Exploring the Effect of Automation Failure on the Human's Trustworthiness in Human-Agent Teamwork

Nikki Bouman

October 20, 2022

Exploring the Effect of Automation Failure on the Human's Trustworthiness in Human-Agent Teamwork

by

Nikki Bouman

in partial fulfilment of the requirements for the degree of

Master Computer Science
in Computer Science and Engineering

at the Delft University of Technology
to be defended publicly on November 7th 2022

Responsible Professor:	Prof. dr. Catholijn M. Jonker
Supervisors:	Dr. Myrthe L. Tielman
	Carolina Jorge
	Dr. Jie Yang (external)

The work in this thesis was made in the:

Interactive Intelligence
Department of Intelligent Systems
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

TU Delft

Abstract

Collaboration in teams composed of both humans and automations has an interdependent nature, which demands calibrated trust among all the teammembers. For building suitable autonomous teammates, we need to study how trust and trustworthiness function in such teams. In particular, automations occasionally fail to do their job, which leads to a decrease in human's trust. However, research has given contradictory statements about the effects of such a reduction of trust on the human's trustworthiness, i.e. human's characteristics that make them more or less reliable to the automation. As such, this study investigates how automation failure in a human-automation teamwork scenario affects the human's trust in the automation and human's trustworthiness towards the automation. We present a between-subjects controlled experiment in which the participants perform a simulated task in a 2D grid-world, collaborating with an automation in a "moving-out" scenario. During the experiment, we measure the participants' trust and trustworthiness regarding the automation both subjectively and objectively. Our results show that automation failure negatively affects the human's trustworthiness, as well as their trust in and liking of the automation. Learning the effects of automation failure in trust and trustworthiness can contribute to a better understanding of the nature and dynamics of trust in these teams, foreseeing undesirable consequences and improving human-automation teamwork.

Acknowledgements

I left this chapter to be written on the last possible day, ending my thesis with a thank you. However, now I am stuck. There are so many people to thank, it would be impossible to call them all by name. Thank you all participants, for just participating, or also making the day more fun by having a cup of coffee with me afterwards, or even finding more participants. Thank you Catholijn Jonker, Myrthe Tielman, and especially Carolina Jorge for being my supervisors, guiding me through my whole thesis and encouraging me to keep thinking and achieving more. Special thanks to Tessa Gagestein, Vanisha Jaggi and Vishala Ramrattansing for clearing my mind with boardgame, video game and movie days. Saving the best for last... I want to thank my parents for being there for me, physically, mentally, and of course financially. You really took away a lot of stress by not forcing anything (I think I was stricter on myself than you would ever be on me). And Steyn Scheffers, thanks for listening to all my worries and complaints, comforting me on stressful days, and believing in me when I failed to do so. I think back to my time studying with joy and pride, but feel relieved that a new part of my life is now starting.

Contents

List of Figures

List of Tables

1

Introduction

Artificial Intelligence (AI) is intelligence demonstrated by machines, which are often referred to as automations (Laurent et al., 2019). The concept of automations was introduced even before Christ. Around the year 400 BC people believed that Talos, a bronze automation, was programmed to walk the coast of Crete three times a day, guarding the island (Kearns, 2016). Following this, many more automations were imagined, such as Leonardo da Vinci's robot, or the Digesting Duck from Jacques de Vaucanson (Wood, 2003).

The creation of automation began with the intention of having it take over a human's task, filtering into general applications. This was "often without being called AI because once something becomes useful enough and common enough it's not labelled AI any more" (Bostrom, 2006)[1]. With the rise of automations, certain human jobs became obsolete, creating widespread worry among people that they would lose their jobs due to automations. However, Licklider (1960) did not see a future where automation would replace humans, but rather in which humans and automations would collaborate. Indeed, automation shows benefits for humans in terms of improved decision-making, performance, and reduced workload (Parasuraman et al., 2000).

A common example where humans and automations need to collaborate, is in Search and Rescue groups (e.g. Blitch, 1996; Guznov et al., 2016). Here, robots and humans work together to try to find and rescue humans from a dangerous area, taking advantage of the differences between the human's and automation's strength. Humans, for example, are better at recognising danger, whereas automations are better at remembering large amounts of data (Bradshaw et al., 2011). Human-automation teams operating in life-saving situations are becoming increasingly common. There are surgical automation that allow surgeons to focus on the complex aspects of a surgery (e.g. Laurent et al., 2019), and image recognition software that assist doctors in diagnosing patients with skin diseases (e.g. Wei et al., 2018; Aggarwal, 2019). In such teams, trust between the teammates is essential for the successful functioning, since trust connects similar interests and pro-team behaviour, and creates behavioural norms that encourage collaboration (Groom & Nass, 2007).

However, trust is not a simple concept. Literature has focused on exploring trust in human-automation teams, particularly looking into the differences between human-human and human-automation trust (e.g. Jian et al., 2000; Centeio Jorge et al., 2021; Groom & Nass, 2007), how this trust can be optimised (e.g Webber, 2008; Groom & Nass, 2007; J. D. Lee & See, 2004), and which factors reduce this trust (e.g. Robinette et al., 2017; Madhavan et al., 2006; Falcone & Castelfranchi, 2004).

Unfortunately, automations occasionally fail to perform as expected. This can be, for example, due to a developed bias (e.g. Dastin, 2018; Victor, 2016; Hern, 2018; Levin, 2016), incorrect advice (e.g. Lohr, 2021; Beardsworth & Kumar, 2019), or a malfunctioning, putting bystanders in life-threatening or even life-taking situations (e.g. McCausland, 2019; Matyszczyk, 2016; Yadron & Tynan, 2016). Research found that automation failure has a significant impact on a person's trust. Consequently, that person has a significantly lower level of trust in it in subsequent interactions (Robinette et al., 2017). It is proposed that such a reduction in trust can result into the trustor himself becoming less trustworthy (Falcone & Castelfranchi, 2004; Tullberg, 2008). This is not

[1] Retrieved from CNN, accessed on 07/10/2022.

certain, as research also proposed that there might not be an influence of trust on the trustor's trustworthiness (Salem et al., 2015). As this is a contradiction, and as these studies are not performed on human-agent teamwork, this forms the foundation of our research. As such, in this work we explore the effect of automation failure on the human's trustworthiness.

1.1 Research question

This thesis aims to answer the following question:

What is the effect of automation failure on the human's trustworthiness in human-automation teamwork?

To answer this question, we designed an experiment in which a human participant collaborates with an automation on a task, during which the automation starts to fail. During and after this experiment, we track the human's behaviour and ask them to fill out a questionnaire, capturing their trust in the automation and their own trustworthiness in the process.

1.2 Chapter overview

The remainder of this thesis is divided into five more chapters. We look at research that is related to automation, trust, trustworthiness and teamwork. In chapter 3 we explain the experiment design and process for the participants of the experiment. It contains a detailed description of the reasoning behind the experiment's design and setup. Subsequently, in chapter 4 we analyse the data that is collected during this experiment. Several factors will be tested, after which the results are discussed in chapter 5. Additionally, this chapter highlights the shortcomings of this study, along with several recommendations for future work on this topic. We end this thesis in chapter 6 with our conclusion.

2

Background and related work

This chapter covers the background and related work that revolves around the research question. The basic concepts of trust and trustworthiness are covered, and a model of trust is presented. All factors of this model are explained, after which the chapter transitions into human-automation teams, the trust within those teams and the effect of automation failures on this trust.

2.1 Trust

Trust is a social construct that originates from interpersonal relationships (Dagli, 2018). In this thesis, trust is defined as the willingness of a party to be vulnerable to the actions of another party (Mayer et al., 1995). We refer to the trusting party as the "trustor", and the party being trusted as the "trustee". Trust is based on the expectation that the trustee will perform a particular action important to the trustor, irrespective of the ability to monitor or control the trustee. This implies a situation in which an individual is vulnerable, and their vulnerability rests with the actions, behaviours, or motivations of another individual (Wagner et al., 2018).

Trust is a subjective attitude of the trustor, which involves the perceived trustworthiness of the trustee (Centeio Jorge et al., 2021). This is visualised in Figure 2.1, where the factors of perceived trustworthiness lead to the degree of trust. In this model, trust is shaped by the propensity to trust and the factors of perceived trustworthiness. Trust and risk directly influence the decision to engage in trust.

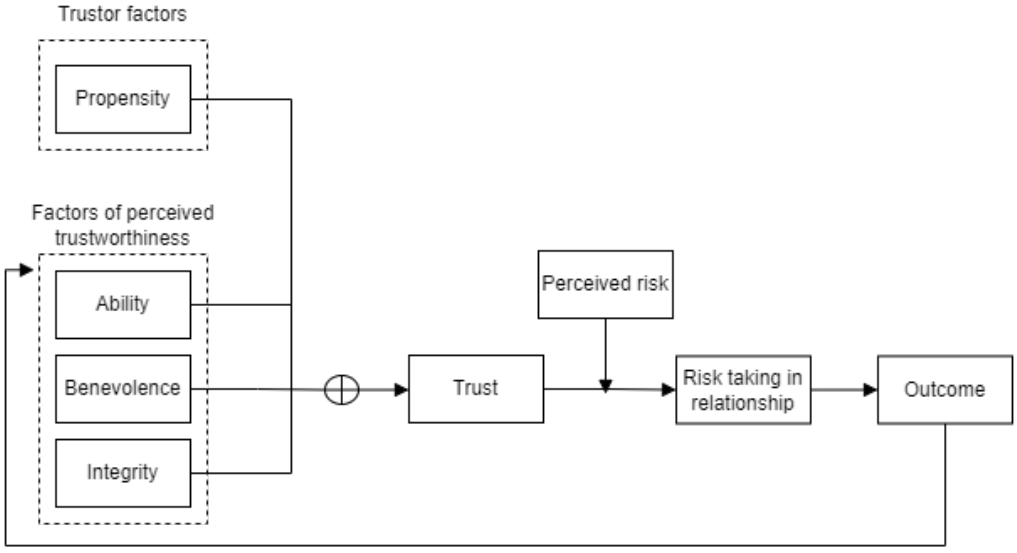


Figure 2.1: Model of trust from Mayer et al. (1995).

2.1.1 Trustworthiness

Trustworthiness can be seen as an objective property of the trustee. It is defined as the extent to which an actor has the ability to execute relevant tasks, is benevolent towards its teammates, and demonstrates integrity (Mayer et al., 1995). Ability refers to the skills and knowledge that enable one to have influence within some specific domain. Benevolence is the trustor's belief in the trustee's desire to do good on behalf of the trustor (wanting to help). Lastly, integrity is the trustor's belief that the trustee adheres to a set of principles that the trustor finds acceptable. Consider the difference between asking a good friend to paint your wall, and employing a professional painter. The painter has better abilities for the task, but the friend will be more benevolent, since they are helping you as a friend, not for payment. Moreover, you will most likely know how much integrity your friend has, but this is difficult to observe in someone you just met. However, your expectations might be high, since the painter is getting paid for their work and wants to maintain their status. The trustee's ability, benevolence and integrity are the factors that constitute their trustworthiness. The trust that the trustor has in them changes with the trustor's perception of the trustee's trustworthiness.

