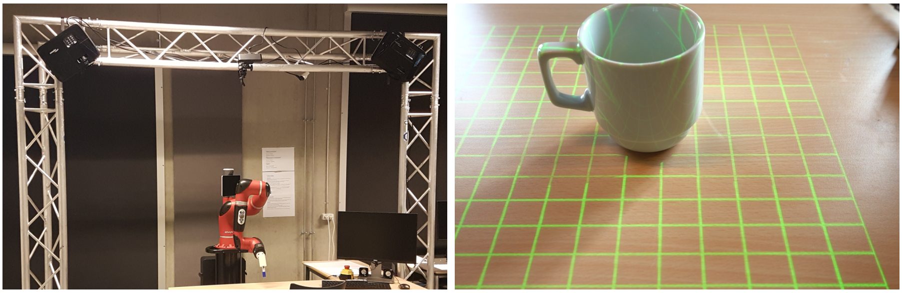


Fig. 3.1: Left: Full HRC cell setup with Sawyer robot, projectors and infrared camera. © 2020 IEEE [6]. Right: A coffee cub sitting under the dual-AR projection, preventing the cub from occluding either side of the AR overlay.

used in the experiment documented in Paper E [4].

Because of lockdown restrictions in 2020 I also designed a VR HRC test cell with the goal of enabling test subjects to participate remotely if they had the right equipment. The VR environment was designed to allow the same types of communication as the AR setup. The main method of communication used in this setup was text displayed on the work surface, mirroring the projection-based AR. The VR test setup is shown in Figure 3.2.

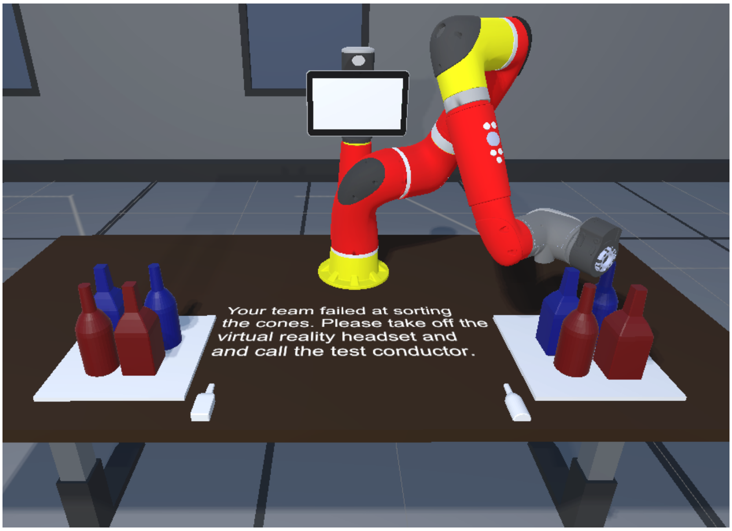


Fig. 3.2: The virtual environment and virtual robot used to test mistake explanations for trust repair. The bottles had to be sorted with round bottles on the right and square bottles on the left. The space between the two bottle areas allows for displaying text for communicating with the participant [7].

3.1.1 Evaluating AR Displays for Task Communication

I started my research by evaluating the use of AR as a method of system communication in the context of industrial manufacturing, as it allows for displaying task-critical information inside the work area without being obstructive. In order to determine what AR display type should be used going forward, I did an experiment, comparing four display types in a task based on a hypothetical situation fitting the meat production context. This was based on the scenario that a computer vision system would detect foreign elements or impurities on or below the surface of a piece of meat, and the operator would have to remove it as it is being shown to them on an AR display.

To represent this scenario in the test I used a tray of sand, as it allows participants to dig into it easily, yet it can easily be rearranged and reused for other participants. To evaluate which AR display type is the most suitable for the task I asked participants to poke into the sand with a nail and hit a point projected into it using one of four displays. The nail was mounted to a tracked HTC Vive controller, and the participants were asked to poke at the position as quickly and precisely as possible, then confirm their hit with a second controller.

I based the four AR display types tested on the categorizations by Elia et al. [2] and Kruijff et al. [9] as well as the setup used by Charoenseang & Tonggoed [1] and Fang et al. [3]. These being a head-mounted display (HMD), a handheld device, projection-based AR and a top-down camera feed presented on a monitor. In my case I used a see-though glasses-like HMD, as opposed to one with displays and a camera feed, and the handheld devices was mounted on an adjustable arm, as the participants would be using both hands in the task. The displays I evaluated are shown in Figure 3.3.

On the AR displays the targets were shown as red dots. The displays also showed a green grid along the surface of the sand to help participants see where the targets are in relation to it. This also allowed for depths cues like occlusion and motion parallax for the HMD and handheld display. Because neither the projection-based AR nor the top-down camera feed on the monitor allowed for depth cues, the target positions were displayed with a number indicating their depth under the surface in millimeters. In hindsight, for more valid comparisons the depths should also have been displayed for the other two displays. Still, based on the results and observations this would likely not have made a difference on the decision.

Results showed that the projection-based AR display was among the highest in accuracy and the lowest in task completions times along with high Standard Usability Scores. A significant challenge in using the HMD and the mounted handheld display is that they are dependent on visual tracking, because they are non-static, which makes them highly vulnerable to occlusion when the operator reaches in front of their cameras. For the mounted handheld display I was able to position it such that it could have all tracking markers and the entire work area within its camera's field of view (FOV), making it more viable.

The HMD was the worst performing display due its small display area withing the operator's field of vision in additions to the low FOV of its mounted camera. The small display area where the AR overlay was visible required the operator to search around the surfaces of the sand, as they could not observe it all at once. This, in turn, made tracking more difficult as the mounted camera would have to maintain

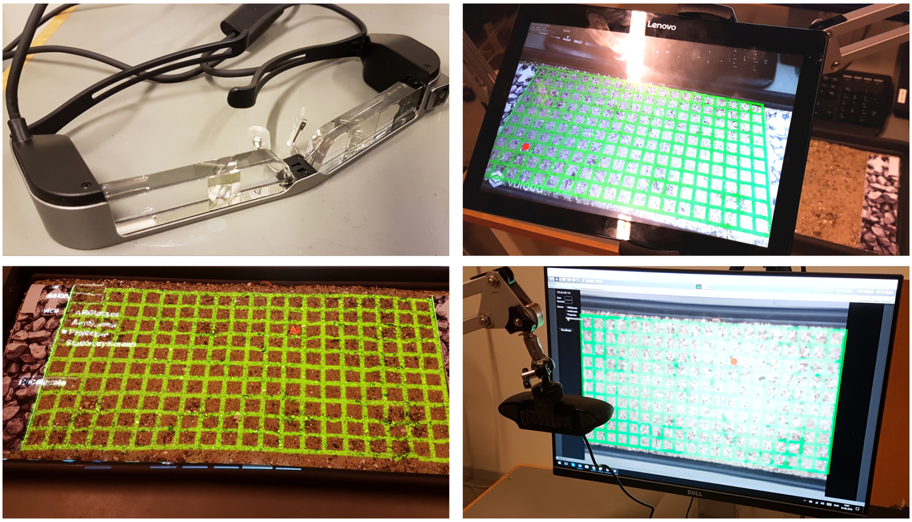


Fig. 3.3: The four AR display types evaluated in the experiment. Top-left: Epson Moverio Bt-300 see-through HMD with camera. Top-right: Android tablet with back-camera feed on adjustable arm. Bottom-right: Projected overlay on the surface of the sand. Bottom-right: Top-down camera feed with overlay on monitor. © 2019 Springer, Cham [5]

or resume tracking as it would be moving between markers, making the search even more difficult. For better conditions I could have placed more markers along the surface, but it may not be as feasible for the manufacturing context. Alternatively, the tracking solution would have to be separate from the headset, so that the AR overlay would be fed to the HMD according its position. The experiment and results are documented in detail in Paper B [5].

3.2 Effects of Robot Mistake Explanation

To evaluate the utility of system communication and transparency in HRT repair based on the collaboration context and output modalities I had been working with up to this point, I did a study in collaboration with the Lab for Human-Centered Artificial Intelligence at Augsburg University. The objective of the study was to test types and methods of explaining an error made by the robot during a shared task that could be recreated in the projection AR test cell. The experiment was performed using the VR test setup.

Our initial research question was focused on the output modalities used to explain the robot's error to the operator. We performed a preliminary study using an online survey, featuring videos of a virtual robot to illustrate it committing two types of errors. In a scenario where the robot had to sort differently shaped bottles and put them at either end of a table, one video would show the robot sorting them incorrectly and another video showed the robot knock over a bottle while performing the task otherwise correctly. As a between-subject condition, after watching each robot error the participants would receive an explanation of the error in either text projected onto the table or from synthesized speech. The failed sorting was explained as a computer vision error while knocking over bottle was explained as the motion planning system not being calibrated properly.

While we expected the participants to request more detailed explanations, perhaps using visual aids that could be implemented using projection AR, most participants requested additional details on how to correct the error in the system. We therefore decided to evaluated whether adding details on how to solve the problem would improve trust repair after the robot error. With this change in focus we limited the scope of the experiment and only tested with the robot sorting bottles incorrectly and the explanations were only delivered in text.

The sorting test from the preliminary study was adapted to be a collaborative task where the robot and operator had to work together as a team to sort the bottles, assigning the bottles to each team member by color, red or blue. The virtual environment and setup used for the experiment is shown in Figure 3.2. Every participant would perform the task with the robot twice, and report their trust toward robot after each task using Schaefer's 14-item robot trust scale [10]. Measuring HRT and performing experiments are explored in detail in the Chapter 4. In the first task the robot would sort its bottles correctly, and the participant would be informed that the team succeeded when they both had finished. In the second test the robot would sort them incorrectly and once they were both finished, the participant would be informed that the team had failed. Depending on the test group, some participants would receive no additional information about the failure or the error, others would be informed that the system had performed a computer vision error, resulting in the robot's error, while last group would also be told that improving lighting conditions would fix the problem. After the experiment each participant would fill out a Explanation Satisfaction Scale [8] (ESS) post-test.

After testing with 30 participants results of Schaefer's trust scale showed a significant decrease in trust between the first and second task, suggestion that the robot acting counter to the shared objective is an effective trust dampener. The scale did not, however, show significant difference in trust levels after the robot error between the levels of explanation. The post-test ESS did, however, show significant difference in regards to trust between the no-explanation condition and the two other conditions, meaning participants found the explanations useful for informing whether they could trust the robot. Despite this we can not assume that this is transferable to trust in the robot, especially considering the ESS was administered after the trust assessment questionnaires. Further studies are required to learn how trust in different construct within the system affect on another regarding the user's perception of them. This study is documented in detail in Paper C [7].

3.3 Summary

In this chapter I presented our research in human-system communication for HRC. This included our evaluation of suitable AR display technologies, the design and of

AR HRC cell used for the HRT experiments and our test of mistake explanations as trust-repairing actions.

- Based on accuracy, task completions times and usability scores, I would work with projection-based AR going forward.
- Our AR HRC cell features two projectors mounted on an aluminium rig to allow for two overlapping projections, minimizing occlusion. It also allows for mounting a top-down camera or sensor.
- After testing levels of explanation after the robot performed an error, providing an explanation to why the robot made the error had no trust-repairing effect.

References

[1] S. Charoenseang and T. Tonggoed, "Human–robot collaboration with augmented reality," in *International Conference on Human-Computer Interaction*. Springer, 2011, pp. 93–97.

[2] V. Elia, M. G. Gnoni, and A. Lanzilotto, "Evaluating the application of augmented reality devices in manufacturing from a process point of view: An AHP based model," *Expert Systems with Applications*, vol. 63, pp. 187–197, nov 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0957417416303505

[3] H. Fang, S. Ong, and A. Nee, "A novel augmented reality-based interface for robot path planning," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 8, no. 1, pp. 33–42, 2014.

[4] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *The 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) IEEE International Symposium on Robot and Human Interactive Communication*. IEEE Computer Society Press, 2019.

[5] ——, "Testing Augmented Reality Systems for Spotting Sub-Surface Impurities." Springer, Cham, aug 2019, pp. 103–112.

[6] ——, "Human-robot trust assessment using motion tracking & galvanic skin response." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

[7] K. Hald, K. Weitz, E. André, and M. Rehm, "Trust repair with virtual robot using levels of mistake explanation," in *Under review at IFIP Conference on Human-Computer Interaction – INTERACT 2021*. ACM, 2021.

[8] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," 2019.

[9] E. Kruijff, J. E. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," in *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, oct 2010, pp. 3–12. [Online]. Available: http://ieeexplore.ieee.org/document/5643530/

[10] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

Chapter 4

Trust Assessment

In this chapter I present my research into human-robot trust (HRT) assessment based on the robot trust and trust calibration definitions presented in Section 1.3. This includes four empirical studies documented in Papers D through G, but I have split the chapter into sections pertaining to the base topics of my research approach. These topics are how I measure the ground truth in HRT, how I disrupt operator expectations with trust-dampening actions and how I use different tracking methods to measure physical reactions when the expectations are disrupted.

4.1 Measuring Ground Truth

As outlined Section 2.1, Schaefer [7] created two versions of a HRT scale that was both more sensitive and valid than previous trust in automation scales, one with 40 items and one with 14 items. I used the 14-item scale in the experiment documented in Paper F [4], but because even the short version of scale would take a long time to fill out repeatably in a series of tasks, I decided to investigate quicker alternative ways of trust reporting.

For two experiments, documented in Papers D [1] and E [2], I used an Android tablet to administer a shortened version of Hoffman's measure [6], using only three items pertaining to robot predictability and operator feeling of safety. The three items were formulated as the following statements:

- I trust the robot to do the right thing at the right time.
- I felt safe working next to the robot.
- The robot's reaching motions was surprising.

The statement referring the robot's reaching motion is mainly relevant in the experiment documented in Paper D [1], as the robot would reach toward the participants with a wooden baton. This question should be changed to fit the robot's actions in other experiments. To get higher granularity and with the aim of getting quicker and more intuitive responses from the participants, the three question were formulated as statements and the participants had to state their level of agreements using a sliding scale with option to simply touch the scale to move the marker, allowing for rapid responses. The combined trust score is the average score for the three questions where disagreement yield zero and agreement yields one, with the statement that the robot's motion was surprising having its value inverted. The questionnaire is shown in Figure 4.1.

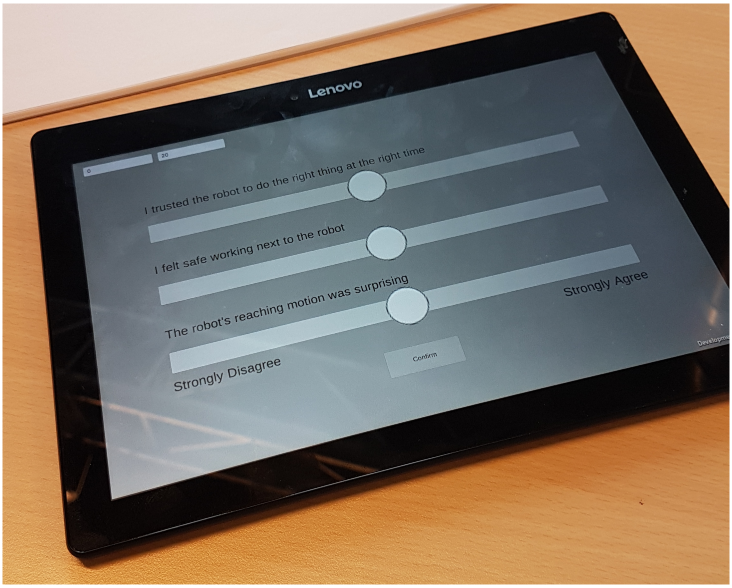


Fig. 4.1: The tablet used to administer the short trust scale questionnaire. The sliding scales let the participants quickly report their attitude toward the robot as level of agreement with the three statements [1]

Due to the lock-down in 2020, I decided to experiment with HRT assessment in virtual reality (VR), allowing for tests without bringing people into the laboratory. I designed it to be used as part of data crowd-sourcing, where participants use it at home if they have the required VR equipment. Because of this it would have to be brief, compared to my other experiments, to avoid people quitting part-way through. This is especially a risk due to the test design where participants perform repetitive tasks with the intend of having them lose count, so I can disrupt their expectations halfway through. Doing this when the participant is alone instead of in a laboratory setting likely has a higher risk of participants losing patience and quitting, especially if they do not notice the trust-dampening action in the experiment. To make the test brief and also support input modalities for as many VR setups as possible, the questionnaire is stripped down to one questions on whether the participants feels safe while close to the virtual robot, which is reported on a seven-point scale according to agreement. This scale in the virtual environment is shown in Figure 4.2. The design and an in-lab evaluation of this measure are documented in Paper G [3].

While these small and rapid questionnaires have not been verified on their validity

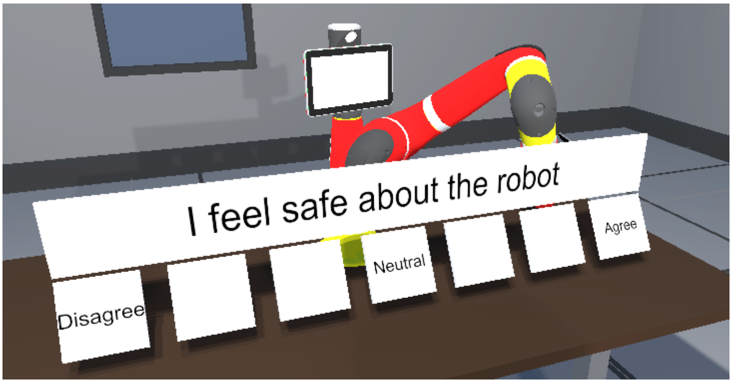


Fig. 4.2: The virtual robot, environment and seven-point scale used by participants to rate their agreement to the statement, that they felt safe about the virtual robot [3].

compared to Schaefer's [7] scales in a comparative experiment, in all experiment they yielded results expected when the robot would perform trust-dampening action in the more volatile conditions, such as speeding up or performing wrong actions. The scores throughout the experiments show a drop in trust at disruption of expectations, often followed by slow trust recovery. The trust scores throughout the experiments for effective trust-dampening actions are shown in Figure 4.3. Still, it would strengthen the conclusions to perform a study verifying the validity of the measures.

A potential weakness to the trust scales is that the scores have to be within set minimum and maximum values. Based on participants' perceptions of the robot and the scales, different participants may report trust at the extremes of the scales at different levels of perceived robot trustworthiness. This is especially concerning when analysing delta trust scores after the robot performs a trust-dampening action, as a higher trust level before the action allows for a greater drop in trust. To check if there is an effect of high trust levels on delta trust after trust-dampening, I collected all the trust scores before and after for conditions in each of my experiment that showed significant drop in trust. Normalizing the scores and performing linear and logarithmic regression on trust delta scores dependent on trust score before yielded R-squared values of 0.08 for both. This suggests the effect of trust levels before trust dampening can not be proven, though it should still be considered when planning future studies.

4.2 Disrupting Operator Expectations

Throughout my experiments I have tested various ways of disrupting the participants' expectations toward the robot during their test sessions. Among the robot-related factors affecting trust development listed by Hancock et al. [5] are reliability and predictability, and in the experiments documented in Papers D through G I would manipulate these.

The common approach in the experiments is to have the participants perform a repetitive collaborative task with the Sawyer robot, in-between witch they would report their trust toward the robot. The robot's behaviour would be identical for the first half of the tasks. During these tasks I would expect to see an increase in trust toward the robot, as the participants are getting used to it and getting confident with the task. Then the robot would change behaviour to disrupt participant expectations, often yielding a drop in trust. For tests on changes in speed the robot would maintain the new behaviour for the rest of the tasks, often resulting in gradual recovery of trust. The statistical analyses for the four experiments are summarized in Table 4.1, comparing reported trust immediately before and after the trust-dampening actions occurred.

In the experiment documented in Paper D [1] the robot would hold a wooden baton that it would hand over to the test participant. Depending on the condition the robot would start at 25, 50 or 75 percent maximum movement speed for the first half of the tasks, after which the speed would be increased by 25 percentage points. An additional test condition was whether there would be a work surface between the robot and operator. There was an error in the analysis in Paper D, so despite what is documented, having the work surface between the robot and participant did not yield any significant difference, and the increase in speed was only effective when the robot went from 25 to 50 movement speed. This may be due to the proportionality of the increases in speed, as going from 25 to 50 percent doubles the speed while going from 50 to 75 percent only increases it by 50 percent. The trust scores from the significantly effective conditions from this experiment are shown in Figure 4.3, Section A.

In later experiments I would test conditions where the robot would either increase or decrease its movement speed. In the experiment documented in Paper E [2] the robot would hold a felt pen and draw a square on a piece of paper held down to a table by the test participant. in the VR experiment in Paper G [3] the robot would move toward the participant holding a plate with the letter A or B for the participant to read and report. Both of these experiments only showed significant changes in reported trust when the robot increased speed, suggesting that feeling of safety has a higher effect than robot predictability in these cases. As in Paper D, there is an error in analysis in Paper [2], yet the conclusion is the same with correct analysis. The trust scores for the speed-increase conditions from these experiments are shown in Figure 4.3, Sections B and C.

With increase in robot speed proven to affect HRT, for the last close-proximity HRC experiment I tested different types expectation disruptions. In the experiment documented in Paper F [4] the robot and participant each had to move a set of colored cones from side of a table to the other, repeatably. The participant would move the red cones while the robot would move the blue cones. For every task an arrow projected on the work surfaces indicated to which side the cones should be moved. In this experiment I tested the effect of robot dependability in addition to predictability by simulating two different error types. The first error type is designed to affect predictability, as the robot changes its movement trajectory and moves closer to the participant than previously. For the other error I changed dependability by having the robot initially move a cone to right place, after which it would move it back in the wrong direction. The participant had the option to stop the robot if they felt it acted

Table 4.1: Summary of statistical analyses in HRT experiments. Testing the difference in reported trust before and after a trust-dampening action, using t-tests for parametric data and Wilcoxon rank sum tests otherwise.

Test	*t*	*w*	*df*	p-value
Baton Handover Experiment [1]				
Increasing speed: 75% - 100%	1.39		16.27	.18
Increasing speed: 50% - 75%		57		.62
Increasing speed: 25% - 50%	4.99		16.29	< .01
Collaborative Drawing Experiment [2]				
Increasing speed		380		< .01
Decreasing speed		260		.09
Virtual Reality Experiment [3]				
Increasing speed		78		.03
Decreasing speed		45		.69
Collaborative Sorting Experiment [4]				
Irregular movement		71		.12
Wrong action	3.50		16.95	< .01

against the objective. The robot would then move the cone back to correct position and continue the task. Test results only showed a significant drop in trust when the robot moved the cone to the wrong position. An additional condition in this experiment was whether the participant was instructed to move the cones simultaneously with the robot, or if they were taking turns, moving one cone at a time. This condition showed no significant effect. The trust scores for the wrong-action condition from this experiment are shown in Figure 4.3, Section D.

4.3 Testing Body Tracking & Apprehension

The first body tracking experiment was performed using an Orbbec Astra RGB-Depth (RGB-D) camera mounted on a light stand and positioned facing the participants with the robot between them. In each task the robot would reach toward the participant with a wooden baton, handing it over to them, increasing the movement speed without warning halfway through the test. A participant retrieving a baton from the robot gripper is shown in Figure 4.4.

This experiment utilized the skeleton tracking software compatible with the camera, allowing me to easily distinguish and record the tracking data from the participants' limbs separately. In the analyses I focused on the positions and movements of the participants' hands and head, as I expect the extremities of the body to move the most during apprehensive behavior to increase their distance from the robot. The analyses showed that only the non-dominant hand's movement could be correlated with the reported trust level, which may have been due to them still reaching for the wooden baton out of habit with their dominant hand. The study is documented in detail in Paper D [1].

There are restrictions to the tracking setup with skeleton tracking, as it required a few meters of distance to the participant for proper tracking, and since it has to face

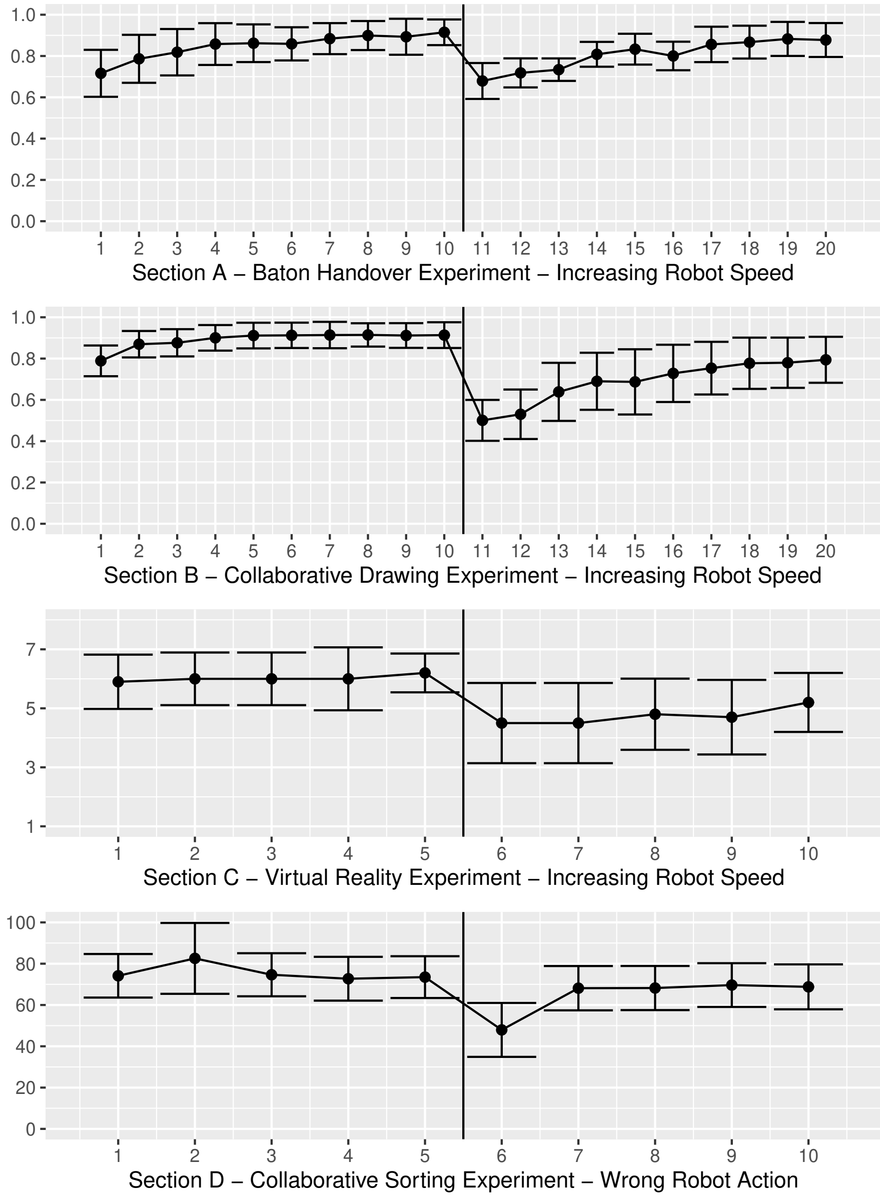


Fig. 4.3: Average trust scores with confidence intervals throughout the tasks for conditions with effective trust-dampening actions in the experiments in Papers D through G. The vertical lines show the midway point where the trust-dampening action occurred.

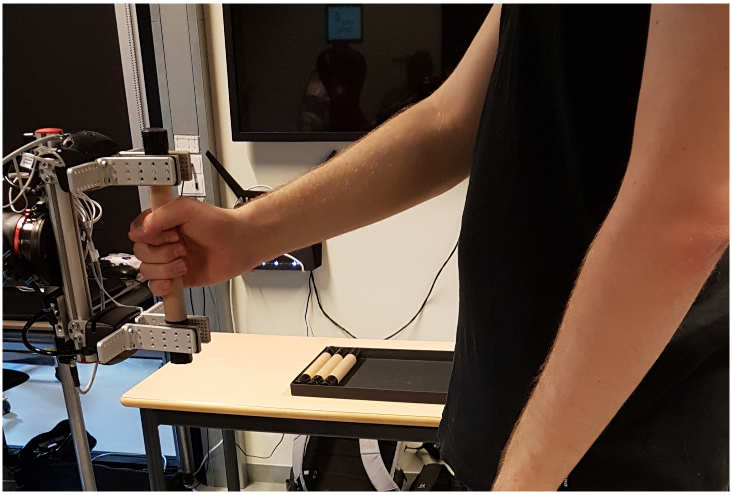


Fig. 4.4: The participant grabbing the wooden baton from the robot. © 2019 IEEE [1].

them front-on, the robot will have to be in-between them, occluding the participant. Because of this I proceeded with the top-down approach to operator tracking, which also informed the design of the AR HRC cell setup described in Section 3.1.

With the RGB-D camera mounted at the top of the aluminium rig, pointed down at the operator, I use the depth frame to track their approximate position and posture. First, a snapshot of the empty test environment is subtracted from the recorded frames. Each frame is then aggregated by each horizontal line of pixels, averaging all non-zero depth values, because zeros are pixels with no reflected infrared light. Looking at the values we can get an impressions of how the operator is physically distributed within the frame, as the values closest to the camera can be assumed to be the top of the head and the difference between the positions closest and farthest from the robot gives an impression of their posture, and if they are sitting or standing upright or leaning away or toward the robot. Examples of this tracking is shown in Figure 4.5. The design is documented in detail in Paper E [2].

I used this tracking method in two experiments. The first involved the participant sitting in front of the robot, holding down a piece of paper as the robot would draw a square on it. This scenario is shown in Figure 4.6. Although, the reported trust in the robot matched my hypotheses when the robot performed a trust-dampening action, there was no correlation with operator movement or proximity to the robot afterwards. This experiment is documented in detail in Paper E [2]. For the last close-proximity HRC experiment I had the participants do the standing task of moving cones across a table to see if more freedom of movement would yield more movement when disrupting their expectations. The setup for the test and the standing task is shown in Figure 4.7. However, as with the previous experiment, even though certain disruptive robot actions yielded changes in trust, there was no consistent correlation with movement or proximity. The last HRC experiment is document in Paper F [4].

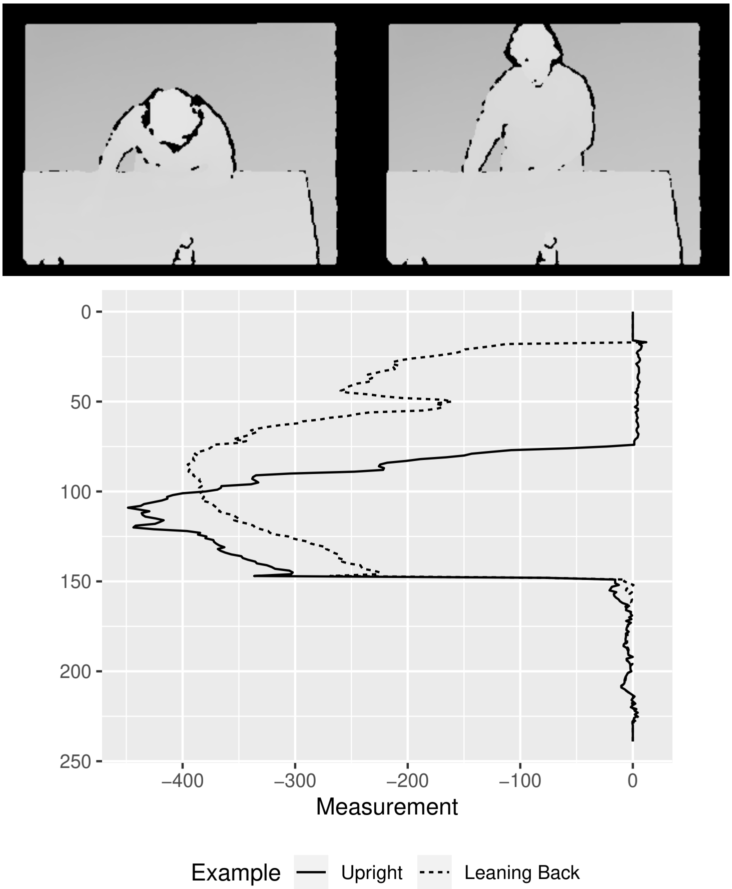


Fig. 4.5: Images from the top-down IR camera. Left: User sitting upright. Right: User leaning backwards. Below are the groups of aggregated tracking data based on the samples. © 2020 IEEE [2].

Fig. 4.6: A test participant holding down a piece of paper as the robot draws on it. © 2020 IEEE [2].

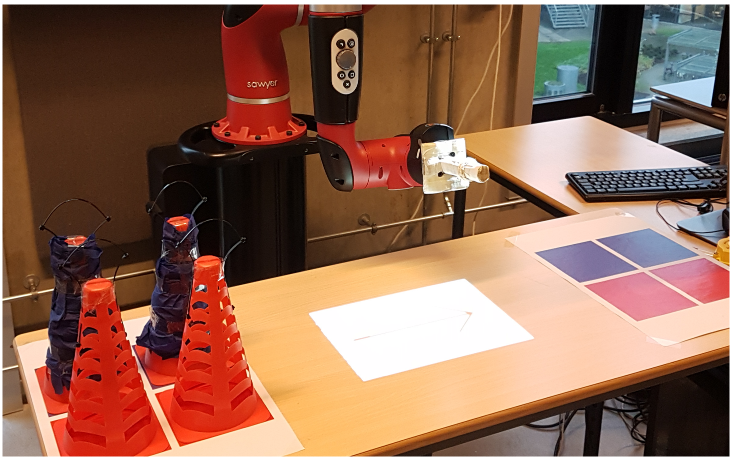


Fig. 4.7: The test setup where the participants would stand in front of the robot and repeatably move the colored cones from one side of the table to the other [4].

4.4 Summary

In this chapter I described my experiments into HRT assessment documented in Papers D through G, focusing on measuring reported trust, disrupting operator expectations and using two different types of body tracking.

- I designed a rapid trust reporting tool to be used between repeated tasks in close-proximity HRC. While the trust measure in my studies lets me retain hypotheses regarding gradual build of trust, drop in trust after disruption of expectations followed by gradual trust recovery, it would highly benefit from a verification study, comparing it to a standardized trust scale.
- I tested changes in robot speed, irregular movements and mistakes in task execution to affect operator trust. Results showed that increasing movement speed and mistakes in task execution caused decreases in operator trust, while irregular movements and decreasing speed had no effect.
- In testing body tracking to assess trust through physical signs of apprehension I found no consistent correlations between reported trust and operator motions or proximity to the robot during or after trust-dampening robot actions.

References

[1] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *The 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) IEEE International Symposium on Robot and Human Interactive Communication*. IEEE Computer Society Press, 2019.

[2] ——, "Human-robot trust assessment using motion tracking & galvanic skin response." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

[3] ——, "Human-robot trust assessment in virtual reality experiment for data crowdsourcing," in *Under review at 2021 International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[4] ——, "Human-robot trust assessment using top-down visual tracking after robot task execution mistakes," in *Under review at IFIP Conference on Human-Computer Interaction – INTERACT 2021*. ACM, 2021.

[5] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[6] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," Tech. Rep., 2013. [Online]. Available: http://guyhoffman.com/publications/HoffmanRSS13Workshop.pdf

[7] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

Chapter 5

Conclusion & Future Work

In my research I have worked toward enabling robot-augmented production by addressing the challenge of maintaining appropriate operator trust during close-proximity human-robot collaboration (HRC). I focused on real-time trust assessment with the goal to enable adaptive robot behaviour and real-time human-robot trust (HRT) calibration. My first step was enabling system communication using a non-obstructive display type for the manufacturing setting. To answer my first research question, *"How can we enable communication between the operator and the system controlling the robot using augmented reality (AR)?"*, i performed a usability study of different types of AR devices, in which projection-based AR was the most accurate and usable. I then designed the system communication for future studies around projection-based AR [3].

To answer my second research question, *"How can we use system communication in trust-repairing actions to increase HRT?"*, I performed a study in collaboration with the Lab for Human-Centered Artificial Intelligence at Augsburg University. After performing a user experiment using a virtual robot in virtual reality (VR) we found that the robot performing an action going counter to the shared objective significantly decreased HRT, but providing an explanation to the error did not result in less of a decrease, even though the explanations were rated higher in regards to informing whether they could trust the robot.

For the third research question, *"How can we measure HRT throughout repeated close-proximity HRC tasks?"*, I designed a digital solution for rapid reporting of trust toward the robot using a touch screen application. While this trust scale measured scores consistent with my hypotheses regarding decrease in operator trust after disruptive robot actions, a comparative study verifying it against standardized trust scales [9] would be highly beneficial.

My approach to real-time trust assessment was based on the hypothesis that decreases in trust are measurable in the operator's movements and their preferred proximity to the robot, as HRT is based on the willingness to engage in a vulnerable situation with a robot [1, 8]. To test this hypothesis I ran a series of experiment where the participants would perform repeated tasks with a Rethink Robotics Sawyer robot, during which the robot performs a trust dampening action, disrupting participant expectation with the purpose of invoking a decrease in trust. Between each task the participant would rate their trust towards the robot in a HRT scale. As for my fourth research question, *"How can we lower the operator's trust toward the robot through trust dampening actions?"*, increasing robot movement speed without warning or having the robot perform actions counter to the objective significantly lowered reported trust [2, 4, 6].

I then tested the correlation between the trust scores and the tracked movements during trust-dampening actions. While some of the trust-dampening actions did yield decreases in reported robot trust I found no correlations between the trust and the movements or proximity of the participants. As such, for my fifth research question, *"How can we correlate body tracking as signals of physical apprehension with measured HRT throughout repeated close-proximity HRC tasks?"*, at the current stage of my research, it is inconclusive. Lastly, for the final research question, *"How can we perform HRT assessment experiments in VR?"*, from the experiments I performed, trust score results were consistent with experiments with a real robot. This was the case both when verifying the VR test setup [5] and evaluating mistake explanations as trust-repairing actions [7].

For future work, although initial analyses did not yield useful correlations, I have gathered a large amount of tracking data, so further analyses may yield beneficial insights. Possibly, a machine learning approach can be useful to parse the tracking data. Alternatively, rather than the current top-down RGB-depth camera approach, where the depth frames are aggregated by averages for tracking of operator positions and posture, I can research skeleton detection algorithms suitable for the setup.

Another consideration for future research is the implications of performing HRC experiments on trust in laboratory environment. A considerable challenge to the validity of my approach is that participants may realize that the robot will make a mistake or otherwise disrupt trust at some point in the test, based on the questions they answer between tasks. This may be helped by instead using between-subjects experiments with a control that would not experience trust dampening actions and only having participants rate trust once. This would, however, yield much fewer data points per participants. Alternatively, we have to carefully consider how we frame the HRC experiments and how it affects the participants' expectations. For example, in my experiments there were no framing to explain why the participants would perform tasks with the robot. As such, the participant's perception of the situation is just that they are in an experiment run by the test conductor. We can assume that the participant's trust in the robot is influenced by their trust in the test conductor, because the conductor is the one controlling what occurs during experiment, unless they assume the robot errors are real. On the other hand, framing the experiment as a quality assurance test of the robot, introducing the possibility that robot is fallible, may more effectively yield the desired results.

Lastly, for future research more focus may be necessary on whether participants actually recognize robot errors or irregular behaviour, depending on the context. As could be seen in the VR trust repair [7] and cone sorting [6] experiments, having the participants actively performing tasks simultaneously with the robot may affect their ability to keep part of their attention on the robot. Future experiments should include an element pertaining to the degree to which participants recognized robot errors or irregular movements.

References

[1] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[2] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *The 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) IEEE International Symposium on Robot and Human Interactive Communication*. IEEE Computer Society Press, 2019.

[3] ——, "Testing Augmented Reality Systems for Spotting Sub-Surface Impurities." Springer, Cham, aug 2019, pp. 103–112.

[4] ——, "Human-robot trust assessment using motion tracking & galvanic skin response." 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

[5] ——, "Human-robot trust assessment in virtual reality experiment for data crowdsourcing," in *Under review at 2021 International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[6] ——, "Human-robot trust assessment using top-down visual tracking after robot task execution mistakes," in *Under review at IFIP Conference on Human-Computer Interaction – INTERACT 2021*. ACM, 2021.

[7] K. Hald, K. Weitz, E. André, and M. Rehm, "Trust repair with virtual robot using levels of mistake explanation," in *Under review at IFIP Conference on Human-Computer Interaction – INTERACT 2021*. ACM, 2021.

[8] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, jan 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50{_}30392

[9] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

Part II

Papers

Paper A

Shared-Space Human-Robot Collaboration - Literature review on experiments in collaborative robots in close-proximity with human operators

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper is under review in
Springer Journal, Autonomous Robots.

The layout has been revised.

Abstract

To get an overview of the research done into shared-space human-robot interaction we performed a literature review of relevant papers from the IEEE, ACM and Springer databases. We filtered 2097 papers from our initial search based on the criteria that they had to document a human-robot experiment where they share a space without any physical safety barriers, yielding 125 papers published between the years 2000 and 2019. We have written up overviews of years of publication and the growths of bodies of research regarding motivations, objectives, data collected and more. Initial findings suggest that human-robot collaboration for industrial manufacturing have been a major driving factor for research, with initial focus on efficiency, but with a growing interest in worker safety in recent years. We have also constructed a conceptual hierarchy of the types of roles for the robots in collaboration context.

1 Introduction

The field of human-robot collaboration is rapidly evolving with the next step being robotic collaboration partners sitting within reach of a human operator. With the next stage in the development of human-robot interaction (HRI) being shared-space and close-proximity human-robot collaboration (HRC), it is important to know findings and methods that have been used so far. Work is being done to implement collaborative robotics in many context and for many purposes. These include, but are not limited to, collaborative industrial assembly, relieving workers of strenuous and repetitive tasks, assisting the elderly or people otherwise in need of aid, as well as robots for collaborative transport of heavy or bulky objects.

Our goal is create a review and make an overview of the field of research into shared-space human-robot collaboration (SHRC). We define shared space as a human operator working with a robot in a shared environment with no physical barriers between them for safety or otherwise. Collaboration is defined as when the operator and robot are not only working towards the same goal, but also interacting with the same object or objects at the same time. Examples of this would be collaborative lifting or positioning and handing over objects.

For this literature review we go through 125 relevant papers from the previous two decades and categorize them in terms of goals, technology, methods and findings among others. We then aggregate the data and create and overview of the body of research throughout the years and with reflections on the field today and where it may go in the near future. We also analyse the practical collaboration type as well as the social roles of the robot as a partner in SHRC and write a conceptual hierarchy in terms of the types of robot influence on the collaboration.

2 Procedure

In this section we describe the methods used for gathering and sorting sources as well as the filtering and coding processes for the literature review.

2.1 Initial Search

The first collection of literature is gathered by searching the publication databases of Springer, ACM and IEEE with the search terms "Human-Robot Collaboration", "Human-Robot Cooperation", "Human-Robot Coordination" and "Human-Robot Team". We limit the search to these three databases as they cover the largest robotics conferences, and our main focus is on the boundaries of the research field. These search results are initially filtered to remove doubles and eliminate all articles of less than 3 pages of length. While "Human-Robot Interaction" is also a common keyword, in addition to covering collaborative manipulation of objects, it also commonly describes interaction with social robots. To not significantly lengthen the filtering process, this keyword was omitted.

2.2 Manual Filtering

After the initial search, the articles are manually filtered based on relevance to SHRC. To be included in the review the articles must describe an experiment using human test subjects sharing a space with a robot as they do a collaborative task. In this context "sharing a space" means that the human and a robot subjects must be in a proximity that allows physical contact, excluding them being in separate rooms or separated by a barrier, transparent or otherwise. For a robot to be relevant it must be moving at least one limb during the collaboration, whether it is for physical manipulation of an object or for communication. This excludes communication exclusively through audio-visuals from a monitor or otherwise. For a task to be considered collaborative the human and robot must interact with the same object or group of objects. This does not exclude sorting tasks. These features must be indicated in the title, abstract or keywords in order to be considered.

2.3 Coding

After filtering out irrelevant sources, we read through the remaining papers, focusing on specific elements and topics. With our focus being on procedures of experiments in SHRC these topics are stated author motivations, research objectives, their definition of HRC, technology used aside from the robot, task types in the experiment, operator-robot proximity, types of data gathered and overall results and conclusion from the authors. To do the coding, we identify the topics relevant for each category and add the labels to our data sheet if there are no fitting labels already. For each publication we add a binary indicator for whether each label in each category applies to the paper. This means that multiple levels in each category can apply to the same paper. For example, a robot used in an experiment can be classified as both a robot arm and a mobile robot. In addition, we classify the experiments as either technical test or HRI test based on whether the purpose is to verify the performance of new technology or if it also investigates human factors, respectively.

3 Findings

The initial search described in Section 2.1 yielded 2097 results from between 1992 and 2020 after removing doubles and articles less than 3 pages of length. After manually filtering the papers according to the guidelines in Section 2.2 we had a total of 125 paper for the review.

3.1 Topics

After reading through the papers and coding the relevant topics we have gathered the items listed in the tables below. During the coding process the papers were categorized as documenting either technical tests or human-centered tests. Technical tests focus on verifying the design and implementation of an HRC-related solution while still including at least one human participant, where human-centered tests focus on measuring human factors in HRC. The number of publications and test types throughout the years is shown in Figure A.1. We can see that technical HRC experiments have had the earliest publications in the early two-thousands with steady increase since then, while human-centered experiments had a high increase in publications in 2015 to comprise more than half the total HRC publications.

The stated motivations of the authors are listed in Table A.1 and the release years and accumulative totals are shown in Figure A.2. While some times only stated in the abstract, these are usually the first thing in the introduction and explain the intended long-term purpose of the research. Most often, they do not state an explicit purpose or context outside the research objective itself and are thus labelled as such. The most common motivation for HRC experiment is improving efficiency in human-robot manufacturing teams, followed by ensuring operator safety. Developing robot adaptation is often motivated by enabling obstacle avoidance during robot motion and often overlaps with operator safety and trust assessment.

The research objectives are listed in Table A.2. They are often stated explicitly in the introduction, but they can also be inferred by through the technology used and the data collected in the paper for a more complete picture. As such, testing visual tracking is a frequent objective, as it is often a part of the the system design. Objectives often overlap, and as intention recognition is often a step towards other objectives, such as motion or collaboration planning, it has a high frequency.

The types of HRC in the experiments are listed in Table A.3 and the release years are shown in Figure A.3. The label of shared work space is for experiments that did not involve the human and robot touching each other or a shared object. These usually involved collaborative sorting or the robot serving an item by placing it near the operator.

In Table A.4 the technologies used in addition to the main robot are listed, split between hardware and software. These include both technologies that enable the HRC and ones that are used for data collection. Force-torque sensors are often featured, and the sensors they use are usually the one build into the robot itself. The depth camera category includes infra-red cameras and ultra-sonic sensors, whether in single or multi-sensor configurations. Similarly, the RGB camera category includes single and multi-camera solutions. Projected augmented reality is grouped separately from

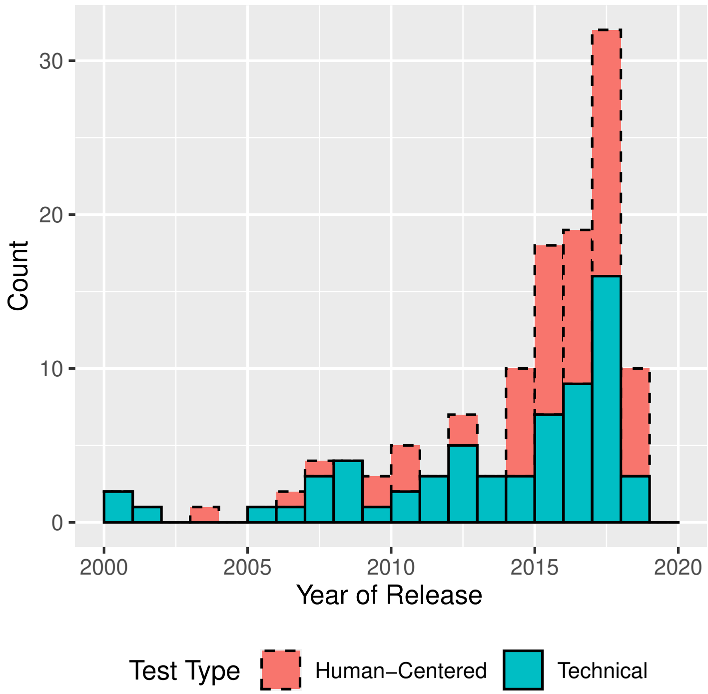


Fig. A.1: Number of SHRC-relevant papers published throughout the early two-thousands, colored according to test type.

other visual interfaces, which are usually on regular monitors.

The robot types are listed in Table A.5. Mobile robots and multiple robot labels always overlap with one of the other categories. The lift category is robots that use a scissor lift mechanism, rather than a jointed robot arm.

The HRC tasks performed in the experiments are listed in Table A.6. Many of these are also listed as collaboration types while others overlap, such as pick and place and pointing tasks both being shared work space collaboration.

The human-robot proximity is rarely documented with units of measurement, so in Table A.7 it is categorized mainly based on collaboration context. The release years and accumulated totals are shown in Figure A.4. Table distance means that the robot was either mounted to the work surface or was at the opposite side of it. Holding small and large objects means the operator and robot were separated by an object they were simultaneously touching as part of the collaboration. Small objects were usually no larger than a handball and large objects would be a table or a rod of about one meter in length.

The data types collected in the experiments are listed in Table A.8, split between objective and subjective measurements. The sensor data label is for internal system signals for primarily technical tests. The label miscellaneous questionnaires is used

Table A.1: Frequency of stated motivations in the papers.

Motivations	Freq.
Nothing stated [1–56]	56
Efficient manufacturing [56–90]	35
Operator safety [27, 32, 38, 43, 45, 48, 61–64, 67, 84, 87, 91–105]	28
Robot adaptation [30, 33, 39, 47, 50, 56, 57, 69, 76, 86, 89, 98, 104–115]	24
Disability aid [95, 103, 116–119]	6
Trust assessment [120–122]	3
Combating labor shortage [94, 123]	2
Robot maintenance [124]	1
Household robots [125]	1

when they are not standardized tests, such as the System Usability Score.

3.2 Test Results

Most of the papers focus on verifying implementation of novel SHRI solutions, most reporting success while others describe the weaknesses in their design or refined their problem outline. In these publications the results and conclusions are closely related to their research objectives and technologies used as listed in the Tables above. Still, several experiments are exploratory in nature and their results can reflect HRI in general.

Coban and Gelen [45] found that human-robot teams performed better than robots on their own. Several authors found that operators preferred autonomous and proactive robots over reactive ones [35, 68]. Baraglia et al. found that proactive robots yielded better teamwork fluency [68]. Han and Yanco [54] applied proactive robotics and handover task, where enabling the robot to the detect operator grasp effort patterns and act accordingly significantly improved the experience and efficiency. The preference for proactive robots partners is consistent with research showing that requiring repeated verbal instructions towards the robot has a negative effect on operator satisfaction [13, 17].

In researching human-robot group dynamics and collaboration planning, both Giuliani and Knoll [13] as well as Gombolay et al. [65] found that the operators were willing to cede planning authority to the robot partner. This may be a sign of operators trust in the system as a whole. Still, Gombolay et al. [65] also concluded that the operators still value human partners over robots in groups collaborations. Dragan et al. [21] found that motion planning that communicates robot intent improves collaboration, and Vannucci et al. [37] found that operators believe robot motions influence

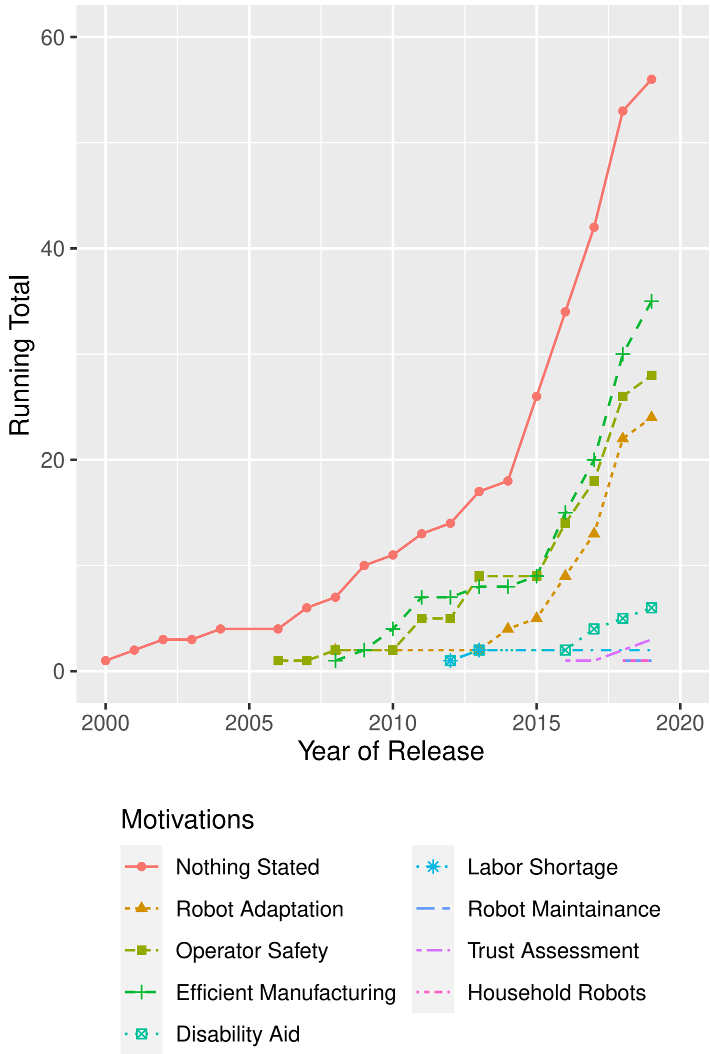


Fig. A.2: Accumulative total of research motivations in the papers.

their actions, both of which are consistent with other research by Giuliani and Knoll. They found that when the robot partner is assigned either supportive of instructive role, the operator will adopt the opposite role [13, 17]. Also, Laursen et al. [24] found that operator behaviour was affected by the robot's ability to provide negative feedback. While Fischer et al. [23] found that operators prefer robot with eyes utilize social gaze, Zheng et al. [26] found that having the robot shifting gaze between end goal and operator face during a handover task did not improve timing, but it was perceived as better communication. Lastly, Reinhardt et al. [36] found that operator trust towards

Table A.2: Frequency of research objectives, stated or inferred from the data collected.

Research Objectives	Freq.
Intention recognition [11, 15, 16, 22, 29, 31, 34, 35, 37, 38, 40–42, 46, 48, 53, 54, 56, 57, 64–66, 68, 69, 73, 74, 77, 80, 85, 88–90, 94, 97, 98, 102, 104–116, 118, 125]	51
Visual Tracking Test [10, 12, 19, 20, 22, 25, 29, 30, 32, 33, 36, 38, 40, 43–48, 56, 61, 63, 64, 67, 70, 74, 76, 77, 79, 83, 85, 91–93, 95–97, 99–101, 105, 108, 111, 112, 115, 117, 118, 123]	48
Motion planning [2, 6, 10, 12, 15, 31, 34, 35, 37, 41, 42, 48, 54, 56, 57, 63, 67, 73, 74, 76, 77, 88, 90, 91, 94, 96, 97, 99, 102, 105, 110, 115]	32
Collaboration planning [10, 33–35, 38–41, 43, 47, 48, 58, 65, 66, 68, 69, 71, 83, 84, 104, 105, 114, 125]	23
Collaborative heavy lifting [1–4, 6, 12, 14, 18, 22, 28, 47, 55, 56, 71, 79, 84, 104, 125]	18
Test control system [5, 8, 9, 27, 30, 50, 79, 111, 114]	9
Performance assessment [45, 51, 58, 62, 66, 67, 71, 78, 83]	9
Test usability [5, 49, 51, 70, 72, 82, 83, 97]	8
Operator safety [49, 58, 81, 93, 96, 100]	6
Test robot social behavior [13, 17, 23, 24, 26, 75]	6
Assess mental workload [49, 58–60, 66]	5
Test gesture recognition [25, 89, 90, 94]	4
Test speech recognition [69, 80, 102, 123]	4
Operator trust assessment [83, 120–122]	4
Robot design framework [52, 86, 87, 119]	4
Assess operator perception [28, 36, 37]	3
Test robot maintenance system [103, 124]	2
Test tactile system [7]	1
Test robot motion patterns [21]	1

the robot was enhanced if it used submissive motion cues, rather than dominant ones.

Lots of work have been done in measuring operator mental strain and perceived safety during SHRC, both of which are critical to successful HRC [75]. Several researchers found that low robot motion speed [59, 122] and non-straight motions [59] reduced mental strain and that the operator should be informed of the robot motion speed before the robot moves [59, 60]. The benefit of communication on percep-

Table A.3: Frequency of collaboration types used in the experiments in the papers.

Collaboration Type	Freq.
Simultaneous manipulation [1–7, 12, 18, 19, 22, 28, 30, 31, 33, 43, 44, 47, 50, 55, 56, 75, 79, 84, 100, 101, 104, 107, 109, 110, 114, 117, 125]	33
Cooperative assembly [10, 13, 15, 17, 23, 25, 35, 38, 45, 48, 51, 52, 58, 61–66, 81–83, 85–87, 90, 93, 97, 102, 106, 113, 120, 121]	33
Shared work space [13, 24, 34, 36–38, 42, 49, 59, 60, 67–70, 72–74, 76–78, 88, 91, 96, 99, 105, 112, 115, 122]	28
Handover [16, 20, 21, 26, 32, 40, 41, 46, 53, 54, 63, 65, 71, 80, 89, 94, 95, 98, 111, 118, 119, 123]	22
Collaborative transportation [2, 3, 6, 11, 12, 14, 28, 55, 92, 98, 104]	11
Assisted lifting [8, 43, 47, 79, 101, 116]	6
Simultaneous tool use [5, 27, 29, 104, 108, 124]	6
Assisted robot control [8, 9, 57]	3
Dressing the operator [39, 103]	2

tion of safety has also been shown using projection-based augmented reality overlays [97]. Also, Kato et al. [60] found that increasing operator-robot distance can help lower mental strain, while Bergman and van Zandbeek [122] found that short motion stopping distances to the operator would increase mental strain [122]. Despite other researchers finding that low motions speed and longer operator distances reduced mental strain, Reinhardt et al. [36] found that operator trust did not correlate with robot predictability.

4 Reflections

In this section we present our subjective reflection regarding the evolving trends in the field of SHRC based on our findings. Based on our review we want to determine how robot partners and their roles can be conceptualized based on their use in human-robot experiments, both as practical and social collaborator. Lastly, we present our prediction on how SHRC trends will evolve in the near future.

4.1 Research Trends

While the first occurrence of an SHRC experiment we found was in the year 2000, the first database result on HRI, HRC and teams was in 1992. During the filter process we found that HRI was often used to describe research into purely social robots, rather than manipulators. Manipulator meaning that robot was build to physically

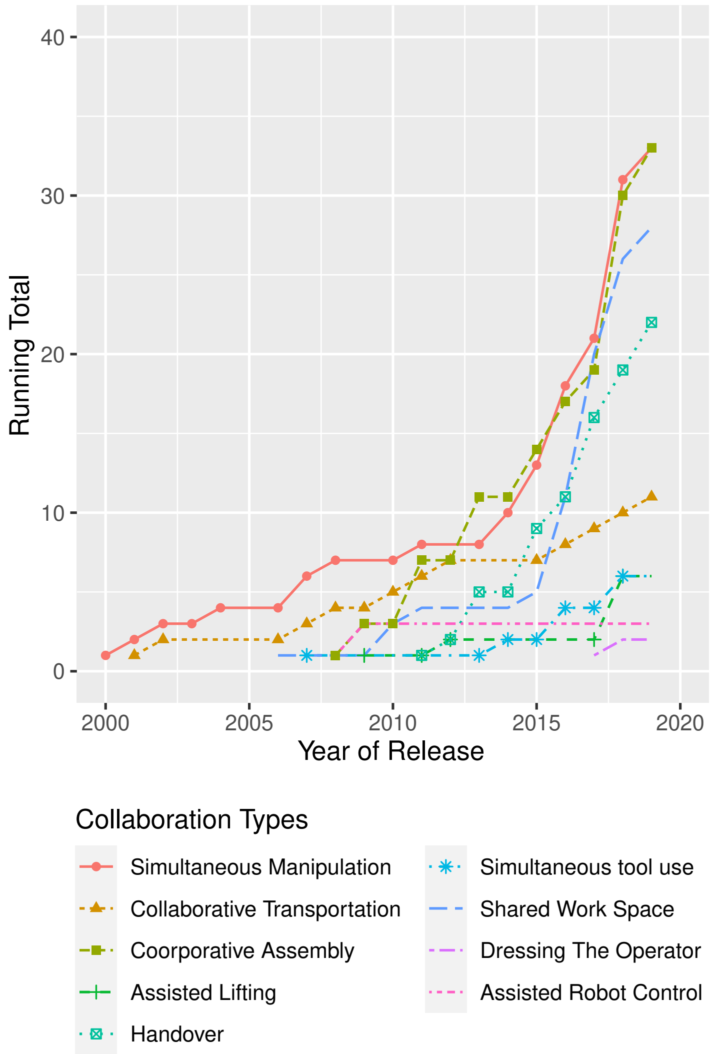


Fig. A.3: Accumulative total of collaboration definitions in the papers.

manipulate an object. This does not include robot dancing partners, which is also a widely researched topic. HRC often refers to remote control or collaboration, often developed for search and rescue operations in earthquake scenarios.

Looking at Figure A.1 we see that the earliest publications focused on the technical aspects, rather than assessing human factors. It was not until 2014 the number of human-focused papers started matching the number of technical papers. This suggests either an increase in interest in the one or two years leading up to this point or that the technological or ethical requirements for safe SHRC had been met to a wider

Table A.4: Frequency of technologies used in the designs or experiments in the papers. Hardware is listed at the top, software at the bottom.

Technology	Freq.
Force-torque sensors [1–6, 8, 9, 12, 14, 18, 27–31, 33, 41, 50, 54, 55, 63, 76, 79, 84, 91, 95, 101, 104, 107, 109, 114, 117, 124, 125]	35
Depth sensors [2, 12, 16, 25, 33, 35, 36, 48, 53, 56, 63, 64, 68, 71, 76, 77, 85, 92–97, 99, 102, 110, 117, 118, 123]	29
Marker tracking [10, 15, 20, 22, 29, 32, 37, 57, 67, 73, 74, 79, 83, 100, 101, 105, 108, 111, 112, 115]	20
RGB cameras [13, 17, 19, 30, 38, 43, 44, 46, 47, 58, 61, 62, 75, 91]	14
Physiological measurements [29, 40, 45, 59, 60, 89, 101, 102, 108]	9
Inertial sensors [34, 40, 48, 85, 90, 92]	6
Pressure sensors [78, 101, 116]	3
Projected Augmented Reality [49, 70, 72]	3
Robot-mounted controller [8, 9, 28]	3
Visual interfaces [82, 83]	2
Audio interfaces [82]	1
Tactile sensors [7]	1
Probabilistic state machine [3, 11, 16, 22, 30, 40, 42, 62, 64–66, 68, 72, 96, 98, 108–110]	18
Machine learning [15, 33, 39, 46, 51, 56, 85, 89, 99, 103, 113, 121, 124]	13
Speech recognition [13, 17, 41, 46, 57, 69, 80, 94, 102, 106, 123]	11
Gesture recognition [25, 90, 94]	3
Multi-robot coordination [6]	1
Face tracking [26]	1
Gaze tracking [69]	1

extend, making the technology more widely available for research.

Looking at the stated motivation in publications in Figure A.2, the three most stated are enabling HRC in manufacturing, enabling robot adaptation and ensuring operator safety. With the sudden increase in publications going from 2015 to 2016, it suggests a growing interesting in SHRC as part of industrial manufacturing in position where the operator's function cannot be fulfilled by a robot alone. However,

Table A.5: Frequency of robot types used in the experiments in the papers.

Robot Types	Freq.
Single-arm robot	77
[1, 3–11, 15, 16, 18, 21, 23, 25, 27–30, 32, 33, 36, 38, 42, 45, 51–53, 57–63, 65–70, 72–74, 76–79, 82, 86–97, 99, 101–104, 108–110, 112, 120–125]	
Dual-arm robot	38
[2, 12, 13, 17, 20, 26, 31, 34, 35, 39–41, 43, 46, 48–50, 54, 56, 64, 75, 80, 81, 83–85, 98, 100, 105–107, 111, 113–115, 117–119]	
Humanoid	6
[14, 22, 24, 37, 47, 55]	
Mobile robot	5
[2, 3, 6, 12, 65]	
Multiple robots	3
[6, 10, 84]	
Lift	2
[71, 116]	
Three-digit robot	2
[19, 44]	

standardized ergonomic scores have only been measured twice in SHRC experiments [81, 100], suggesting that the interest in operator safety has initially been in regards to avoiding physical hazards from the robot itself, rather than reducing health risks from working conditions. The two papers addressing labor shortage were not related to manufacturing, as may be expected, but rather to address shortages of nurses for handing over tools during surgery [94, 123]. Also starting around 2016 is a growing body of research into robots as aids to the disabled or citizens otherwise in need of assistance.

Very little research has been done in communication methods between the robot and the operator aside from the movement of the robot itself, with regular monitors only being used in two publications [82, 83] and augmented reality being used in three [49, 70, 72]. This suggests a significant gap in research of user-centered design concerning testing SHRC.

4.2 Collaboration in SHRC

Based on the collaboration and task types in papers, listed in Tables A.3 and A.6, we can categorize the conceptual levels of SHRC. Following are our categories based on the literature review.

No physical collaboration

In this category the robot and operator are not touching each other or a shared object at the same time. This type of collaboration uses the robot for communication only, such as in the case of pointing or holding a projector for augmented reality. The only example of a robot used for pointing is by Hoffman and Breazea [57].

Table A.6: Frequency of robot task types used in the experiments in the papers.

Task Types	Freq.
Pick and place [5, 11, 13, 23, 24, 34, 36–38, 42, 49, 63, 68–70, 72–74, 76, 78, 88, 97, 98, 103, 105, 115, 119]	27
Cooperative assembly [10, 15, 17, 25, 32, 35, 45, 48, 51, 52, 58, 61, 62, 66, 81–83, 86, 87, 90, 93, 98, 102, 113, 120, 121]	26
Handover [11, 16, 20, 21, 26, 40, 41, 43, 46, 53, 54, 63, 65, 71, 80, 89, 95, 98, 106, 111, 118, 119, 123]	23
Object positioning [1–4, 6, 8, 12, 18, 28, 31, 34, 43, 79, 91, 104, 107, 109, 111, 112, 114, 125]	21
Manipulation [7, 30, 33, 39, 43, 47, 50, 56, 70, 75, 84, 100, 110, 117]	14
Object transportation [2, 3, 6, 12, 14, 28, 55, 91, 104, 109]	10
Holding [18, 19, 22, 44, 79, 92, 94, 101, 106, 111]	10
Tool use [27, 29, 104, 108, 124]	5
Lifting [56, 116]	2
Move to position [77, 122]	2
Drawing [5]	1
Pointing [57]	1

Turn-taking

In this category the robot and operator physically interact with a shared object, but never at the same time. Examples of this are solo object positioning tasks and pick-and-place tasks as listed in Table A.6.

One passive collaborator

The next level is when the robot and operator have physical contact, either directly or through a shared object, but one or the other remains mainly stationary during the interaction. An examples of this is the type collaborative assembly where the robot hold the object while the operator performs the assembly task. The handover task is somewhere between this level and the next, depending on the implementation, as the giver often stationary while waiting for the receiver to take the object.

Table A.7: Frequency of operator-robot proximity in the experiments in the papers.

Operator-Robot Proximity	Freq.
Table distance [11, 13, 16, 17, 21, 23–26, 34–38, 41, 45, 46, 48, 49, 52, 54, 58–63, 65, 66, 68–70, 72–75, 80–82, 85, 87, 88, 90, 93, 94, 96–99, 102, 105, 110, 112, 113, 115, 118, 119, 121, 123, 125]	60
Holding small object [2, 5, 10, 15, 18, 20, 29, 32, 40, 43, 44, 53, 56, 57, 67, 77, 83, 86, 89, 91, 92, 95, 100, 103, 104, 109, 111, 120, 122]	29
Holding large object [1, 3, 4, 6, 7, 12, 14, 19, 22, 30, 31, 33, 47, 50, 55, 71, 101, 104, 107, 108, 114, 117]	22
Touching [5, 8, 9, 27, 28, 76, 79, 116, 124]	9
Side-by-side [42, 51, 84, 106]	4
Beyond table [58]	1
Operator-mounted [78]	1

Concurrent collaboration

The majority of paper in the literature review focus on concurrent collaboration, wherein the robot and operator touch each other or a shared object at some point in the task. These can be simultaneous manipulation, object positioning, transportation and assisted lifting among others. Handover tasks can be considered concurrent collaborations in the cases where the robot acts proactively according to the operator's actions to improve the interaction.

4.3 Robot Collaborator Roles

When planning the filtering process we already limited our definition of SHRC to require shared space with no dividing barriers between operator and robot and manipulating a shared object. Still, reading through the papers fitting this description, robots have been put in different roles in relation to the operator. Following is our outline of these roles.

Robot as a smart tool

This type of collaboration often involves assisted lifting of heavy objects, but can also be the robot equipped with a power tool. This kind of collaboration is characterized by moving the robot end-effector physically while the robot follows the motion or adjusts to optimize the process. Papers featuring collaborative tool use and heavy lifting are listed in Table A.6.

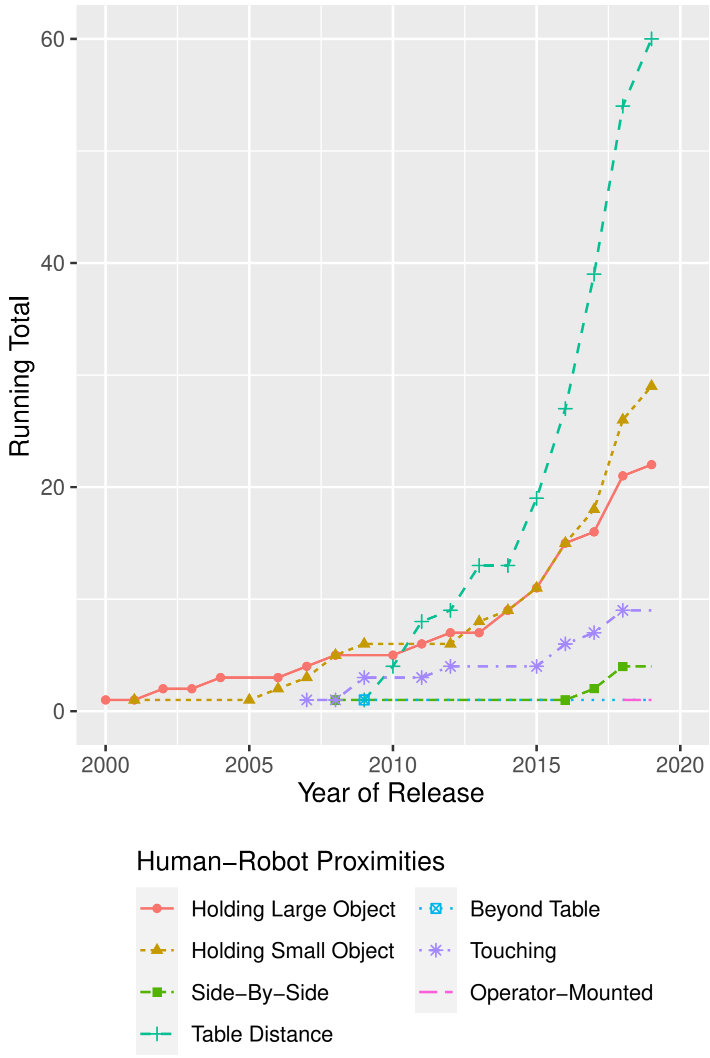


Fig. A.4: Accumulative total of human-robot proximities in the papers.

Robot as a robot

In this case the robot works and reacts according to the pre-programmed process, objects and environment, rather than reacting to the operator. At most, the robot will adjust movement to avoid collisions. This can involve pick-and-place operations at pre-determined positions or picking up and holding an object so the operator can perform assembly before the robot takes the object away.

Table A.8: Frequency of data collected in the experiments in the papers. Objective measurement are the top, subjective measures at the bottom.

Types of Data Collected	Freq.
Error rate/count [5, 11, 14, 27, 32–34, 38, 40, 46, 47, 50, 51, 55, 62, 63, 65, 67, 69, 70, 72, 76, 77, 79, 80, 84, 89, 94–96, 104, 105, 109–111, 113, 117, 118]	38
Observations [1–4, 6, 10, 12, 19, 20, 23, 26, 36, 43–45, 48, 53, 56, 61, 68, 85, 86, 92, 99, 105, 106, 111, 114, 119, 124, 125]	31
Sensor metrics [2, 3, 5, 6, 8–10, 14, 18, 19, 22, 28–31, 40, 45, 55, 61, 89, 91, 92, 101, 104, 107–109, 116, 117, 124, 125]	31
Task completion time [4, 5, 7, 12, 13, 16, 21, 37, 41, 45, 51, 54, 57, 58, 62, 65–68, 71, 77, 78, 83, 93, 94, 103, 106, 110]	28
State-machine probability metrics [11, 20, 22, 30, 36, 46, 48, 59, 64, 67, 69, 74, 77, 80, 90, 94, 96, 98, 111]	19
Machine learning metrics [15, 39, 42, 54, 56, 85, 99, 102–104, 113–115]	13
Task step count [13, 17, 68, 72, 123]	5
Physiological measurements [58–60, 69]	4
Operator intervention count [24, 49, 88, 121]	4
Command count [13, 17]	2
Ergonomics score [81, 100]	2
Operator gaze duration [26]	1
Operator reaction time [24]	1
Robot-operator proximity [110]	1
Misc. questionnaires [4, 17, 21, 23, 25, 26, 28, 35–37, 49, 52, 54, 65, 66, 68–70, 72, 73, 75, 82, 87, 97, 100, 102, 110, 112, 120, 122]	30
Stress assessment [49, 52, 53, 58, 66, 83]	6
Trust score [21, 83, 121]	3
Usability score [49, 83]	2

Robot as assistant

As opposed to the robot as a robot, what can lift the robot to the level of assistance is enabling movement adjustment to react directly to the operator's actions. This requires that robot can track the actions of the operator as well as keep track of the shared objective, allowing for operator intention recognition. An example of this the proactive robot tested by Han and Yanco [54].