2.1.2 Appropriate trust

Calibrated trust between teammates means that someone's perceived trustworthiness of a teammate matches the teammate's actual trustworthiness (de Visser et al., 2020). This is often referred to as "appropriate trust". Over-trust refers to trust that surpasses the trustee's ability, benevolence or integrity, whereas under-trust means that the trust falls short of these factors of the trustee (J. D. Lee & See, 2004). When there is appropriate trust, there is no under-trust or over-trust (Centeio Jorge et al., 2021).

2.1.3 Propensity to trust

One of the factors that affect the degree of trust that the trustor has in a trustee is the propensity to trust. Propensity to trust is a factor that affects the likelihood that the trustor will trust. It can be thought of as the general willingness to trust others (Mayer et al., 1995). It influences how much trust the trustor will have in the trustee, before the trustor knows the details of the trustee. Mayer et al. (1995, p. 716) proposes that "The higher the trustor's propensity to trust, the higher the trust for a trustee prior availability of information about the trustee". Propensity to trust is included in Figure 2.1 to affect the degree of trust.

2.1.4 Risk and reward

In this thesis, we see risk as an influence on the decision to trust, rather than directly influencing the degree of trust itself (see Figure 2.1). It is possible that the trustee somehow betrays the trustor's trust, whether this is done deliberately or not, thus the act of trusting someone carries the risk of being betrayed. Mayer et al. (1995, p. 726) states that "the perception of risk involves the trustor's belief about likelihoods of gains or losses outside of considerations that involve the relationship with the particular trustee". Trustors do not only evaluate the risks, but also the rewards that come from an engagement. This is incorporated in the model from Johnson and Bradshaw, which is based on the model from Mayer et al., and can be seen in Figure 2.2. However, trust can only be evaluated with respect to the context in which and the method by which the action is being performed by a specific trustee (Johnson & Bradshaw, 2021).

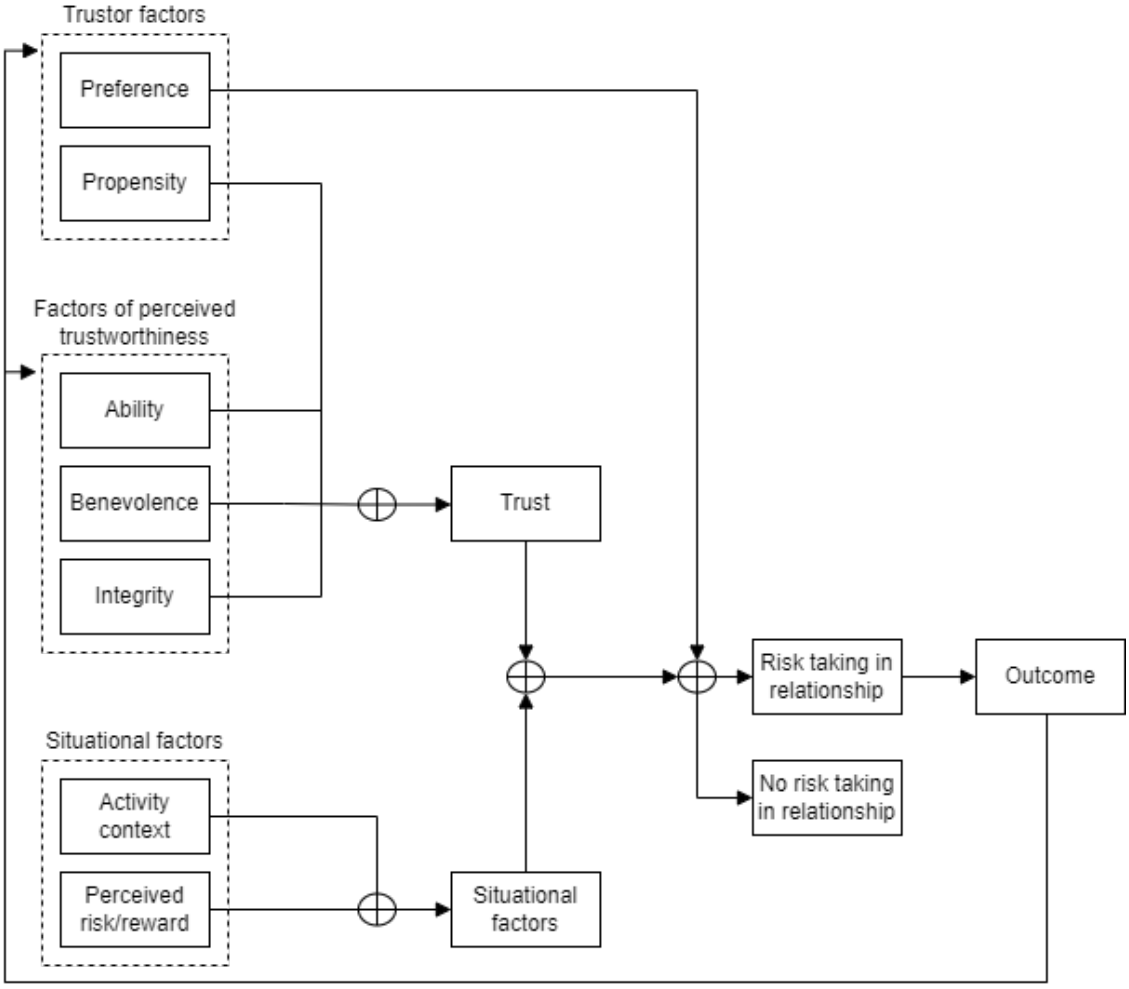


Figure 2.2: Model of trust from Johnson and Bradshaw (2021), extended from Mayer et al. (1995).

2.1.5 Activity context

"Do you trust them?" must be followed by "trust them to do what?" (Mayer et al., 1995). This is activity context, another situational component that influences the degree of trust. For example, you might trust someone to look after your dog, but not to look after your children. It is the same person that you have to trust, but in a different context. The details of the work and the methods by which it is performed matter (Johnson & Bradshaw, 2021). To be more explicit, one must answer the *what*, *who*, and *how* to assess trust (Johnson & Bradshaw, 2021); trust *who* to do *what* and *how* will they do it? This is also incorporated into the new trust model depicted in Figure 2.2.

It is suggested that an assessment of ability does not generalise across dissimilar tasks or situations (Mayer et al., 1995). The ability to accomplish the task must be assessed in every new situation. If the trustor does not assess the capabilities of the trustee in every new situation, the trustor might be trusting the trustee inappropriately.

2.1.6 Preference

The model that Johnson and Bradshaw have created (Figure 2.2) includes another factor not found in the original model (Figure 2.1) from Mayer et al.: a preference. This is because the trustor may trust the trustee, but still decide not to engage with them based solely on preference. "For example, nobody doubts the efficacy of automatic shifting mechanisms of today's cars, yet some people still choose to manually shift for the pleasure of it." (Johnson & Bradshaw, 2021, p. 383).

2.2 Trust in human-automation teams

A human-automation team is a team that consists of at least one human and one automation. Human-automation teams share knowledge, depend on each other's output, and work together on common functions (Chen & Barnes, 2014). This thesis defines automation as "any sensing, detection, information-processing, decision-making, or control action that could be performed by humans but is actually performed by a machine." (Moray, Inagaki, & Itoh, 2000, p. 1). People interact with automations on a daily basis. Consider a Google Assistant, self-driving car, or robot vacuum cleaner as examples. Automations are increasingly being developed as partners rather than tools (Klein et al., 2004), allowing humans to focus on their own tasks and strengths and covering their weaknesses.

Successful technologies take advantage of the differences between the human's and automation's strengths, as human reasoning has different characteristics than algorithmic reasoning (Chen & Barnes, 2014). For example, algorithms may only achieve limited accuracy, but they outperform humans because of their consistency (Kahneman & Klein, 2009), making them more suitable for tasks that are too repetitive, fast, or dangerous for humans to perform (Kohn et al., 2021). For example, it was found that in military command and control decision-making teams, human-automation teams performed significantly better than teams consisting of solely humans (Fan et al., 2010).

To maintain credibility and performance, frequent interaction with the members of a team is considered as an important element of team effectiveness, since it builds a relationship with the other members of the team, resulting in greater trust (Webber, 2008). Trust between teammates is essential for the successful functioning of a team (Groom & Nass, 2007).

2.2.1 The difference of trust in humans and automations

Jian et al. (2000) propose that human-human relationships are conceived differently from human-automation relationships, because an assessment of distrust in another human seems more negative than in an automation. For example, benevolence is about interpersonal relationships, meaning it might not develop in human-automation relationships in the same way it does for human-human ones (Centeio Jorge et al., 2021). Furthermore, there is symmetry to interpersonal trust, in which the trustor and trustee are each aware of the other's behaviour, intents, and trust (Deutschi, 1960). However, there is no such symmetry in the trust between humans and automations (J. D. Lee & See, 2004). Trusting something that is unable to trust and to feel guilt or betrayal proves to be difficult for humans (Groom & Nass, 2007). It has been shown that even the propensity to trust humans differs from the propensity to trust automations (Madhavan & Wiegmann, 2004).

Studies suggest that people perceive automations as more credible sources of information than humans (J. Lee & Moray, 1992; Wright et al., 2016). However, humans also tend to rely on their own decisions, even when provided with feedback that their performance was inferior to that of the automation (Dzindolet et al., 2002). Moreover, humans have a tendency to blame the automation for negative outcomes (Morgan, 1992; Frieainan, 1995), while being reluctant in giving credit to the automation (Madhavan & Wiegmann, 2007). The less a user trusts the automation, the sooner they will intervene in its progress of a task (Olsen & Goodrich, 2003). Trust depends on the timing, consequences, and expectations associated with failures of the automation (J. D. Lee & See, 2004).

2.2.2 The effect of automation failure

Research shows that a single error from an automation strongly affects a person's trust (Robinette et al., 2017). A mistake made by automation will cause a person to have a significantly lower level of trust in it in subsequent interactions (Robinette et al., 2017). High expectations in the automation result in a steeper decline in trust in case of an automation failure than it would in case of a human error (Madhavan et al., 2006). In other words, humans expect automations to have a near perfect performance, causing them to pay too much attention to errors made by automations (Dzindolet et al., 2002), whereas they do not expect their human partners to be perfect. As Falcone and Castelfranchi state: "To every trustee's failure corresponds a reduction of the trustor's trust towards the trustee itself" (Falcone & Castelfranchi, 2004, p. 3).