Robot as collaborator

For the robot to considered a collaborator it must have the same capabilities of the robot assistant while also enabling proactive planning and actions, as opposed to purely reactive actions. This often creates an overlap of HRC and social robotics, as dominant and submissive roles can occur between the robot and operator as researched by Giuliani and Knoll [13] and Gombolay et al. [65].

4.4 Future Trends

With no sign of a plateau in the growth of the body of research relevant to SHRC in manufacturing, this is likely to continue in the foreseeable future. It is, however, likely that there will be an increased focus in operator ergonomics in order to limit long-term injuries as well as focus on visual communication, using augmented reality, virtual reality, monitors or others, due to comparably small body of research. Also, progress in SHRC in manufacturing along with growing interest in HRC for disabled citizens, applying the knowledge from the former is likely to results in growing body of research in the latter.

References

[1] T. Takubo, H. Arai, and K. Tanie, "Virtual nonholonomic constraint for human-robot cooperation in 3-D space," *IEEE International Conference on Intelligent Robots and Systems*, vol. 1, pp. 300–305, 2000.

[2] K. Kosuge, H. Kakuya, and Y. Hirata, "Control algorithm of dual arms mobile robot for cooperative works with human," *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 5, pp. 3223–3228, 2001.

[3] T. T. and A. H. andTanie K., "Control of mobile manipulator using a virtual impedance wall," *Proceedings 2002 IEEE International Conference on Robotics and Automation*, pp. 3571–3576, 2002.

[4] S. Yigit, C. Burghart, and H. Wörn, "Co-operative carrying using pump-like constraints," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 4, pp. 3877–3882, 2004.

[5] V. Duchaine and C. M. Gosselin, "General model of human-robot cooperation using a novel velocity based variable impedance control," *Proceedings - Second Joint EuroHaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, World Haptics 2007*, pp. 445–451, 2007.

[6] Y. Hirata, Y. Matsuda, and K. Kosuge, "Handling of an object in 3-D space by multiple mobile manipulators based on intentional force/moment applied by human," *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, 2007.

[7] K. Suwanratchatamanee, M. Matsumoto, and S. Hashimoto, “Human-machine interaction through object using robot arm with tactile sensors,” *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN*, pp. 683–688, 2008.

[8] T. Tsumugiwa and R. Yokogawa, “Object inherent dynamics based motion control in human-robot cooperative task system,” *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3377–3382, 2009.

[9] T. Tsumugiwa, Y. Watanabe, and R. Yokogawa, “Robot motion control using mechanical load adjuster with motion measurement interface for human-robot cooperation,” *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, pp. 467–472, 2009.

[10] P. Gil, J. Pomares, S. Puente, F. A. Candelas, G. J. García, J. A. Corrales, and F. Torres, “A cooperative robotic system based on multiple sensors to construct metallic structures,” *International Journal of Advanced Manufacturing Technology*, vol. 45, no. 5-6, pp. 616–630, 2009.

[11] M. Awais and D. Henrich, “Human-robot collaboration by intention recognition using probabilistic state machines,” in *19th International Workshop on Robotics in Alpe-Adria-Danube Region (RAAD 2010)*. IEEE, jun 2010, pp. 75–80. [Online]. Available: http://ieeexplore.ieee.org/document/5524605/

[12] J. Stückler and S. Behnke, “Following human guidance to cooperatively carry a large object,” *IEEE-RAS International Conference on Humanoid Robots*, pp. 218–223, 2011.

[13] M. Giuliani and A. Knoll, “Evaluating supportive and instructive robot roles in human-robot interaction,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7072 LNAI, pp. 193–203, 2011.

[14] P. Evrard and A. Kheddar, “Human-Humanoid Co-working in a Joint Table Transportation,” *2012 International Conference on Social Robotics*, pp. 357–366, 2012.

[15] S. Nikolaidis, P. Lasota, G. Rossano, C. Martinez, T. Fuhlbrigge, and J. Shah, “Human-robot collaboration in manufacturing: Quantitative evaluation of predictable, convergent joint action,” *2013 44th International Symposium on Robotics, ISR 2013*, 2013.

[16] K. P. Hawkins, Nam Vo, S. Bansal, and A. F. Bobick, “Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration,” in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*. IEEE, oct 2013, pp. 499–506. [Online]. Available: http://ieeexplore.ieee.org/document/7030020/

[17] M. Giuliani and A. Knoll, “Using Embodied Multimodal Fusion to Perform Supportive and Instructive Robot Roles in Human-Robot Interaction,” *International Journal of Social Robotics*, vol. 5, no. 3, pp. 345–356, 2013.

[18] M. Gil, S. Lee, M. S. Kang, K. Shin, and C. S. Han, “Sensor device for intuitively manipulating the heavy material-handling robot,” *4th Annual IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2014*, pp. 527–530, 2014.

[19] Y. Yamakawa, K. Kuno, and M. Ishikawa, "Human-robot cooperative task realization using high-speed robot hand system," *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, vol. 2015-August, pp. 281–286, 2015.

[20] A. M. Bestick, S. A. Burden, G. Willits, N. Naikal, S. S. Sastry, and R. Bajcsy, "Personalized kinematics for human-robot collaborative manipulation," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-December, pp. 1037–1044, 2015.

[21] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[22] W. Sheng, A. Thobbi, and Y. Gu, "An Integrated Framework for Human–Robot Collaborative Manipulation," *IEEE transactions on cybernetics*, vol. 45, no. 10, pp. 2030–2041, 2015.

[23] K. Fischer, L. C. Jensen, F. Kirstein, S. Stabinger, Ö. Erkent, D. Shukla, and J. Piater, "The Effects of Social Gaze in Human-Robot Collaborative Assembly," *International Conference on Social Robotics*, vol. 1, pp. 61–71, 2015.

[24] C. Ø. Laursen, S. Pedersen, T. M. B, and O. Caprani, "The Positive Effect of Negative Feedback in HRI Using a Facial Expression Robot," *International Workshop on Cultural Robotics*, vol. 9549, no. January 2017, pp. 107–130, 2015. [Online]. Available: http://link.springer.com/10.1007/978-3-319-42945-8

[25] M. Wongphati, H. Osawa, and M. Imai, "Gestures for Manually Controlling a Helping Hand Robot," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 731–742, 2015.

[26] M. Zheng, A. J. Moon, E. A. Croft, and M. Q. Meng, "Impacts of Robot Head Gaze on Robot-to-Human Handovers," *International Journal of Social Robotics*, vol. 7, no. 5, pp. 783–798, 2015.

[27] H. Bo, D. M. Mohan, M. Azhar, K. Sreekanth, and D. Campolo, "Human-robot collaboration for tooling path guidance," *Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics*, vol. 2016-July, pp. 1340–1345, 2016.

[28] J. Schmidtler and K. Bengler, "Size-weight illusion in human-robot collaboration," *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 874–879, 2016.

[29] L. Peternel, N. Tsagarakis, and A. Ajoudani, "Towards multi-modal intention interfaces for human-robot co-manipulation," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-November, pp. 2663–2669, 2016.

[30] A. Ghadirzadeh, J. Bütepage, A. Maki, D. Kragic, and M. Björkman, "A sensorimotor reinforcement learning framework for physical human-robot interaction," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-November, pp. 2682–2688, 2016.

[31] B. Nemec, N. Likar, A. Gams, and A. Ude, “Bimanual human robot cooperation with adaptive stiffness control,” *IEEE-RAS International Conference on Humanoid Robots*, pp. 607–613, 2016.

[32] J. Hwang, M. Yang, I. H. Suh, and T. Kwon, “Real-time grasp planning based on motion field graph for human-robot cooperation,” *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-November, pp. 1025–1032, 2016.

[33] M. P. Sbanca and G. L. Mogan, “Human Robot Collaboration for Folding Fabrics Based on Force/RGB-D Feedback,” *Advances in Intelligent Systems and Computing*, vol. 371, pp. 47–57, 2016. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84951320510{&}partnerID=tZOtx3y1

[34] R. Graham, J. Das, and D. Lucas, “On Planning and Task Achievement Modalities for Human-Robot Collaboration,” *Experimental Robotics*, vol. 109, pp. 819–839, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-319-23778-7

[35] R. Schulz, P. Kratzer, and M. Toussaint, “Building a bridge with a robot: A system for collaborative on-table task execution,” *HAI 2017 - Proceedings of the 5th International Conference on Human Agent Interaction*, pp. 399–403, 2017.

[36] J. Reinhardt, A. Pereira, D. Beckert, and K. Bengler, “Dominance and movement cues of robot motion: A user study on trust and predictability,” *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, vol. 2017-January, pp. 1493–1498, 2017.

[37] F. Vannucci, A. Sciutti, M. Jacono, G. Sandini, and F. Rea, “Adaptation to a humanoid robot in a collaborative joint task,” *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, vol. 2017-January, pp. 360–365, 2017.

[38] D. Riedelbauch and D. Henrich, “Coordinating flexible human-robot teams by local world state observation,” *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, vol. 2017-January, pp. 1000–1005, 2017.

[39] R. Silva, M. Faria, F. S. Melo, and M. Veloso, “Adaptive indirect control through communication in collaborative human-robot interaction,” *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-September, pp. 3617–3622, 2017.

[40] L. Chen, H. Wu, S. Duan, Y. Guan, and J. Rojas, “Learning human-robot collaboration insights through the integration of muscle activity in interaction motion models,” *IEEE-RAS International Conference on Humanoid Robots*, pp. 491–496, 2017.

[41] A. Roncone, O. Mangin, and B. Scassellati, “Transparent role assignment and task allocation in human robot collaboration,” *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1014–1021, 2017.

[42] S. Nikolaidis, S. Nath, A. D. Procaccia, and S. Srinivasa, “Game-Theoretic Modeling of Human Adaptation in Human-Robot Collaboration,” *ACM/IEEE International Conference on Human-Robot Interaction*, vol. Part F127194, pp. 323–331, 2017.

[43] K. Murakami, S. Huang, H. Sumi, M. Ishikawa, and Y. Yamakawa, "Towel-Like Object Alignment with Human-Robot Cooperation and High-Speed Robotic Manipulation," *2018 IEEE International Conference on Robotics and Biomimetics, ROBIO 2018*, pp. 772–777, 2018.

[44] Y. Yamakawa, Y. Matsui, and M. Ishikawa, "Development and Analysis of a High-speed Human-Robot Collaborative System and its Application," *2018 IEEE International Conference on Robotics and Biomimetics, ROBIO 2018*, pp. 2415–2420, 2018.

[45] M. Coban and G. Gelen, "Realization of human-robot collaboration in hybrid assembly systems by using wearable technology," *2018 6th International Conference on Control Engineering and Information Technology, CEIT 2018*, pp. 1–6, 2018.

[46] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, "Interactive Text2Pickup Networks for Natural Language-Based Human–Robot Collaboration," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3308–3315, 2018.

[47] M. Murooka, S. Nozawa, M. Bando, I. Yanokura, K. Okada, and M. Inaba, "Simultaneous Planning and Estimation Based on Physics Reasoning in Robot Manipulation," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3137–3144, 2018.

[48] K. Darvish, B. Bruno, E. Simetti, F. Mastrogiovanni, and G. Casalino, "Interleaved Online Task Planning, Simulation, Task Allocation and Motion Control for Flexible Human-Robot Cooperation," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 58–65, 2018.

[49] Z. Materna, M. Kapinus, V. Beran, P. Smrž, and P. Zemčík, "Interactive Spatial Augmented Reality in Collaborative Robot Programming: User Experience Evaluation," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 80–87, 2018.

[50] B. Nemec, N. Likar, A. Gams, and A. Ude, "Adaptive Human Robot Cooperation Scheme for Bimanual Robots," *Advances in Robot Kinematics 2016*, pp. 371–380, 2018.

[51] R. Kalpagam Ganesan, Y. K. Rathore, H. M. Ross, and H. Ben Amor, "Better Teaming Through Visual Cues: How Projecting Imagery in a Workspace Can Improve Human-Robot Collaboration," *IEEE Robotics and Automation Magazine*, vol. 25, no. 2, pp. 59–71, 2018.

[52] V. Lipovaya, F. Duarte, and F. Lima, *Evaluation of Human-Robot Interaction Quality: A Toolkit for Workplace Design*. Springer International Publishing, 2018. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-96071-5{_}153

[53] D. Vogt, S. Stepputtis, B. Jung, and H. B. Amor, "One-shot learning of human–robot handovers with triadic interaction meshes," *Autonomous Robots*, vol. 42, no. 5, pp. 1053–1065, 2018. [Online]. Available: https://doi.org/10.1007/s10514-018-9699-4

[54] Z. Han and H. Yanco, "The Effects of Proactive Release Behaviors during Human-Robot Handovers," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2019-March, pp. 440–448, 2019.

[55] D. J. Agravante, A. Cherubini, A. Sherikov, P. B. Wieber, and A. Kheddar, "Human-Humanoid Collaborative Carrying," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 833–846, 2019.

[56] Y. Cui, J. Poon, J. V. Miro, K. Yamazaki, K. Sugimoto, and T. Matsubara, "Environment-adaptive interaction primitives through visual context for human–robot motor skill learning," *Autonomous Robots*, vol. 43, no. 5, pp. 1225–1240, 2019. [Online]. Available: https://doi.org/10.1007/s10514-018-9798-2

[57] G. Hoffman and C. Breazeal, "Achieving fluency through perceptual-symbol practice in human-robot collaboration," *HRI 2008 - Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction: Living with Robots*, pp. 1–8, 2008.

[58] J. T. C. Tan, F. Duan, Y. Zhang, K. Watanabe, R. Kato, and T. Arai, "Human-robot collaboration in cellular manufacturing: Design and development," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, oct 2009, pp. 29–34. [Online]. Available: http://ieeexplore.ieee.org/document/5354155/

[59] M. Fujita, R. Kato, and A. Tamio, "Assessment of operators' mental strain induced by hand-over motion of industrial robot manipulator," *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 361–366, 2010.

[60] R. Kato, M. Fujita, and T. Arai, "Development of advanced cellular manufacturing system with human-robot collaboration-assessment of mental strain on human operators induced by the assembly support," *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, pp. 355–360, 2010.

[61] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, may 2011, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/5942335/

[62] F. Duan, J. T. C. Tan, and T. Arai, "A new human-robot collaboration assembly system for cellular manufacturing," *Proceedings of the 30th Chinese Control Conference, CCC 2011*, pp. 5468–5473, 2011.

[63] S. Haddadin, "Towards the Robotic Co-Worker," *Springer Tracts in Advanced Robotics*, vol. 90, pp. 195–215, 2011.

[64] H. Ding, M. Schipper, and B. Matthias, "Collaborative behavior design of industrial robots for multiple human-robot collaboration," *2013 44th International Symposium on Robotics, ISR 2013*, vol. 49, no. 0, pp. 1–6, 2013.

[65] M. C. Gombolay, R. A. Gutierrez, S. G. Clarke, G. F. Sturla, and J. A. Shah, "Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams," *Autonomous Robots*, vol. 39, no. 3, pp. 293–312, 2015.

[66] S. M. Rahman, Z. Liao, L. Jiang, and Y. Wang, "A regret-based autonomy allocation scheme for human-robot shared vision systems in collaborative assembly

in manufacturing," *IEEE International Conference on Automation Science and Engineering*, vol. 2016-November, no. Cmmi, pp. 897–902, 2016.

[67] R. Hayne, R. Luo, and D. Berenson, "Considering avoidance and consistency in motion planning for human-robot manipulation in a shared workspace," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 3948–3954, 2016.

[68] J. Baraglia, M. Cakmak, Y. Nagai, R. Rao, and M. Asada, "Initiative in robot assistance during collaborative task execution," *ACM/IEEE International Conference on Human-Robot Interaction*, vol. 2016-April, pp. 67–74, 2016.

[69] C.-M. Huang and B. Mutlu, *Anticipatory Robot Control for Efficient Human-Robot Collaboration*, 2016. [Online]. Available: http://www.eyetracking-glasses.com

[70] R. S. Andersen, O. Madsen, T. B. Moeslund, and H. B. Amor, "Projecting robot intentions into human environments," *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pp. 294–301, 2016.

[71] J. Kinugawa, A. Kanazawa, and K. Kosuge, "B-PaDY: robot co-worker in a bumper assembly line," *ROBOMECH Journal*, vol. 3, no. 1, 2016.

[72] H. Zhu, V. Gabler, and D. Wollherr, "Legible action selection in human-robot collaboration," *2017 26th IEEE International Symposium on Robot and Human Interactive Communication*, 2017.

[73] V. Gabler, T. Stahl, G. Huber, O. Oguz, and D. Wollherr, "A game-theoretic approach for adaptive action selection in close proximity human-robot-collaboration," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2897–2903, 2017.

[74] T. Arne, F. Py, J. Pinto, M. A. Silva, T. A. Johansen, J. Sousa, and K. Rajan, "Hybrid Human Motion Prediction for Action Selection Within Human-Robot Collaboration," *2016 International Symposium on Experimental Robotics*, 2017.

[75] S. R. Fletcher and P. Webb, "Leader-Follower Strategies for Robot-Human Collaboration," *A World with Robots*, vol. 84, pp. 159–169, 2017. [Online]. Available: http://link.springer.com/10.1007/978-3-319-46667-5

[76] L. Gracia, C. Perez-Vidal, D. Mronga, J. M. de Paco, J. M. Azorin, and J. de Gea, "Robotic manipulation for the shoe-packaging process," *International Journal of Advanced Manufacturing Technology*, vol. 92, no. 1-4, pp. 1053–1067, 2017.

[77] X. Zhao and J. Pan, "Considering Human Behavior in Motion Planning for Smooth Human-Robot Collaboration in Close Proximity," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 985–990, 2018.

[78] K. Nakabayashi, Y. Iwasaki, S. Takahashi, and H. Iwata, "Experimental Evaluation of Cooperativeness and Collision Safety of a Wearable Robot Arm," *RO-MAN 2018 - 27th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1026–1031, 2018.

[79] L. Roveda, N. Castaman, S. Ghidoni, P. Franceschi, N. Boscolo, E. Pagello, and N. Pedrocchi, "Human-Robot Cooperative Interaction Control for the Installation of Heavy and Bulky Components," *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pp. 339–344, 2018.

[80] J. Brawer, O. Mangin, A. Roncone, S. Widder, and B. Scassellati, "Situated Human–Robot Collaboration: predicting intent from grounded natural language," *IEEE International Conference on Intelligent Robots and Systems*, pp. 827–833, 2018.

[81] B. Busch, M. Toussaint, and M. Lopes, "Planning Ergonomic Sequences of Actions in Human-Robot Interaction," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 1916–1923, 2018.

[82] J. Höcherl, M. Schmargendorf, B. Wrede, and T. Schlegl, "User-Centered Design of Multimodal Robot Feedback for Cobots of Human-RobotWorking Cells in Industrial Production Contexts," *50th International Symposium on Robotics, ISR 2018*, pp. 442–449, 2018.

[83] B. Sadrfaridpour and Y. Wang, "Collaborative Assembly in Hybrid Manufacturing Cells: An Integrated Framework for Human–Robot Interaction," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 3, pp. 1178–1192, 2018.

[84] Y. C. Peng, D. S. Carabis, and J. T. Wen, "Collaborative manipulation with multiple dual-arm robots under human guidance," *International Journal of Intelligent Robotics and Applications*, vol. 2, no. 2, pp. 252–266, 2018. [Online]. Available: https://doi.org/10.1007/s41315-018-0053-y

[85] J.-j. Cabibihan, R. Chellali, C. Wing, and C. So, *Predicting the Target in Human-Robot Manipulation Tasks*. Springer International Publishing, 2018, vol. 1. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-05204-1{_}34

[86] M. Winckler and K. Kuusinen, *Towards Tool-Support for Robot-Assisted Product Creation in Fab Labs*. Springer International Publishing, 2018, vol. 1. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-05909-5{_}14

[87] D. Schulze Kissing, C. Bruder, N. Carstengerdes, and A. Papenfuß, *Comfort Design in Human Robot Cooperative Tasks*. Springer International Publishing, 2019. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-02053-8{_}79

[88] G. Bartels, D. Beßler, and M. Beetz, "Episodic Memories for Safety-Aware Robots," *KI - Künstliche Intelligenz*, vol. 33, no. 2, pp. 123–130, 2019. [Online]. Available: https://doi.org/10.1007/s13218-019-00584-3

[89] W. Wang, R. Li, Z. M. Diekel, Y. Chen, Z. Zhang, and Y. Jia, "Controlling Object Hand-Over in Human–Robot Collaboration Via Natural Wearable Sensing," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 1, pp. 59–71, 2019.

[90] P. Neto, M. Simão, N. Mendes, and M. Safeea, "Gesture-based human-robot interaction for human assistance in manufacturing," *International Journal of Advanced Manufacturing Technology*, vol. 101, no. 1-4, pp. 119–135, 2019.

[91] A. &. C. Benites, "Velocity control for safe robot guidance based on fused vision and force/torque data," *Trabajo Infantil*, vol. 53, no. 9, pp. 1689–1699, 2006.

[92] J. A. Corrales, F. A. Candelas, and F. Torres, "Hybrid tracking of human operators using IMU/UWB data fusion by a Kalman filter," *HRI 2008 - Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction: Living with Robots*, pp. 193–200, 2008.

[93] J. J. Hu, C. N. Huang, H. W. Wang, P. H. Shieh, and J. S. Hu, "Safety-based human-robot collaboration in cellular manufacturing: A case study of power protector assembly," *2013 International Conference on Advanced Robotics and Intelligent Systems, ARIS 2013 - Conference Proceedings*, pp. 28–31, 2013.

[94] M. G. Jacob, Y. T. Li, and J. P. Wachs, "Surgical instrument handling and retrieval in the operating room with a multimodal robotic assistant," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2140–2145, 2013.

[95] M. Bdiwi, A. Kolker, J. Suchý, and A. Winkler, "Automated Assistance Robot System for Transferring Model-Free Objects From/To Human Hand Using Vision/Force Control," *International Conference on Social Robotics 2013*, pp. 40–53, 2013.

[96] F. Dittrich and H. Woern, "Robot activity adaptation for safe human-robot collaboration based on probabilistic risk modeling," *Proceedings of the 18th Mediterranean Electrotechnical Conference: Intelligent and Efficient Technologies and Services for the Citizen, MELECON 2016*, pp. 1–6, 2016.

[97] D. Liu, J. Kinugawa, and K. Kosuge, "A projection-based making-human-feel-safe system for human-robot cooperation," *2016 IEEE International Conference on Mechatronics and Automation, IEEE ICMA 2016*, pp. 1101–1106, 2016.

[98] A. M. Zanchettin and P. Rocco, "Probabilistic inference of human arm reaching target for effective human-robot collaboration," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-September, pp. 6595–6600, 2017.

[99] Y. Wang, X. Ye, Y. Yang, and W. Zhang, "Collision-free trajectory planning in human-robot interaction through hand movement prediction from vision," *IEEE-RAS International Conference on Humanoid Robots*, pp. 305–310, 2017.

[100] B. Busch, G. Maeda, Y. Mollard, M. Demangeat, and M. Lopes, "Postural optimization for an ergonomic human-robot interaction," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2017-September, pp. 2778–2785, 2017.

[101] W. Kim, J. Lee, L. Peternel, N. Tsagarakis, and A. Ajoudani, "Anticipatory Robot Assistance for the Prevention of Human Static Joint Overloading in Human–Robot Collaboration," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 68–75, 2018.

[102] W. Wang, R. Li, Y. Chen, Z. M. Diekel, and Y. Jia, "Facilitating Human–Robot Collaborative Tasks by Teaching-Learning-Collaboration From Human Demonstrations," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 640–653, 2019.

[103] G. Canal, G. Alenyà, and C. Torras, "Adapting robot task planning to user preferences: an assistive shoe dressing example," *Autonomous Robots*, vol. 43, no. 6, pp. 1343–1356, 2018. [Online]. Available: https://doi.org/10.1007/s10514-018-9737-2

[104] M. Khoramshahi and A. Billard, "A dynamical system approach to task-adaptation in physical human–robot interaction," *Autonomous Robots*, vol. 43, no. 4, pp. 927–946, 2018. [Online]. Available: https://doi.org/10.1007/s10514-018-9764-z

[105] S. Choi, K. Lee, H. A. Park, and S. Oh, "A Nonparametric Motion Flow Model for Human Robot Cooperation," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 7211–7218, 2018.

[106] P. F. Dominey, G. Metta, F. Nori, and L. Natale, "Anticipation and initiative in human-humanoid interaction," *2008 8th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2008*, pp. 693–699, 2008.

[107] B. Nemec, A. Gams, M. Deniša, and A. Ude, "Human-robot cooperation through force adaptation using dynamic motion primitives and iterative learning," *2014 IEEE International Conference on Robotics and Biomimetics, IEEE ROBIO 2014*, pp. 1439–1444, 2014.

[108] L. Peternel, T. Petrič, E. Oztop, and J. Babič, "Teaching robots to cooperate with humans in dynamic manipulation tasks based on multi-modal human-in-the-loop approach," *Autonomous Robots*, vol. 36, no. 1-2, pp. 123–136, 2014.

[109] B. Whitsell and P. Artemiadis, "On the role duality and switching in human-robot cooperation: An adaptive approach," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, no. June, pp. 3770–3775, 2015.

[110] S. Pellegrinelli, H. Admoni, S. Javdani, and S. Srinivasa, "Human-robot shared workspace collaboration via hindsight optimization," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-November, pp. 831–838, 2016.

[111] G. J. Maeda, G. Neumann, M. Ewerton, R. Lioutikov, O. Kroemer, and J. Peters, "Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks," *Autonomous Robots*, vol. 41, no. 3, pp. 593–612, 2017.

[112] M. Lee Chang, R. A. Gutierrez, P. Khante, E. Schaertl Short, and A. Lockerd Thomaz, "Effects of Integrated Intent Recognition and Communication on Human-Robot Collaboration," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3381–3386, 2018.

[113] E. C. Grigore, A. Roncone, O. Mangin, and B. Scassellati, "Preference-Based Assistance Prediction for Human-Robot Collaboration Tasks," *IEEE International Conference on Intelligent Robots and Systems*, pp. 4441–4448, 2018.

[114] B. Nemec, N. Likar, A. Gams, and A. Ude, "Human robot cooperation with compliance adaptation along the motion trajectory," *Autonomous Robots*, vol. 42, no. 5, pp. 1023–1035, 2018.

[115] R. Luo, R. Hayne, and D. Berenson, "Unsupervised early prediction of human reaching for human–robot collaboration in shared workspaces," *Autonomous Robots*, vol. 42, no. 3, pp. 631–648, 2018.

[116] H. Nozawa, S. Lee, J. I. Kim, and K. Inoue, "Human-robot cooperation for heavy object manipulation based on sensing pneumatic pressure in air cushions," *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, pp. 391–396, 2012.

[117] K. Saxena, R. Labuguen, R. P. Joshi, N. Koganti, and T. Shibata, "A study on human-robot collaboration for table-setting task," *2017 IEEE International Conference on Robotics and Biomimetics, ROBIO 2017*, vol. 2018-January, pp. 183–188, 2017.

[118] G. W. B, L. Ren, and J. S. Dai, "Intention-Based Human Robot Collaboration," *International Conference on Intelligent Robotics and Applications*, vol. 1, no. October, pp. 71–83, 2017.

[119] L. Ciuccarelli, A. Freddi, S. Iarlori, S. Longhi, A. Monteriù, D. Ortenzi, and D. Proietti Pagnotta, *Architecture for cooperative interacting robotic systems towards assisted living: A preliminary study*. Springer International Publishing, 2019, vol. 544. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-05921-7{_}38

[120] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, 2016.

[121] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with Trust for Human-Robot Collaboration," *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 307–315, 2018.

[122] M. Bergman and M. van Zandbeek, "Close Encounters of the Fifth Kind? Affective Impact of Speed and Distance of a Collaborative Industrial Robot on Humans," *Human Friendly Robotics*, pp. 127–137, 2019.

[123] J. P. Wachs, "Robot, pass me the scissors! How robots can assist us in the operating room," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7441 LNCS, pp. 46–57, 2012.

[124] L. Peternel, N. Tsagarakis, and A. Ajoudani, "A Method for Robot Motor Fatigue Management in Physical Interaction and Human-Robot Collaboration Tasks," *IEEE International Conference on Intelligent Robots and Systems*, pp. 2850–2856, 2018.

[125] M. Cevzar, T. Petrič, and J. Babič, "Shared Control for Human-Robot Cooperative Manipulation Tasks," *Mechanisms and Machine Science*, vol. 49, no. June, pp. 797–804, 2018. [Online]. Available: http://link.springer.com/10.1007/978-3-319-61276-8

Paper B

Augmented Reality Technology for Displaying Close-Proximity Sub-Surface Positions

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper has been published in the
IFIP Conference on Human-Computer Interaction – INTERACT 2019 Vol. 11747,
pp. 641–659, 2019.

Abstract

When designing human-system collaboration to assist in strenuous manual tasks we need to develop methods of communication between the system and the human. In this paper we are evaluating augmented reality (AR) technologies for displaying task-relevant information when the target is on a work surface for a typically standing work operation. In this case we are testing AR interfaces for displaying sub-surface positions. To do this we compare four types of AR interfaces, a head-mounted see-through display, a mounted see-through display, top-down surface projection and graphical overlays on a static monitor. We performed the experiment with 48 participants. Data analyses show significant difference between the AR interfaces in terms of task completion times and user satisfaction with the projection-based display being the fastest and most satisfying to the participants.

1 Introduction

Repetitive strenuous movements involved in production work can lead to musculoskeletal diseases in the long term [1]. This issue can be addressed by introducing assistive and collaborative systems that can relieve some of the strain. This also has the potential to increase productivity. In order to do this we need to develop communication methods between the system and the human, from here referred to as the operator. The communication from the system to the operator in the production context will involve conveying the details pertaining to the current task, and for this study we seek to utilize augmented reality (AR) interfaces for this purpose.

The studies and development are done in the context of industrial meat production in which employees stand on the meat processing lines. In this case the term operator will refer to a single production employee collaborating with an instance of the system's assisting agents. The system will be assisting in tasks involved in sequential meat processing where each employee performs one task on each piece of meat for a period of time. These tasks include positioning the meat, cutting it down to size, trimming fat layers or picking out impurities. For this test we focus on the latter task and develop methods for the system to show the position of an impurity, whether it be on or under the surface of the meat, that it can be addressed and removed by operator.

Because the operators will be working in close proximity with potentially hazardous hardware it is critical that task-relevant information can be displayed non-obstructively to the operator in order to communicate the current objective of the system. This will allow the operator to anticipate the actions of the system, leading to safer collaboration and improve human trust in the system as the communication is developed further.

We focus on evaluating AR interfaces since these can be used hands-free. The goal is to evaluate and compare four different types of AR interfaces in terms of effectiveness, ergonomics and user acceptance when showing sub-surface positions in an opaque mass acting as the analog for the meat. The four types of AR interfaces are a head-mounted see-through display, a tablet-based see-through display mounted to an adjustable arm, top-down surface projection and graphical overlays on a static monitor. The experiment plans and procedure have previously been presented in a poster abstract at Human-Work Interaction 2018 [2], while this paper is more accurate.

2 Related Research

The communication from the system to the operator in the manufacturing context will involve conveying the details pertaining to the current task. Novak-Marcinin et al. [3] define augmented reality-aided manufacturing (ARAM) as the overlap of AR-aided robot control, AR-aided testing, AR-aided assembly and AR-aided transport and storage. The experiment in this paper is to evaluate interfaces for AR-aided assembly, because of the meat production context, where it will be used to aid production staff.

Regarding preliminary evaluation of AR devices, Elia et al. [4] proposed a 4-step model to be applied in specific manufacturing processes. The fist step is a multi-criteria analysis for ranking the most effective AR systems for the purpose, which is the current stage of this project. The ranking is performed by comparing the hardware options in terms of output modalities, reliability, responsiveness and agility. The ranking is done using pair-wise comparison followed by analysis and ranking of the AR devices. Elia et al. [4] categorize types of AR hardware as head-mounted displays (HMD), handheld devices, projectors and haptic force feedback systems. The second step is obtaining a judgment matrix using pair-wise comparison between criteria, followed by evaluation of local weights and consistency of comparison in step 3, with final ranking of devices as step 4.

Kruijff et al. [5] classified potential issues with AR caused by a combination of the environment of use, capturing the environment, the method of augmentation, the types of display device and user. They also point out whether these issues are predominant with particular display types which are categorized similarly to Elia et al. [4]: Head-mounted displays (video see-through or optical see-through), handheld mobile devices or projector-camera system (stationary or mobile). Relevant issues for this study include wearable see-though displays having limited field of view (FOV) and vergence-accommodation conflict for virtual objects and surface-based distortions for projector-based setups.

For this experiment we consider the environment of use be recreating the relevant working conditions pertaining to freedom of movement and posture allowed in a standing task. The four categories outlined by Elia et al. and Kruijff et al. have all been considered for the experiment. However, seeing as all of them are primarily visual aids as opposed to haptic force feedback systems, the latter is not included in this experiment. Since the context of the study allows for handheld devices to be implemented in combination with existing production tools, we would consider haptic feedback as a possible addition to visual augmented reality, so it may be evaluated as an addition at a later stage.

Human-system collaboration enabled by AR has been studied previously, often in the context of human-robot collaboration. However, often these tests have not been performed in the context of a close-proximity task with the user standing at a table. Green et al. [6] tested a human-system collaboration system utilizing an HMD where the operator was sitting at a table. However, the headset used in this case, rather than being a see-through AR display, was an eMagin Z800 headset using OLED displays with the augmented video fed from a mounted webcam. With the potential to have the webcam pointed downwards toward the table the operator would be relieved from bending their neck to look directly down at the tracking markers. The paper does not specify any angle adjustment in the implementation. This solution is not considered for this comparison due to the potential hazard or limiting the operators field of view as opposed to see-through HMDs which allow the user to still see outside of the display field. Even-though not pertaining to a task specifically, Vogel et al. [7] proposed using projection-based AR to show an outline of a safe working area in relation to the collaborative system.

Schwerdtfeger et al. [8] go into depth describing the projected AR, specifically using lasers. They points out the cons of HMD AR devices, those being narrow field of view, limited resolution, swimming effect, multiple focus planes as well as eye fatigue. While laser-projected AR can address some of these issues it is limited to displaying information on surfaces in the environment and the image must be distorted to compensate for environment geometry and viewing angle, whether the projector is head-mounted or stationary. In addition, it also introduces the challenge of occlusion by either the user or other objects. In order to avoid surface distortion for this test we use an even surface for this comparison.

Swan et al. [9] studied how depth perception is affected while using AR devices in that subjects tended to underestimate distance in AR when they are projected at less than 23 meters distance to the user, after which the bias switches to overestimation. Comparing this to short distance error, Singh et al. [10] estimated an error of -5.5 cm at most for distances less than 50 cm. From these finding we should expect our participants to underestimate the target distances in our test. However, we can not know if this is true when the user can judge distance in relation to a real surface as opposed to judging a target hanging in the air.

Similarly to showing sub-surface positions, augmented reality has previously been used to imitate x-ray vision. Avery et al. [11] emphasized that when showing the content beyond the surface using an graphical overlay it should include an edge overlay representing the surface as a depth cue, using occlusion as a depth cue so the object does not appear to float in front of the surface. We are implementing the same method in our test applications by projecting a graphical grid overlay on the surface of our meat analog.

The main contribution of this experiment to the fields of human-robot collaboration and ARAM is in the comparison between the AR devices, but the significance is in the environment and conditions it will be utilized, as comparing the systems in a low-distance setup while standing at a work surface has rarely been done.

3 AR Devices

We are testing four different types of AR devices. Three of them are based on three of the types outlined by both Elia et al. [4] and Kruijff et al. [5]; head-mounted see-through display from here referenced as HMD, see-though mobile display and projection-based AR. In addition, we are comparing a video feed on a monitor augmented with graphical overlays because this is currently a typical way of displaying information in meat production settings. The software for the four AR devices are developed using Unity 3D. Similar to all of the versions is that a green grid is projected aligned with the surface of the sand, similarly to what is described by Avery et al. [11].

The targets are shown as a red sphere or circles at 10 mm in diameter projected into the sand. Due the nature of the devices the display methods for the targets differ between the displays. Because the image for the top-down projector and the stationary screens are limited to 2D projections on the surface of the sand, the red dot is shown along with a number indicating the depth of the target in millimeters. We considered using similar labels for HMD and tablet devices, but because the user will see the target projected from different angles depending on their viewing angle as opposed to always seeing it from the top and directly down. Both the HMD and tablet show 3D-rendered images, allowing for occlusion and motion parallax as occlusions by the green grid as cues, which is not possible for the remaining displays without using head-coupled perspective.

Because of the varying nature and performances of the displays, accurate calibration between all devices proved very difficult. Because of this, this study focuses on the accuracy spread between each device. In practice this means that the accuracy for each device is measured by the offsets from the median offset from the targets for each participant. By doing this we assume that the median hit is an accurate hit as adjusted to the user.

3.1 Head-Mounted See-Through Display

The HMD using an Epson Moverio Bt-300 which is an Android-based device and is equipped with 0.43 inch wide panel 720p displays at a 30 Hz refresh rate. The software is implemented using Unity 3D with the Vuforia AR plugin. The HMD is shown in Figure B.1. The impurities are projected into the sand on the glasses while also utilizing the overlay grid.

Using Vuforia along with the build-in 5 megapixel camera on the right-hand side of the headset the system is tracking using the AR markers at either end of the surface of the sand. The goal is to avoid the participants occluding the trackers by only having to use one hand for the tasks, leaving the tracker on the opposite side exposed. The feed from the camera itself is not displayed on the HMD, only the grid targets are displayed. The overlay is manually offset and rotated to best fit the surface of the sand. This process includes adjusting the rendering FOV to fit the display area of the glasses. In this case the FOV is set to 21 degrees, despite the manufacturer advertising 23 degree FOV for the device.

Since we are not able to perform eye tracking using the hardware to determine the convergence point, rendering the position in stereo would leave the user with difficulty focusing. Because of this the target is not rendered in stereo, but rather as a 2D overlay similarly to the see-through display, and when analyzing the data the participants' dominant eye must be taken into consideration.

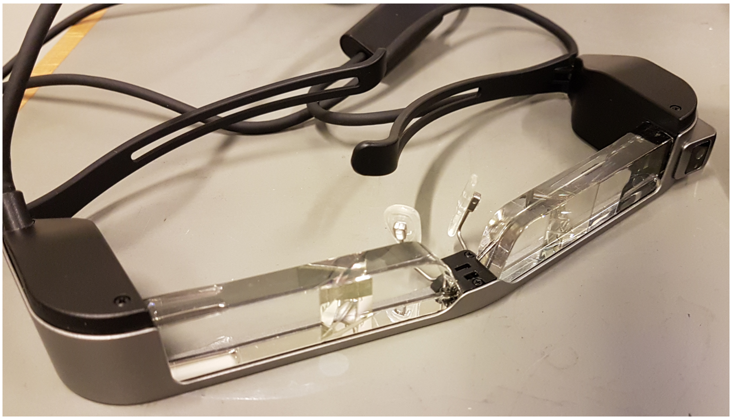


Fig. B.1: The Epson Moverio Bt-300 glasses used for the experiment. [2]

3.2 Mounted See-Through Display

Similarly to the HMD, the see-through tablet display is running Android and Unity with Vuforia. The display is shown in Figure B.2. The tablet is mounted to an adjustable stand, so it can be positioned according the user's height while oriented to show the entire surfaces area of the sand and all the targets in frame. For this test the tablet is manually adjusted to each subject and remains stationary where an AR tracking marker is visible.

We focus on the marker opposite the user's dominant hand to prevent occlusion. The participant also has to be able to reach underneath the tablet and their hands are visible throughout the test sessions and they can coordinate their movement with the targets in the camera feed. Since the tablet remains mostly stationary during each session, Vuforia is mainly used when the tablet is initially adjusted for the user, this device is less susceptible to errors and latency introduced by Vuforia.

3.3 Top-Down Projection

This device is set up using a projector mounted to a tripod with a 3D printed adapter and pointed down towards the sand. The projection is shown in Figure B.3, bottom left. The projection is adjusted using a piece of paper with a grid printed on it matching the green grid overlay. The projected overlay features the green grid and the targets are shown as red dots with depth in millimeters shown next to them.

3.4 Stationary Screen

The monitor setup utilizes a 25 inch PC monitor with an aspect ratio of 16 by 10. The monitor is shown in Figure B.4, bottom right. We use a Logitech C920 on an adjustable mount pointed straight down to the surface of the sand. Similarly to the top-down projection setup, the camera feed on the monitor features the green grid overlay and

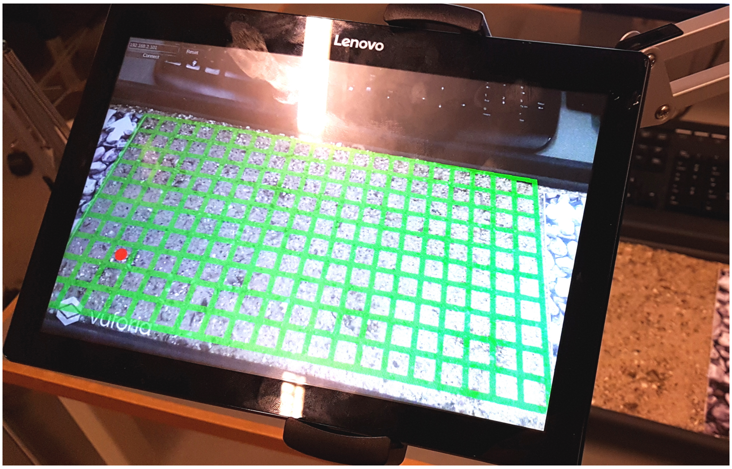


Fig. B.2: The arm-mounted see-through display setup with impurity projected into the sand as a red dot assisted by a grid overlay aligned with the surface. [2]

Fig. B.3: The grid projected onto the sand with the targets shown as a red dot. [2]

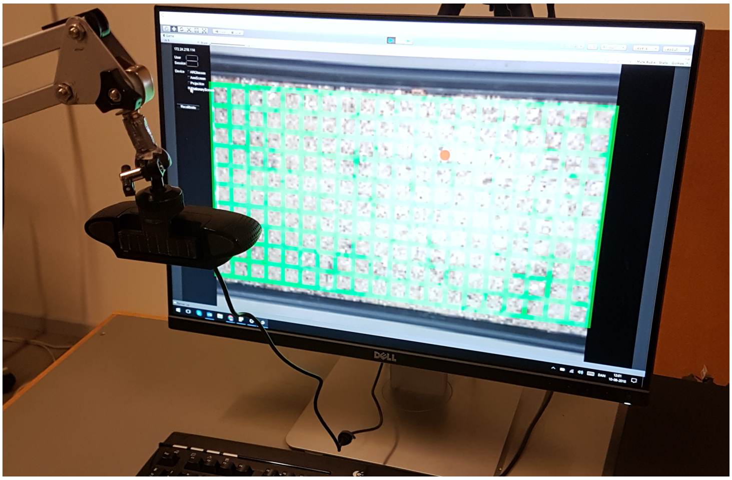


Fig. B.4: The monitor and top-down camera. [2]

the targets are displayed along with label indicating their depth. Similarly to the tablet setup, the user will be seeing their hands in the video feed during the test sessions. The camera position is calibrated using the printed grid and by placing an HTC Vive controller in the video feed and matching its position with the representation on the monitor.

4 Setup

The experiment is performed using a tray of sand as the analog for a cut section of meat, as it allows the test subject to poke into it with a tool to address the impurities that will be displayed in it using the AR devices. We use dry loose sand with low density so allow easy entry and to prevent visible entry points to stay throughout a test session. The sand also allows for the surface to be smoothed out by hand.

The participant's performance is measured using an HTC Vive controller with a nail mounted to the bottom as shown in Figure B.5. Using the six degrees of freedom tracking capabilities of the controller the participants are using the tip of the nail to poke into the body of sand as closely as they can while holding the controller in their dominant hand. A second Vive controller is held in the non-dominant hand and the participants will use a button on it to confirm when they have hit a target. Confirming with the non-dominant hand prevents shaking the tip of the nail during a button press.

The HTC Vive tracking space is running using a Windows 10 PC which also acts as the host for the test application, developed in Unity 3D. In order to show the target position on the HMD and the mounted tablet, both of which run Android, the PC acts as network host and sends target positions to the devices acting as clients over a wifi connection throughout the test sessions.

Fig. B.5: The needle used for the experiment, made from a 3D printed mount and a nail attached with a bolt in the loop designed for the wrist strap.

The surface position of the sand is calibrated in the tracking space using the tracked nails on the Vive controller along with a printed piece of calibration paper which also fits in-between the tracking markers used for the HMD and mounted tablet. The tracking markers are positioned at either side of the testing area with the goal of having at least one visible to the tracking camera regardless of where the user is looking and to prevent occlusion of the trackers. The full test setup is shown in Figure B.6.

5 Experiment

Each participant is introduced to the topic at the beginning of the experiment and they are asked to sign a consent form, followed by a questionnaire pertaining to their age, sex, height, dominant hand and dominant eye. In cases where participants do not know their dominant eye, it is determined with the Miles Test [12].

5.1 Pointing Tasks

Each participant performs a set of pointing tasks with each AR device. The order of the devices is counterbalanced between participants in order to counter bias and initial confusion about the tasks with the first devices. With four systems we have 24 possible combinations. We do each combination twice to reach a total of 48 participants.

While wielding the HTC Vive controller the participant is asked to point the tip of the nail in the center of the target as they appear, as quickly and precisely as they can. To do this they will be asked to have the HTC Vive controller rest on the side of their dominant hand between the thumb and index finger as shown in Figure

Fig. B.6: The full setup for the AR test [2].

B.5. When the participant believes they have reached the target center, they must confirm by pressing the trackpad on the controller in the non-dominant hand. To prevent accidental double-presses there is a delay of one second from when a target is confirmed to when the next target is shown, during which the confirm button is disabled. This also allows the participants to return their hand to a natural position close to their body, but they are not instructed to do so.

The sand is held in a rectangular plastic tray and has a depth of 5 centimeters. The targets are shown at depths of 0, 5 and 10 millimeters, meaning that participants can not expect to drive the nail to the bottom of the tray and get a precise hit. At each of the three depths the targets are distributed on two rows and four columns with 10 centimeters of spacing in both dimensions, so the participants will get to both reach across and away from the center of their body. With a 24 targets per device per user, we get a total of 4608 samples. The rectangular shape for the tray and the sand is appropriate for the experiment considering that at the points of fine operation in meat production, such as picking out impurities, the meat has been cut down to these shapes.

For each target we measure the task completion time from when a new target appears until the participant confirms a hit. This includes time spend searching for the targets. The accuracy is measured as the offsets from the target center to the tip of the nail at confirmation for the three dimensions individually as well as the absolute distance from target to tip. After using each of the four devices the participants get to evaluate the device in terms of acceptance and ease of use with System Usability Scale (SUS) [13].

5.2 Hypotheses

With the experiment we aim to prove that there are significant differences in the effectiveness and user experience between the AR display types, specifically in terms of the following hypotheses:

- H1: Spotting and hitting sub-surface targets when using different AR devices will yield different task completion times.
- H2: Hitting sub-surface targets using the different AR devices will yield different hit accuracy.
- H3: The different AR devices have different usability based on the Standard Usability Score.

6 Results

As described in Section 3, the four devices can be split into two categories, the AR glasses and mounted tablet showing depth through perspective, occlusion and parallax while the remaining are limited to 2D overlays. In this section these categories are described as the 3D and the 2D enabled devices, respectively.

The experiment was performed with 48 participants, average age 24 years, ranging between 20 and 34. Nine participants were female, ten were left-handed. 21 reported having left-eye dominant, 25 reported right, one reported both eyes to be dominant and one could not be determined. Eight participants had positive eyeglass prescriptions, ten had negative prescriptions and two had unknown prescriptions.

6.1 Data Analyses

Performing analyses of variance with significance threshold at .05 shows significant difference among the AR devices in task completion times when comparing them individually ($F(2,3926) = 333.3$, $p < .001$) and when comparing the 3D and 2D enabled devices ($F(1,3926) = 124.7$, $p < .001$) with Tukey's HSD post-hoc analysis showing significant difference between all devices. The projection-based display yielded the lowest average time (2.3 seconds), followed by the stationary screen (3.55 seconds), the mounted see-through display (4.5 seconds) and the AR glassed (9.92 seconds) had the longest average task time. A summary of the task time data is shown in the box plot in Fig. B.7. These results confirm the first hypothesis.

To analyze the accuracy between AR devices, in order to compensate for dominant eye and calibration inaccuracies between test participants, the samples are corrected according to their per-session median hit value. By doing this we make the assumption that the median offsets are for a precise hit, making the median values zero for the data sets used. Figure B.8 shows the per-session median hit values for each condition and participants. The high median offsets for depth suggest a calibration error.

Performing a multiple analysis of variance, dependent values being the hit offsets on three separate dimension, there is significant difference between the four AR devices ($F(2,3926) = 6$, $p < .01$) as well as when comparing the 3D and 2D enabled

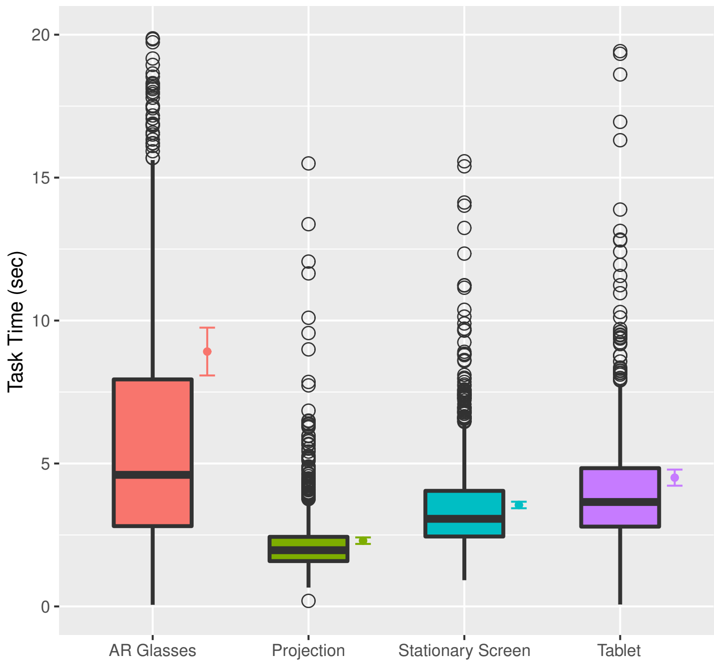


Fig. B.7: Box plot of the task completion times between the four AR interfaces. Next to each box plot are their mean values and confidence intervals.

devices ($F(1,3926) = 3$, $p < .002$), confirming the second hypothesis. The handedness of the user did not yield significant difference in hit offsets.

Looking at hit spread, the AR glasses have the largest average standard deviation at 51 millimeters in target hit offsets for all three dimensions. In comparison, the stationary screen, projection and tablet displays have averages of 10, 28 and 30 millimeters, respectively.

Investigating the difference in sideways offsets in relation to the user for the four devices ($F(2,3926) = 3.36$, $p < .05$), a Tukey's HSD post-hoc analysis shows no significant difference between any pair of devices, suggesting the difference depends on whether the interfaces were 3D or 2D enabled ($F(1,3926) = 9.26$, $p < .003$). Similarly for offsets going towards or away from the body is only significantly difference between the two groups ($F(1,3928) = 8.27$, $p < .005$). The average offsets along with standard deviations along the surface are shown in Figure B.9.

For the depths offsets as well there is only significant difference between the two categories ($F(3,3926) = 3.4$, $p = .017$) and all of the average offsets are within one millimeter of each other. The depth offsets are shown in Figure B.10.

The SUS scores show significant difference between the four devices ($F(3,186) = 92.03$, $p < .001$) with Tukey's HSD post-hoc analyses showing significance between all four devices. There is not significant difference in user acceptance when comparing

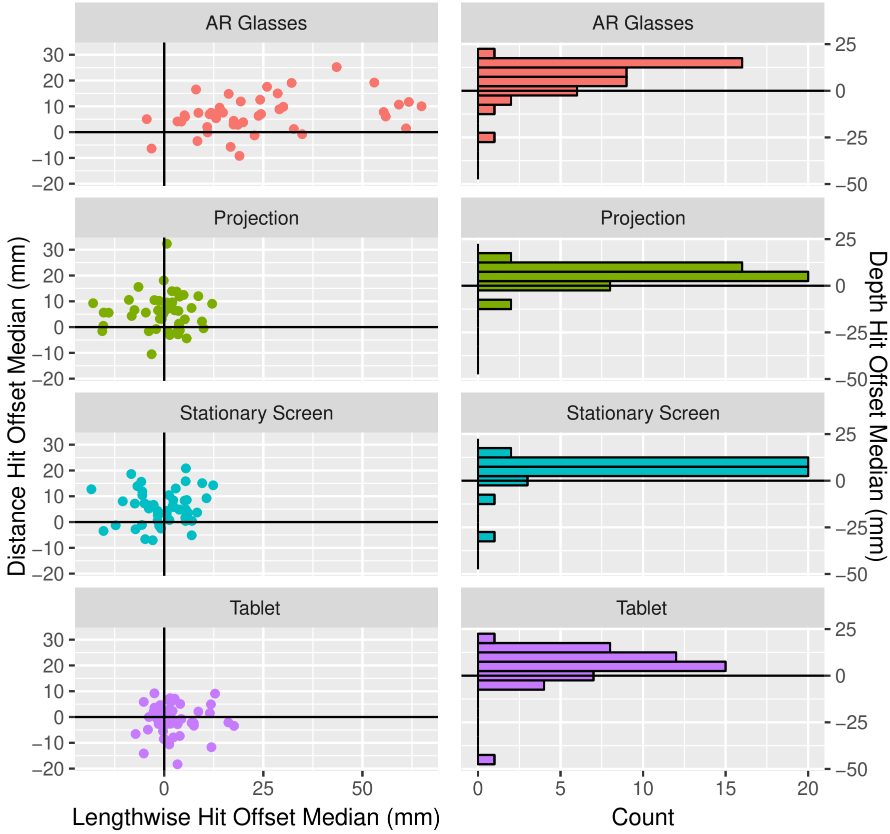


Fig. B.8: Per-session surface median hit offset values for each participant (left) and histogram for depth offset medians in bins of 5 mm (right) for each condition.

2D and 3D devices. However, two of the four devices do not have normally distributed scores according to the Shapiro-Wilk normality test, the AR glasses ($p = .062$) and the mounted see-through display ($p = .16$), making the third hypothesis harder to retain. The SUS scores are summarized in Figure B.11.

6.2 Observations

The AR glasses had the longest overall task completion time. The long task times are likely due to the narrow FOV, for multiple reasons. Firstly, the FOV does not allow for the user to see the augmentation overlay on the entire surface at once, requiring additional search time to the task. Despite them being able to see all of the surface at once, the AR glasses only cover a small segment of their FOV, creating a letterbox effect. Despite the mounted tablet display tracking and showing targets in similar way,

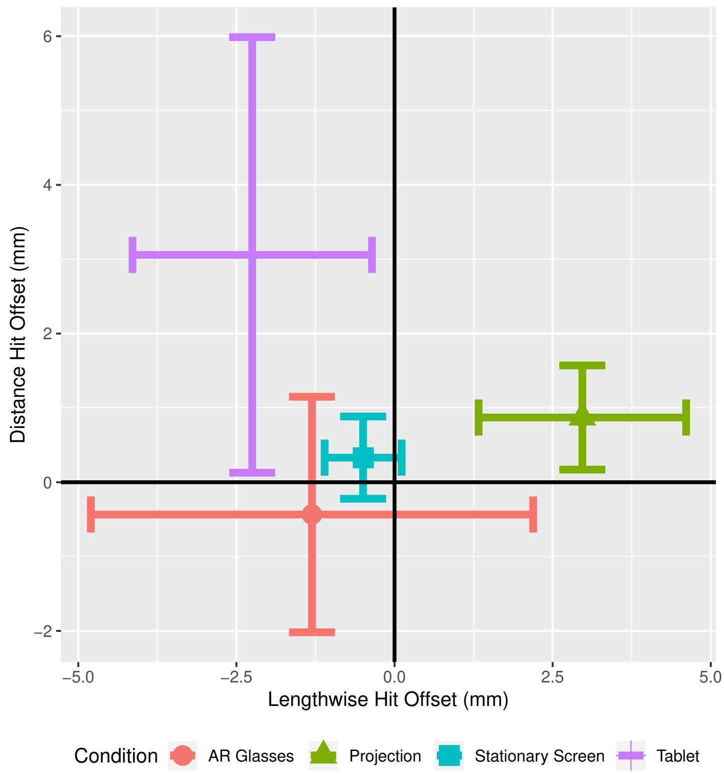


Fig. B.9: The average offsets along with standard deviations along the surface between the four AR devices.

its position and FOV did allow for user to see the entire surface at once, eliminating the need for search time.

The limited FOV of both the display and the camera used for the tracking in combination with the short distance to the surface of the sand made it difficult for the participants to inspect the entire surface area while also keeping the tracking marker in view of the camera. This in combination with errors and latency introduced by Vuforia made a sub-par user experience.

In addition, the glasses' FOV does not allow the user to glance downwards, leading them to turn their head downwards to an uncomfortable degree as they leaned in over the tray, while also having to turn their head to look around and search for targets. A few test participants commented on this, stating that using the glasses was starting to give them neck pains. Many participants who tried other devices after the glasses would comment out loud on how much easier it was.

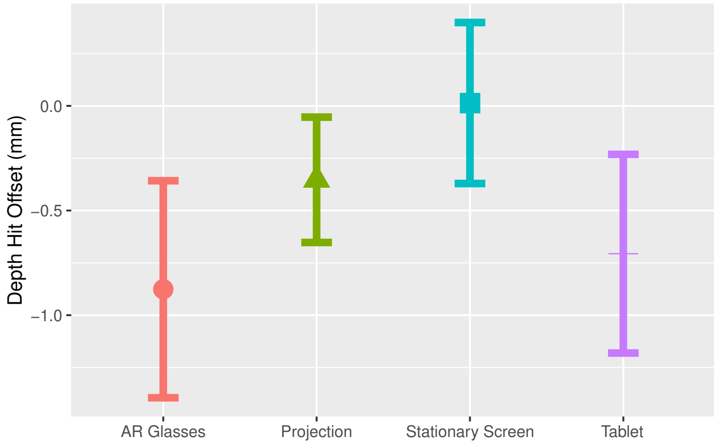


Fig. B.10: Average depth offsets between the four AR devices along with confidence intervals.

7 Discussion

Despite the accuracy of the AR devices proved significantly different, it is hard to define a concrete set of tendencies, seeing as significance differs between devices and axes and the average hit positions differ as seen in Figure B.9. Also, the average hit accuracy in depth, though significantly different, are within one millimeter of one another, making it less relevant in a real-world context. It does show, however, that the projection based AR and the monitor were the two devices with the lowest standard deviations along the surface.

The analyses show significant differences when comparing the 2D and 3D enabled interfaces, both for task completion time and accuracy. For the latter this grouping is more consistently significant, because the post-hoc analyses show no significant difference in pairwise comparison. This may be due to both of them being static and showing the positions from a top-down perspective as opposed to the see-through display which required interpretation of perspective and occlusion as depth clues and the glasses that required constant tracking as the user moves, introducing noise. The tracking noise would also be introduced for the mounted displays in a real-world setting as the display would be moved around. Nevertheless, the inaccuracy for the projection and the monitor can likely be fixed with hardware and software adjustment, where the remaining devices have the challenge of tracking and depths communication.

The SUS results show that the projection-based AR and the monitor-based AR were the only two devices with averages scores reaching above the standard cut-off point of 68. The analysis of variance was used despite not all of the groups being normally distributed because due to the nature of the SUS scale where groups will be tailed in different directions dependent on their average position on the scale, making them hard to fit in any statistical model.

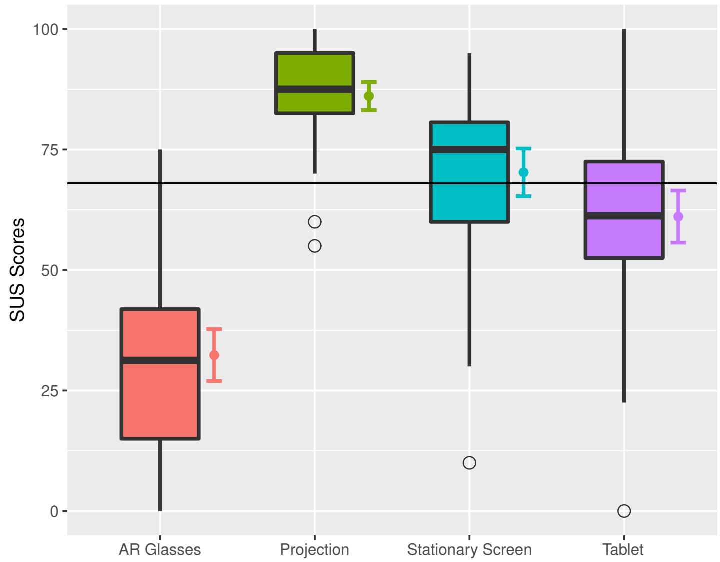


Fig. B.11: Box plot of the SUS scores between the four AR interfaces and a line at the cut-off value of 68. Next to each box plot are their mean values and confidence intervals.

The low task completion time combined with the high SUS score for the projection-based AR system is likely due to the direct connection between the display and the target, the interfaces being on the target itself. Eliminating the requirement of coordinating hands with a display offset from the target seems to make the interface more accessible. This is despite the problem with projection-based AR that the user will occlude the projection when interacting with it.

The issues with the AR glasses described in Section 6.2 make the Epson Moverio non-viable for near-distance tasks. An alternative would be the technique used by Green et al. [6] where the user's viewing direction would be shifted downwards, compensating for the FOV. However, whether this approach, occluding part of the user's FOV during a potentially hazardous task, would work is uncertain. It would be beneficial to repeat the experiment using a HMD specifically designed for the context of close-proximity tabletop operations. This would involve expanding the display FOV, allowing the user to search by scanning with their eyes rather than turning their head straight at the target. Another required feature is eye-tracking in order to properly implement stereo display.

It is worth considering after this experiment whether 2D and 3D enabled AR devices are comparable in terms of accuracy and user experience due to differences in affordances. Because the projection-based overlay and the screen are static displays, the only tracking involved in a real-life scenario is of the meat and the targets as the subject is moved around work surface. As such, these two displays potentially introduce less tracking noise, which is also likely to have affected the test results.

The results and conclusions to this experiment are mainly valid in the simulated context and would benefit from repeating in a setup with real meat. In that case, the results and observations illustrate the limitations of the AR glasses as implemented in this experiment, which should be addressed before they are assessed with real meat, either by different hardware or tracking solutions.

8 Conclusion

This paper presents a comparative study of the usability of four types of AR displays for showing sub-surface impurities in meat by having the participants point to targets inside an analog made of sand. The four display types are wearable AR glasses, a mounted see-through display, projection-based AR and a monitor displaying a top-down video feed with graphical overlay. The goal is to determine suitable interfaces for augmented meat production.

After performing the evaluation with 48 participants our three hypotheses were retained with significantly different task completion time, accuracy and user acceptance depending on the AR display type used. Data analysis indicates that projection-based AR yields the second-lowest variance in combination with the lowest task completion times and the highest SUS score, making it the most suited for the task with the mounted see-through display and stationary screen being viable alternatives, while the AR glasses showed to be non-viable for near-distance tasks as they were implemented for the experiment.

Even though the research is aimed at the meat production industry, the results are relevant to any industry that utilized manual processes while standing at a work surface. Whether the results would pertain to sitting tasks at a desktop might require further testing due to the difference in distance and postures.

As stated in Section 7, for further comparison between the devices, the limitations of some of the devices have to be addressed with more suitable hardware and tracking solutions before they are tested further in a real-world context. That is if the 2D and 3D-enabled devices can be considered comparable from an accuracy perspective due to the difference in affordances. Even so, at early stages of development it is relevant to compare them in terms of the usability and user acceptance that stems from the different affordances.

References

[1] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of Electromyography and Kinesiology*, vol. 14, no. 1, pp. 13–23, feb 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641103001251?via{%}3Dihub

[2] K. Hald, M. Rehm, and T. B. Moeslund, "Testing Augmented Reality Systems for Spotting Sub-Surface Impurities." Springer, Cham, aug 2019, pp. 103–112.

[3] J. Novak-Marcincin, J. Barna, M. Janak, and L. Novakova-Marcincinova, "Augmented Reality Aided Manufacturing," *Procedia Computer Science*, vol. 25,

pp. 23–31, jan 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S187705091301209X

[4] V. Elia, M. G. Gnoni, and A. Lanzilotto, "Evaluating the application of augmented reality devices in manufacturing from a process point of view: An AHP based model," *Expert Systems with Applications*, vol. 63, pp. 187–197, nov 2016. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0957417416303505

[5] E. Kruijff, J. E. Swan, and S. Feiner, "Perceptual issues in augmented reality revisited," in *2010 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, oct 2010, pp. 3–12. [Online]. Available: http://ieeexplore.ieee.org/document/5643530/

[6] S. A. Green, J. G. Chase, X. Chen, and M. Billinghurst, "Evaluating the Augmented Reality Human-Robot Collaboration System," in *2008 15th International Conference on Mechatronics and Machine Vision in Practice*. IEEE, dec 2008, pp. 521–526. [Online]. Available: http://ieeexplore.ieee.org/document/4749586/

[7] C. Vogel, M. Poggendorf, C. Walter, and N. Elkmann, "Towards safe physical human-robot collaboration: A projection-based safety system," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, sep 2011, pp. 3355–3360. [Online]. Available: http://ieeexplore.ieee.org/document/6094550/

[8] B. Schwerdtfeger, D. Pustka, A. Hofhauser, and G. Klinker, "Using laser projectors for augmented reality," in *Proceedings of the 2008 ACM symposium on Virtual reality software and technology - VRST '08*. New York, New York, USA: ACM Press, 2008, p. 134. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1450579.1450608

[9] J. Swan, M. Livingston, H. Smallman, D. Brown, Y. Baillot, J. Gabbard, and D. Hix, "A Perceptual Matching Technique for Depth Judgments in Optical, See-Through Augmented Reality," in *IEEE Virtual Reality Conference (VR 2006)*. IEEE, pp. 19–26. [Online]. Available: http://ieeexplore.ieee.org/document/1667622/

[10] G. Singh, J. E. Swan, J. A. Jones, and S. R. Ellis, "Depth judgment measures and occluding surfaces in near-field augmented reality," in *Proceedings of the 7th Symposium on Applied Perception in Graphics and Visualization - APGV '10*. New York, New York, USA: ACM Press, 2010, p. 149. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1836248.1836277

[11] B. Avery, C. Sandor, and B. H. Thomas, "Improving Spatial Perception for Augmented Reality X-Ray Vision," in *2009 IEEE Virtual Reality Conference*. IEEE, mar 2009, pp. 79–82. [Online]. Available: http://ieeexplore.ieee.org/document/4811002/

[12] W. R. Miles, "Ocular dominance demonstrated by unconscious sighting." *Journal of Experimental Psychology*, vol. 12, no. 2, p. 113, 1929.

[13] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

Paper C

Trust Repair of Virtual Robot Using Levels of Mistake Explanation

Kasper Hald, Katharina Weitz, Elisabeth André, Matthias Rehm

The paper is under review in
IFIP Conference on Human-Computer Interaction – INTERACT 2021

The layout has been revised.

Abstract

Human-robot collaboration in industrial settings is an increasing research field in robotics. When working together, robot mistakes are an important factor to decrease trust. It is unclear whether explanations help to restore human-robot trust after a mistake. In our study, we investigate whether system explanations as a trust-repairing action after a robot makes a mistake in a collaborative task is helpful. Our pilot study revealed that users are more interested in solutions to errors than they are in just why the error happened. Therefore, in our main study, we evaluated three levels of mistake explanations (no explanation, explanation, and explanation with solution) after a robot in VR made a mistake in executing a shared objective. After testing with 30 participants we found that the robot making a mistake significantly affects trust toward the robot, compared to it completing the task successfully. While participants found the explanations helpful to trust or distrust the robot, the levels of the explanation did not lead to an increase in trust towards the robot after a mistake. In addition, we found no significant impact of explanations on self-efficacy and the emotional state of the participants. Our results show that explanations alone are not sufficient to increase human-computer trust after robot mistakes.

1 Introduction

The collaboration between humans and machines in the industrial setting is becoming more and more realised due to the enormous progress in the fields of robotics and machine learning. Through the use of machine learning methods like Deep Reinforcement Learning, it is possible for robots to interact autonomously in industry tasks and to adapt dynamically to the demands placed on them. In addition, a intuitive usage and interaction by humans become more and more common. However, the more natural the handling of robots in industry becomes, the more demands humans place on them. If these demands are not met, human-robot collaboration (HRC) can be disrupted. In addition to reduced trust and frustration, this can have serious consequences such as accidents and production losses [1, 2]. To enable successful HRC it is important that we can maintain human-robot trust (HRT), especially when working with a robot at close proximity. To this end we investigate the capabilities of system communication with the human collaborator to perform trust-repair through explanation in cases where the robot makes a mistake during the execution of a shared objective. We base the research on the context of a shared task where the human and robot have to move a collection of objects within a shared tabletop work space. To integrate system communication with non-obstructive output modalities we base the design of the communication system on projection-based augmented reality (AR), so that messages can be displayed directly on the work surface. To sum up, we investigate how we can use mistake explanations after a robot mistake as a trust-repairing action in order to maintain trust during close-proximity collaboration. Rather than implementing the communication system using real hardware, we test our prototype iterations using computer-generated demonstrations and virtual reality (VR) testing environments. With our work, we make the following contributions:

- We give insights about requirements and expectations of end-users towards robot explanations after mistakes

- We present a VR setup to research robot-mistakes in close-proximity tasks
- We report results about the impact of different levels of explanations after robot-mistakes on trust, explanation satisfaction, self-efficacy, and emotional state of end-users
- We discuss the challenges using explanations in HRC

2 State of the Art

2.1 Human-Robot Trust & Robot Mistakes

De Visser et al. [1] use a definition of trust in the context of HRC as the human's willingness to engage in a situation characterized by vulnerability with another party based on their expectation toward that party. In this context the other party is the robot. In a meta-analysis Hancock et al. [2] categorized the constructs that affect the operator's perception of the robot into human-related, robot-related, and environmental factors. The Robot-Related factors are further split into performance-based and attribute-based factors, covering how the robot performs or behaves and how the robot looks or where it is, respectively. Looking at performance-based robot-related factors, reliability, dependability, and predictability have significant effect HRT. They also outline the importance of appropriate trust levels toward the robot in HRC, as too much trust may lead to dangerous situation as a result of misuse, whereas too little trust may lead to the robot not being utilized optimally. Schaefer [3] developed two HRT scales based on the operator's perception of the robot's characteristics, performance, predictability, and more. The long scale have 40 questions while the shorter version has 14 questions. Kessler et al [4] compared these scales to a standardised scale of trust in automation with conflicting results, suggesting that the two scales evaluated different factors, making them not interchangeable.

In testing robot dependability and its effects on trust, Salem et at. [5] found that a home companion robot would be perceived as less trustworthy after making a mistake, even though the mistake did not significantly affect participants' willingness to follow the robot's instruction. In addition to the factors outlined by Hancock et al. [2], HRT has also been shown to be affected by the general transparency of the system controlling the robot. Boyce et al. [6] compared three transparency conditions in a simulation. Higher levels of transparency yielded higher trust measured using a modified automation trust scale. Due to the scale used one has to consider whether the trust pertains to the simulated robot or the communication system. Comparing decision explanations for a robot in a simulated reconnaissance mission, Wang et al. [7] found that low-ability robots gained more trust from explanation, as opposed to no explanation, whereas high-ability robots did not gain trust from them. When testing a robot that would assign blame after a mistake occurred, Kaniarasu & Steinfeld [8] found that people would be annoyed when the robot blamed them, but they trusted the robot less if it kept blaming itself.

On the importance of the presence of the robot, as we are testing using VR simulations, both Wainer et al. [9] and Bainbridge et al. [10] compared a co-located robot with a remote robot presented on a screen, and both found that the co-located robot was significantly favoured. However, Duguleana et al. [11] found, when comparing HRI with a real robot and with one presented in immersive VR, participants reported high engagement toward the virtual robot and rated it at 7.8 out of 10 in realism, relative to the real robot.

2.2 Explanations in Human-Robot Interactions

Evidence suggests that a lack of transparency, with respect to the decisions of an autonomous agent, might have a negative impact on the trustworthiness of a system, which in return hurts the overall user-experience [12, 13].

The reemerging research field of explainable artificial intelligence (XAI) [14] investigates approaches to address this problem. Current research on XAI is mainly dealing with methods to explain the decisions of deep neural networks (e.g., [15–17]). Various promising approaches have meanwhile been developed for these use-case (the interested reader is referred here to works of e.g., [18, 19]). In the field or human-robot interaction, different XAI approaches are discussed to gain insights in behaviour and goals of robots (e.g., the work of [20]).

Alongside the question of how explanations can be generated, the research field of XAI is also concerned with the question of how explanations can be communicated to users. In particular, communicating explanations to end-users is a challenge here, as they need to interact with the system (e.g., a robot) but have no knowledge how the system works. The work of Wang et al. [21] shows that explanations to end-users about a well working robot increases transparency, trust, and performance in human-robot interactions. But robots also make mistakes and are not free of errors. When an error occurs, without an explanation end-users are often unable to understand how the error arose, how to fix it, and how to avoid it in the future. This leads to performance losses as well as distrust [22]. But even with explanations, less accurate autonomous systems lead to a decrease of trust in robots abilities, and success of the task [21]. Therefore it is critical to investigate, whether it is possible to repair trust in the system and if so, which aspects of an explanation are relevant to increase trust.