Falcone and Castelfranchi and Tullberg found that when a person has a reduction in trust in someone, their own trustworthiness towards that person is decreased. This is found in human-human studies (Tullberg, 2008) or in multi-agent studies based on human-human theories (Falcone & Castelfranchi, 2004). In contrast, it is suggested that a reduction in trust might not influence the trustor's trustworthiness (Salem et al., 2015), found in a human-automation non-collaborative setting. However, this study claims to have found a significant difference in trust in their two conditions (one with automation failure and one without), while there might not be. They have performed a Mann-Whitney U-test, resulting in a U-value of 129.5 with 40 participants. In order to have a significant test result, the U-value with 20 participants in each of the conditions should be equal to or lower than 127. This means that they might not have found a difference in trust in the automation, where they also did not find a difference in the human's trustworthiness.

If, despite this possible invalid test result, the study from Salem et al. is significant, then our study examines whether there are similar outcomes in a collaborative setting. If the study is not significant, then there are no studies done on the effect of a reduction in trust, or specifically automation failure, on the human's trustworthiness in a human-automation setting. In this thesis, we conduct a study on this, filling the scientific gap on this part of the trust dynamics in human-automation teams.

3

Methodology

The aim of this thesis is to find the effect of automation failure on the human's trustworthiness in human-automation teamwork. It is known that automation failure will cause a person to have a significantly lower level of trust in it in subsequent interactions, but the effect of it on that person's trustworthiness is unknown. To find an effect, a study is designed and conducted in a collaborative setting. This chapter elaborates the design choices for this experiment and provides an overview of the participants, materials, measurements and procedure.

3.1 Hypothesis

As stated in subsection 2.2.2, there is contradicting research regarding the effect of automation failure. Falcone and Castelfranchi (2004) and Tullberg (2008) believe that the effect will be negative, based on human-human theories, while Salem et al. (2015) believe that there might not be any effect, studied in a human-automation setting. However, the study in a human-automation setting might not have found a reduction in trust at all, leaving room for further research.

In this thesis, we believe that the trustor's trustworthiness decreases when automation failure occurs. We speculate this because we believe that at least benevolence and integrity would significantly decrease if the trustee fails to perform the collaborative task. This results into the following hypothesis:

> *Automation failure has a negative effect on the human's trustworthiness in human-automation teamwork.*

3.2 Experimental Design

The experiment in this thesis employs a fixed experimental between-subjects design. All participants are assigned to one of the two experimental conditions: either the one with automation failure (experimental group), or the one without (control group). The independent and controllable variable is automation failure, with the dependent variable being the human's trustworthiness. For the experiment, the participant performs a simulated task on the computer, collaborating with an automation. This task is executed twice, with a questionnaire before and after each game. An overview of the whole procedure is given in Figure 3.1.

In short, the experiment consists of an informed consent, a pre-test, the first game, a mid-test, the second game, and a post-test. The condition to which a participant is assigned makes a difference in the second game only, as this is where the automation will either fail or act the same as before. The pre-test includes general questions about the participant and their propensity to trust automation. The mid-test and the post-test contain identical questions, since we want to measure the difference between the two games. These questions attempt to assess the human's trust in the automation and the human's trustworthiness. This design allows us to measure the repercussions of automation failure on trust and trustworthiness.

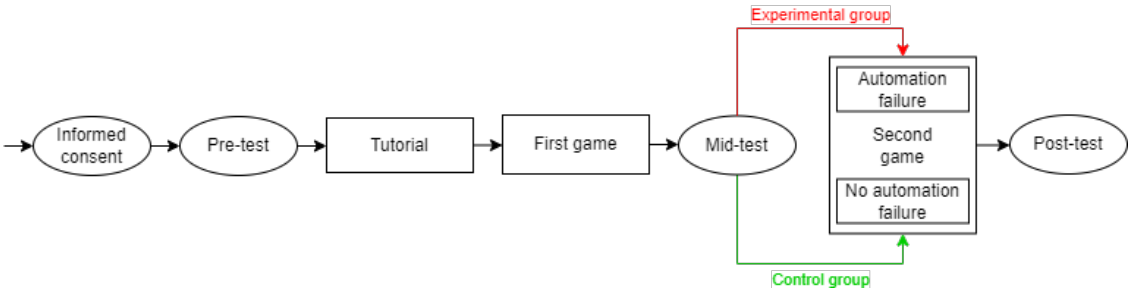


Figure 3.1: The procedure for each participant, with the second game depending on the group they were assigned to (control group or experimental group).

3.3 Participants

The aim was to recruit at least 52 participants, as this was the minimum sample size needed when trying to achieve a large effect size, large power and small error probability while still aiming for a feasible amount of participants. Eventually, 54 participants were recruited for the experiment, resulting in 27 participants per group ($d = 0.944$, $\alpha = 0.05$, $1 - \beta = 0.92$). Participants range from 18 to 69 years old. All participants report normal or corrected-to-normal vision and normal colour vision. Participation is voluntary, and no rewards are given. The study obtained ethical approval from the Human Research Ethics committee of the Delft University of Technology (approval ID: 2303).

The participants are split in half to fill both conditions. They are assigned to either one of the conditions based on their answers to the pre-test (age, gender, and gaming experience), as we try to balance the two groups. Table 3.1 provides an overview of the participants' characteristics.

	Control group		**Experimental group**	
	Amount	*Percentage*	*Amount*	*Percentage*
Men	11	40.7%	10	37%
Women	16	59.3%	17	63.0%
Other	0	0%	0	0%
Daily gamers	2	7.7%	2	7.4%
Weekly gamer	4	15.4%	6	22.2%
Monthly gamers	5	19.2%	7	25.9%
No gamers	15	57.7%	12	44.5%
18-29 years old	20	74.1%	24	88.9%
30-39 years old	4	14.8%	1	3.7%
40-49 years old	0	0%	0	0%
50-59 years old	2	7.4%	1	3.7%
60-69 years old	1	3.7%	1	3.7%
70+ years old	0	0%	0	0%

Table 3.1: An overview of the characteristics of the participants.

3.4 Materials

Calendly is a service that allows participants to book an appointment for the experiment. The scheduled experiment is conducted using a computer running the MATRX software. This shows a two-dimensional simulated collaborative task: moving boxes to the desired location, which is inspired by the video game Moving Out[1]. The participant's trust in and liking of the robot and their own perceived trustworthiness are measured in between the games using a questionnaire in Qualtrics. This section further elaborates these three concepts.

3.4.1 MATRX

MATRX[2] stands for Human-Agent Teaming Rapid Experimentation. It is a Python package designed for human-agent team research. It provides a basic user interface in a two-dimensional grid-world with human controlled agents, autonomous agents, and the possibility of teams. This gives the developer a basic structure to implement their experiment in. For this thesis, the MATRX core version 2.1.2 is used. MATRX requires the use of Flask[3] (version 2.0.2 is used). This is Python-based (version 3.9) web framework. This package allows the MATRX interface to be displayed on your localhost. We run the experiment on a Windows computer with an Intel Core i7-6700HQ CPU @ 2.60GHz processor and 8GB RAM.

3.4.2 Qualtrics

Qualtrics[4] has developed a survey software, frequently used by the Delft University of Technology. People can use this software to build and distribute their own online surveys, as well as collect and analyse the results. In this thesis, we use Qualtrics for the questionnaires, which are further elaborated in subsection 3.6.1. During the experiment, the participant is requested to complete this questionnaire, of which the response is saved by Qualtrics.

3.4.3 Calendly

Calendly[5] is a free online appointment scheduling tool. We use it in this thesis to allow participants to schedule an appointment for the experiment. This provides us with an overview of all planned experiments, as well as the opportunity to send a reminder to the participant.

3.5 Task

The goal of the task that is programmed using MATRX is to collaboratively move boxes to the correct location in what we call the "dropzone" within the time restriction. There are two agents in the field: the human (controlled by the participant) and the robot (the automation). There are three types of boxes, which determine whether they can be carried alone or together. The team score increases for every box delivered correctly into the *dropzone*. The robot behaves differently in the control group than in the experimental group, which is further explained in this section after all the game's aspects and design choices are elaborated.

[1] store.epicgames.com/en-US/p/moving-out
[2] matrx-software.com
[3] flask.palletsprojects.com
[4] qualtrics.com
[5] calendly.com

3.5.1 Game design

The game contains multiple elements that need to be further elaborated. Two of the most important aspects are the human, controlled by the participant, and the robot, which is an automation. Further aspects are places, items or indicators in the game. This section describes all these aspects of the game.

The human

The participant in the experiment can move the human in the game (on the left in Figure 3.2) by using the 'WASD' or arrow keys. In this game, there are boxes that need to be placed in the correct location, which will be further explained in this chapter. They can lift a box alone with 'L', or together with the automation by pressing 'H' (see Figure 3.3 for how the agents are displayed when carrying a box). If the robot is not near the human when pressing 'H', an exclamation mark appears, indicating that the human wants to carry it together with the robot (see Figure 3.3). This exclamation mark represents the agent asking for help, and disappears when the call for help is answered, or when the agent moves away from the box. Finally, the human can use the 'P' key to place a box that it is currently carrying.

The robot

The robot (on the right in Figure 3.2) moves around on its own, carrying and placing boxes alone, or asking for help with a box. In short, the robot would go through the following steps:

1. Check the dropzone, which box is next?
2. Find the (closest) box of that type.
3. Walk to the box.
4. (Ask for help with the box, depending on its type)
5. Carry the box.
6. Walk to the corresponding place in the dropzone.
7. Drop the box.

During the steps in this process, it would constantly check the following:

- Does the human call for help?
- Has the next box according to the dropzone changed? (e.g. a box has broken, or the human has placed the box already)

The robot starts with looking at the dropzone, checking which box is next in line to be delivered in the correct order. After seeing what type of box is next to be delivered, the robot searches the field for the closest box of that type. It then walks to that box, and depending on the type (further explained in the subsection "Boxes"), it asks for help or carries it alone. When the robot asks for help, a red exclamation mark appears (as shown on the left in Figure 3.3). When it asked for help, the robot waits for the human for about ten seconds. If the human did not come to help, the robot will check which box is the next to be delivered, and walk to the closest box of that type again. It will go back to the box it wanted to carry in the first place after that box has been carried to the correct location, or the waiting time for help has exceeded again. Once the robot is carrying

a box, it walks to the corresponding dropzone location and puts the box in its destination. If the human and the robot are carrying the box together, the human is in charge of walking while the robot waits for the drop. When the box is dropped, the robot starts the process steps over again, looking at which box is next up to be delivered to the dropzone.