3 Pilot Study

The scope of our work is to investigate HRT in an interaction scenario in that the robot makes a mistake. In the pilot study we conducted, we first wanted to investigate whether different *explanation modalities* (i.e., *textual* or *auditory*) are preferred by participants. In addition, we varied the *type of error*:

- *Colour vision error*: To illustrate the colour vision error, the robot shown is moving a bottle on incorrect shape. The explanation given was:"A computer vision error occurred. The system did not successfully distinguish the shapes in the current lighting conditions."
- *Calibration error*: Here the robot knocked over one of the cones while moving the bottles. The explanation given was: "A calibration error occurred. The motion planner did not properly compensate for the robot's momentum."

The pilot study was conducted to guarantee, that the different explanation modalities and type of error did not significantly differ in their impact on trust. Furthermore, we wanted to gain insights whether the given explanations were sufficient enough and whether/which additional information participant find helpful. In more detail, we we formulated the following hypotheses:

- **H1**: After being presented with a robot mistake in videos of a virtual robot and a given modality of explaining the mistake, the user can describe the nature of the mistake accurately.
- **H2**: There will be no difference between the modality of explanation (i.e., textual and auditory) regarding likeability and performance of the robot.
- **H3** There will be no difference between the modality of mistake (i.e., calibration error and colour vision error) regarding likeability and performance of the robot.

To answer these hypotheses, we used a between-subjects design for the modality of explanation (i.e., textual or auditory), meaning that every participant saw online one of the explanation modalities. For the two different robot mistakes (i.e., colour vision error and calibration error), we used a within-subjects design. Here, every participant saw both types of errors during the study.

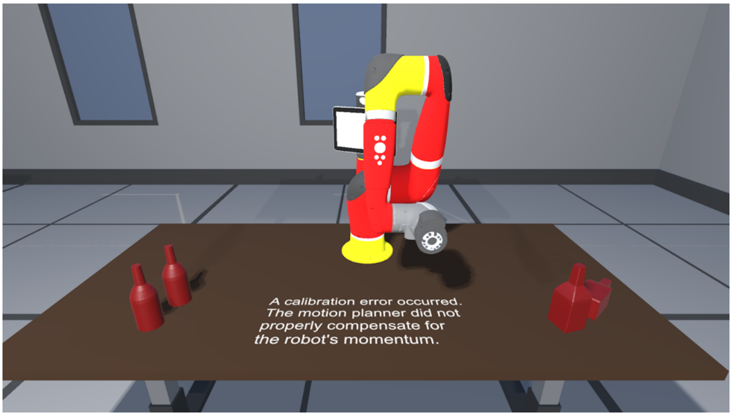
3.1 Procedure

The pilot study took place as an online questionnaire. Within this questionnaire, the participants were shown a series of videos of a virtual robot modeled after the Rethink Robotics Sawyer[1] model. This robot had the task of sorting bottles based on their shape.

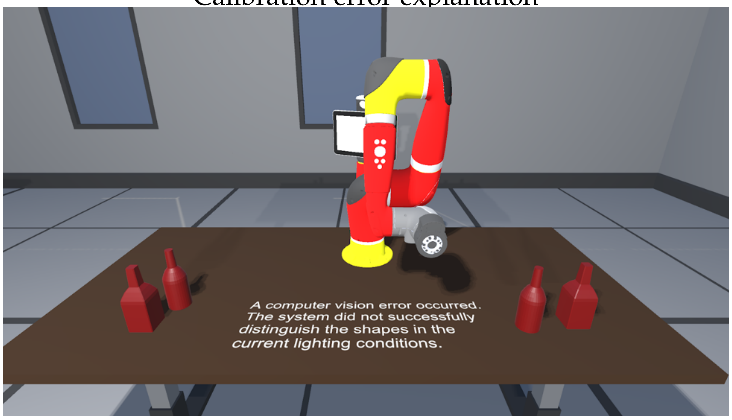
- **First video:** The first video showed the robot successfully completing the sorting task. Then, the participants rated the performance of the robot and their impression of the robot. They were then asked to briefly describe the robot, its behaviour, and the task it was performing.
- **Second video:** The second video showed the robot performing the same task again, but making a mistake (i.e., computer vision or calibration error). The participants then answered the same questions about the robot's performance and their impression. After that, they were asked to briefly describe what the difference was from the previous video.
- **First Explanation:** Subsequently, they were shown an explanation of the previously seen mistake (i.e., textual or auditory explanation). The textual explanation modality being shown in Figure C.1. Next, the participants had to answer several questions about the explanation shown.
- **Third video:** After answering these questions, they were shown a third video, also of the robot making a mistake.
- **Second explanation:** Here, again, an explanation was shown to them afterwards and the participants had to evaluate it.

[1]https://www.rethinkrobotics.com/sawyer/

At the end of the online study, participants had to provide some demographic information (e.g., age, gender) as well as their knowledge and attitude towards AI and XAI.



Calibration error explanation



Colour vision error explanation

Fig. C.1: Textual explanation modality. Two robot errors were explained during the pilot study: a *calibration error* and a *computer vision error*.

3.2 Evaluation Methods

To gain insights of the user's impressions regarding the robot errors and the explanation modalities, we used different scales.

Performance To evaluate the perceived robot performance, we asked the participants after every video to rate the performance of the robot, using a 7-point Likert scale (1= not good, 7= very good).

Likeability Similar to the measurement of the perceived robot performance, we asked the participants after the no-error video as well as after each explanation, how much the liked the robot and if they wanted to work with the robot.

Explanation Quality To measure the quality of the presented explanations, we used two items of the Explanation Satisfaction Scale (ESS), proposed by [23]. Here we asked the participants (1) whether the explanations helped to trust the robot and (2) whether they helped to understand how the robot worked. In addition, we asked two general questions regarding the explanations, i.e., "Have you learned anything because of the explanation?" and "Was the explanation easy to understand?". We also asked for free-form feedback. Here we wanted to know from the participants which parts of the explanation were easy/not easy to understand, whether they would have needed more/additional information and which one and why the explanation was not helpful (when participants answered the "Have you learned anything because of the explanation?"question with yes).

In addition, at the end of the pilot study we collected personal information from participants as well as their knowledge and attitudes toward AI and XAI.

3.3 Participants

In our pilot study, 20 people between 21 and 54 years (M = 29.3, SD = 7.47) participated. 11 of them were male, 9 were female. All participants had heard about the term AI, but only 9 of them had heard about XAI.

3.4 Results

Rating of Robot Performance & Likeability

To compare the variables likeability and performance between the no-failure robot and the two error conditions, we conducted paired t-tests. Here, the performance of the no-failure robot was perceived significantly higher compared to the calibration error robot, $t(19) = 9.20$, $p = < .001$, $d = 2.06$ (large effect) as well as the colour vision error robot, $t(19) = 9.11$, $p = < .001$, $d = 2.04$ (large effect).

Similar results were found for the likeability. The no-failure robot was liked significantly more compared to the calibration error robot, $t(19) = 3.66$, $p = .002$, $d = 0.8$ (large effect) as well as the colour vision error robot, $t(19) = 3.84$, $p = .001$, $d = 0.86$ (large effect).

Rating of Explanation Quality

Results for different robot errors To get a general impression of the explanation quality, we asked the participants whether they had learned something because of the explanation and whether the explanation was helpful or not. Here, we found that 14 participants stated that they had learned something from the calibration error explanation. 17 participants stated they had learned something from the computer vision error explanation.

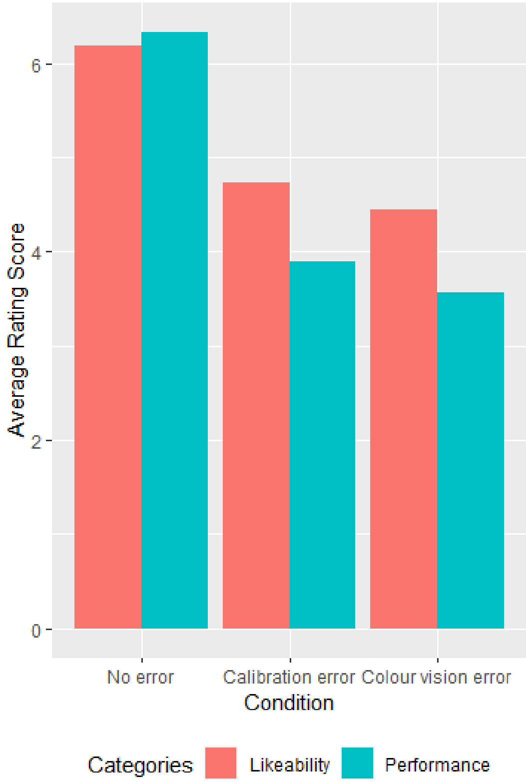


Fig. C.2: Rating of the robot in the no-error and the two error conditions.

To evaluate the explanation quality in more detail, we used two items ("help to trust or distrust the robot" and "help to understand how the robot works") proposed by [23].

Conducted paired t-test revealed that the computer vision error explanation helped more to trust or distrust the robot compared to the calibration error explanation, $t(19) = -2.77$, $p = 0.012$. For the understanding of the inner workings of the robot, no difference between the two error explanations were found, $t(19) = -0.89$, $p = 0.38$.

Besides the quantitative feedback of the participants, we also analysed the qualitative free-form feedback. Here, participants mentioned for computer vision error, that it would be helpful to add information how to solve the error (e.g., information whether the error occurred because the lightning was to dark or too bright). For the calibration error, participants mentioned that the explanation was too technical and they would have needed more information how to fix the error or how to calibrate the robot correctly to avoid similar errors in future.

Results for different explanation modalities Besides the comparison of the impact of the different robot errors on the explanation impression of the participants, we also analysed the impression of the different explanation modalities (textual vs. auditory). Here we found no significant differences between the conditions (see Table C.1).

Table C.1: No significant differences in explanation quality between the two different explanation modalities (textual vs. auditory) for both types of robot error. Trustworthy refers to "help to trust or distrust the robot", Understandable refers to "helps to understand how the robot works".

Type of Robot Error	Explanation Quality	*t*	*df*	p-value
calibration error	trustworthy	18	-0.94	.36
	understandable	18	-1.40	.18
computer vision error	trustworthy	18	-0.87	.39
	understandable	18	-0.50	.62

3.5 Discussion

From the pilot study, it became apparent that people rated the robot significantly worse in terms of its performance and likeability when it made a mistake. The general study design in terms of trust repair (comparing trust of a correct working robot and a robot who makes an error) was therefore maintained for the final study.

Based on the pilot study, it appeared that the explanation for the calibration error was too technical for end-users without experience in robotics. These resulted in significant lower trust rating and was mentioned by participants in the free-form feedback. We therefore decided to use only the computer vision error in the final study and to generate explanations for it. Inspired by the free-form feedback, we also decided to use 3 different levels of error explanation: (1) no explanation, (2) explanation of error source, (3) explanation of error source, and a possible solution. Since we did not find any huge differences regarding the modality of explanation (textual or auditory), we decided not to compare these factors in the final study. Due to better comparability, we decided to use only textual explanations.

4 Experiment

To ensure high fidelity of system communication to the participants we opted to test HRC and mistake explanation using VR, rather than implementing and testing with a real robot and projection-based AR overlays. This also increased the test rate, as we could test with multiple participants at once, the only limit being the number of VR hardware setups. Based on the results from the pilot study, where the participants asked for more solution-oriented explanations rather than technical ones, we decided to define and test different explanation levels. The *first level* is an explanation to why the robot made the error, while the *second level*, in addition, explains how to solve the problem causing the error. We compare these two levels as trust-repairing actions after a robot mistake along with a control condition, where no explanation is provided, the user is only told that the robot failed the task. Our hypotheses are as follows:

- **H1:** Providing an explanation after a robot makes a mistake will yield higher levels of trust toward the robot than providing no explanation.
- **H2:** Providing different levels of explanation after a robot makes a mistake will yield different levels of trust toward the robot.
- **H3:** Adding solution-oriented details to robot mistake explanations will yield higher operator trust than explanations without them.

4.1 Virtual Environment

The test was performed using HTC Vive VR headsets and Vive Wand 6 degrees-of-freedom controllers. The virtual environment consisted of an office environment with desks and office chairs with participants being situated in an isolated corner of the room. Within reach of the participant is a desk with the robot mounted on top. The robot is a model after the Rethink Robotics Sawyer robot. On the table is also a white square platform at either side of the robot with a little copy of the bottles involved in the test shown next to them, indicating which shapes of bottles have to be put where. The task involves sorting bottles by whether they have a round base or a square base. At startup there are four bottles on each of the platforms, two red and two blue on each, and both have one bottle of each color that does not match the shape. This means that when the test starts both the participant and the robot have to switch two bottles between the platforms to complete the shared objective. Between the two platforms is room to display text to convey instructions and explanations to the participants. The text is displayed on the surface, similarly to a projected AR overlay. The participants are able to pick up the red bottles by moving a controller withing 20 cm of their center and pressing the trigger. Letting go of the trigger releases the bottle, and they drop straight down as they cannot be thrown. In the case that a bottle is dropped on the floor, rather than requiring the participant to pick it back up, it will be moved back to its initial position. The test setup and robot in the virtual environment is shown in Figure C.3.

4.2 Procedure

After reading the experiment information and signing a consent, the participant was given instructions to how to complete the test by the test conductor. The participant was informed that they would perform a collaborative task with a virtual robot and that they would be given instructions via the text displayed on the table. It was emphasized that they should read the instructions carefully, before they were told to put on the VR headset. The participant was introduced to the task by the text display. They were told that robot was their teammate and that they were only supposed to move the red bottles while the robot moved the blue ones as they sorted the bottles

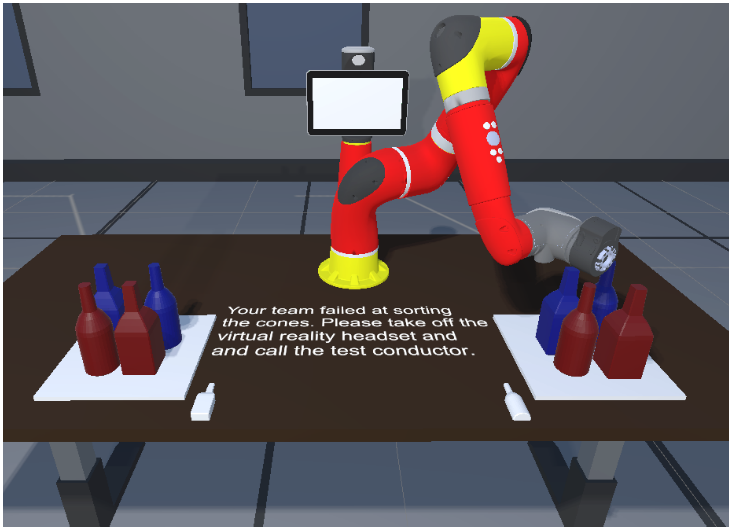


Fig. C.3: The virtual test setup featuring the robot, bottles, their platforms and indicators and the display text on the desk surface.

according to the small white bottles shown next to their white platforms. The participants proceeded through the text instruction using the Menu button at the top of the Vive wands. Before starting the task the participants were told how to move the bottles and they were told to try it.

When the participant was told to press the Menu button to start the task and they proceeded to press it, the robot would start moving the blue bottles. If the participant moved the bottles before they started the task, the bottles were moved to their starting position when the task began. The task was completed when the participant had sorted their bottles and the robot was done moving its bottles. In the first task the robot moved the bottles successfully, and the participant was presented with this message on the table: *"Your team succeeded at sorting the bottles. Please take off the virtual reality headset and and call the test conductor"*. When they took off the headset, they were presented with the 14-item version of the Schaefer HRT questionnaire [3]. Once the participant had completed the questionnaire, they were instructed to put the headset back on and follow the instructions.

Once they had put the headset on again, the display told them to start the task again by pressing the Menu button. In the second test the robot would make a mistake. Rather than switching a round-base and square-base bottles between their platforms, sorting them correctly, it would switch two round-base bottles, leaving two blue bottles in their wrong positions. The task ended once the participant had completed their half of the task correctly and the robot had stopped moving. The participant was then presented with this message displayed on the table: *"Your team failed at sorting the bottles"*. If the participant was testing the condition with no explanation of the mistake they were immediately presented with the text, *"Please take off the virtual reality headset and and call the test conductor"*. If the participant was testing the condition where they were given an explanation, they were presented with the message, *"A computer vision error occurred. The system did not successfully distinguish the bottles"*, before being told to take the headset off. Lastly, if the participant was in the condition with solution-oriented details, in addition the previously mentioned explanation they were presented with the message, *"Better lighting conditions will help with successful sorting"*, before being told to take the headset off. Once they had taken the headset off the participants was presented with another HRT questionnaire as well as additional post-test questionnaire, which they were told to fill out outside the laboratory.

4.3 Evaluation Methods

To evaluate the participants' impression during and after the VR task, we used the following scales.

Trust. During and after the VR task, we presented the 14-item version of the Schaefer HRT questionnaire [3] at the end of each task.

Explanation Satisfaction. We used the Explanation Satisfaction Scale (ESS), proposed by Hoffman et al. [23] to measure the participants' subjective satisfaction with the kind of information (no explanation, explanation, or explanation with solution) that we presented after the robot mistake.

Emotions. We used items for the subscales *anger*, *happiness*, *anxiety*, and *relaxation* of the Discrete Emotions Questionnaire (DEQ) [24] to evaluate the participants feelings after the VR task.

Self-efficacy. We used two items to measure the self-efficacy towards the robot. For this, we used a variation of the item proposed by Bernacki et al. [25] (i.e., "How confident are you that you would successfully interact with a robot like this one in the study in the future" and "How confident are you that you could solve a robot error like this one in the study in the future?").

4.4 Participants

30 participants between 21 and 31 years (M = 24.0, SD = 2.30) took part in our experiment. Of these 11 were female and 19 male. 29 of the participants had heard of the term AI, but only 4 had heard of the term XAI.

5 Results

5.1 Trust Scores

The participants answered an HRT questionnaire after completing each sorting task with the robot, the first one being successful, while in the second task the robot would make a mistake. With all data groups being parametric, performing a pair-wise t-test showed significant difference in HRT scores between the first and second task, whether no explanation ($t(15) = 5.3$, $p < .001$), the base explanation ($t(18) = 7.0$, $p < .001$) or solution-oriented explanations ($t(17) = 4.7$, $p < .001$) were provided.

However, performing a one-way ANOVA showed no significant effects of the explanations nor the type of explanation on the HRT scores after the second task ($F = (2, 27) = .23$, $p = .79$), nor on the delta of HRT scores between tasks ($F(2, 27) = .17$, $p = .84$). The average trust scores with confidence intervals are shown in Figure C.4.

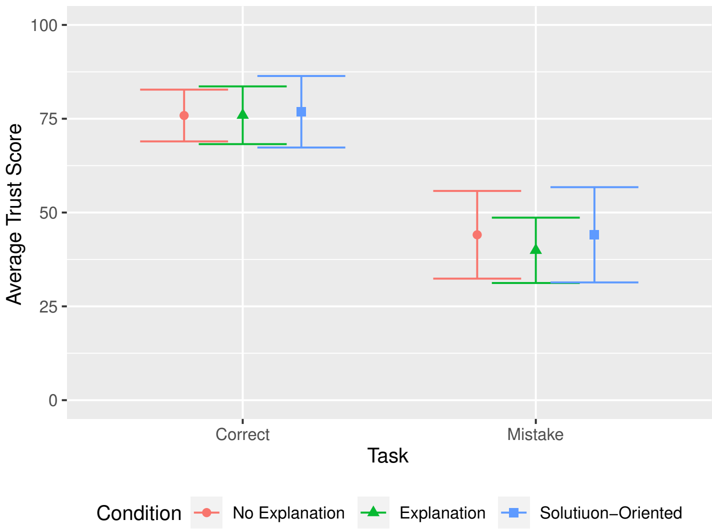


Fig. C.4: The average HRT scores and confidence intervals for the first and second HRC task between explanation conditions.

5.2 Post-Test Questionnaire

Explanation Satisfaction, Trust, and Self-efficacy

After the VR experiment, all participants answered the post questionnaire including questions about their explanation satisfaction[2], their general impression of the robot and their self-efficacy towards the robot. To evaluate these variables between the three conditions, we conducted a one-way MANOVA. Here we found a significant statistical difference, Wilks' Lambda = 0.59, $F(10, 42) = 2.86$, $p = .008$. The following ANOVA revealed that only the variable trust showed a significant differences between the conditions, $F(2, 25) = 5.92$, $p = .008$.

[2]We calculated an overall explanation satisfaction value and used in addition the item for the helpfulness of explanation to trust or distrust the robot as a single variable

To determine the direction of this difference between the three conditions, we used post-hoc comparisons[3]. We found the following differences:

- The participants' impression of helpfulness of the explanation to trust / distrust the system were significantly higher in the explanation & solution condition compared to the no explanation condition t = -3.73, p = .002, d = 1.67 (large effect).
- The participants' impression of helpfulness of the explanation to trust / distrust the system were significantly higher in the explanation condition compared to the no explanation condition t = 2.49, p = .04, d = 1.13 (large effect).

Emotional state

To evaluate possible differences in the emotional state of participants between the three conditions, we conducted a one-way MANOVA for the emotion categories happiness, anger, anxiety, and relaxation. Here we found no significant statistical difference, Wilks' Lambda = 0.84, $F(8, 46) = 0.50$, $p = .84$.

6 Discussion

6.1 Main Findings

Based on the analyses of the trust scores we have to reject all three hypotheses. While all three conditions yielded significant decreases on reported HRT based on the scales, providing explanations to the error, with or without suggested solutions, showed no significant difference in trust, suggesting no trust-repairing effect. While the ESS trust score showed that participants found the given explanations helpful to decide whether to trust or distrust the robot, this subjective impression of the participants was not reflected in their trust ratings during the VR task. Despite the effect of the helpfulness of the explanations to trust or distrust the robot, this trust can not be assumed to be transferable to trust in the robot, especially as scales for trust in automation and HRT are not interchangeable [4]. The explanations in our study also did not increase participants' self-efficacy, meaning that they did not feel more confident to interact with the robot in the future. Nevertheless, the ESS trust score can be seen as a first indicator that explanations might support trust-recovery in HRC, but that an explanation alone is not enough to recover trust after a robot-mistake, even when participants retrospectively rate the explanation as helpful. The effectiveness of explanations seems to depend on various aspects. One important variable is the scenario of the task. Compared to our VR task, Nikolaidis et al. [26] found out that in their study (a physical human-robot collaboration task), explanations greatly increased human trust to take robot's suggestions. Another important variable is the emotional presentation of the explanation. The affect in how an explanation is presented to the user plays a role in the effectiveness of the explanation [27, 28]. Affective feedback given

[3]We used the Holm correction for multiple testing to adjust the p-values for all post-hoc tests we calculated.

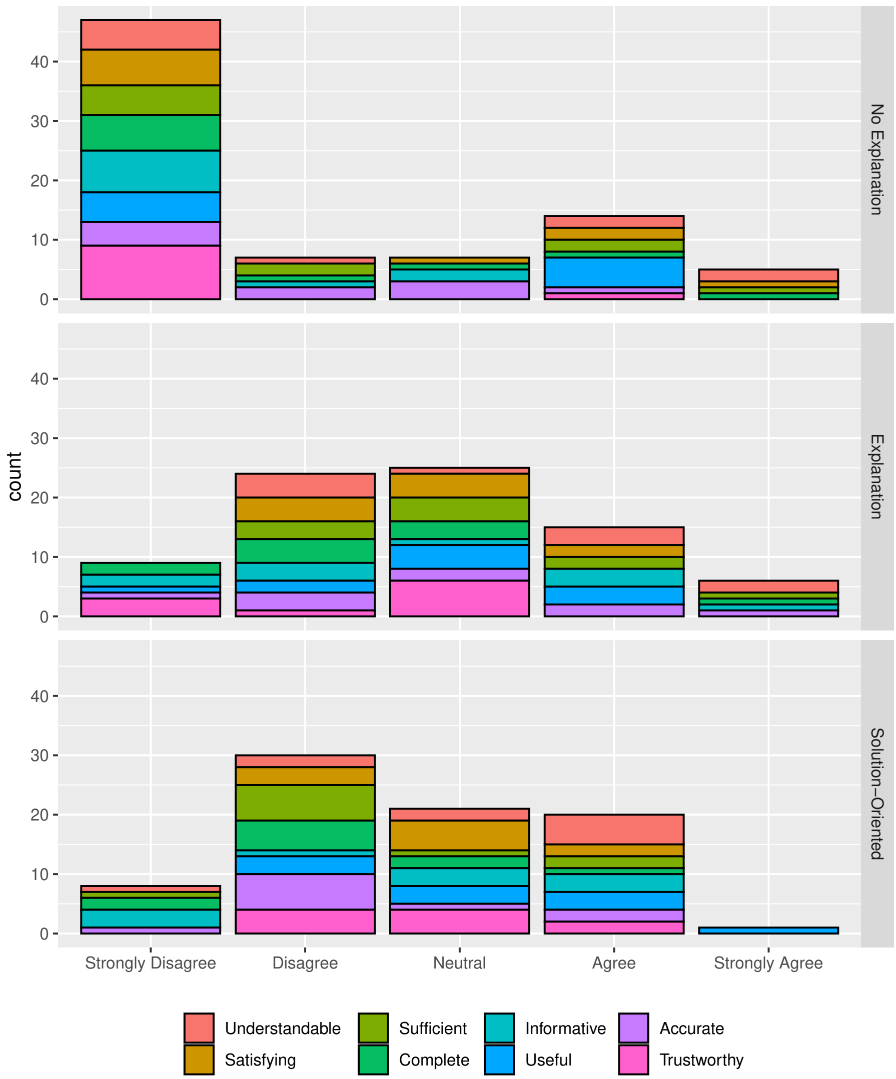


Fig. C.5: Distribution of the items of the Explanation Satisfaction Scale between the three different explanation style conditions.

by a robot leads to a more positive user impression [29, 30]. The work of Robinette et al. [31] propose that the apology of a robot after an error increases trust in the user.

6.2 Limitations

The results may have been affected by the participants' understanding of the collaborative task. Some participants seemed to have difficulty with the task, as they would often move a bottle matching the shape of the bottle moved by the robot, rather than following the instructions and sorting bottles according to the indicators on the table. The difficulty understanding the task may affect the participants' perception of the robot's mistake and the explanation by extension. If the participants do not understand the task, when told that the team failed the task, they may not think to inspect the robot's work and recognize its mistake, which can affect their perception of the explanations. Lastly, having the participants perform tasks simultaneously with the robot may affect how attentive they can be toward the robot and whether they can critically inspect the robot's work during the task. In future experiments the instructions should be clearer or the bottles should be distinguishable by more factors than their shapes while still indicating which should be moved by the robot or the participant.

6.3 Future Work

For future studies it would be valuable to explicitly ask participants about how their perception of the system communication affect their perception of the physical robot. In addition, investigating whether there is a separation between the physical robot and its operating system and communications in the participants mental model. Considering participants showed higher trust toward the explanations relative to the robot, they may consider the robot and system as two separate entities.

To make the explanations for HRC more effective, the recommendations of Kunkel et al. [32] and Weld et al. [33], among others, should be considered for further studies. Kunkel et al. [32] point out that richer explanations are preferred by users. In addition, Weld et al. [33] recommend interactive explanations. Here, the robot could be provide answers to follow-up questions and actions (e.g., giving more details, changing the vocabulary, attempting to correct the error), leading to a more social process of explanation.

7 Conclusion

We set out to investigate how we can utilize system communication and mistake explanation to maintain trust in a collaborative robot after it makes a mistake. In our conducted pilot study we found that end-users preferred less technical explanations with a greater emphasis on how to solve the error more useful. Using a VR testing environment for our main study, we evaluated three levels of explanations after the robot made a mistake in executing a shared objective in sorting a set of bottles by shape in collaboration with our participants. After comparing the conditions (no explanation, explanation of robot error, and explanation of error with solution-oriented details) with 30 participants we found no significant effects on their trust toward the robot. While participants found the explanations helpful to trust or distrust the system, we can not assume this trust to be transferable to the robot or the robot's operating system. Future studies should consider the participants' understanding of the shared task with the robot, ensuring that they recognize the nature of the robot's mistake and gain the most from the explanations. In addition, special consideration should be put into investigating the participants' mental model of the interactions between the robot, the system and the explanation system to gain understanding to which construct the trust is placed.

References

[1] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[2] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[3] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[4] T. T. Kessler, C. Larios, T. Walker, V. Yerdon, and P. Hancock, "A comparison of trust measures in human–robot interaction scenarios," in *Advances in human factors in robots and unmanned systems*. Springer, 2017, pp. 353–364.

[5] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn, "Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust," in *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2015, pp. 1–8.

[6] M. W. Boyce, J. Y. Chen, A. R. Selkowitz, and S. G. Lakhmani, "Effects of agent transparency on operator trust," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, 2015, pp. 179–180.

[7] N. Wang, D. V. Pynadath, and S. G. Hill, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 109–116.

[8] P. Kaniarasu and A. M. Steinfeld, "Effects of blame on trust in human robot interaction," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 850–855.

[9] J. Wainer, D. J. Feil-Seifer, D. A. Shell, and M. J. Mataric, "The role of physical embodiment in human-robot interaction," in *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2006, pp. 117–122.

[10] W. A. Bainbridge, J. Hart, E. S. Kim, and B. Scassellati, "The effect of presence on human-robot interaction," in *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2008, pp. 701–706.

[11] M. Duguleana, F. G. Barbuceanu, and G. Mogan, "Evaluating human-robot interaction during a manipulation experiment conducted in immersive virtual reality," in *International Conference on Virtual and Mixed Reality*. Springer, 2011, pp. 164–173.

[12] K. Stubbs, P. J. Hinds, and D. Wettergreen, "Autonomy and common ground in human-robot interaction: A field study," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 42–50, 2007.

[13] J. D. Hoffman, M. J. Patterson, J. D. Lee, Z. B. Crittendon, H. A. Stoner, B. D. Seppelt, and M. P. Linegang, "Human-Automation Collaboration in Dynamic Mission Planning: A Challenge Requiring an Ecological Approach," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 23, pp. 2482–2486, 2006.

[14] D. Gunning, "Explainable artificial intelligence (xai)," *Defense Advanced Research Projects Agency (DARPA)*, 2017.

[15] A. Heimerl, K. Weitz, T. Baur, and E. Andre, "Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–1, 5555.

[16] K. Weitz, T. Hassan, U. Schmid, and J.-U. Garbas, "Deep-learned faces of pain and emotions: Elucidating the differences of facial expressions with the help of explainable ai methods," *tm-Technisches Messen*, vol. 86, no. 7-8, pp. 404–412, 2019.

[17] T. Huber, K. Weitz, E. André, and O. Amir, "Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps," *CoRR*, vol. abs/2005.08874, 2020. [Online]. Available: https://arxiv.org/abs/2005.08874

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[19] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[20] R. Sheh, ""why did you do that?" explainable intelligent robots," in *AAAI Workshop-Technical Report*, 2017, pp. 628–634.

[21] N. Wang, D. V. Pynadath, and S. G. Hill, "The impact of pomdp-generated explanations on trust and performance in human-robot teams," in *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, 2016, pp. 997–1005.

[22] D. Holliday, S. Wilson, and S. Stumpf, "User trust in intelligent systems: A journey over time," in *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 2016, pp. 164–168.

[23] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable AI: challenges and prospects," *CoRR*, vol. abs/1812.04608, 2018.

[24] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions," *PloS one*, vol. 11, no. 8, p. e0159915, 2016.

[25] M. L. Bernacki, T. J. Nokes-Malach, and V. Aleven, "Examining self-efficacy during learning: variability and relations to behavior, performance, and learning," *Metacognition and Learning*, vol. 10, no. 1, pp. 99–117, 2015.

[26] S. Nikolaidis, M. Kwon, J. Forlizzi, and S. Srinivasa, "Planning with verbal communication for human-robot collaboration," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 3, pp. 1–21, 2018.

[27] J. Klein, Y. Moon, and R. W. Picard, "This computer responds to user frustration: Theory, design, and results," *Interacting with computers*, vol. 14, no. 2, pp. 119–140, 2002.

[28] R. W. Picard and J. Klein, "Computers that recognise and respond to user emotion: theoretical and practical implications," *Interacting with computers*, vol. 14, no. 2, pp. 141–169, 2002.

[29] H. Hastie, P. Dente, D. Küster, and A. Kappas, "Sound emblems for affective multimodal output of a robotic tutor: A perception study," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 256–260.

[30] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Modelling empathic behaviour in a robotic game companion for children: an ethnographic study in real-world settings," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, 2012, pp. 367–374.

[31] P. Robinette, A. M. Howard, and A. R. Wagner, "Timing is key for robot trust repair," in *International conference on social robotics*. Springer, 2015, pp. 574–583.

[32] J. Kunkel, T. Donkers, L. Michael, C.-M. Barbu, and J. Ziegler, "Let me explain: Impact of personal and impersonal explanations on trust in recommender systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[33] D. S. Weld and G. Bansal, "Intelligible artificial intelligence," *ArXiv e-prints, March 2018*, 2018.

Paper D

Proposing Human-Robot Trust Assessment Through Tracking Physical Apprehension Signals in Close-Proximity Human-Robot Collaboration

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper has been published in the
28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp. 1–6, 2019.

Abstract

We propose a method of human-robot trust assessment in close-proximity human-robot collaboration involving body tracking for recognition of physical signs of apprehension. We tested this by performing skeleton tracking on 30 participant while they repeated a shared task with a Sawyer robot while reporting trust between tasks. We tested different robot velocity and environment conditions with an unannounced increase in velocity midway through to provoke a dip trust. Initial analysis show significant effect for the test conditions on participant movements and reported trust as well as linear correlations between tracked signs of apprehension and reported trust.

1 Introduction

In the context of industrial repetitive manual production collaborative robots can be implemented to help relieve strenuous activities and help prevent long-term deceases [1]. However, working in close proximity to an industrial robot may cause the collaborating worker, from here on referred to as the operator, to feel unsafe, depending on the nature of the work. For this experiment we define trust towards the robot as the combination of feeling physically safe around it and a being able to predict the robot's actions in the context of the shared task.

For example, when working in industrial meat production, implementing robots can help in cutting or in flipping or positioning meat for easier operation. Because the robot will be equipped with either sharp knifes or powerful gripping tools, there is a high risk that the operator will feel unsafe working near it, decreasing efficiency.

Our long-term goal is to develop computer vision and body tracking methods that can be implemented non-obstructively in the production context and interpret the operator's trust towards the robot based on movement correlated with the robot's actions and the current task. In the future this can be used to have the robot adapt its movement patterns, velocity and secondary communication, such as information through audio-visual interfaces, according to the comfort of the operator to maintain their trust throughout the collaboration.

This paper documents an early experiment on operator posture during a human-robot collaboration task by utilizing an infrared camera for body tracking. We compare operator posture and motions and how they are affected when the robot motions are suddenly changed and whether they are different when a work surface is between the robot and operator.

2 State of the Art

Human-robot collaboration is an emerging field within HRI, as can be seen from dedicated workshops and tracks at HRI conferences [2]. A particular challenge in HRI is the asymmetrical communication capabilities of the operator and the robot, creating a need for a multi-level coordination between them to achieve successful interaction: The communication level, the physical level, the social level and the task level. The communication and physical levels are the main focuses of this study. The physical level because the motions of the robot influence the trust of the operator and the communication level because the system interpreting the operator's attitude is part of implicit communication. The task level will be relevant in later studies involving varying tasks, where this experiment involves a simple repeated task.

When describing trust in automation Lee & See [3] wrote the definition of trust as the attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability. To use this definition it requires elaboration to include the appropriateness of the trust, defined as the relationship between the true capabilities of the agent and the level of trust. In addition one must define the influence of context, the goal-related characteristics of the agent, and the the cognitive processes that govern the evolution of trust level.

Lee & See [3] proposed that humans develop trust in automation based on a combination of analytic, analog and affective processing of external information and internal believe. Analytic processes are cost-benefit analyses while analog processes refer to the assessment of category memberships. Affective processes refer to responses in confirmation or violation of implicit expectations.

In addition to trust assessment, these processes can be utilized in implicit communication with a system, which was proven to improve performance in human-computer interaction [4] [5]. Rani el at. [6] framed implicit human-robot communication as affective communication, where the affective state is interpreted by the robot system and informs the robot's next action in combination with sensed operator actions. It was shown that the operator's feeling of safety and comfort was improved if the robot is adaptive to their movements by incorporating human-aware motion planning systems.

Hancock et al. [7] performed a meta-analysis on factors affecting trust in human-robot interaction. They found the main influence to be robot characteristics, in particular robot performance, with environmental factors moderately affecting trust. With this experiment we aim to develop methods for inferring the changes in trust through reactive body postures as a result of unexpected changes to expected behaviours, thereby violating the operator's explicit expectations.

Dragan et al. [8] tested trust in robots dependant on the robot's motion patters using Hoffman's metrics for fluency in human-robot collaboration [9], where the trust in the robot is evaluated using a composite measure. Their results showed that a robot with predictable motions, rather than purely functional motions, was more accepted by the operator. Predictable motions meant that the robot would move as expected by the participant, rather than only planning to avoid collisions which would lead to non-intuitive movement patterns. The post-study questionnaires were administered after each participant had experienced all three conditions and includes questions pertaining to trust. The delay and lengths of the questionnaires may have affected the qualitative measurements. While we take inspiration from this method we aim to avoid the effect of the delay in this experiment.