While carrying a box alone, the robot constantly checks to see if the destination is still empty. The robot does this because the human can be quicker, placing the same type of box on the desired location before the robot is able to. The robot also continuously checks what the human is carrying. If the robot discovers that it is carrying the same type of box as the human, before it has crossed the safezone, it places the box in the safezone, trying to keep the collaboration as smooth as possible. Lastly, the robot constantly checks whether the human asks for help. If they do, and the robot is not currently carrying anything, it immediately goes to the human. If the robot is carrying something, it first places the box in the destination, and then goes to the human, if they are still asking for help.

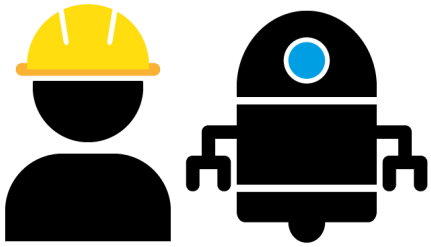


Figure 3.2: The two agents in the game, respectively the human and the robot.

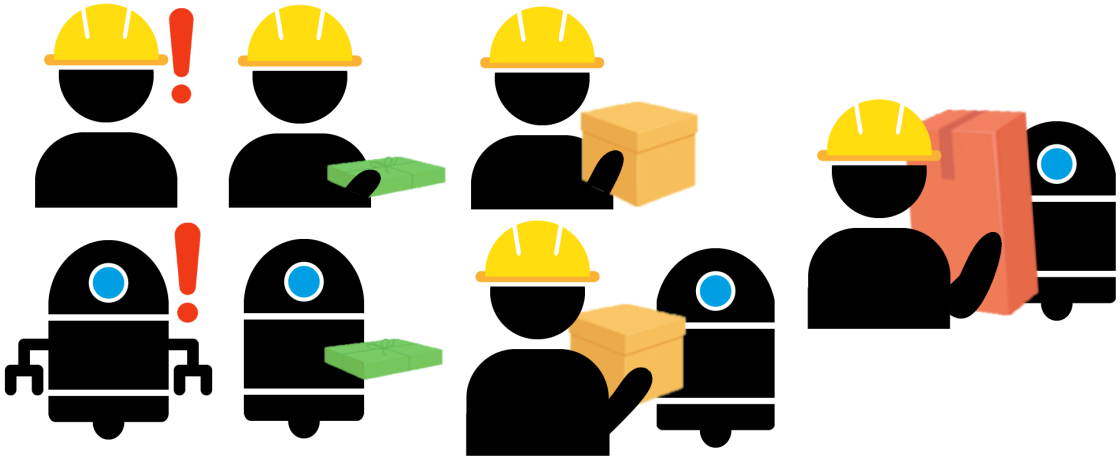


Figure 3.3: The agents and all the actions that involve a different image (asking for help, carrying a box alone, and carrying a box together respectively).

Boxes

There are three types of boxes that can be lifted and moved: light, medium and heavy (see Figure 3.4). The light box (recognisable by its green colour and small size) can only be carried by the human or the robot separately. The medium box (recognisable by its yellow colour and medium size) can be carried alone by the human or together with the robot. However, if the human chooses

to carry it alone, they will be walking thirty times slower than usual. Lastly, there is a heavy box (recognisable by its red colour and big size), which can only be carried together. Table 3.2 provides a brief overview of the various boxes. All boxes can break when placed incorrectly, indicated by dents in the box and a darker colour (see the lower row in Figure 3.4), which is discussed later in this chapter.

When a box is being carried together by the human and the robot, the human is in control over their movement. When the human decides to drop the box, the robot follows this action and both agents are displayed separately again, with the box that they held now also on the ground, out of their hands.

The decision for these types of boxes is made because we want to make the human and the robot depend on each other as much as possible, highly favouring collaboration, while at the same time making it possible for the human to stop the collaboration. This way, we can easily observe the human's behaviour and intentions, resulting in the ability to study the human's trustworthiness.

The robot and the human depend on each other because they at least have to carry all the heavy boxes together, if they want to obtain the maximum amount of points (further explained in the subsection "Score"). This forces collaboration between the agents. Moreover, a box that is being carried together can break (explained later in this chapter), making them even more dependent on each other. The medium box provides the human with the choice to not collaborate with the robot, lowering the human's trustworthiness. Lastly, the light boxes exist to try to balance the collaboration, such that the human may decide to focus solely on the light boxes whenever they do not want to collaborate with the robot. For example, we see the act of ignoring the robot's calls for help as a decrease in the human's benevolence towards the robot, thus showing a decrease in the human's trustworthiness.

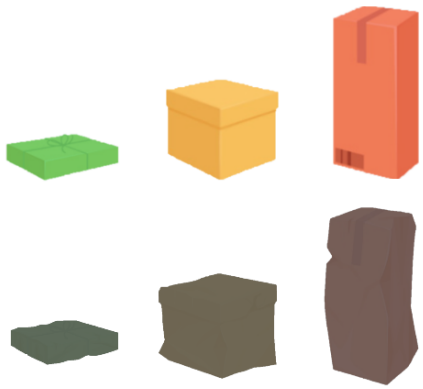


Figure 3.4: The three types of boxes that are in the game (light, medium, and heavy respectively). The top row shows them in their initial and intact state, whereas the bottom row shows their broken form.

	Human	Robot	Human and robot
Light	X	X	
Medium	X[6]		X
Heavy			X

Table 3.2: The different types of boxes (light, medium, heavy) and how they can be carried.

The dropzone

The dropzone is the line of more transparent boxes above the black fence, as can be seen in Figure 3.5. This is where the boxes in the field need to be delivered. When placing a box on the corresponding slightly transparent version, that box cannot be picked up again. For example, in Figure 3.5, a green and red box have already been placed, indicated by the first two boxes in the dropzone being normal opacity. If a box is placed in the wrong dropzone place (not on the same type of box), then the agents can lift it up again to move it to the correct spot. When boxes in the field break (explained in the next section), the box that is next in line in the dropzone of the order of boxes from left to right is also shown as broken, indicating that that box does not need to be delivered any more (see Figure 3.5, the third box in the dropzone).

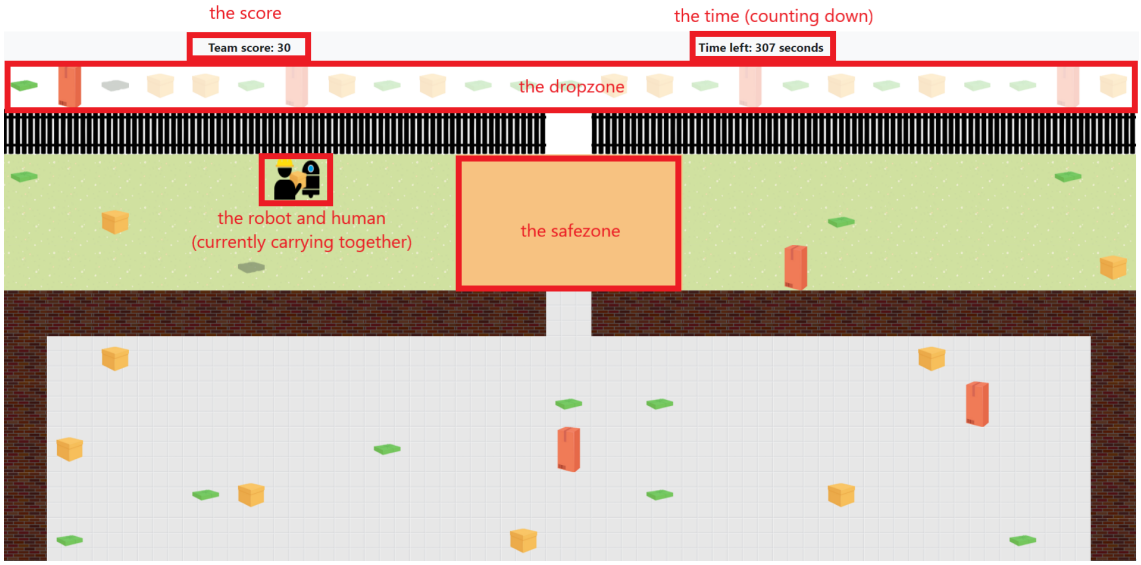


Figure 3.5: A screenshot of the game, including information about the different agents and zones.

The safezone

Boxes need to go to the dropzone, but it is possible for the human to accidentally pick up a box that is not the next one in line. For this reason, the safezone was created. In this zone (indicated in orange, in between the wall and fence openings) boxes can be safely placed without breaking it. All boxes that are placed outside the drop- or safezone break. The option to break boxes made for a way for the human to deliberately break boxes, decreasing their trustworthiness. Breaking boxes also creates an option for the robot to fail.

Score

The game keeps track of the team score. Each box that is correctly placed in the dropzone contributes ten points to the score, regardless of the type of box. Boxes can be placed in the dropzone in any sequence, but delivering them from left to right (without skipping any) gives the team five extra points per box. To ease the decision to stop the collaboration, all boxes add the same amount of points to the score. Moreover, the five extra points they gain for following the

[6]Thirty times slower movement speed.

sequence nudges the human towards collaborating with the robot. Making boxes worth different points could make the extra points inconsiderable.

In this experiment, there are twenty-five boxes located in the dropzone (twelve light, eight medium, five heavy). When a box is broken, the extra five points can still be received for the box next to it. This means that the human can choose to purposely break a box, thus skipping it, without losing the extra points. This also creates a way to make it evident to see that the human's trustworthiness has decreased, for example, if the human decides to only break heavy boxes.

To emphasise the concept of collaboration, the use of a team score is chosen rather than individual scores. The extra points awarded to the team for placing a box in the correct order is given to compel the participant to stick to the order. In other words, the extra points are given to force the user to lift all the types of boxes. Without forcing the order, there is no particular reason for the user not to carry all the green boxes on its own first.

Time

A time restriction is added to force the user into making a decision to complete the task as quickly as possible. For example, the human will notice that the robot is failing to do their job, so because of time constraints, the human would not try to carry all the medium and heavy boxes with the robot, hoping that the robot will not drop them, but rather aim for the light boxes to be sure of the delivery. Aside from this reasoning of experimental design, the time restriction makes the experiment more convenient in practical terms, since people with less gaming experience would possibly take longer to finish the task.

3.5.2 Automation failures

If the participant is in the experimental group and currently playing the second game, then the robot has to show faulty behaviour. This failure should be a performance-related factor (e.g. reliability, false alarm rate, failure rate, etc.), since those were found to be better predictors of trust development than attribute-related factors (e.g. robot personality, anthropomorphism, etc.) (Hancock et al., 2011). Therefore, the focus was to let the robot fail in terms of their performance. This consists of breaking boxes, placing them in the wrong location in the dropzone, or picking up a box that is not the next up box according to the dropzone sequence.