Other measurements have been used for trust assessment in the context of HRI. In studying implicit human-robot communication Rani et al. [6] successfully tested an affect recognition system utilizing physiological measurement during interaction with a remote robot. Freedy et al. [10] assessed human trust by recording operator intervention in remote collaboration with an unmanned ground vehicle depending on perceived competency. While these measurements are relevant, the context of manual production work often requires the use of both hands, which does not allow for directly operating or interrupting the robot, and physiological measuring devices may prove too intrusive or obstructing for daily operations.

Body tracking and posture have previously been used to enable and secure safety in human-robot interaction, but they are rarely used as an attitude assessment tool. For ensuring safe human-robot collaboration Morato et al. [11] implemented multiple Kinect sensors while Tan & Arai [12] used a triple-camera setup. Both used computer vision for skeleton tracking with Morato et al. working with standing work while Tan & Aria focused on sedentary work. While we use a different type of depth camera for this experiment the goal is to develop methods for tracking and trust assessment for both standing and sitting operators.

3 Inferring Apprehension

We propose a method of trust assessment through tracking the operator's physical reactions, postures and movements and interpreting them as signs of apprehension in relation to the collaborating robot and the context of the shared task. The features of these movements can be categorized as by Aghajan et al. [13] for non-symbolic interpreting of gestures. These features include plane of gesture, closeness, radius, tempo, velocity, force, gesture frequency and quantity as well as constricted versus expansive and jerky versus smooth motions.

When analyzing the tracking data, signs of apprehension can be divided into two categories, those being unforeseeable changes in motion patterns and reactive motions following predictable patterns. The first category involves changes in physical movement patterns as a result of sudden changes in trust toward the robot that the system is not programmed to categorize. Recognizing these requires a period of movement in relation to the shared task that are consistent to a degree where changes can be recognized in real time. An example of this could be the operator changing hand trajectory to have a longer radius from the robot end effector as an results of unexpected changes in robot behaviour.

The category of predictable movement patters is based on human recognition of physical apprehension that the system is programmed to recognize. Examples of these would be changing proximity to the robot, leaning back or retracting hands to avoid expected danger. Additionally, reactions can involve hesitation in proceeding with the task. The analysis for this experiment is based on measuring this category of reactions. Specifically, we are looking at changes in proximity from the robot for the tracked points at the participant' head, spine, hands and elbows, as well as changes in task completion time.

4 Evaluation

The purpose of the experiment is to induce varying levels of trust towards the robot by affecting the operator's feeling of safety. In a collaborative task we measure the operator's motion patterns as a robot arm approaches them at different velocities. We test the velocities in three conditions, with and without a work-surface between the robot and operator, making a total of six test conditions. These conditions are tested between subjects. The hypotheses of the experiment are:

- **H1:** The participants' reported trust toward the robot is significantly affected by the velocity of the robot arm's movements.
- **H2:** The participants' reported trust toward the robot is significantly affected by having a work surface between the participants and the robot.
- **H3:** The participants' movements and posture are significantly affected by the velocity of the robot arm's movements.
- **H4:** The participants posture are significantly affected by having a work surface between the participants and the robot.
- **H5:** The participants' movement patterns correlate with their reported levels of trust toward the robot at the velocity and work surface conditions.

4.1 Hardware & Software

The robot used in the experiment is a Sawyer by Rethink Robotics with a set of pneumatic grippers and motions are planned using Intera Studio. The Sawyer robot features animated eyes on a screen during operations by default. These are disabled during the experiment to limit the participants' perception of the robot to its movement patterns. The robot is equipped with a red emergency shutdown button, which is positioned within reach of both the test participant and the test conductor.

The participants' movements are tracked and recorded using Unity, which is also used to control the robot. They are synchronized using a TCP/IP connection. The depth camera is an Orbbec Astra infrared camera which uses the Orbbec Body Tracking SDK. It is mounted on a light stand in front of the participant at a height allowing it to consistently track their upper body.

4.2 Procedure

After signing a consent form the participant is introduced to the robot and the task. The participant is directed to a rectangular area marked on the floor with a size that allows comfortably standing and keeping the participants at a consistent distance to the robot. To prevent participants being startled at the first task, the robot motion is shown once at its lowest velocity.

The task is for the participant to take a wooden baton that is handed to them by the robot. The baton is handed to the robot by the test conductor before the robot hands it over to the participant by moving straight toward them, stopping at a distance where they can grab it with their elbow near their side. This is shown in Figure D.1. The batons used for the test are 25 cm long and have a diameter of 2 cm. The robot holds the baton with a set of pneumatic grippers and the participant retrieves it by pulling it toward themselves, applying at least 20 N of force, once the robot is in its waiting position, after which it returns to the conductor.

The participants are instructed that each time they take a baton they have to turn to their right and put it in a box on a table next to them. This serves to direct their attention to a tablet on the table where they have to report their attitude toward the robot after each task. Inspired by the questionnaires by Dragan et al. [8], the participants are instructed to rate the robot in three categories, by stating agreement to a statement on a scale between strongly agree or strongly disagree. The three statements are:

- I trusted the robot to do the right thing at the right time.
- I felt safe working next to the robot.
- The robot's reaching motion was surprising.

The participant can report their agreement on a linear scale by freely moving or tapping on a slider on the touchscreen. The goal of this is to allow quick responses on a more granular level than on, for example, a Likert scale, hopefully yielding more representative results. The test conductor will manually command the robot to move to the participant when they have return to the marked zone on the floor and are facing the robot.

We compare two independent variable. The first is whether there is a work surface in the form of a table between the participant the robot. The table is positioned such that the end position for the robot gripper is above the table. The second variable is the starting velocity of the robot's motions. This variable has three conditions, each with their own starting velocity. These being 25, 50 and 75 percent of the robot's max velocity, labelled slow, medium and fast velocity, respectively.

The conditions are compared between subjects. Half of the participants will perform the test with the table and half without. Each velocity condition is performed with five test subjects with and without work surface, making a total of 30 participants. Each test involves twenty tasks in total, the first ten being at the starting velocity. In order to provoke the participant's trust level, after the tenth task the robot movement velocity is increased with 25 percentage points without warning. With this approach we expect to see an increasing level of trust throughout the first ten tasks with a change in posture and attitude at the velocity change, followed by readjustment to the robot. The conditions are tested between subjects to avoid participants anticipating what will happen. Throughout the entire test the depth camera is recording skeleton tracking data at 30 samples per second, which is logged along with a label for the robot's current state.

5 Results

The experiment was performed with 30 participants at the ages between 21 and 28 at an average of 24. Participants included 22 men and 8 women, 27 right-handed and 3 left-handed.

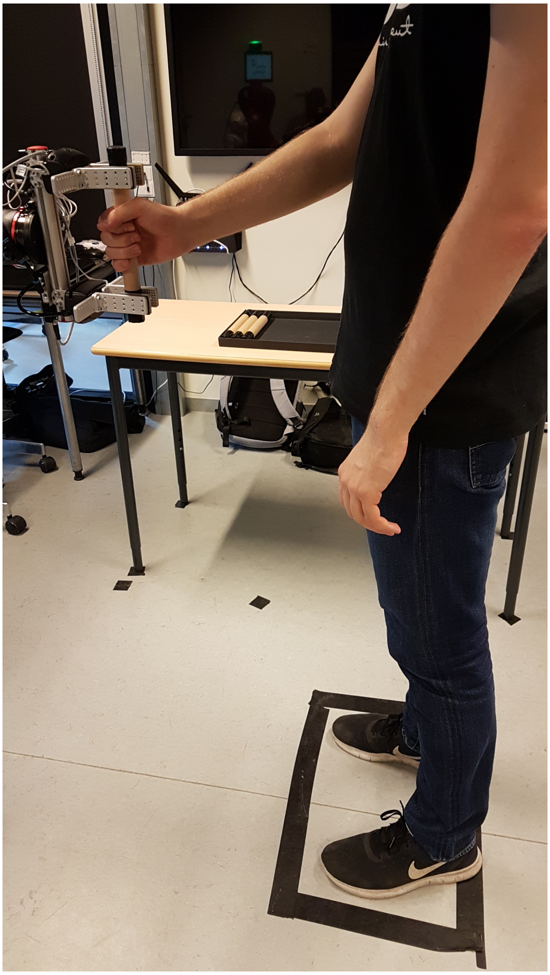


Fig. D.1: A test participant standing with the marked zone on the floor, grabbing the baton from the robot in the condition with no table in-between.

The questionnaire results are aggregated by inverting the third question to weigh positively toward robot trust. Figure D.2 shows the average answers for all participants between tasks and conditions. The vertical lines marks the step up in velocity. We see gradual increase in trust for the first ten task for all groups with dips after the velocity increase being most pronounced in the slow velocity group.

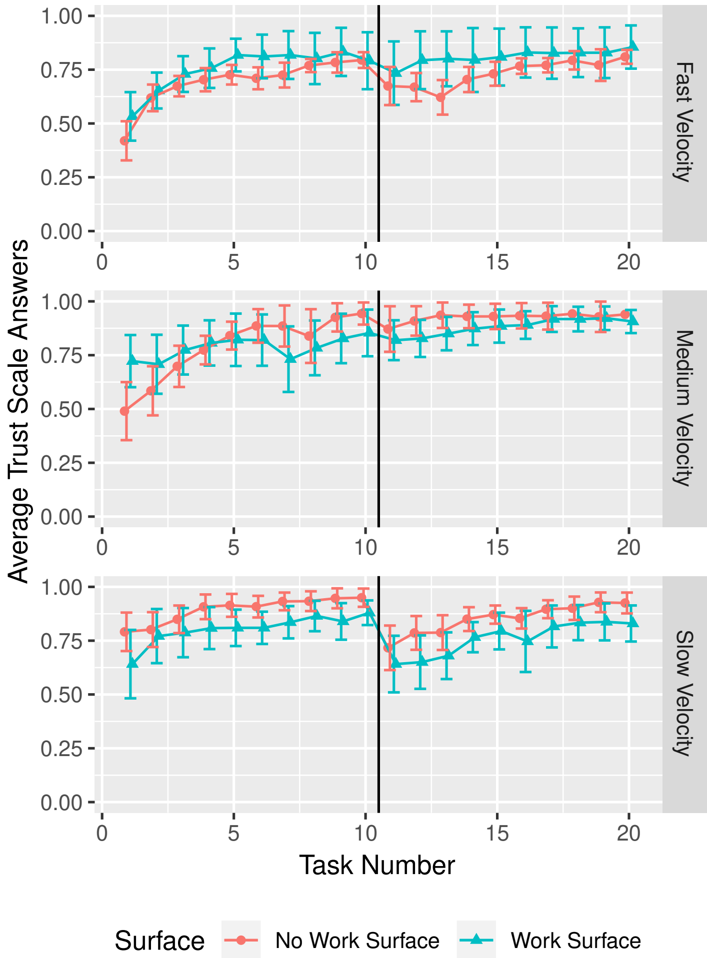


Fig. D.2: Average and confidence intervals for the questionnaire answers of each task between the test groups with a line marking the change in robot movement velocity.

To test H1 and H2, and evaluate the trust reporting method, we perform a multivariate analysis of variance with significance threshold at .05 on the three answers, treating the task number and whether a task is before or after the step up as a within subject variables and the velocity and work surface conditions as between subject variables. Keep in mind that these velocity groups include the initial velocity and the stepped-up velocity, meaning that every participant experience two robot velocity settings. Performing the test for all twenty tasks shows significant differences depending on velocity condition ($F(2,572) = 17.13$, $p < .001$), work surface ($F(1,572) = 4.64$, $p < .01$), task number ($F(1,572) = 2.72$, $p < .05$) as well as the two halves of the test ($F(1,572) = 7.60$, $p < .001$) as well as interactions between several factors.

In order to test the gradual increase in participant trust throughout the first tasks visible in Figure D.2, we repeat the analysis on the first ten tasks only. This shows significant difference between task numbers ($F(1,286) = 2.84$, $p < .05$). Performing an analysis of variance on the three answers separately shows that task number is significant for two out of three; trusting the robot to do the right thing and the right time ($F(1,286) = 3.99$, $p < .05$) and on the robot's motion being surprising ($F(1,286) = 7.54$, $p < .01$). Performing the test on the final ten tasks yields similar results with significant difference between tasks ($F(2,286) = 7.75$, $p < .001$), which is also true when testing the answers separately. This suggests that the participants found the robot more predictable as the experiment went on with a dip in predictability in the middle followed another gradual increase, as would be expected.

Repeating the analysis, while focusing on the difference in before and after the step up in velocity, only the interaction between the velocity and task number ($F(2,46) = 2.58$, $p < .05$) show statistical significance, and testing the three questions separately shows that task number ($F(1,46) = 7.48$, $p < .01$) and interaction between task number and velocity ($F(2,46) = 3.79$, $p < .05$) only yield significant difference in regards to the robot motions being surprising.

The significant effect of the test conditions suggests we can retain H1 and H2. Additionally, the trends regarding gradual gain in trust throughout the tests lends confidence in this method of measuring human-robot trust.

For testing H3 and H4 we perform a similar multivariate analysis of variance while adding participant handedness as a between subject variable. We are specifically testing delta motions towards or away from the robot within the first two seconds after the robot starts moving towards the participant, measured 30 samples per seconds. Running the analysis across all tasks and tracked bones shows significant statistical difference for velocity conditions ($F(2,35167) = 24.66$, $p < .001$), work surface conditions ($F(1,35167) = 7.43$, $p < .001$), task number ($F(1,35167) = 4.64$, $p < .001$), first and second half of the test ($F(1,35167) = 2.00$, $p < .05$) and handedness ($F(1,35167) = 20.54$, $p < .001$) in addition to the interaction between multiple factors.

Testing the effect of the step up in velocity by testing the tracked motions for all tracked bones before and after showed no significant difference between tasks, suggesting no significant gradual changes. Similarly, neither testing the first and last ten tasks alone showed no significant difference as tasks proceeded. However, aiming to remove noise, limiting the multivariate analysis to only head and hands tracking show significant difference for the tasks before and after the step up in velocity ($F(1,3506) = 4.73$, $p < .01$), while there is still no significant difference between tasks throughout the first and latter half of the test. The average movement toward and away from the robot throughout the tests is illustrated in Figure D.3.

In addition to tracking motions we analyze the time it takes for the participants to grab the baton after the robot has arrived in order to gauge their reactions. Testing across all tasks show significant statistical difference for task number ($F(1,571) = 20.37$,

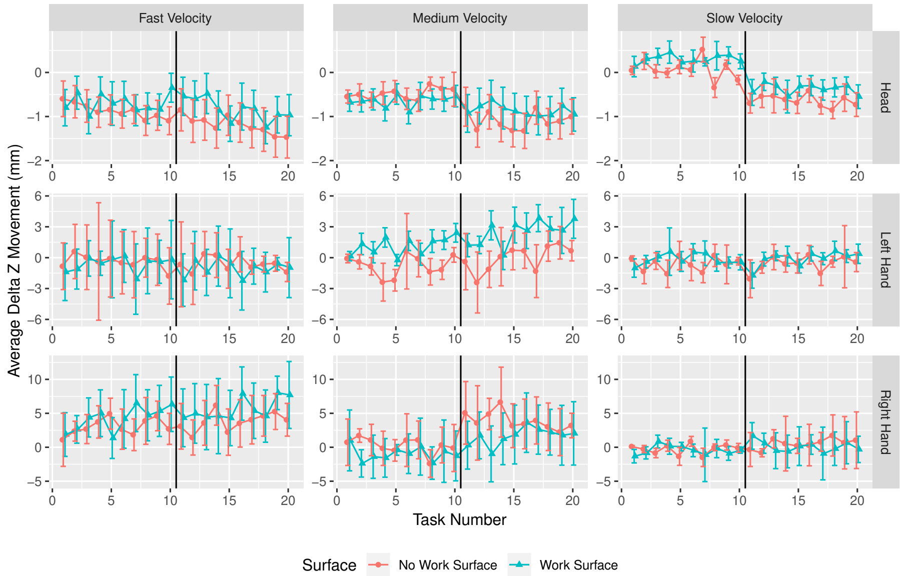


Fig. D.3: Average delta motion towards and away from the robot of the participants' head and hands for the first two seconds after the robot starts moving, negative being away from the robot. Includes confidence intervals.

$p < .001$), the first and second half of the test ($F(1,571) = 4.67$, $p < .05$), the interaction between velocity conditions and work surface ($F(1,571) = 34.88$, $p < .001$) and between velocity conditions and task number ($F(1,571) = 5.49$, $p < .01$). Comparing times before and after the step up in velocity showed no significant difference, while task number were statistically significant through the first half of the test ($F(1,286) = 13.83$, $p < .001$), but not in the latter half. The average times to grab the batons are shown in Figure D.4. The analysis results indicate that we can retain H3 and H4 since the conditions have significant effect on the movements and reaction of the participants.

To test H5 we perform Pearson's product-moment correlation pair-wise for linear correlation between the three trust questions and the delta bone positions toward and away from the robot within the first two seconds of the robot moving towards the participant. For the head motion there is only correlation for the first question ($p < .001$), while the movement of the non-dominant hand correlates to all three questions ($p < .05$, $p < .01$, $p < .0001$), respectively, and movement of the dominant hand correlates with none. This can be explained by the participants reaching out for the baton in different ways, while leaving the non-dominant hand at the side. Also, with the participant anticipating the baton they are likely to engage in the reaching motion in spite of the level of trust, leaving the opposite side of the body to move with apprehension. The correlation between the reported trust indicators and the participant movements suggest we can retain H5.

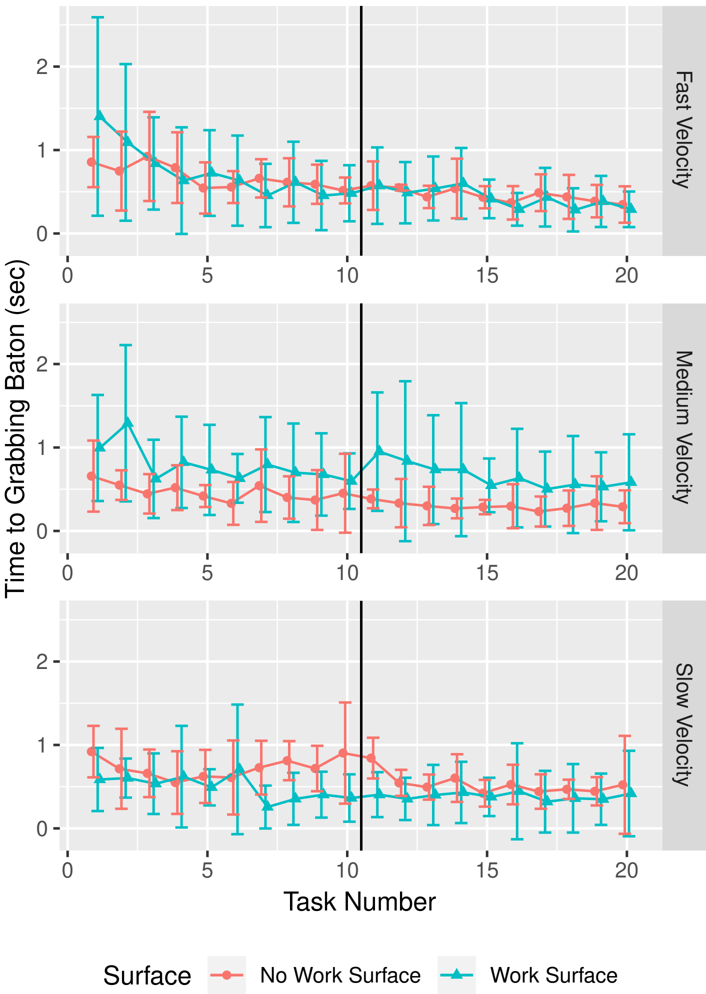


Fig. D.4: The average time and confidence intervals for participants to retrieve the baton after the robot finished moving in each task between the test groups with a line marking the change in robot movement velocity.

6 Discussion

Despite the results showing significant effects of the test conditions on the reported trust indicators and the participants' reactive movements, the extend is only that they have an effect, rather than the nature of the effects, specifically. Extended machine learning analysis including all tracking points is needed to fully interpret the data and develop predictive measures of human-robot trust. Also, in regards to H5, even though the analysis showed significant linear correlation in some instances, further analysis is necessary to determine the best fit, whether it be exponential, logarithmic or otherwise.

The questionnaire results roughly ranging from 0.5 to 1, as shown in Figure D.2, suggests that a more aggressive test design may be beneficial to provoke more mistrust in the robot and get a wider range of data. This could involve the robot performing operation between the operator's hands rather than just moving toward their body, lowering the proximity to not just the hands, but the head as well. This may require Ethic's Board approval, seeing as it may be a risk to test participants in case of sudden movements from either the operator or robot causing dangerous collisions.

The novel method of trust assessment through quick reporting between each task will benefit from further verification. Even though the results reflect what would be expected from the test conditions it would benefit from being correlated with additional measurements, such as galvanic skin response or facial expression analysis. Additionally, initial reported trust being around 0.5 may be a side effect or the default slider value being in the center. Alternatively, it may stem from the participant wanting to start the test with values in the middle of the scale to have space for decreases and increases.

Reported trust level will also be highly affected by the context of the experiment as well as the nature of the questions. Because the participants are asked about their trust of the robot, they are very likely to expect a break in that trust. We aimed to counter this with the continuous scale rather than a segmented scale, like the Likert scale, resetting the slider between tasks and requiring each slider to be moved before they can be confirmed, preventing participants from reporting the same value until they knew they noticed a change in robot behaviour.

7 Conclusion

We propose a novel method of human-robot trust assessment through tracking of physical apprehension signal in the context of human-robot collaboration. We test this by performing skeleton tracking during a repeated human-robot collaboration task where we violate the participants expectations of the robot's behaviour midway through, aiming to provoke a drop in robot trust. We test this at varying robot movement velocities as well as with and without a work surface between the robot and operator.

Though the method of assessing operator trust will benefit from further verification, the results support the hypotheses. Participants reported relative levels of trust as expected from the test conditions and initial analyses show correlation between reported trust and reactive movement while collaborating with the robot.

While there are significant effects on participant movement from robot velocity and the presence or absence of a work surface, deeper analyses are required to develop methods of concretely interpreting human-robot trust through physical apprehension with the future goal of real-time trust assessment in human-robot collaboration.

References

[1] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of Electromyography and Kinesiology*, vol. 14, no. 1, pp. 13–23, feb 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641103001251?via{%}3Dihub

[2] B. Hayes, I. Kruijff-Korbayova, M. Sierhuis, J. A. Shah, B. Scassellati, M. C. Gombolay, M. F. Jung, K. Hindriks, J. de Greeff, C. Jonker, M. Neerincx, J. M. Bradshaw, and M. Johnson, "HRI Workshop on Human-Robot Teaming," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts - HRI'15 Extended Abstracts*. New York, New York, USA: ACM Press, 2015, pp. 255–256. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2701973.2714396

[3] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, jan 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50{_}30392

[4] J. Greenstein and M. Revesman, "Two Simulation Studies Investigating Means of Human-Computer Communication for Dynamic Task Allocation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 5, pp. 726–730, sep 1986. [Online]. Available: http://ieeexplore.ieee.org/document/4075638/

[5] P. A. Lasota and J. A. Shah, "Analyzing the Effects of Human-Aware Motion Planning on Close-Proximity Human–Robot Collaboration," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 1, pp. 21–33, feb 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720814565188

[6] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, jan 2004. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0263574703005319

[7] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[8] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," in *Proceedings of the Tenth Annual*

ACM/IEEE International Conference on Human-Robot Interaction - HRI '15. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[9] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," Tech. Rep., 2013. [Online]. Available: http://guyhoffman.com/publications/HoffmanRSS13Workshop.pdf

[10] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *2007 International Symposium on Collaborative Technologies and Systems*. IEEE, may 2007, pp. 106–114. [Online]. Available: http://ieeexplore.ieee.org/document/4621745/

[11] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-Time Human Tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, p. 011006, jan 2014. [Online]. Available: http://computingengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4025810

[12] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, may 2011, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/5942335/

[13] H. Aghajan, J. C. Augusto, and R. L.-C. Delgado, *Human-centric interfaces for ambient intelligence*. Academic Press, 2009.

Paper E

Human-Robot Trust Assessment Using Motion Tracking & Galvanic Skin Response

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper has been published in the
2020 IEEE/RSJ International Conference on Intelligent Robots and SystemsIEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

Abstract

In this study we set out to design a computer vision-based system to assess human-robot trust in real time during close-proximity human-robot collaboration. This paper presents the setup and hardware for an augmented reality-enabled human-robot collaboration cell as well as a method of measuring operator proximity using an infrared camera. We tested this setup as a tool for assessing trust through physical apprehension signals in a collaborative drawing task, where participants hold a piece of paper on a table while the robot draws between their hands. Midway through the test we attempt to induce a decrease in trust with an unexpected change in robot speed and evaluate subject motions along with self-reported trust and emotional arousal through galvanic skin response. After performing the experiment with forty participants, we found that reported trust was significantly affected when robot movement speed was increased. The galvanic skin response measurement were not significantly different between the test conditions. The motion tracking method used in this study did not suggest that subjects' motions were significantly affected by the decrease in trust.

1 Introduction

Human-robot collaboration (HRC) is becoming increasingly relevant as we are seeing the long-term consequences of repetitive manual production work [1], such as industrial meat production, in which the staff are especially affected [2]. Musculoskeletal diseases not only affects the quality of life for those suffering, but also makes production work less attractive for potential staff. We are working towards enabling close-proximity HRC, allowing the robot to relieve the collaborating worker, from here referred to as the operator, of heavy and repetitive actions while keeping the operator safe and feeling secure. To this end we are researching methods of assessing the operator's level of trust towards the robot partner in a non-obtrusive way in real-time.

We use the same definition of HRC as Herrmann and Leonards [3], where the robot and the operator works on the same component at the same time. In an HRC task where the operator has to use both hands for manipulation of objects, the system needs hands-free and non-obstructive methods of human-robot interaction (HRI). The goal is to develop a HRC cell, a collaborative human-robot work space, that enables both implicit and explicit communication from the operator to the robot using computer vision while displaying task-relevant information to the user with augmented reality (AR). Specifically, in this report we propose and evaluate a method of measuring and recording the operator's proximity to the robot with a depth camera setup with a very small footprint, allowing the robot system to adapt accordingly. We aim to develop methods for inferring the changes in operator trust through reactive body postures as physical apprehension signals from changes in the robot's behaviours as a violation of the operator's explicit expectations. The full prototype and test setup is shown in Figure E.1. The long-term goal is to develop non-obstructive solution that will fit into a production context, allowing robot system to interpret the operator's trust towards the robot based on proximity tracking. Using cross-referencing with the current shared objective, the aim is to have the system use the information to adapt to

Fig. E.1: Full HRC cell setup with Sawyer robot, projectors and infrared camera.

the operator by adjusting movement patterns or secondary communications methods, such as AR.

2 Background

While body and posture tracking have been used and assessed in enabling safe HRC, it is rarely utilized for real-time trust assessment. Morato et al. [4] used a setup of multiple Kinect sensors for ensuring safe HRC in an environment for standing work while Tan & Arai [5] used a triple-camera setup for sedentary HRC. Both used skeleton tracking algorithms. Similarly, Hald et al. [6] used skeleton tracking for proximity tracking and trust assessment, showing correlation between user proximity and attitude towards the robot. In order to limit the physical footprint of our setup to allow for close rows of HRC cells we use a single depth camera pointed downwards, not allowing for effective skeleton tracking. Our long-term goal is to develop methods for tracking and for both standing and sedentary HRC.

Lee & See [7] defined trust in automation as the expectation that the agent will achieve their goal in a situation characterized by uncertainty and vulnerability. This definition requires elaboration in order to include whether the trust is appropriate, which is derived from the relationship between the capabilities of the agent and the level of trust. Additionally, we have to consider the influence of the automation context as well as the goal-related characteristics of the agent. Lee & See proposed that trust in automation is created through a combination of analytic, analog and affective processes of external information and internal believe. We use these characteristics of trust in automation to define human-robot trust in HRC. We focus on trust as a the attitude towards the robot in the moment, rather than necessarily as a results of long-term interaction.

This research is relevant as enabling implicit communication has been shown to improve human-robot interaction [8] [9]. Also, Rani et at. [10] showed that human-aware motion planning systems improve the feeling of safety and comfort when used to adapt to the operator.

In a meta-analysis, Hancock et al. [11] found that trust factors in HRI were mainly influenced by the robot's characteristics, in particular its performance, while environmental factors had moderate effect. Dragan et al. [12] tested the effect of a robot's motion pattern on human trust using Hoffman's metrics for fluency in human-robot collaboration [13] and found that predictable motions were more accepted by the operator than purely functional motions. Hoffman's metrics were collected using a post-test questionnaire after the participants had been through all three conditions which includes questions pertaining to trust. In order to avoid the potential effects of delaying the questionnaire, we derived a shorter version of the trust metrics to be administered throughout the experiment.

Rani et al. [10] successfully used physiological measurement in an affect recognition system in the context of interacting with a remote robot. While physiological measurement might prove intrusive during daily operations, for our experiment we will use them to help verify our assessment methods.

3 Proximity Tracking

The setup for the human-robot collaboration cell is shown in Figure E.1 and consists of a roughly two meter by two meter aluminum rig equipped with two projectors and an infrared (IR) camera. A seat, a work surface and the robot are arranged in the center of the rig. The dual-projection setup, with projectors positioned at either side and at a roughly forty-five degree angular offset from the work surface, enables projection-based AR as an output modality and is hard to fully occlude when reaching across the surface, as long as the projections are calibrated to match. The IR camera, an Orbbec Astra, specifically, is mounted at the top-center of the rig and pointed downwards toward the user. These types of camera allow for skeleton tracking, but this requires a lot of distance while facing the front of the user. This is problematic in this setup with the user seated and facing the robot, so we have designed the cell to rely on the IR images only. The camera has a resolution of 320 by 240, but the outer edges never receive light, leaving the useful resolution at around 277 by 213.

3.1 Data Processing

In order to infer the proximity and posture of the user we make an aggregate of the IR camera frame to see how they reflect the light along the vertical axis of the frame. Examples of the infrared frames are shown in Figure E.2 where a user is shown sitting upright and leaning back.

The first step of processing the frame is saving an empty background average of the work environment. The per-line average is from all the non-zero pixels in the line, as black pixels are areas with no reflected light and are considered noise. Subtracting

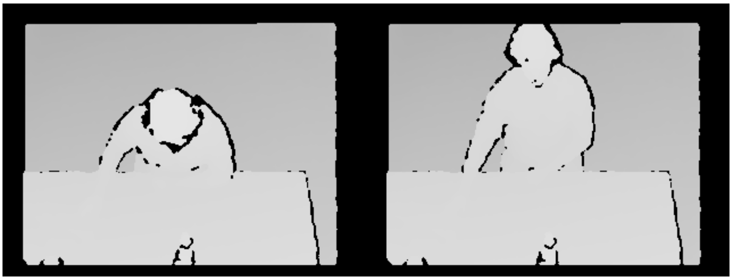


Fig. E.2: Images from the top-down IR camera. Left: User sitting upright. Right: User leaning back.

the background averages from the per-line averages with user in the frame shows how much closer the user is to the camera than the background. Examples of these measurements are shown in Figure E.3. In the example there is a visible difference in overall magnitude and spread, which can be used to infer the proximity and posture of the user by looking at the bounds and peaks of the curve. However, as can be seen in the examples, having the work surface so close to the camera compared to the floor makes the measurements less sensitive in that part of the frame as the difference, such as from the user's arms, is lowered.

4 Experiment

The objective of the experiments is to evaluate the top-down proximity tracking system as a tool to infer operator trust during close-proximity collaboration. To do this we have designed a collaborative drawing task in which the operator has to position a piece of A4 paper in a space on a table in front of the robot which is marked using the AR rig. The operator's role is to hold the paper down to the table as the robot, equipped with a felt pen in a 3D printed mount, moves in and draws a square on the paper, between the operator's hands. Midway through the experiment the robot's movement speed is changed without warning, changing pattern in order to provoke a decrease in trust. The aim is to determine if the operators proximity correlates with their trust towards the robot, which is assessed from participant arousal inferred from galvanic skin response (GSR) and with self-reporting using a questionnaire. In addition, we investigate whether the measures are affected by the operator's ability to hear the robot, as different movement speeds produce different motor noise, which may affect the operator's perception of the robot. For assessing the proximity tracking we test the following eight hypotheses:

- **H1:** The participants' reported trust toward the robot is significantly affected by changing velocity of the robot arm's movements and whether they are wearing ear protection.
- **H2:** The participants' movements are significantly affected by changing velocity of the robot arm's movements and whether they are wearing ear protection.

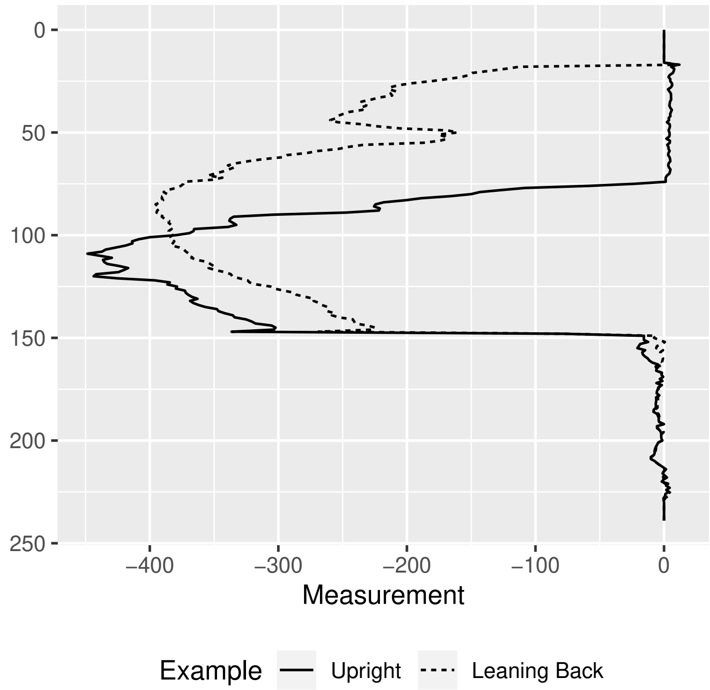


Fig. E.3: Depth measurements aggregated from the top-down IR camera with background subtraction of a user, sitting upright and leaning backwards.

- **H3:** The participants' GSR response is significantly affected by the velocity of the robot arm's movements and whether they are wearing headphones.
- **H4:** The participants' movements and proximity to the robot correlate with their reported levels of trust toward the robot.

4.1 Procedure

At the beginning of the test the participants are presented with a printed consent form and description of the experiment. After signing the consent form the participants are sat at a table with the Sawyer robot facing them. They are then introduced to the task procedure: First they have to take a piece of paper from a stack on their left, which they position in a marked space on the table. Once the paper is positioned and they hold it down, the robot moves from its resting position and draws a square on the paper. This motion is activated manually by the test conductor. This is demonstrated during the introduction using the robot speed the participants starts with. This is to help the participant feel comfortable at the beginning of the test. In addition to marking the area for the paper, the AR rig is also used to show the lines that the robot is going to draw. This is done with projected red lines, which the robot will draw over. This is to help inform the participants of the current status of the robot and when they can safely let go of the paper. Figure E.4 shows a participants sitting in front of the robot, holding the paper as the robot draws along the projected red line.

Fig. E.4: A participant sitting in front of the robot, holding the paper down. The paper is held in the area marked with a projected white rectangle, fitting an A4 sheet of paper. The robots holds the felt pen in a custom-designed 3D printed mount that is fitted with foam. The robot draws along the red line which is projected on the paper. To the participant's left is the stack of paper they can grab from and the tablet used to answer the questionnaire throughout the test.

Once the drawing is complete, the participant has to put the paper off to their right, after which they have to report their attitude and trust towards the robot using a questionnaire on a tablet to their left. The participants are instructed to state agreement to three statement on a scale between strongly disagree or strongly agree on sliding scales, yielding scores between 0 and 1, respectively. Based on the Hoffman's metrics on human-robot trust [13] the three statements are:

- I trusted the robot to do the right thing at the right time.
- I felt safe working next to the robot.
- The robot's reaching motion was surprising.

The participants are told to grab a new piece of paper and repeat the task until the test is over. Before the test, electrodes are attached to the back of the participants shoulder, opposite their dominant hand to limit disturbances, for measuring GSR with a Bluetooth-enabled device strapped to their upper arm. The GRS device infers the level of arousal in the participant by measuring the electric conductivity across the skin between the attached electrodes.

The task is repeated a total of twenty times, and after the first ten repetitions the robot movements speed is changed. In order to determine if the participants reacts to an increase in speed or rather to just a change in speed, half the participants start at a slow speed while the other start at high. The low speed is at a ratio of 0.2 of the robot's highest speed and the high speed is a ratio of 0.4. The test conditions, whether the participant is wearing ear and the robot's beginning speed, are counter-balanced with ten participants for each combination of conditions, leading to a total of forty participants.

5 Results

A total of forty subjects participated in the experiment. 34 were male and 6 were female, 34 were right-handed, 6 were left-handed, and ages ranged from 21 to 28 age with and average age of 23.5 years.

Figure E.5 shows the aggregated questionnaire for each task, grouped by condition along with confidence interval. The aggregates are the average responses to the three questions, each answered on a scale between 0 and 1, where the weight of the last question is inverted, so that that participant disagreement with the statement that the robot's motion was surprising will counts positively towards trust. The vertical lines marks the midway through the tests where the speed ratio was either increased or decreased. For the groups that started with the slow speed ratio we see that reported trust started high throughout the first half of the test, followed by a dip in trust when the speed was increased, as we would expect. The reported trust is then gradually recovered throughout the later half of the test. These effects are recognizable, though less pronounced, for the participants who started with high speed with a speed decrease in the middle. Still, this groups starts with lower trust towards the robot, which gradually builds up towards the halfway point.

Running the Shapiro-Wilk test on the reported trust scores, the tracking data and the GSR revealed that most data groupings are not normally distributed. As such, all the data is treated as non-parametric. To evaluate hypothesis one we run a Wilcoxon rank sum test on the reported trust scores immediately before and after the change in robot movement speed. The test shows significant difference for participants for whom speed was increase, both with ($W = 88$, $p < .01$) and without ear protection ($W = 100$, $p < .01$), while not significant for participants who experienced a decrease in speed.

To compare the robot start speed and ear protection conditions we compare the pairs of trust scores for the tenth and eleventh tasks, separately. The Wilcoxon rank sum test showed significant difference in trust scores between the speed conditions after the speed change, both with ($W = 11$, $p < .01$) and without ear protection ($W = 12$, $p < .01$), while there were no significant differences before the speed change. There were no significant effects from the ear protection in any condition. From these results we can retain hypothesis one in that unforeseen changes in robot movement speed affects reported trust, but only for increases in speed, and the ability to hear the robot motors has insignificant effect.

The participants' movement and proximity between the conditions are shown in Figure E.6 and Figure E.7. The proximity is measured as the position of the participant's highest point, usually the top of the head, along the vertical axis of the depth image, measured in pixels. We measure the participants' movement reactions to the robot as the delta changes in proximity within the first second of the robot moving to draw. Figure E.6 shows the average delta movement among participants for each task, showing whether the participants overall moved away or towards the robot.

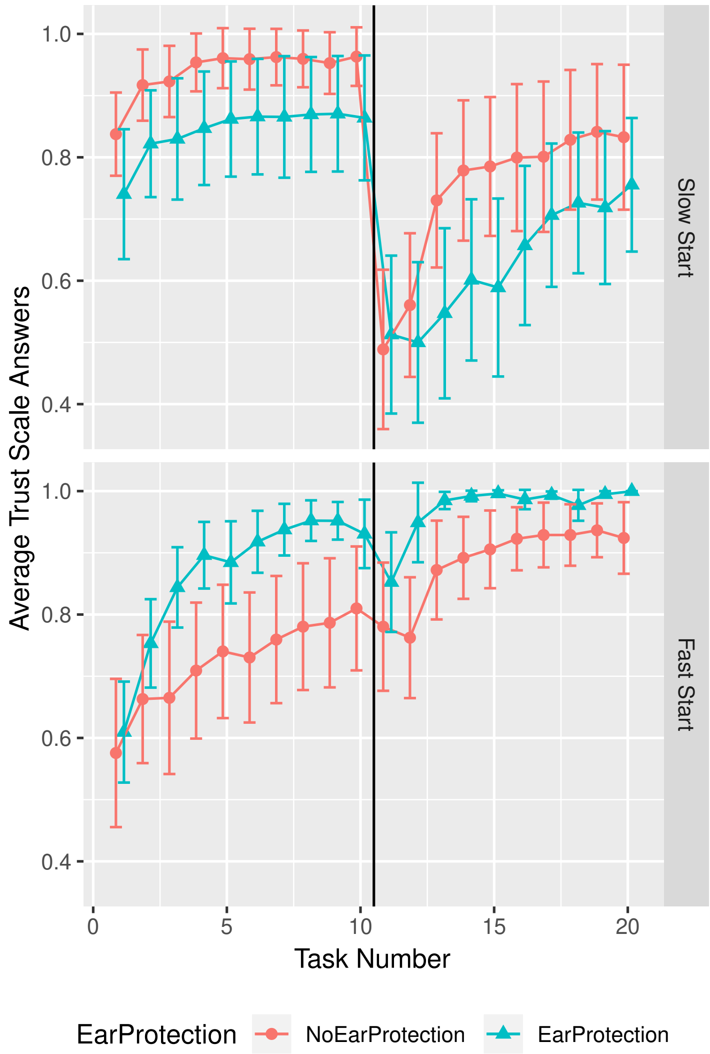


Fig. E.5: Average reported trust as aggregated from the questionnaire answers throughout the test along with confidence interval.

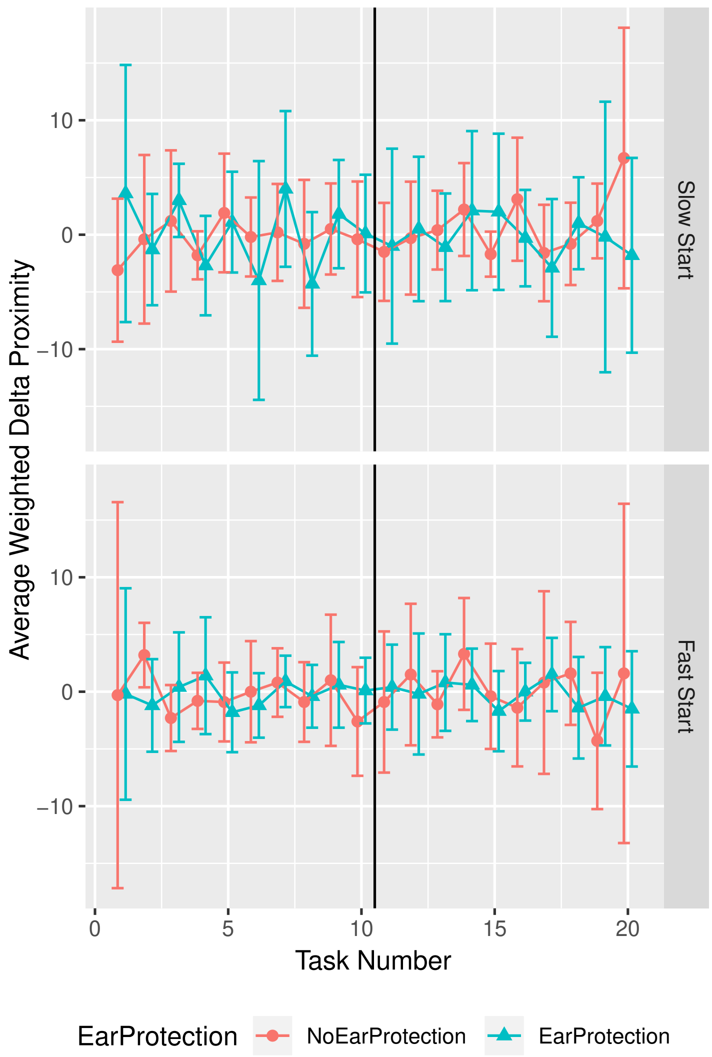


Fig. E.6: Average participant motion for the first second of robot movement weighted by movement direction along with confidence interval. The vertical lines mark the midway point of the test where the robot speed ratio changes.

Figure E.7 shows the average absolute. The difference is that this is the total movement of the participant, regardless of whether they are moving away or towards the robot, showing how much the participants move in general. Looking at both Figures, we see no obvious tendencies among the conditions, whether it being in the first or later half or right before and after the changes in robot speed ratio.

When comparing both the delta and absolute movement between conditions similarly to how we did with the reported trust scores, the Wilcoxon rank sum test yield no significant differences, regardless of data groupings. Due to the lack of significant difference and the inconsistencies we cannot reject the null hypotheses for hypothesis two.

Figure E.8 shows the average normalized GSR measurement between participants and split between conditions. After noise removal the data is normalized by fitting the range of reading for each participant between zero and one. As such, the confidence intervals increase as the tests go on as most participants start the test with high resistance across the skin, placing them close to the value one, with the resistance decreasing at different rates throughout the tests.

To test the curves of the normalized GSR we perform a Wilcoxon rank sum test between the four conditions at the midpoint of the experiment. This yielded no significant differences. As such we cannot reject the null hypotheses for hypothesis three. Lastly, performing both Kendall's and Spearman's rank correlations, neither the weighted delta movements nor the absolute movements showed significant correlation with the reported trust scores, meaning we cannot reject the null hypothesis for hypothesis four.

6 Discussion

The trust score results indicate a decrease in trust towards the robot, as is to be expected from the experiment design, where trust were only significantly affected by increases in speed. However, our goal to verify it as a measurement using GSR as a measure of arousal was not met in this study. This may be due to the measurements not being sensitive enough for the arousal experienced, but it may also be a flaw in the procedure, as the electrodes may not have had enough time to warm up and level out before starting the tasks. Future experiments will be started with a warm-up period as well as a period for taking baseline measurements.

The method of motion tracking used in this experiment is not an effective indicator of trust through physical apprehension signals. This may be due to low sensitivity from the low camera resolution, but it may also be an issue with the nature of the task, in that having the participants be sedentary and holding down the paper may inhibit movement. Alternatively, only looking at the movement of the peak position of the user is not enough to indicate movement or apprehension signals. In follow-up studies we will investigate alternative data processing methods that takes advantage of all the data we're collecting. In addition, aggregating the IR frame data along the horizontal axis may reveal more physical signals. We can also look into operator proximity to the robot throughout the test as they place the paper in order to infer trust. In addition, it could be beneficial to design a shared task where the participant

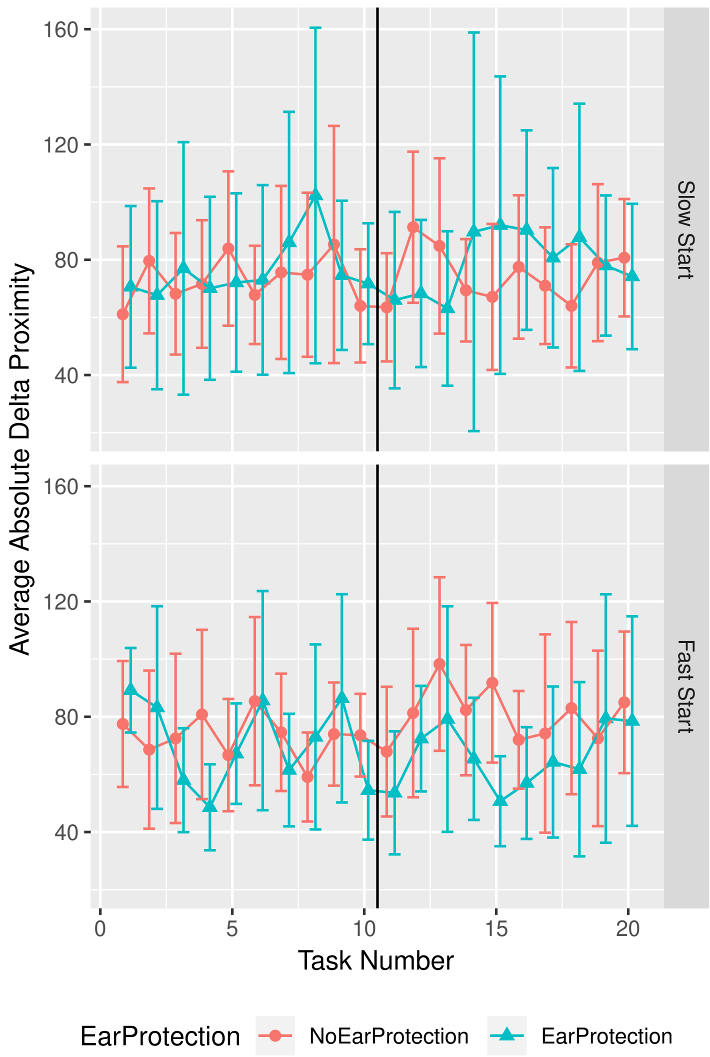


Fig. E.7: Average absolute participant motion for the first second of robot movement along with confidence interval. The vertical lines mark the midway point of the test where the robot speed ratio changes.

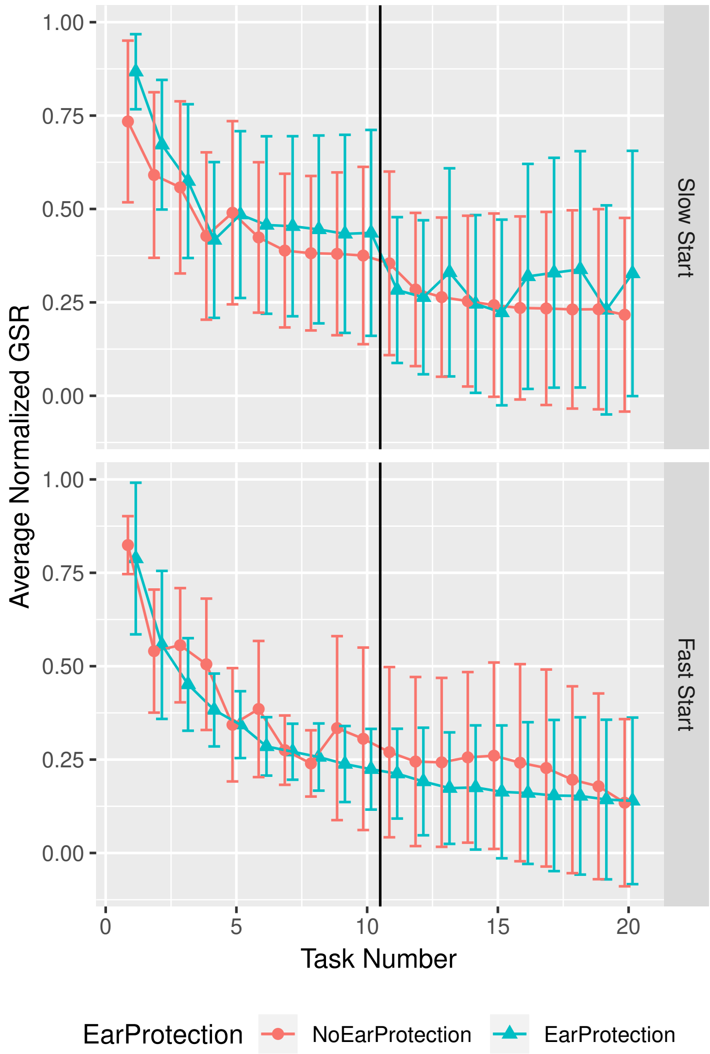


Fig. E.8: Average normalized GSR at the beginning of each task along with confidence interval. The vertical lines mark the midway point of the test where the robot speed ratio changes.

is standing, allowing for more motion as they are not pinned down by having to sit in a chair.

For this experiment we focused on changes in robot speed for inducing a decrease in trust, but for real world application it would be valuable to look into more subtle signs of system error. One possibility with our current setup would be to either remove or alter the AR projection that shows what the robot is doing, which may induce more uncertainty and thereby mistrust in the system. A problem with changing the speed is that the participants may expect that something unexpected may happen due to the experiment context, so simulating subtle system errors may do less to immediately affirm their suspicion. With a longer experiment it would be interesting to look into the recovery in trust that we see in the later half of the test. As the participants know that changes can happen, they may expect a break in movement patterns to happen again, which may affect trust recovery.

7 Conclusion

We aimed to develop a system to assess human-robot trust in real time during close-proximity HRC. Using a top-down IR depth camera we aggregated the frame data to measure the operator's proximity to the robot in order to infer trust towards the robot from physical apprehension signals.

We tested this setup as a tool for assessing operator trust based on reactions to a sudden change in robot movement speed in order to provoke a disruption of expectations. We looked at both increases and decreases in speed, as well as participants with and without ear protection to see if motor noise from the robot on speed changes have any effect. To determine the effects on operator trust we assessed the subjects attitudes towards the robot using self-reporting through questionnaires and emotional arousal from GSR. After performing the experiment with forty participants, we found that reported trust towards the robot was significantly affected when the robot's movement speed was unexpectedly increased. This was not the case for speed decreases. Wearing ear protection did not yield any significant difference, suggesting little effect from the motor noises. The GSR measurement were not significantly different between the test conditions, which may be due to an insufficient warm-up period. The analyses for the motion tracking method used in this study did not suggest that the participants' motions were significantly affected by a decrease in trust. The method was based on tracking the position of the point of the participant closest to the camera. For future studies we will work with the data we collected to design a data processing method that takes better advantage of the amount of data collected in order to obtain a more sensitive motion measure. We will also re-design the shared task, allowing the participant to stand up, allowing for more motion.

References

[1] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of Electromyography and*

Kinesiology, vol. 14, no. 1, pp. 13–23, feb 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641103001251?via{%}3Dihub

[2] Arbejdstilsynet Denmark, "Anmeldte erhvervssygdomme 2012-2017," 2017. [Online]. Available: https://at.dk/media/4814/bilag-b-tabeller-med-anonymitet-final.pdf

[3] G. Herrmann and U. Leonards, *Human-Humanoid Interaction: Overview*. Dordrecht: Springer Netherlands, 2016, pp. 1–16.

[4] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-Time Human Tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, p. 011006, jan 2014. [Online]. Available: http://computingengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4025810

[5] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, may 2011, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/5942335/

[6] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *The 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) IEEE International Symposium on Robot and Human Interactive Communication*. IEEE Computer Society Press, 2019.

[7] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, jan 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50{_}30392

[8] J. Greenstein and M. Revesman, "Two Simulation Studies Investigating Means of Human-Computer Communication for Dynamic Task Allocation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 16, no. 5, pp. 726–730, sep 1986. [Online]. Available: http://ieeexplore.ieee.org/document/4075638/

[9] P. A. Lasota and J. A. Shah, "Analyzing the Effects of Human-Aware Motion Planning on Close-Proximity Human–Robot Collaboration," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 1, pp. 21–33, feb 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720814565188

[10] P. Rani, N. Sarkar, C. A. Smith, and L. D. Kirby, "Anxiety detecting robotic system towards implicit human-robot collaboration," *Robotica*, vol. 22, no. 1, pp. 85–95, jan 2004. [Online]. Available: http://www.journals.cambridge.org/abstract{_}S0263574703005319

[11] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[12] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[13] G. Hoffman, "Evaluating Fluency in Human-Robot Collaboration," Tech. Rep., 2013. [Online]. Available: http://guyhoffman.com/publications/HoffmanRSS13Workshop.pdf

Paper F

Human-Robot Trust Assessment Using Top-Down Visual Tracking After Robot Task Execution Mistakes

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper is under review in
IFIP Conference on Human-Computer Interaction – INTERACT 2021

The layout has been revised.

Abstract

With increased interest in close-proximity human-robot collaboration in production settings it is important that we understand how robot behaviors and mistakes affect human-robot trust, as a lack of trust can cause loss in productivity and over-trust can lead to hazardous misuse. We designed a system for real-time human-robot trust assessment using a top-down depth camera tracking setup with the goal of using signs of physical apprehension to infer decreases in trust toward the robot. In an experiment with 20 participants we evaluated the tracking system in a repetitive collaborative pick-and-place task where the participant and the robot had to move a set of cones across a table. Midway through the tasks we disrupted the participants expectations by having the robot perform a trust-dampening action. Throughout the tasks we measured the participant's preferred proximity and their trust toward the robot. Comparing irregular robot movements versus task execution mistakes as well simultaneous versus turn-taking collaboration, we found reported trust was significantly decreased when the robot performed an execution mistake going counter to the shared objective. This decrease was higher for participant working simultaneously as the robot. The effect of the trust-dampening actions on preferred proximity was inconclusive due to unexplained movement trends between tasks throughout the experiment. Despite being given the option to stop the robot in case of abnormal behavior, the trust-dampening actions did not increase the number of participant disruptions for the actions we tested.

1 Introduction

The field of human-robot interaction (HRI) has expanded to include close-proximity human-robot collaboration (HRC) as several manufacturing industries are looking to implement human-robot teams where possible in order to assist or relieve production staff of strenuous or repetitive tasks. An important element in close-proximity HRC, where the human, from here referred to as the operator, works with the robot on a shared objective with no safety barriers, is that the operator feels safe and can trust the robot. An appropriate level of trust is very important as under-trust increases the risk of under-performance in human-robot teams as it can lead to dis-use of the robot, and over-trust can lead to accidents through dis-use [1].

With our research we aim to contribute to the field of HRC by working to enable real-time human-robot trust assessment, allowing us to maintain trust at an appropriate level through trust calibration. Either trust-repairing or trust-dampening actions from the robot can be used to correct the operators level of trust [1]. In this experiment, specifically, we evaluate different trust-dampening action along with a top-down depth camera tracking system designed to track position and posture with the goal to use physical signals of apprehension as a sign of lowered trust toward the robot. The tracking setup is shown in Figure F.1. While the system is designed to track both posture and proximity, the scope of the experiment is to evaluate the utility of changes in operator proximity to the robot for trust assessment. In a repeated collaborative pick-and-place task we test two types of trust-dampening actions from the robot. One action to affect predictability of the robot and one to affect dependability. In addition, we test the effects of the operator working simultaneously with the robot compared to taking turns with the robot. In this paper we make the following contributions:

- We give insight into the effects of two kinds of trust-dampening actions as well as simultaneous collaboration versus turn-taking collaboration on human-robot trust during repetitive HRC tasks.
- We present our top-down depth camera tracking system and evaluate its utility to observe changes in operator proximity to the robot after trust-dampening actions.
- We present the results and analyses of trust questionnaires and proximity tracking from an experiment with 20 participants.
- We discuss some of the challenges involved in laboratory experiments on human-robot trust assessment.

Fig. F.1: The human-robot collaboration test setup with Rethink Robotics Sawyer robot as well as depth camera and projectors mounted to an aluminium rig, allowing for operator tracking and projection-based augmented reality, respectively.

2 State of the Art

Malik and Bilberg [2] developed a reference model for the terminologies used to describe HRI. They outlined five levels of engagement in HRC, the first level being when the robot is isolated from the operator, such as in a cell. The coexistence level is when there are no barriers between operator and robot, but they do not share a work space. When the parties share a work space, but only occupy it interchangeably it is defined as the synchronization level, whereas in the cooperation level they can occupy the space simultaneously. The collaboration level of HRC is defined as when the operator and the robot manipulate an object at the same time. In this experiment we test effects of the synchronized and cooperation levels, as one conditions has them work in turns while the other has them work simultaneously.

Trust in HRC is commonly defined as the willingness of a participant to engage in a situation involving being vulnerable with another party based on their perceived capability and intent, the other party being the robot and its operating systems. Is is important for safe and efficient collaboration to maintain an appropriate level of trust in the robot. De Visser et al. [1] outlined calibrated trust as the operator's perceived trustworthiness of the robot matching a measure of the robot's objective trustworthiness, which is based on its capability to complete its objective. They also stated that if the operator trusts the robot too much it may lead to misuse, increasing the risk of accident, while too little trust may lead to disuse, causing losses in productivity. In a meta-analysis Hancock et al. [3] categorized the elements that can affect operator trust in the robot, splitting them up between operator-related, robot-related and environment-related factors. The robot-related factors were further grouped into performance-based and attribute based factors with dependability, reliability, predictability and failure rates among the performance-based factors.

In an experiment with a drink-serving robot Dragan et al. [4] found that human-robot trust was significantly affected by the robot's movement path with participants preferring predictable motions designed to appear natural, rather than motions calculated to be optimal by the motion planning system. Bergman & Zandbeek [5] tested speed and stopping distance as a robot moved toward the participants and found them to significantly affect trust and argued that they could be utilized as communicative cues from the robot. We similarly base the trust-dampening action in our experiment on disrupting predictability and perception of dependability.

Schaefer [6] developed two versions of a trust assessment questionnaire. A long version with 40 items and a short version with 14 items. In both versions participants rate the likelihood of the robot completing its objective, making errors and being dependable among others. In a comparative study Kessler et al [7] had participants rate a robot using both Schaefer's scale and a trust in automation scale, finding conflicting results, suggesting that robot trust and automation trust scales are not interchangeable. As an alternative measure of human-robot trust, Freedy et al. [8] inferred trust based on the number of participants interventions as they observed an unmanned ground vehicle. To assess participant trust throughout our experiment we use Schaefer's 14-item questionnaire as well as giving the them the option to interrupt the robot during the collaboration.

Computer vision have previously been studies for its utility in safety system in close-proximity HRC, implementations including depth camera [9] and multi-camera [10] setups. As for computer vision for human-robot trust assessment, Sadrfaridpour et al. [11] tested hand-tracking using makers to infer operator trust based on changes in working speed in an HRC assembly task using a neural network.

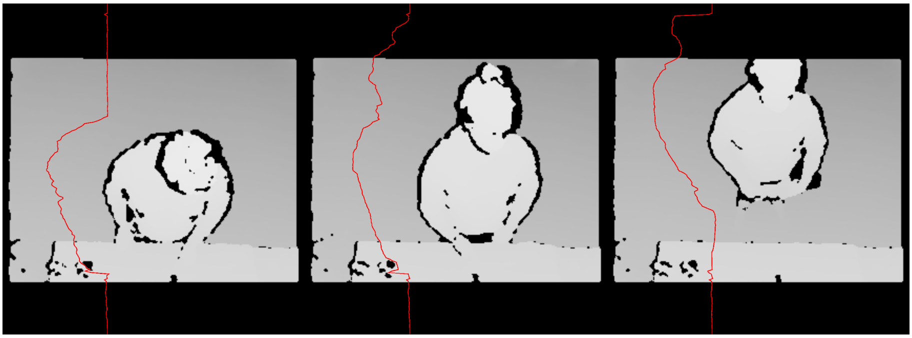


Fig. F.2: Example of the top-down infrared tracking with the red lines being the aggregations of the horizontal lines on the frames. To the left the operator is stood upright close to the robot with the red having a narrow spread near the bottom. In the middle frame the operator is standing in the same position but leaning back, yielding a wider spread across the frame. In the right-most frame the operator is a step further away from the robot while standing upright, giving a narrow range at the top of the frame.

3 Top-Down Tracking System

The posture and position tracking is implemented using an Orbbec Astra depth camera mounted to an aluminum rig above the work area and pointing downwards, as shown in Figure F.1. Before beginning the tracking a background image is taken with no operator present. The background image pixel values are subtracted from the frames during tracking. To track the position and posture of the operator, each frame of the infrared camera feed is aggregated by the horizontal lines, where the mean value of the pixels in each line is logged, excluding zero values as those are positions where no infrared light is detected due to occlusion. Examples of infrared camera frames and aggregated lines are shown in Figure F.2. The lines with mean values below zero are where the operator is present, as the distance to the camera is shorter at those position. Using the range of horizontal lines with the operator present we can infer the operator's posture as they will be more spread out across the image when filmed from above, though evaluation of this is beyond the scope of this experiment.

4 Experiment

To evaluate the tracking system and its utility for trust assessment we tested it in a close-proximity HRC scenario where the participant has to move a set of cones from one side of a table to the other in collaboration with the robot. The task is repeated ten times, midway through which during the sixth task we introduce a trust-dampening action from the robot. In a between-subject experiment we test two types of trust-dampening actions, pertaining to the robot's dependability and predictability, respectably. To test predictability, one the sixth repetition of the task the robot will move in a different path, moving in an arch closer to the participant. To test dependability the robot will make a mistake in executing the task by picking up and moving the wrong cone. In addition, in another between-subject condition we test the effect of having the collaboration be turn-based, where the robot moves one of its cones first, then the participant, or have them move the cones at the same time. This totals four between-subject conditions. This is because it may affect how safe the participants feels working with the robot and how attentive they are toward it during the task. The test setup is shown in Figure F.3. We are testing the following hypothesis:

- **H1:** Reported trust in a close-proximity HRC task is significantly affected by the robot changing its movement patterns.
- **H2:** Reported trust is significantly affected by the robot making a mistake in task execution and going counter to the shared objective.
- **H3:** The chance of the participant interrupting the robot is significantly increased by the robot performing a trust-dampening action.
- **H4:** Participants' preferred proximity to the robot is significantly affected by the robot performing a trust-dampening action.
- **H5:** Reported trust and participants' preferred proximity are significantly affected by whether the collaboration is simultaneous or turn-taking.

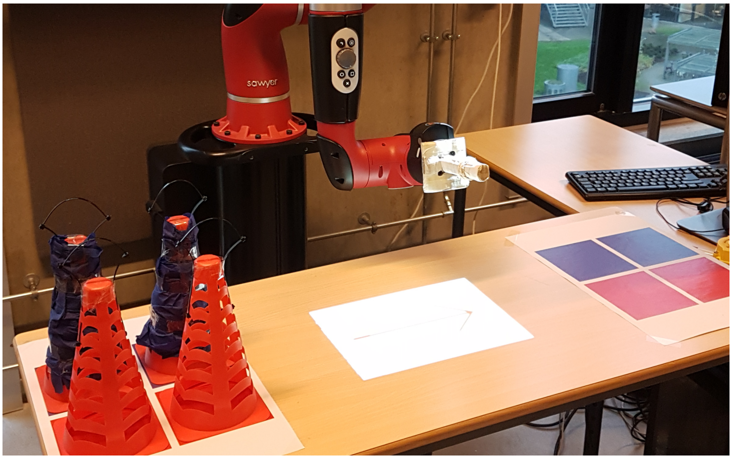


Fig. F.3: The test setup where the participant stand in front of the robot and repeatably move the colored cones from one side of the table to the other.

4.1 Procedure

In the laboratory, first, the participant is introduced to the robot and collaborative task. They are told that they, along with the robot, have to move a set of four cones from one side of the table to the other. The participant has to move the red cones while the robot moved the blue cones, and they can only move one cone at a time. The participant is not allowed to touch the blue cones. They have to move the cones to their colored squares at the other side of the table, shown in Figure B.6. The correct side of the table is always indicated with an arrow projected onto the table as shown in Figure F.4. When participants in the simultaneous collaboration condition start moving the cones, the robot starts moving its cones at the same time, while in the turn-taking condition the robot moves one cone first and then waits the until the participant is finished moving theirs before moving the second one. The robot's turn is started remotely by the test conductor behind the participant where they do not see it. The remote activates pre-programmed movements. The robot move the cones using a rod, shown in Figure F.4, that goes through plastic loops attached to the top of the cones. This to avoid having to use a gripper which would require a load compressor, which may startle the participant when it activates to restore pressure. After completing each task the participant is asked to fill out Schaefer's [6] 14-item trust questionnaire before beginning the next task.

When the robot performs the action of moving the blue cones it first picks up the one closest to the opposite side of the table and moves it to blue square farthest away in a straight line, allowing the second cone to be moved to its place without collision. In the first five tasks the robot will complete the task with no error, but at the sixth task it will perform a trust-dampening action, depending on the test condition. If the participant is testing the irregular movement condition, when moving the first cone the robot will move it in an arch toward the participant, but otherwise placing it in the correct position. If they are testing the condition with mistake in execution, the robot would initially move the first cone correctly, but it would then move back where it was. Afterwards the robot moves the cone to the correct position again and then proceeds with the second cone and completes the task. At the beginning of the test the participant is informed that they can press the space bar on a keyboard on the table in front of them to pause the robot in case it does something it is not supposed to. It is counted whenever they pause the robot, and they are asked by the test conductor why they did it.

4.2 Participants

We performed the experiment with 20 participants, five in each condition. Our participants included four female and 16 males, four were left-handed, 16 were right-handed, average age 23, standard deviation 1.9 years.

5 Results

The average trust scores throughout the tasks are shown in Figure F.5 with the average delta scores between tasks shown in Figure F.6. We can see a decrease in trust after the trust-dampening action, marked with the vertical black lines, followed by a slow recovery of trust.

When analysing the data we start with multi-variate analyses based on the tasks and conditions, followed by pairwise comparisons to gain further insight. Table F.1 shows a summary of the statistical analyses performed on the trust scores and participant proximity before and after the trust-dampening actions as well as analyses on

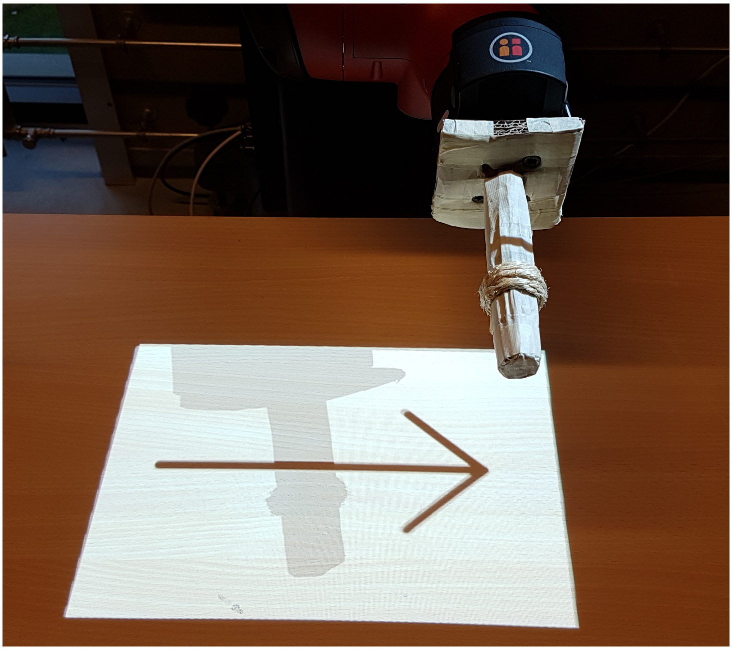


Fig. F.4: The rod mounted to the robot for picking up and moving the cones. The augmented reality overlay is made with two overlapping projections from the projectors mounted at either side of the aluminum rig and pointed diagonally down, as shown in Figure F.1, making the projection harder to occlude.

the differences in decrease between conditions. Analysing the trust scores before and after for each of the four conditions separately only show significant change when the robot performs a wrong actions. While a Kruskal-Wallis analysis showed significant effect from the conditions on trust scores and changes in delta trust going from the fifth to the sixth task, pairwise analyses showed that delta scores were not significantly affected by whether the robot and participant took turns or worked simultaneously. When we compare the trust delta scores between collaboration types after the robot performs a wrong action, simultaneous collaboration shows significantly higher decrease in trust than turn-taking collaboration. Also, comparing the delta trust scores by the mistake types shows significantly larger decrease in trust when the robot performs a wrong actions than it performs irregular movements. This is only the case, however, when the robot and participant are performing the task simultaneously. These results may also be due to the decreased sample size for the more granular analyses. We reject the first hypothesis, as changing the movement pattern did not affect trust, while we retain the second hypothesis, as the robot performing the wrong action did affect trust.

Participants pressed the button to interrupt the robot 12 times between all of them. Only six of these interruptions happened during the robot's trust-dampening actions; once for irregular movement and turn-taking, three times for wrong action during turn-taking and twice for wrong action during simultaneous collaboration. As there was an equal number of interruptions during trust-dampening actions as during regular robot operations we reject the third hypothesis.

Table F.1: Summary of statistical analyses of the reported trust and participant proximity before and after a trust-dampening actions. Multi-variate analyses are performed with ANOVA for parametric data and Kruskal-Wallis otherwise. We test the differences between before and after as well as the difference between effects of conditions using pair-wise t-tests for parametric data and Wilcoxon rank sum tests otherwise.