Overall, the robot breaks eight boxes during the game (two light, four medium, two heavy). The emphasis lies on the medium boxes, since they can optionally be carried alone or together. Four boxes are delivered in the wrong place, which are always light boxes, since the robot is not in control when carrying the medium and heavy boxes. Lastly, three boxes are collected out of order. This can be any type of box, but if it is not a light box, the robot merely asks for help at the 'wrong' box.

3.6 Measurements

To observe how the human's trustworthiness evolves when the automation fails, we need a way to measure their trustworthiness. We do this via a questionnaire (subjective measurements) and by observing the human's behaviour (objective measurements).

3.6.1 Subjective measurements

There are three main categories shown in the model from Johnson and Bradshaw (Figure 2.2) that each need to be measured to understand the reasoning of the participant: trustor factors, factors

of perceived trustworthiness, and situational factors. This section describes how these factors are measured with the use of a questionnaire.

Factors of perceived trustworthiness

The first category from the model is the *factors of perceived trustworthiness*. Asking the participant to self-report their own level of trust is extremely common within this field of research (Hancock et al., 2011). Many existing questionnaires to measure the perceived trustworthiness of another agent exist (e.g. Cahour & Forzy, 2009; Adams, Bruyn, Houde, & Angelopoulos, 2003; Madsen & Gregor, 2000; Merritt, 2011; Singh, Molloy, & Parasuraman, 1993). Hoffman et al. (2018) discuss and review several. Their paper concluded to a final questionnaire, adapting many items from Merritt (2011). Since this author has more usable scales on other factors that we want to measure (which will be discussed in the next paragraphs), we decided to use her scale to measure the factors of perceived trustworthiness.

Merritt (2011) has evaluated her scale in an experiment in which participants had to use a fictitious automated weapon detector with the task to screen luggage. The Chronbach's alpha ranged from $a = 0.87$ to $a = 0.92$. The participant could answer to the statements in a 5-point Likert-type response scale ranging from *strongly disagree* to *strongly agree*. The statements were stated from the human's perspective, for example focusing on whether the human thinks they could rely on the robot. Since this automation was used for advice, we have to alter the statements to fit the context of our task, changing it to the robot and its ability to deliver boxes. The statements for this category are included in both the mid-test and the post-test, and can be found in Appendix A (number 15 and 19).

Own perceived trustworthiness

The most essential concept we want to measure is the *human's own perceived trustworthiness*, since this is a significant aspect in our research question. To maintain consistency in the questionnaire, we decide to use the same scale as the factors of perceived trustworthiness. The only difference is the subject, shifting from the robot to the human. (e.g. "I have confidence in the actions of the robot" becomes "The robot can have confidence in my actions".)

During the pilot of this study, we found that this particular part of the questionnaire did not quite capture what we aimed for. When the participant was asked to elaborate their answers, they said that, overall, the robot could for example still have confidence in their actions. This made us realise that the questions should be more specific. The header was changed to state that the following statements were about the second game, and the statement was changed to for example "The robot was able to have confidence in my actions", making it past tense.

The next part of the pilot showed that a ceiling effect was occurring. Remembering that not only Likert scales but also sliding scales were often used for self-reports (Kohn et al., 2021), we decided to change this scale to a slider, providing more granularity. Moreover, the statements were exaggerated (e.g. "The robot was able to have complete confidence in my actions"), making it less tempting to fully agree with the statement.

Trustor factors

Another one of the categories is the trustor factors, consisting of *preference* and *propensity*. Merritt et al. (2013), who developed the trust scale that was mentioned in the preceding paragraphs, has also constructed a propensity to trust scale. This scale contains questions concerning how likely the participant is to trust an automation without knowing the details of the automation. The participant can answer in a 5-point Likert-type response scale ranging from *strongly disagree* to

strongly agree. The complete scale can be found in Appendix A (question 14). We did not alter any questions from this scale.

The other factor in this category is preference. It is difficult to formulate questions regarding preference, since the participant generally has no former experience with the specific task ahead. Merritt has developed a third scale that measurements liking, which is defined as "the degree to which the user feels positively toward the automated system." (Merritt, 2011, p. 358). Since someone having a preference can be defined as "having a greater liking for one alternative over another", this scale could come close to what we want to measure. Moreover, if we would not include this questionnaire, it would be the only part of Merritt's questionnaire that we are not including. We therefore decide to include the liking scale in the experiment, appearing both in the mid-test and post-test. This scale contains statements about the human's feelings towards the automation, e.g. wishing the robot wasn't around, which could be answered in a 5-point Likert-type response scale. It is slightly altered to fit the context of our task (changing the automation in the questions to 'the robot'). The altered scale can be viewed in Appendix A (questions 16 and 20).

Situational factors

The last category of factors from the trust model are the situational factors, consisting of *activity context* and *perceived risk/reward*. Since this is a category that is included for assessing different situations, and we are only interested in the difference between the two conditions, this was not necessary to include in the questionnaire. The only context changing within the experiment, is the difference in automation failure or no automation failure, which is being controlled for.

Strategy

A factor that was added to the questionnaire is the *strategy* of the participant. Knowing their strategy gives more insight into the decisions they made and possibly why their trustworthiness does or does not change. For example, a study found that participants developed a preference for less demanding tasks (Botvinick & Rosen, 2009). If such a thing is the case in our experiment, it would be convenient to know and take into account with the analysis. Moreover, by letting the participant read these possible strategies after the first game, they often realise what is actually possible during the game (e.g. during a pilot one of the participants said to understand why boxes can be broken, after reading the strategy about skipping boxes without losing the extra points). This will stimulate them to think about their actions, and make faster decisions if they encounter automation failure.

3.6.2 Objective measurements

The downside of self-report measurements is that they require interruption of the task, or, if administered at the end of the task, subject to memory failures and the participant's bias (Kohn et al., 2021). Furthermore, self-report results do not consistently and perfectly align with actual trust behaviour (Kohn et al., 2021). To compensate for this, the participant's behaviour in the game is logged. With this, we cannot acquire a trustworthiness level equal to reality, as there is only so much we can observe, but we can reason what it means to be trustworthy in this specific experiment. We will go over the factors that will give us a proxy of the objective trustworthiness.

Benevolence towards the robot shows that you want to help the robot, and is one of the three factors of trustworthiness. In this experiment, wanting to help the robot be observed by counting how many times the human would respond to the call for help from the robot. We will log:

- Participant answered to request for help from the robot.

Cooperation with the robot is another factor that shows trustworthiness and can be observed in this experiment. Being cooperative here means that medium and heavy boxes should be carried together without breaking, calls for help should be answered with actions of helping, and the participant should ask for help as well. For this, we will add to the log:

- Participant asked for help.
- Participant broke a box.
- Participant carried a box alone.
- Participant and robot carried a box together.

The types of boxes are also registered with each action, making a distinction between carrying a medium or a heavy box together. These objective measurements allow for a comparison of the behaviours in the first and second game.

We wish to observe the ability of the participant. The game keeps track of the score, and logs it. However, this cannot provide us with an indication of the participant's ability, since it is the collaborative score. When the participant is in the experimental group, the robot is manipulating this score, influencing the total score. Although the robot would want to break the same boxes in every experiment, it would depend on the participant on whether this box would actually be broken. For example, if the human always carries medium boxes alone, the robot would not be able to break a single medium box, making a difference of four broken boxes (40 points, not counting the bonus for correct order). We therefore decide to not include the participant's ability when observing the objective trustworthiness.

3.7 Procedure

Each participant follows the same procedure, but sometimes in a different condition, depending on their assignation. A flowchart is given in Figure 3.1 for an overview of the whole procedure. The complete questionnaire can be viewed in Appendix A.

Upon arrival, each participant is asked to read the *informed consent*. The complete informed consent can be viewed in section A.1. It is important that the informed consent does not contain any information about the robot failing, since if automation's failures are known in advance, even after these failures occur, trust is not necessarily adversely affected (Lewandowsky et al., 2000).

Only when they agree to participate, can they continue to the *pre-test*, consisting of questions about their gender, age, gaming experience, and propensity to trust automations. The instructor then asks about their vision, precluding the effects that uncorrected vision or colour blindness might have. At this stage, they are given a number corresponding to the group they would be assigned to.

After this pre-test, they have to complete a *tutorial*. This tutorial is completely self-explanatory by pop-up boxes that explain the rules to the participant. This ensures that every participant is given the same information. During this, the participant is free to ask any questions. After the tutorial, the game starts, which is always a game without automation failure.

When the participant finishes the game, they fill in the *mid-test*. The mid-test contains questions about their trust in and liking of the robot, their own perceived trustworthiness, and their strategy for the game. Upon completion, they enter a game either with or without automation failure, depending on their assignation. This last game is followed by the *post-test*, containing the same questions as the mid-test, and ending the experiment. However, participants can optionally

state why they changed their strategy regarding the first game (mid-test), if they changed their strategy at all.

3.8 Data analysis

The data that results from the subjective measurements, which consisted mostly of Likert or sliding scales, is converted into numerical scores. The objective data that is collected during the two scenarios is analysed by a Python script that counts the occurrences of each action that we wanted to record (e.g. how many times the human broke a box). The numerical data from both the subjective and objective measurements is matched for a complete dataset of each participant. This data is analysed in R using the packages `readxl (1.4.0)`, `ggpubr (0.4.0)`, `rstatix (0.7.0)`, `dplyr (1.0.8)`, `rgl (0.110.2)`, `cluster (2.1.4)` and `ggeffects (1.1.3)`.

4

Results

In this chapter, we share the results from the analysed data. We show whether, and if so, how, the participant's trustworthiness changes differently in the control group compared to the experimental group. Both the subjective and the objective data that is obtained is used for these tests. We then investigate how the trust in and the liking of the automation changes. Before executing a test, the assumptions were checked. Every reported test therefore meets the necessary assumptions. An alpha level of .05 was used for all statistical tests. The altered questionnaire used in the experiment shows a Cronbach's Alpha of $\alpha = 0.87$, showing no significant change in internal consistency.

4.1 Participants

Before we start analysing the data that we retrieved with our measurements, we want to verify that gender, age and gaming experience are evenly distributed over the two groups, having no influence on the results. Participants also had to report on their propensity to trust automations. It is important to see whether this is significantly different when comparing the two groups, since, ideally, we want the same type of participants in each group. Lastly, since the participants all answer questions after the first game, which was the same game for everyone, we want to make sure that there are no significant differences in the two groups regarding their trust in and liking of the robot.

4.1.1 Gender

The gender of the participants is self-reported, where either male, female or other is chosen. This study only contains males and females, of which the distribution over the groups can be viewed in Figure 4.1. We transform the data to numerical (Female = 0, Male = 1). Since the data is not normally distributed, we conduct a Wilcoxon Rank-Sum test. There was no significant difference between the gender in the control group and the experimental group, W = 391, p = .584.