Trust Scores	**Before**	**After**	*t*	*w*	*df*	p-value
ANOVA: F(7, 32) = 4.34						< .01*
Simultaneous / Irregular movement	78.14	73.00	1.18		7.31	.28
Turn-taking / Irregular movement	75.14	62.86	1.30		6.30	.24
Simultaneous / Wrong action	79.29	42.00	2.76		7.44	.027*
Turn-taking / Wrong action	67.71	53.86	2.89		6.92	.024*
Change in Delta Peak Proximity						
Kruskal-Wallis: H = 8.74					7	.27
Change in Delta Mid Proximity						
Kruskal-Wallis: H = 17.49					7	.015*
Simultaneous / Irregular movement	-2.81	9.42		4		.095
Turn-taking / Irregular movement	-1.28	-5.28		14		.84
Simultaneous / Wrong action	-30.54	23.69		0		< .01*
Turn-taking / Wrong action	-7.96	6.23		5		.024*
Delta Trust Scores	**Delta 1**	**Delta 2**				
Kruskal-Wallis: H = 10.086					3	.018*
Simultaneous(1) v. Turn-taking(2)	-21.21	-13.07	-1.03		12.14	.32
Irr. movement(1) v. Wrong action(2)	-8.71	-25.57		83		.012*
Simul.(1) v. Turn(2) : Irr. Movement	-5.14	-12.29	1.089		6.29	.32
Simul.(1) v. Turn(2) : Wrong action	-37.28	-13.86		1		.016*
Irr. movement(1) v. Wrong(2) : Simul.	-5.14	-37.28		24		.016*
Irr. movement(1) v. Wrong(2) : Turn	-12.29	-13.86	0.24		6.25	.82
Delta Peak Proximity						
ANOVA: F(3, 15) = 2.22						0.13
Delta Mid Proximity						
Kruskal-Wallis: H = 7.91					3	.048*
Simultaneous(1) v. Turn-taking(2)	16.55	0.48		82		.015*
Irr. movement(1) v. Wrong action(2)	2.07	14.96		31		.17
Simul.(1) v. Turn(2) : Irr. Movement	9.42	-5.22		82		.15
Simul.(1) v. Turn(2) : Wrong action	23.69	6.23		22		.056*
Irr. movement(1) v. Wrong(2) : Simul.	9.42	23.69		7		.31
Irr. movement(1) v. Wrong(2) : Turn	-5.22	6.23		5		.15

The participants' delta proximity to the robot between tasks are shown in Figures F.7, measured by the middle position of their volume along the vertical axis of the infrared camera frames. The proximity is measured as their average proximity to the robot during the first ten seconds of each task, during which the robot would move its

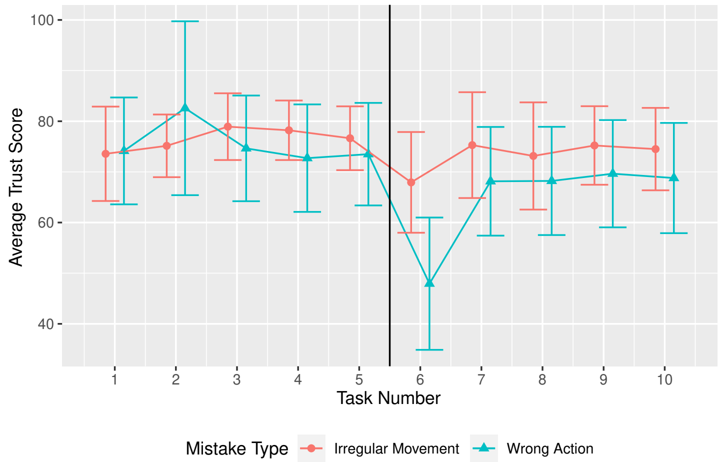


Fig. F.5: Average trust scores with confidence intervals for each task with the trust-dampening action marked with a vertical line.

first cone before making errors. The proximity measured by the position of the peak position of the participant, the top of the head, were not significantly affected by test conditions according to an ANOVA. Kruskal-Wallis analyses on the delta mid positions and on the change in delta positions before and after trust-dampening actions, however, showed significant effect from the test conditions. Pairwise comparisons showed significantly larger increase in distance only when the participant had to perform the task simultaneously with the robot. Comparing collaboration type with each mistake type separately only showed this effect when the robot performed a wrong action. As before this may be due to smaller sample size. Mistake types showed no significant effect on changes in human-robot proximity. Despite the result of the analysis, looking at the trends on delta proximity for simultaneous collaboration, shown in Figure F.7, shows a pattern of participants interchangeably increasing and decreasing proximity to the robot between tasks. It is unclear what causes this trend. Due to the results the fourth hypothesis is inconclusive. We retain the fifth hypothesis in regards to reported trust as there was a higher decrease in trust during simultaneous collaboration than turn-taking when the robot performed a wrong task.

6 Discussion

Based on the results we retain the second hypothesis, that reported trust is significantly affected by the robot making a mistake in execution, while the first hypothesis, pertaining to irregular robot movement is rejected. This suggests that the irregular movement was either not considered counter to the objective or hazardous to the participant, or the participants did not recognize the irregularity. Not recognizing the error may be a matter of attention as the participants were performing their own

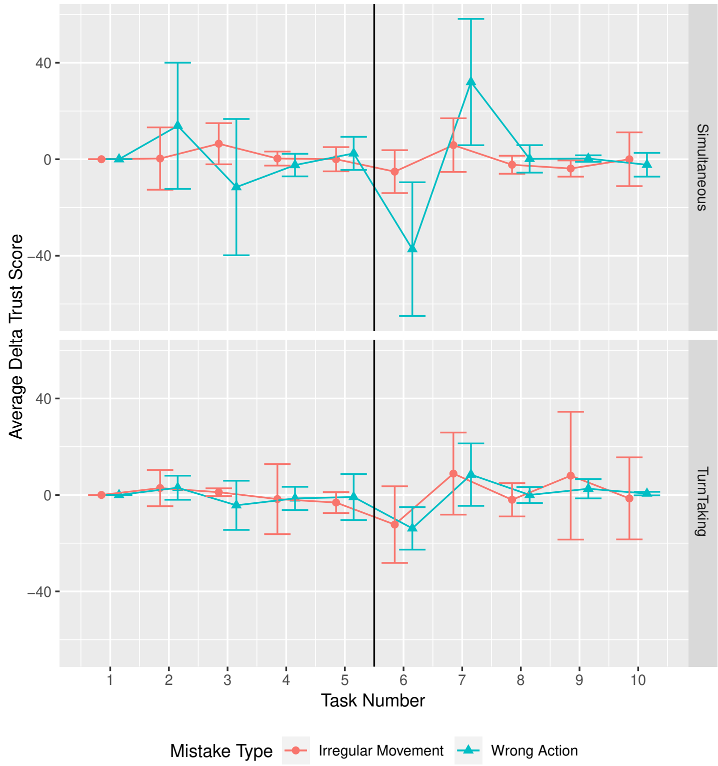


Fig. F.6: Average trust delta scores with confidence intervals for each task with the trust-dampening action marked with a vertical line.

part of the task. It would be beneficial to evaluate the participant's perception of the robot's errors in a test where they only observe the robot. The analysis showed higher decrease in reported trust after a wrong robot action when the participants were doing simultaneous collaboration with the robot than when they were taking turns, which may suggest a higher level of vulnerability in that situation.

Neither type of trust-dampening action increased the chance of the participants interrupting the robot. The absence of interruptions may be due to the participants' interpretation of the instructions. They were told to pause the robot if it did something it was not supposed to. This may have been interpreted as the robot doing something that would bring them in immediate danger or cause damage to the robot itself, rather than just acting counter to the objective. This may also have been due to an inhibition to commit to interrupting the robot if they were not sure if the robot was making a mistake or if they were themselves mistaking. This could be affected by the test conductor's presence in the room. A separate experiment could be conducted where the test ends after the trust-dampening action, so we can ask the participant why they did or did not interrupt the robot and gain some insight into their perception.

While the analysis shows significant effect on participants' preferred proximity

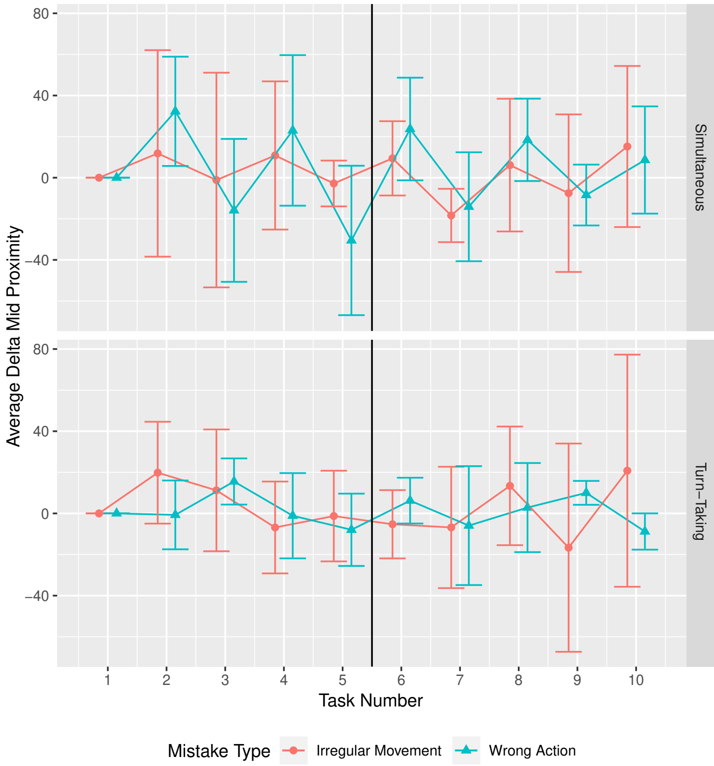


Fig. F.7: Average changes of participant proximity with confidence intervals for each task with the trust-dampening action marked with a vertical line.

from the robot performing a wrong action in the condition where they moved cones simultaneously, looking at the trend throughout the experiment, illustrated in Figure F.7, it becomes unclear. The data shows a trend where the participants would interchangeably move closer to and further away from the robot, the trend lining up so they moved away after the trust-dampening action, yielding a significant change between tasks. Seeing as this trend is consistent throughout the experiment the results become inconclusive. This behaviour may be an effect of the nature of the task where the cones are moved back and forth between the ends of the table. This may have caused the participants to switch between two preferred positions and depending on whether the depth camera, the robot and the table were perfectly aligned, this may have caused the shift in proximity. It may also be due to the smaller sample size when analysis a subset of our participants.

While the tracking system would allow us to monitor movement and changes in posture during the collaboration, it is beyond the scope of this experiment to recognize and isolate the moment when the participants would recognize the trust-dampening actions of the robot. Making an assumption of when they recognize inconsistent behaviour or errors would likely be easier if the robot made sudden changes, like increasing speed and making more motor noise. Isolating when the participant leans away from the robot, rather than increasing their distance from it, would be easier if the participant was discouraged from taking steps. This could be achieved by having the participant stand on a small platform, requiring them to step down to move away, though this would likely be hazardous and not representative of many HRC contexts.

Aside from the challenges specific to this experiment, evaluating human-robot trust in a laboratory context present difficulties in terms of validity. As the participants are aware of the experiment, they are likely to expect something to occur related to the robot, especially when they asked questions pertaining to the dependability of the robot. As such the participant's reported trust in the robot may be affected by their trust in the test conductor if they are present. This can possibly be addressed by framing the collaboration in such a way that trust-dampening actions can occur without being attributed to the experiment procedure itself or to the test conductor. Such a framing could be that the robot is a prototype that needs testing to fix potential errors. In addition, free-form questions pertaining to the participant's perception of the robot and the test scenario would yield valuable insight as to how their trust is affected.

7 Conclusion

We set out to contribute to the fields of HRI and HRC by evaluating our top-down visual tracking system for its utility in real-time trust assessment during close-proximity collaboration. Using a repetitive collaborative pick-and-place task we tested the system in four conditions using two collaboration formats and two methods of dampening trust in the robot while administering human-robot trust questionnaires [6]. After testing the system with 20 participants, five in each condition, we found that trust in the robot was significantly decreased when the robot performed actions that went counter to the shared objective, but not when the robot changed its movement pattern while otherwise performing the task correctly. The decrease in reported trust was significantly higher when the participant and robot performed their shared objective simultaneously rather than taking turns. The trust-dampening actions did, however, not increase the chance that participants would stop the robot during collaboration.

Whether participants' preferred proximity to the robot was affected by trust-dampening actions is inconclusive as the data shows odd movement patterns of switching between going nearer to and further away from the robot throughout the experiment. This may be due to the nature of the shared task where they switched between moving a set of cones from one side of a table to the other. Future experiments should focus on participants' perception of the robot and the trust-dampening actions, as the collaboration may affect their attention towards the robot and how they perceive its actions.

References

[1] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[2] A. A. Malik and A. Bilberg, "Developing a reference model for human–robot interaction," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 13, no. 4, pp. 1541–1547, 2019.

[3] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[4] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[5] M. Bergman and M. van Zandbeek, "Close Encounters of the Fifth Kind? Affective Impact of Speed and Distance of a Collaborative Industrial Robot on Humans," *Human Friendly Robotics*, pp. 127–137, 2019.

[6] K. Schaefer, "The perception and measurement of human-robot trust," 2013.

[7] T. T. Kessler, C. Larios, T. Walker, V. Yerdon, and P. Hancock, "A comparison of trust measures in human–robot interaction scenarios," in *Advances in human factors in robots and unmanned systems*. Springer, 2017, pp. 353–364.

[8] A. Freedy, E. DeVisser, G. Weltman, and N. Coeyman, "Measurement of trust in human-robot collaboration," in *2007 International Symposium on Collaborative Technologies and Systems*. IEEE, may 2007, pp. 106–114. [Online]. Available: http://ieeexplore.ieee.org/document/4621745/

[9] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta, "Toward Safe Human Robot Collaboration by Using Multiple Kinects Based Real-Time Human Tracking," *Journal of Computing and Information Science in Engineering*, vol. 14, no. 1, p. 011006, jan 2014. [Online]. Available: http://computingengineering.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4025810

[10] J. T. C. Tan and T. Arai, "Triple stereo vision system for safety monitoring of human-robot collaboration in cellular manufacturing," in *2011 IEEE International Symposium on Assembly and Manufacturing (ISAM)*. IEEE, may 2011, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/document/5942335/

[11] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and control of trust in human-robot collaborative manufacturing," in *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 115–141.

Paper G

Human-Robot Trust Assessment in Virtual Reality Experiment For Data Crowd-Sourcing

Kasper Hald, Matthias Rehm, Thomas B. Moeslund

The paper is under review in the
2021 International Conference on Robotics and Automation (ICRA), 2021.

The layout has been revised.

Abstract

In this paper we outline some of the challenges of crowd-sourcing data for unsupervised human-robot interaction experiments, as remote testing is a relevant measure in cases where laboratory experiments are not possible or feasible. We designed and implemented a virtual reality application for assessing human-robot trust in a close-proximity human-robot collaboration setting. The experiment involved repeated tasks with a disruptions of participant expectations midway through. After assessing the application and procedure in a laboratory setting, self-reported trust towards the robot was consistent to similar prior experiments, while effect on task performance is inconclusive. While there were no issues with application usability, our remote data submission method can be improved.

1 Introduction

Close-proximity human-robot collaboration (HRC) is becoming increasingly relevant as we are seeing the long-term effects of repetitive manual work in field like manufacturing [1] and industrial production lines, such as musculo-skeletal diseases [2]. When using Herrmann and Leonards' definition of HRC, where a the human and robot work with the same component at the same time, it can be used to relieve some of the strenuous and repeated tasks. However, close-proximity HRC requires trust from the human collaborator towards the robot. Especially in cases like meat production, where the robot will have to be equipped with sharp tools or powerful grippers. In this experiment we aim to assess human-robot trust in close-proximity HRC.

Under circumstances where human-robot laboratory experiments are not possible or feasible, it is relevant to consider how we can perform controlled experiments without the need to gather many test participants in the same location. This is especially relevant for research areas such as human-robot interaction (HRI) where experiments often require the participants to be in the room with the robot. To address these challenges we designed and evaluated a virtual reality (VR) test application, allowing people with consumer-level VR hardware to participate unsupervised from their own homes, enabling us to crowd-source test data.

In our experiment we aim perform a human-robot trust assessment experiment inspired by Hald et al. [3], while adapting it to fit the challenges of creating a VR data crowd-sourcing application. Even-though the goal to enable unsupervised experiments with participants recruited via the internet, we are verifying the procedure in a laboratory setting. The virtual environment used for the experiment is shown in Figure G.1.

2 Background

Gadiraju et al. [4] analyzed and compared the strengths and weaknesses of laboratory experiments compared to crowd-sourced experiments. Strengths of a laboratory experiments include a high level of control of process and environment, possibility

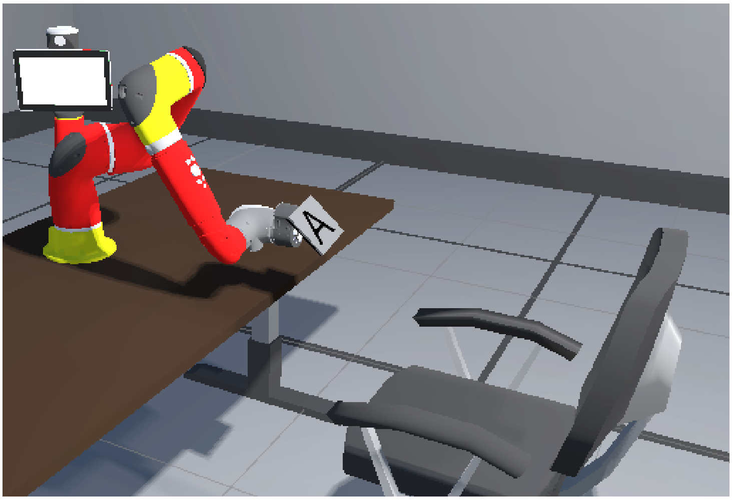


Fig. G.1: The virtual environment used for the experiment, featuring the robot, chair and desk.

of screening of participants, while weaknesses are that they time-consuming and expensive. Crowd-sourced experiments, on the other hand, have the strengths of time-efficiency and ease of access to diverse and representative, while there is less control of environment and less knowledge of participants' background.

Chernova et al. [5] evaluated an experimental procedure for observing reactions to task-specific social behaviours in a virtual robot in a virtual environment, finding very similar patterns of behaviour observed in real-world tasks. Similarly, Matsas and Vosniakos [6] designed a VR training system for HRC in manufacturing tasks, utilizing a Microsoft Kinect for body tracking, showing positive results in regards to the utility of VR applications in HRC simulation. Liu et al. [7] performed two VR HRI experiments, showing that a VR headset significantly improved performance in simulated HRC with the stereo display being a contributing factor. Their virtual environment is similar to the one used in this experiment, as it features a robot arm on a tabletop with the participant sitting at the opposite side of the table.

When assessing human-robot trust, we are working from Lee & See's [8] definition of trust in automation, which is based on the expectation that the agent, in this case the robot, will achieve its goal in a context characterized by uncertainty and vulnerability. Close-proximity HRC can easily involve vulnerability due to the risk of injury due to particularly powerful robots. In our experiment we work with trust in terms of the participant's attitude towards the robot in the current moment, rather than the build-up of trust over a prolonged period of collaboration.

Among the factors affecting trust in HRI are the robot's physical characteristics, its performance, while environmental factors have shown only moderate effect [9]. Additionally, Dragan et al. [10] showed that the robot's motion patterns affect trust, as predictable motions were preferable to motion that is purely functional from a motion-planning perspective.

3 Designing an Unsupervised Experiment

Designing an experiment to be performed unsupervised by a subject in their own home presents a number of challenges in regards to hardware requirements, usability, privacy and in communicating the experimental procedure. The main goal is to make the experiment as accessible as possible to prevent people from losing interest, as there are no repercussions for them to quit the experiment at any moment, as opposed to the social situations that may occur if doing so in a laboratory experiment.

3.1 Hardware Requirements

In this case the hardware requirements for people to participate in the test are a SteamVR-compatible headset and a compatible Windows PC. Requiring what would be considered niche hardware, compared to just a PC with any operating system, will dramatically decrease the number of people available to participate. As such, when recruiting via the internet one should search communities for people with interest in every sub-category of the experiment subjects. In this case, one should link to the experiment in communities for both VR, robots, and experiments and surveys for the chance that a subset of each community are interested in one or more of the other subjects, with an interest pertaining to the hardware requirements being the most relevant.

3.2 Usability

Because there will be no test conductors present and adding a manual or tutorial to the test application will add time to the experiment duration, possible decreasing participant interest, using the test application should be as intuitive as possible. Especially if the experiment does not use common hardware, like mouse, keyboard or gamepad, it can be a benefit to limit the variety of actions required as well as the number of actions the participants can perform at any point. In the case of this experiment, the participants are informed that they will not need the tracked controllers that often come included with a SteamVR-compatible headset. There are more details on this in Section 4.4.

3.3 Privacy Concerns

People may have reservations about downloading and running a program from an unverified source on their home computer, especially if the program has to send data over the internet. To help this issue it should be very transparent how the program works and what data is being send and how. It may encourage trust in the researcher to include contact information in the experiment description.

4 Experiment

In order to test human-robot trust we have designed a repeated task where the virtual robot arm moves toward participant, holding a plate with a letter written on it. The starting and forward positions of the robot is shown in Figure G.2. When the robot finishes its movement, the participant has to confirm whether the letter is an "A" or a "B", after which the robot moves back to its starting position. The letters are shown at random. This task is performed ten time. The first five times the robot moves at one of its two starting speeds, full speed or half speed based on the robot's standard movement planning. After the fifth task, the robot changes to its other movement speed setting without warning with the aim of disrupting the participant's expectations and trust towards the robot. In the experiment we are testing the following two hypotheses:

- **H1:** Self-reported trust in a VR close-proximity HRC task is significantly affected by sudden changes in robot movement speed.
- **H2:** Participant question response time in a VR close-proximity HRC task is significantly affected by sudden changes in robot movement speed.

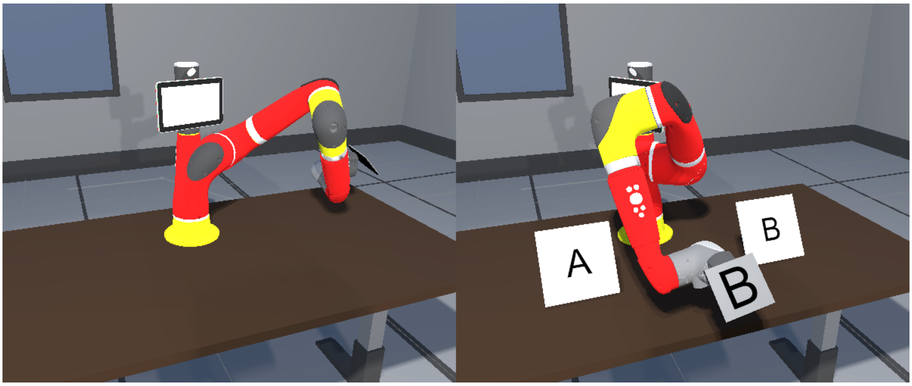


Fig. G.2: Left: The starting position of the robot. Right: The forward position of the robot.

4.1 Virtual Environment

The application is implemented using the Unity 3D game engine and the SteamVR plugin. The virtual environment used for the experiment is an open office area made from the Snap Prototype Office [11] asset pack from the Unity Asset Store. The office environment is used in order to have a simple and recognizable setting, rather than a bare environment, which may be unsettling to the participant.

The environment is based on the sample room from the asset pack, but with the participant situated in a corner isolated from the other tables and with most props removed. This is to prevent distractions, as props might be assumed to be interactive. Within the participants reach inside the VR play area are an office chair and a desk with the virtual robot arm mounted on top of it. The purpose of the office chair is to guide the participant to where they should sit within their play area.

4.2 Virtual Robot

For the experiment we are using a virtual Rethink Robotics Sawyer robot. The robot model is placed at the far side of the desk in front of the participant. The robot's movements are based on the inverse kinematics, size and weight of the real robot. Using a Robot Operating System server on a Linux virtual machine we enable connection between Unity and the real robot. After programming the robot's motions for the experiment, they are executed while the joint positions are streamed to Unity where they are recorded for playback. This records the springy motions occurring when the robot stops, making the virtual robot more realistic. As the motions are recorded, audio of the actuators and springs is also recorded for playback. After recording, the fan noise from the robots cabinet is removed using noise reduction. The cabinet does not appear in the virtual environment.

4.3 User Interface

As the participants are performing the test with the VR applications, they have to press buttons to confirm that they are ready or to select options. To prevent confusing the participants by giving them too many options, they are informed in the experiment description that they do not need any hardware other than a SteamVR-compatible headset; no controllers needed.

The interface in the virtual environment is presented as white boxes floating over the desk, labeled with black text. Each box presents an option and in order to select it, the participant has to focus on it for two seconds. Using a pointer-style input modality, a ray is cast from the tracked headset going in the viewing direction. Where the ray hits a surface, such as on the table, floor, walls or option boxes, a small grey sphere appears, acting as a cursor.

To select an option, the participant has to hold the cursor on the corresponding box for two seconds. As they hold the cursor on the box, eight grey boxes start appearing in a circle formation around it, acting as ticks on a clock face, and the option is selected when the circle is complete.

This interface is communicated to the participants partly by having the cursor be visible at all times and partly through the first option presented in the experiment. The experiment can be presented in either English or Danish, and at the start of the test the participants have to select a language option. The options are presented on two boxes, one labelled "Look here for English" while the other is labelled the same, but in Danish. Using the "look" keyword in combination with visible cursor are used to relay the interface to the participants.

4.4 Procedure

The participants are recruited, brought to the laboratory with the VR hardware and introduced to the main premise of the experiment. They are presented with a PC with a Google Forms page open in a browser and told that they should read the instructions in the survey and perform the test to the best of their ability. They are told that if they have any questions, the test conductor is available. The survey does not require an email address or login to perform the test. This is to maintain participant anonymity to prevent privacy concerns.

The introduction has a Google Drive link to a ZIP file, and the participant is instructed to download the file, unzip it and run the Unity application inside using SteamVR. In the lab setting, the participants are informed that the file is already downloaded and are shown where it is. The participants are informed to return to the survey after completing the test application. In the laboratory experiment we used an HTC Vive VR headset with the Deluxe Audio Strap accessory.

When in the virtual environment, the participant is first introduced to the gaze-directed pointing interface by having them pick the language options, as explained in Section 4.3. Afterwards, instructions about the test itself are shown above the desk, with a "Start" button under it. The language selection and introduction interfaces are shown in Figure G.3. The participant is informed that the robot will show them a letter, and they will have to answer what letter it is. They do this by picking the options "A" or "B", which are selectable at either side of the robot arm, as shown in Figure G.2. After they pick the letter, the robot returns to its outset position. They are also informed that after each task, they have to rate their feeling of safety around the robot. This is done with a Likert scale where they state their agreement to the statement, "I feel safe about the robot". The scale is made from seven selectable boxes, the left-most being labelled "Disagree" and the right-most labelled "Agree" and the middle box being labelled "Neutral". The Likert scale is shown in Figure G.4.

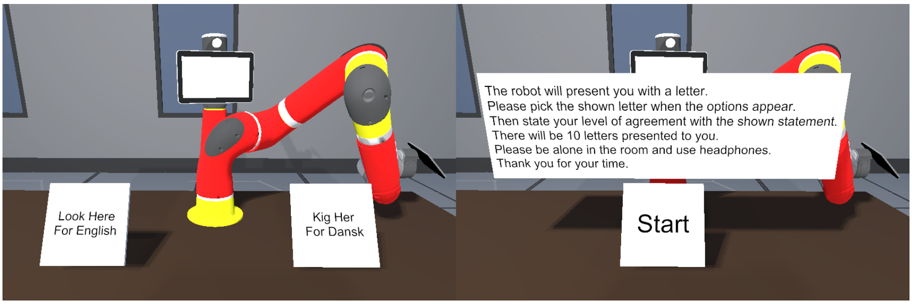


Fig. G.3: Left: The language selection interface. Right: The experiment introduction and start button.

The participant starts with either the slow or fast robot movement setting. When the application is first started, the starting speed is decided by whether the system clock is at an even or uneven second. After the first time, the starting speed will change back and forth between test. This is to attempt to get and even number of samples for each condition in the context of crowd-sourcing data, where the participants are encouraged to have as many people from their household test as possible. They are advised to be alone in the room while testing, though. The first five tasks are performed with the starting speed, after which it changes without warning to the

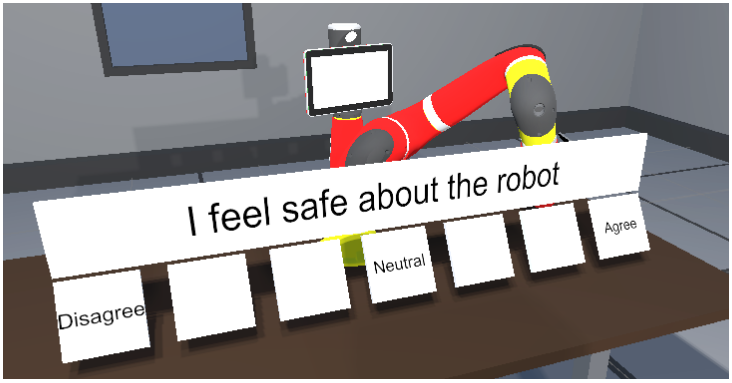


Fig. G.4: The Likert scale interface used in the experiment.

other speed for the last five tasks.

After the last task, the participant is informed that they can find their data log in the "RobotVRTest_Data" folder next to the Unity application executable and that there are further instructions in the survey. The participant is instructed to open the data log in a text editor, such as Notepad, and copy and paste all of the file contents into the first text field of the survey. While this method seems crude in relation to usability and ease of use, it ensures the participant of their privacy and makes transparent what data is sent, as they are able to read it all. They are also presented with the option of sending the log to the researchers via email. After filling in additional details, age, gender and nationality and submitting the survey, the experiment is concluded.

5 Results

While we developed the experiment for crowd-sourcing data, we first need to verify that this is a suitable solution. Therefore we performed the experiment in the laboratory with 20 subjects, 7 female and 13 male, average age 29, raging between 24 and 38. We disinfected the hardware using isopropyl alcohol between each subject.

5.1 Data

The trust scores and task completion times are shown in Figure G.5. Running the Shapiro-Wilk test, the data is not normally distributed when grouped for any of our comparisons, so we perform non-parametric tests. Performing a Friedman test on the repeated trust reports for both groups, starting with slow and fast robot movement, yielded significant difference. The group who started with the slow robot yielded a Chi-squared value of 27.85 ($p < .01$), while the group starting the fast robot had a Chi-squared value of 20.38 ($p < .02$). Analysing the task completions times also yield significant difference for both groups, Chi-squared value of 30.90 ($p < .01$) for the slow robot group, 20.11 ($p < .02$) for the fast robot group.

Performing the Friedman test with post-hoc for the group starting with slow robot movement put task five and six in separate categories, showing a significant effect on reported trust from increasing the movement speed. For the group starting with fast robot movement task five and six were in the same category, suggesting no significant effect on trust when decreasing speed. This means we can retain our first hypothesis as long as it pertains to increasing speed without warning, similarly to prior similar experiments.

The post-hoc test on the task completions times on both movement speed groups yields significant difference between all tasks, making it hard to attribute the changes in completion to the changes in movement speed, rather than task experience or other factors. Because of this results are inconclusive in regards to the second hypothesis.

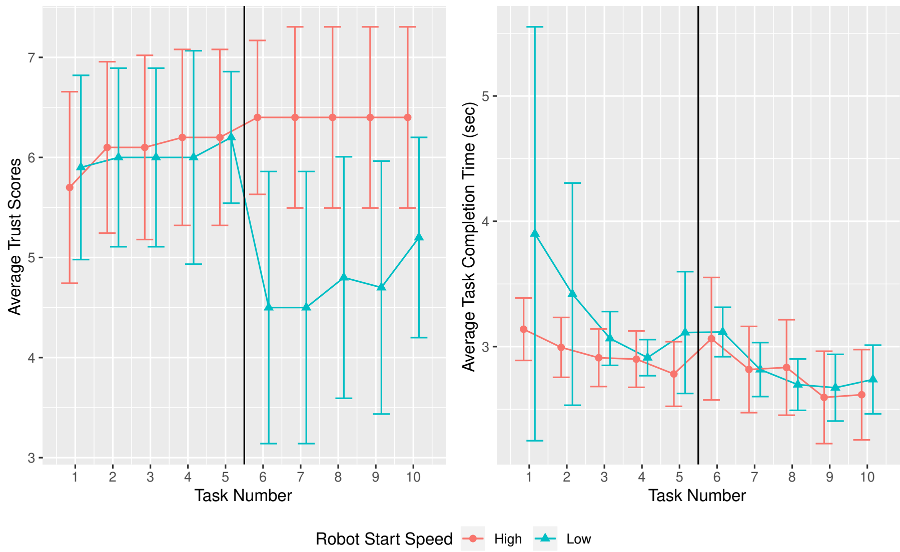


Fig. G.5: Left: The average trust rating after each task. Right: The average task completion time in second for each task. Both include error bars and a vertical line marking the mid-point when the robot movement speed changes.

5.2 Observations

Observing the participants, none had problems with the gaze-directed pointing user interface, as no one spend more than a few seconds before starting to use it, and no questions or comments were made about it.

Several participants asked questions to confirm when they had to start the VR application. Others had trouble opening the log file on Notepad, as it was not the default program for .csv files on the testing computer. Two participants started writing their names in the text field in the survey labelled Data Log, where they were instructed to copy and paste the contents of the log file. This was corrected by the test conductor.

6 Discussion

The trust score analysis suggest a significant effect from changing the robot motion speed. However, due to the context of unsupervised remote testing, the trust questionnaire was shortened to one direct question, compared to previous related experiments, to keep test time short. This may have affected responses, as the participants could suspect something to change to surprise them. Still, the slow increases in trust score shown in Figure G.5 suggest that they were not expecting anything to happen, but they may also have been countering their expectations when reporting, consciously or unconsciously, for the sake of the experiment.

The decision to not label all the buttons in the Likert scale may have given some the impression that some were not selectable. Alternatively, to avoid visual clutter, we could color the buttons something other than the tiles with questions and instructions, to establish without doubt what is interactive and what is not.

Judging by the difficulty some participants had with the data log, it may be worth it to implement a text field in the test applications itself, allowing for easily copying the data log or using a button to add the log to the system clipboard. There is a risk that if a participant has problems with this step and there are no one to assist, that they will abandon the survey at the very end. It could also help to rename the Unity executable file to something more helpful, like "Click Me To Start", rather than the generic application name, to help participants who are unfamiliar with Unity programs, their icons or names.

The issues with participants starting to fill out the survey prematurely or putting their name by mistake may be caused by the laboratory setting, as they may be rushing due to a feeling of pressure. We may assume that if a participant have decided to start the survey on their own accord, they will take their time with it.

Aside from the before-mentioned difficulties, the participants did not generally have issues using the HTC Vive VR headset, even if they stated that they had no experience using them. This suggest that participants will have little to no difficulty if they are performing the test at home with their own hardware.

7 Conclusion

We set out to design a procedure that enables test subjects to participate in HRI experiments from their own home using a VR headset, allowing for data gathering through crowd-sourcing in cases of where laboratory experiments are not possible.

We set up a virtual environment with a virtual Rethink Robotics Sawyer robot arm in order to assess human-robot trust in close-proximity HRC. We disrupt the participants expectations about the robot by having it move towards the participant and changing movement speed midway through the experiments, and we assess the trust throughout the test by having the participant report their feeling of safety after each task.

Verifying the experiment in a laboratory setting with 20 participants, we found that reported trust was significantly affected disrupting expectations, which is consistent with similar experiments with a real robot. The effect on task completion time was inconclusive.

Participant had no problems using the VR headsets during the laboratory experiment, and everyone understood the gaze-directed pointer interface utilized in the test application. Some quality-of-life improvements can be made to the test application pertaining to data logging while maintaining participant anonymity. This will be important in limiting user frustration, increasing the change the participants will complete the experiment when performing the test at home with their own hardware.

Using VR applications for unsupervised experiments into HRI is possible from a usability and validity standpoint, while there are a few improvements to made to our procedure.

References

[1] L. Punnett and D. H. Wegman, "Work-related musculoskeletal disorders: the epidemiologic evidence and the debate," *Journal of Electromyography and Kinesiology*, vol. 14, no. 1, pp. 13–23, feb 2004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1050641103001251?via{%}3Dihub

[2] Arbejdstilsynet Denmark, "Anmeldte erhvervssygdomme 2012-2017," 2017. [Online]. Available: https://at.dk/media/4814/bilag-b-tabeller-med-anonymitet-final.pdf

[3] K. Hald, M. Rehm, and T. B. Moeslund, "Proposing human-robot trust assessment through tracking physical apprehension signals in close-proximity human-robot collaboration," in *The 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) IEEE International Symposium on Robot and Human Interactive Communication*. IEEE Computer Society Press, 2019.

[4] U. Gadiraju, S. Möller, M. Nöllenburg, D. Saupe, S. Egger-Lampl, D. Archambault, and B. Fisher, "Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd," in *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 2017, pp. 6–26.

[5] S. Chernova, N. DePalma, E. Morant, and C. Breazeal, "Crowdsourcing human-robot interaction: Application from virtual to physical worlds," in *2011 RO-MAN*. IEEE, 2011, pp. 21–26.

[6] E. Matsas and G.-C. Vosniakos, "Design of a virtual reality training system for human–robot collaboration in manufacturing tasks," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 11, no. 2, pp. 139–153, 2017.

[7] O. Liu, D. Rakita, B. Mutlu, and M. Gleicher, "Understanding human-robot interaction in virtual reality," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 751–757.

[8] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, jan 2004. [Online]. Available: http://hfs.sagepub.com/cgi/doi/10.1518/hfes.46.1.50{_}30392

[9] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 5, pp. 517–527, oct 2011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0018720811417254

[10] A. D. Dragan, S. Bauman, J. Forlizzi, and S. S. Srinivasa, "Effects of Robot Motion on Human-Robot Collaboration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. New York, New York, USA: ACM Press, 2015, pp. 51–58. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696473

[11] Asset Store Original, "Snaps Prototype Office 3D Environments, Unity Asset Store." [Online]. Available: https://assetstore.unity.com/packages/3d/environments/snaps-prototype-office-137490

SUMMARY

To help enable robot-augmented production, where robots assist production staff to relieve them of repetitive and strenuous tasks, the purpose of the research is enabling real-time human-robot trust assessment by inferring decreases human trust from signs of physical apprehension. To ensure safe and productive human-robot collaboration we have to ensure an appropriate level of trust in the robot, as too much trust can lead to dangerous situations, whereas too little trust can lead to loss in productivity. The main hypothesis is that if the user experiences a decrease in trust, they will increase their distance from the robot by stepping or leaning away from it.

A series of experiments were performed using a Rethink Robotics Sawyer robot and an augmented reality enabled human-robot collaboration cell, using projection to display task critical information within the shared work space. Participants performed repeated tasks with the robot, midway through which the robot would disrupt the participants' expectations in order to decrease their trust towards it. Their movements were assessed using an infrared camera for body tracking to correlate it with decreases in trust.

ISSN (online): 2446-1628
ISBN (online): 978-87-7210-895-7

AALBORG UNIVERSITY PRESS