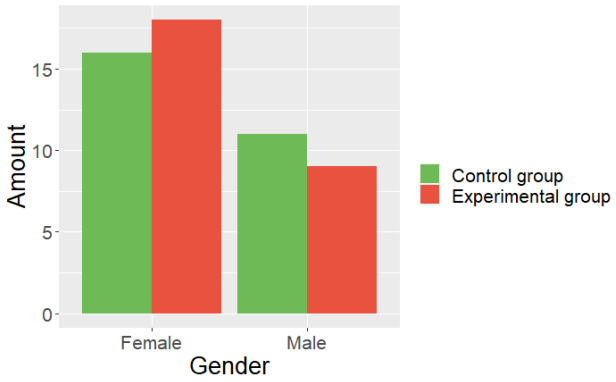


Figure 4.1: The gender distribution per group.

4.1.2 Age

The participants report their age in ranges from 18-29, 30-39, 40-49, 50-59, 60-69, and 70+. For this test, we convert these to integers 1 to 6 respectively. As the data is not normally distributed, a Wilcoxon Rank-Sum test is executed. This finds no significant differences between the ages in the control group and the experimental group, W = 415, p = .197. The distribution of ages is visualised in Figure 4.2.

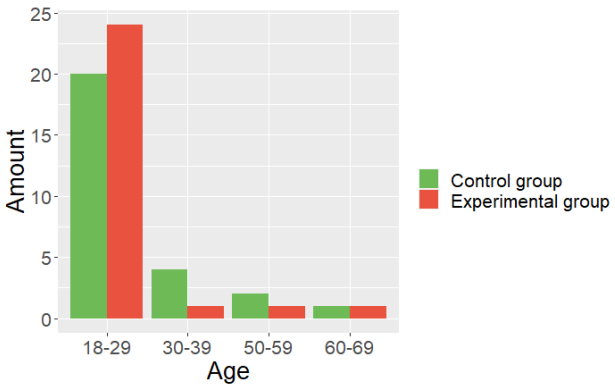


Figure 4.2: The age distribution per group.

4.1.3 Gaming experience

Gaming experience is self-reported by the participants in levels of Daily, A few times a week, A few times a month, and Never (or almost never). These are then converted to integers 1 to 4 respectively. This data, which is visualised per group in Figure 4.3, does not have a normal distribution, which is why a Wilcoxon Rank-Sum test is conducted. This shows no significant difference between the gaming experiences between the control group and experimental group, W = 399, p = .519.

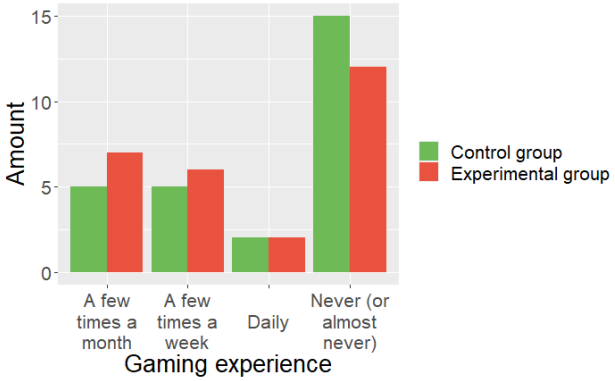


Figure 4.3: The gaming experience distribution per group.

4.1.4 Propensity to trust automation

We want to see whether there were no significant differences between the propensity to trust of the participants in the control group (M = 3.55, SD = 0.79) compared to the participants in the experimental group (M = 3.62, SD = 0.80). The data is viewed in Figure 4.4. An independent t-test shows that there is indeed no significant difference, t(52) = -0.31, p = .755.

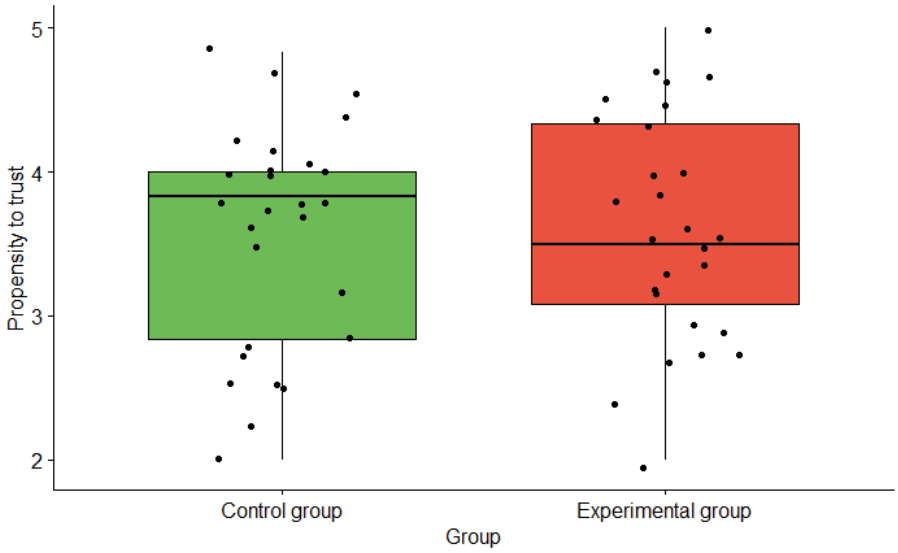


Figure 4.4: A box plot of the propensity to trust per group.

4.1.5 Trust (after the first game)

To make sure that the participants in either group do not show different levels of trust in the robot after the first game, where every participant collaborates with a robot that does not fail, we conduct a test to compare the means of trust after the first game. The data is shown in Figure 4.5. As the data is not normally distributed, we perform a Wilcoxon Rank-Sum test. This shows that there is no significant difference of trust between the control group and experimental group, W = 337, p = .645.

4.1.6 Liking (after the first game)

The last component that needs to be examined for significant differences is the human's liking of the robot. The data, shown in Figure 4.6, does not have a normal distribution, which is why we perform a Wilcoxon Rank-Sum test. The test demonstrates no significant difference between the liking after the first game when comparing the control group with the experimental group, W = 267, p = .092.

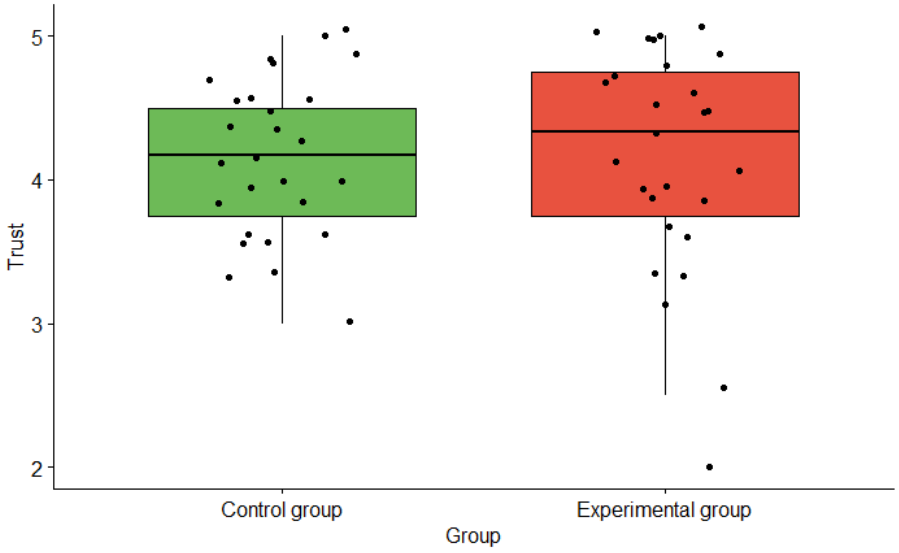


Figure 4.5: A box plot of the participant's trust in the robot after the first game per group.

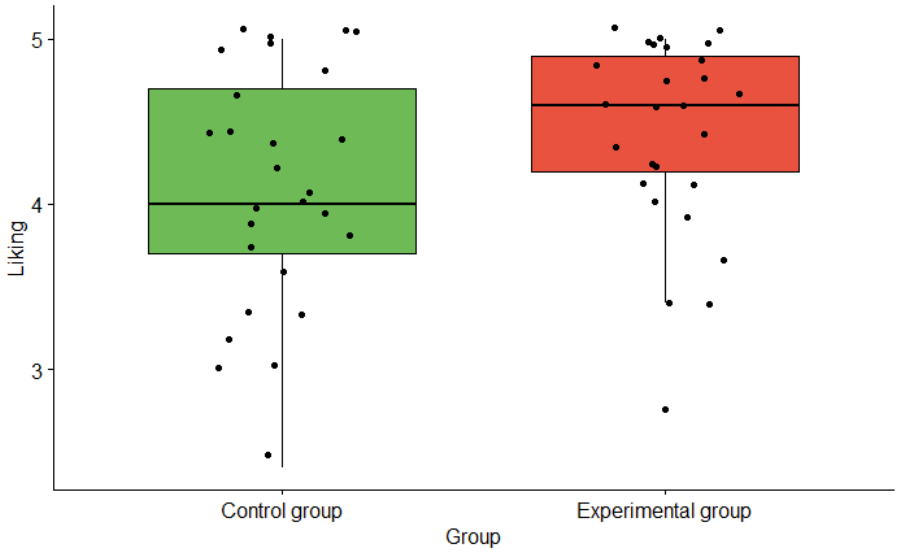


Figure 4.6: A box plot of the participant's liking of the robot after the first game per group.

4.2 Trust in robot

The research question is about the effect of automation failure on the human's trustworthiness. This contains the assumption that automation failure negatively affects the *trust* the human has in the automation. We therefore need to verify that the human's trust indeed has a significant decrease for the experimental group. For this, we first examine whether there is a significant difference in the relative trust that the participants have in the control group compared to the experimental group. The relative trust is calculated by:

$$T_{rel} = T_2 - T_1,$$

where T_1 and T_2 are the trust in robot after the first or second group respectively

The *higher* the relative trust is, the *bigger* the increase of trust from the first game to the second game. The data is visualised in Figure 4.7. For the investigation we conduct an independent two-sample t-test, where the participants in the control group (M = -0.136, SD = 0.55) compared to those in the experimental group (M = -2.70, SD = 0.79) demonstrate a significant difference in the participant's trust in the robot in the two groups, $t(52) = 13.8$, $p < .001$.

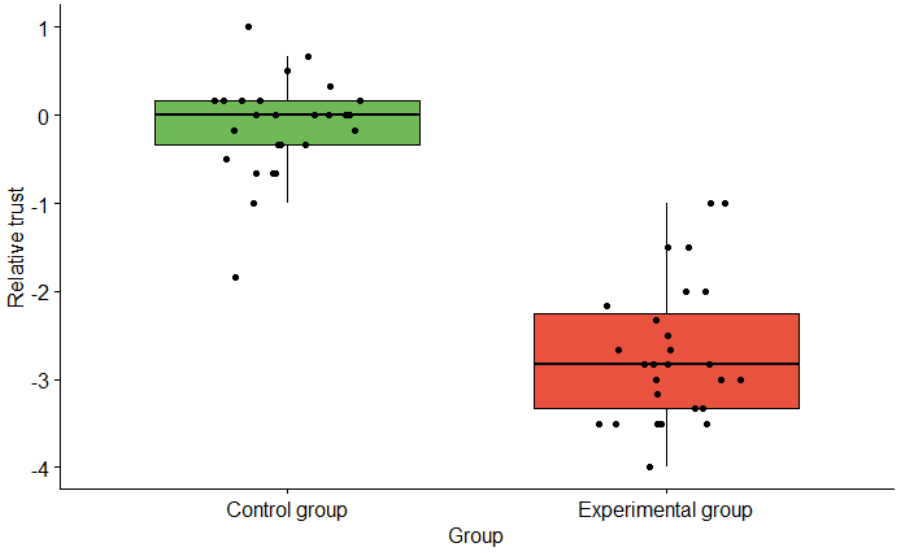
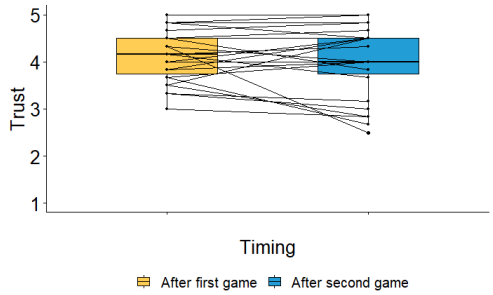
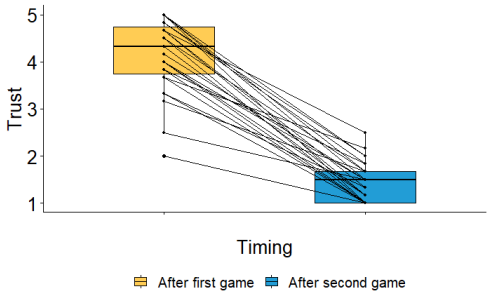


Figure 4.7: A box plot of T_{rel} per group.

Examining this difference, we want to see whether there is an individual difference per group when comparing the trust after the first game with the trust after the second game. This data can be viewed in Figure 4.8. Two paired samples t-tests are performed. There is no significant change in the control group. We do find a significant change in the experimental group when comparing the trust in the first game (M = 4.15, SD = 0.79) with the second game (M = 1.44, SD = 0.42); $t(26) = 17.7$, $p < .001$.



(a) Comparing T_1 and T_2 respectively in the control group.



(b) Comparing T_1 and T_2 respectively in the experimental group.

Figure 4.8: A comparison of the differences in the human's trust in the robot (T) per game in separate groups.

4.3 The human's trustworthiness

In this research, we are interested in how the human's trustworthiness is affected by automation failure. The human's trustworthiness can be split in two categories: the subjective and the objective trustworthiness. We first analyse the subjective trustworthiness, after which we continue with the objective trustworthiness, trying to relate the two with each other.

4.3.1 Subjective trustworthiness

The participants report on their *own trustworthiness* with six questions, answering on a scale from -100 to 100. By taking the average of these six questions, we obtain an estimation of their own perceived trustworthiness. Since these questions are asked after both the first and second group, we particularly want to look at the decrease or increase of their trustworthiness. To obtain this relative trustworthiness, a simple formula is used.

$$TW_{rel} = TW_2 - TW_1,$$

where TW_1 and $TW_2$are trustworthiness after the first or second group respectively

The *higher* the relative trustworthiness, the *bigger* the increase of trustworthiness from the first game to the second game. The collected data per group is visualised in Figure 4.9.

An independent two-sample t-test is performed to compare the change in the participant's subjective trustworthiness in the control group and experimental group. The 27 participants in the control group (M = 11.3, SD = 18.6) compared to the 27 participants who experienced automation failure (M = -15.2, SD = 35.1) demonstrate a significant difference in the participant's own perceived trustworthiness in the two groups, t(40) = 3.47, p = .001.

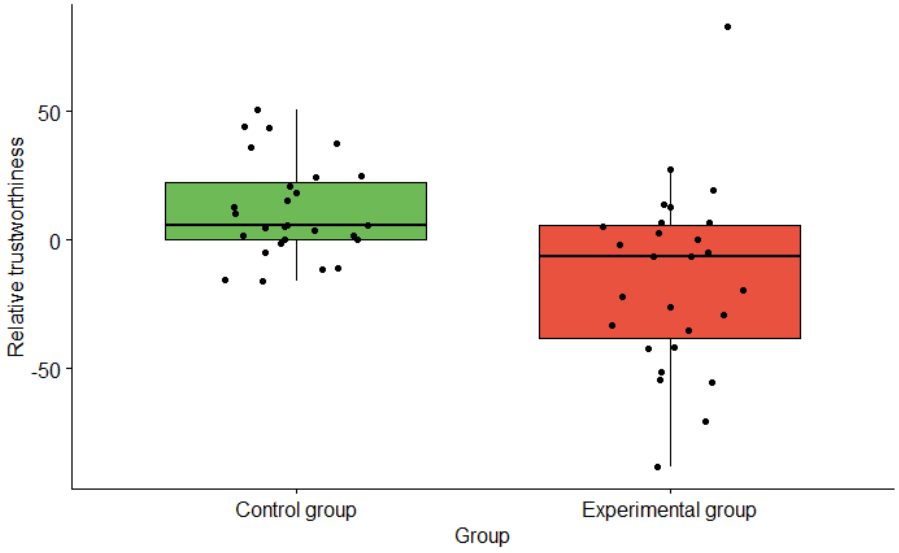
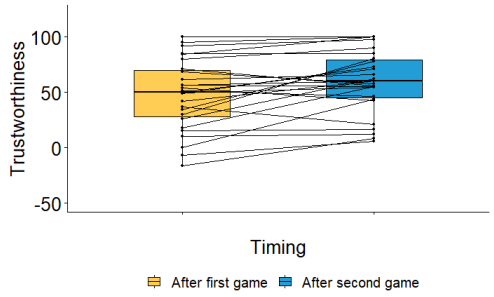
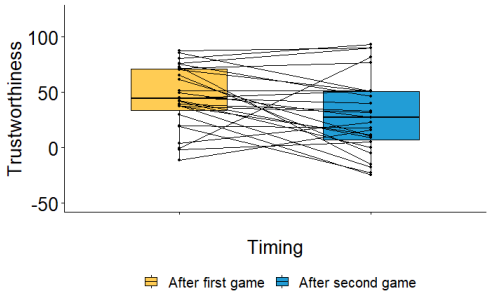


Figure 4.9: The subjective relative trustworthiness of the participants, visualised in a box plot.

To further investigate the difference in subjective trustworthiness from the first to the second game, the data is split in two: one for all the participants in the control group, and one for all the participants in the experimental group. The data of these separate groups is visualised in Figure 4.10. Here we can see that in general, the participant's own perceived trustworthiness increases in the control group, while it decreases in the experimental group. Two paired samples t-tests are performed for examination of these changes. There is a significant difference in the own perceived trustworthiness after the first game (M = 48.0, SD = 31.6) and the second game (M = 59.3, SD = 28.4); t(26) = -3.16, p = .004 in the control groupin terms of an increase. The experimental group also has a significant difference in the own perceived trustworthiness after the first game (M = 45.7, SD = 28.0) and the second game (M = 30.6, SD = 35.4); t(26) = 2.25, p = .033, but for this group it is a decrease.



(a) The difference in the participant's own perceived trustworthiness after the first and second game respectively in the control group.



(b) The difference in the participant's own perceived trustworthiness after the first and second game respectively in the experimental group.

Figure 4.10: A comparison of the differences in the participant's own perceived trustworthiness (TW) in separate groups.

Correlation between trust and trustworthiness

We further investigate this by looking at the correlation between the relative trust and relative trustworthiness. Figure 4.11 shows a plot containing the relative trust against the relative trustworthiness, where Local Polynomial Regression Fitting tries to find a smooth curve per group.

Figure 4.11: A plot of the collected relative trust in combination with relative trustworthiness per group, where the curves are made with Local Polynomial Regression Fitting.

Although we observe no logical line or curve, we do want to see whether we can somehow predict the participant's subjective trustworthiness per group using the relative trust values. Using a model with an interaction effect between the relative trust and the group, it is shown that this statistically significantly predicts the relative trustworthiness ($F(2, 51) = 8.281$, $p < .001$, adj. $R^2 = .216$). Furthermore, the relative trust in the control group does not significantly contribute to this prediction, while the relative trust from the experimental group does ($p < .001$). A plot of the predictions is given in Figure 4.12.

As the results from the prediction shows linearity, we further examine this idea. We conduct a Pearson correlation test to examine the interaction between trust and trustworthiness. This found a positive correlation, $r(52) = .49$, $p < .001$. We observe that there is a medium result for linear correlation.

During examination, we notice that the data points from the control group and experimental group also show that the data from the two groups are segregated. Using K-means cluster analysis, we examine whether this method finds similar clusters. Figure 4.15 shows the result of this analysis, where it has an accuracy of 93%.

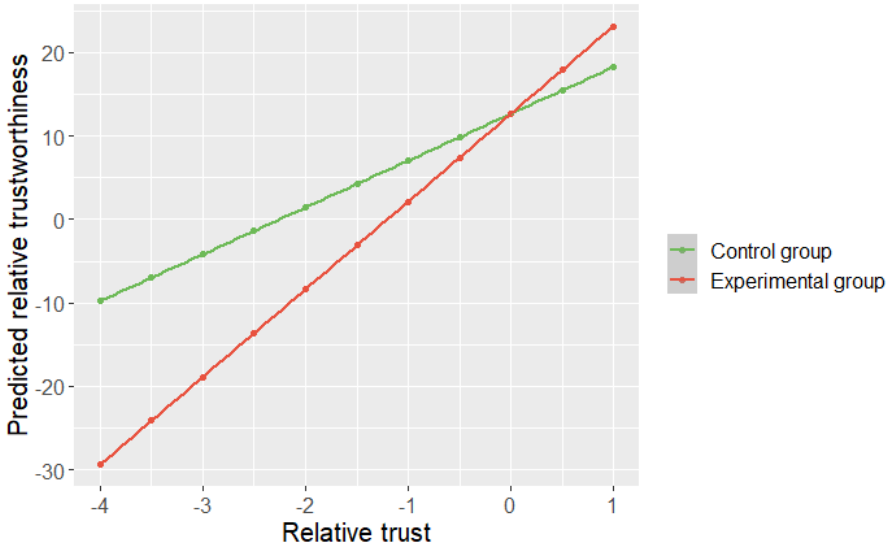


Figure 4.12: Prediction of the interaction effect between relative trust and the groups for relative trustworthiness.

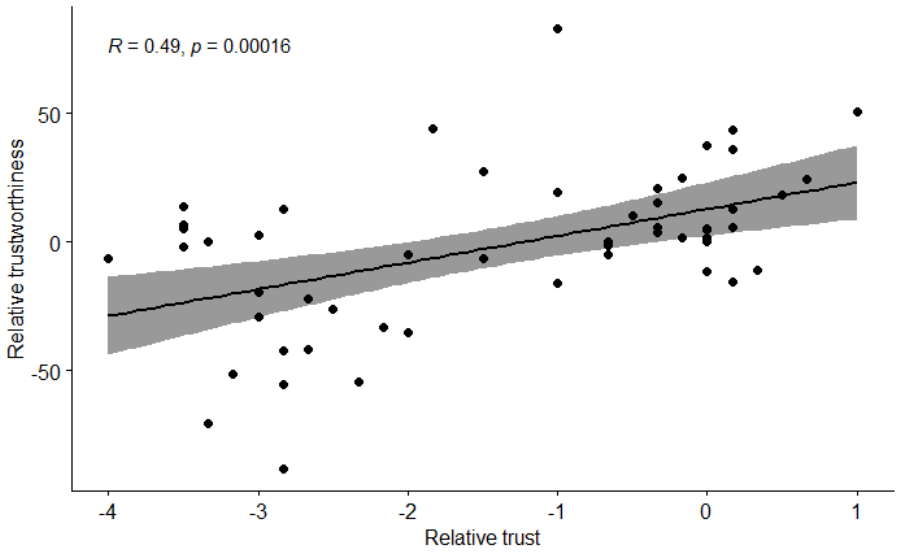


Figure 4.13: Pearson correlation between relative trust and relative trustworthiness.

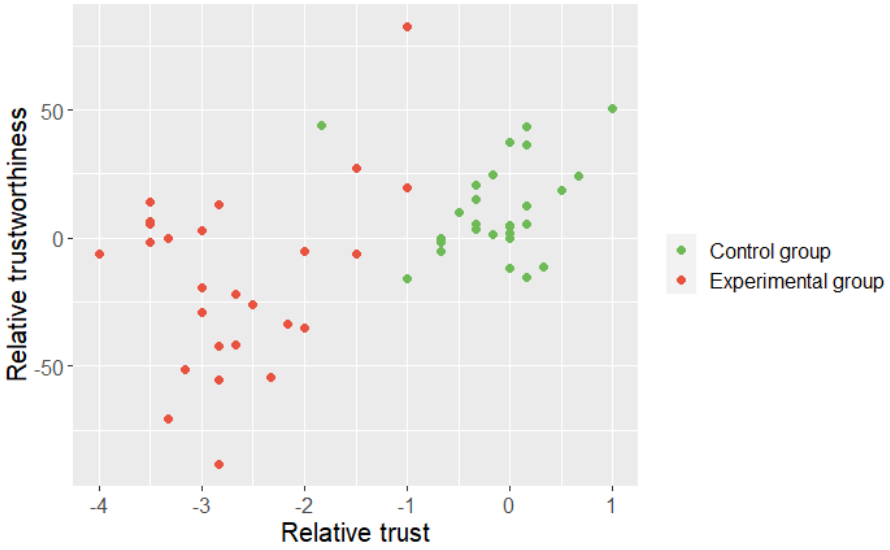


Figure 4.14: Scatter plot of relative trust and relative trustworthiness.

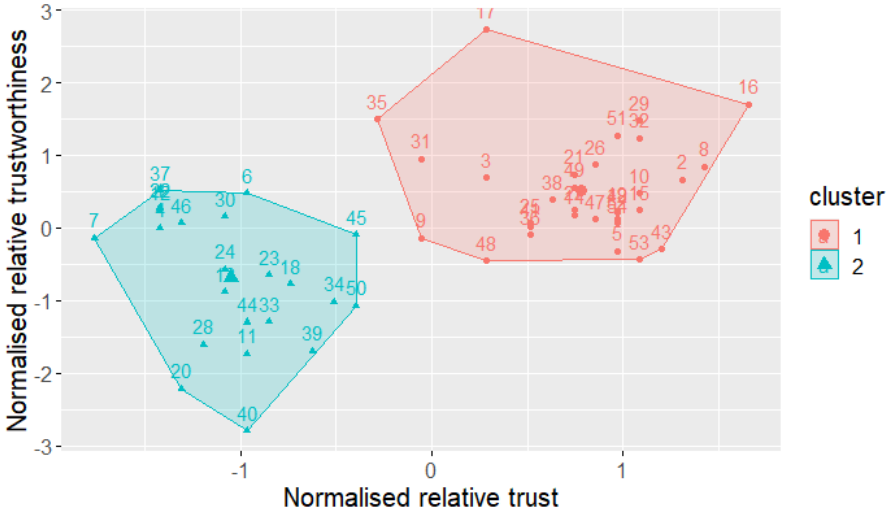


Figure 4.15: K-means cluster analysis of the relative trust with relative trustworthiness.

Subjective trustworthiness per question

The results of the previous section are based on the average of the six questions regarding the participant's own perceived trustworthiness. Since there is a significant change in both groups, we want to further look into which questions were or were not relevant for this change.

An independent t-test is performed on all questions individually, comparing the change in answers per group. However, either the data did not have a normal distribution, or the test showed no significance. Since the independent t-test does show a significance on the average of the

questions combined, we decide to perform a paired t-test for each of the questions. This shows a significant increase for the control group for three out of six questions. Question one about their own competence (first game: M = 47.7, SD = 44.3, second game: M = 64.9, SD = 29.8), t(26) = -2.81, p = .009. Question five about their consistent behaviour (first game: M = 18.4, SD = 58.1, second game: M = 41.7, SD = 40.3), t(26) = -2.39, p = .024. Lastly, question six about the robot being able to rely on them to do their best (first game: M = 67.9, SD = 26.1, second game: M = 77.1, SD = 25.0), t(26) = -2.17, p = .039. Looking at the data from the experimental group, only question four ("The robot was able to depend on me entirely to help it when it asked for help") shows a significance between the first game (M = 55.8, SD = 36.5) and the second game (M = 4.67, SD = 65.8), t(26) = 4.8, p < .001.

4.3.2 Objective trustworthiness

During the experiment, the program takes note of the actions of the participant, as stated in subsection 3.6.2. These factors give an indication of the participant's behavioural trustworthiness. In this section, we go over the recorded factors and analyse them.

Carrying ratio

While doing the experiment, the participant is free to choose how to carry a box. This mostly holds for the medium box, since this is the only box that can be carried both alone or together. However, since the order does not necessarily matter, the participant can still decide to bring in only light boxes, thus only carrying alone. As stated in subsection 3.6.2, the way the participant is carrying the boxes indicates the participant's will to cooperate.

To perform tests on this, we need to take the speed of the participant into account. We cannot compare the carrying behaviour of someone who brought in a total of ten boxes with someone who finished bringing in all twenty-five boxes with time left, since the latter would then automatically be more cooperative for carrying more boxes together. We therefore decide to make it proportional. This has been done by the following formula:

$$CR_x = \frac{CT_x}{CA_x},$$

where CT is the amount of times they carried together,

CA is the amount of times the participant carried alone,

x is the number of the game

Here CR stands for the *carrying ratio*, calculated by how many times the participant carries a box together with the robot or carries a box alone. The closer this ratio is to zero, the more the participant is carrying boxes alone. The overall carrying ratio is calculated by:

$$CR = CR_2 - CR_1$$

Figure 4.16 shows the visualised data, where we can see that the participants from the control group (control group) carry more boxes together than the participants from the experimental group (experimental group). In fact, there is a significant difference between the carrying ratio in the control group (M = 0.9, SD = 1.8) compared to the experimental group (M = -0.9, SD = 3.7), t(38) = 2.31, p = .026.

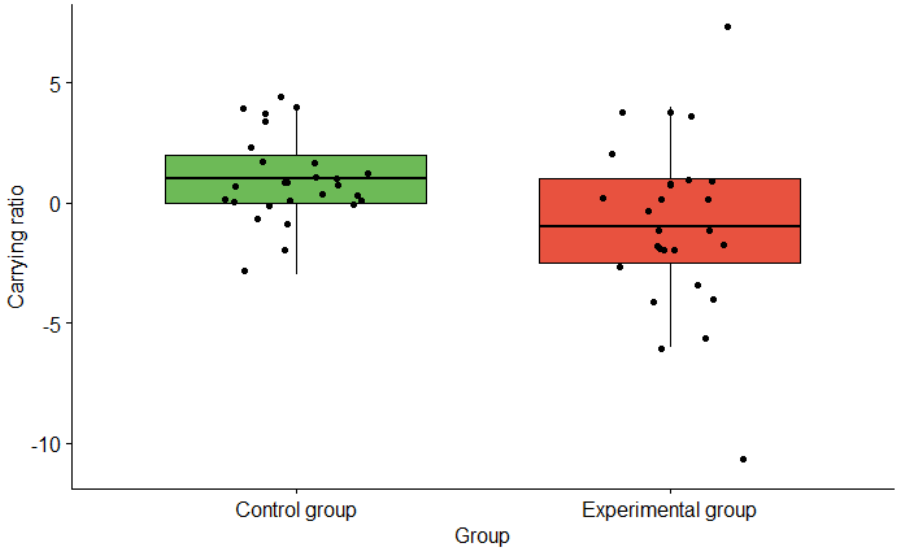
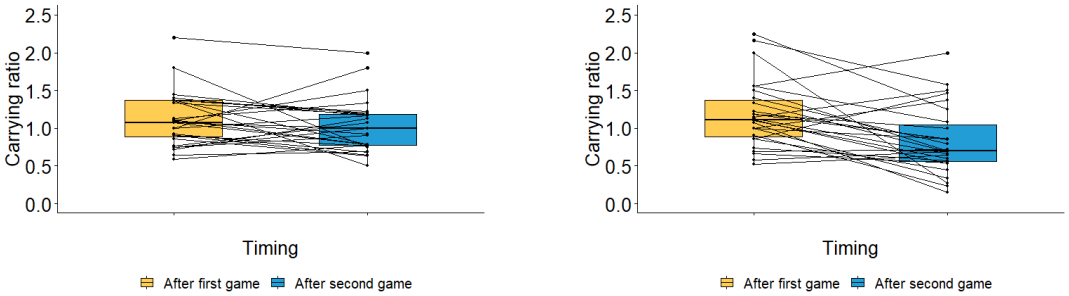


Figure 4.16: CR of the participants visualised in a box plot per group.

Upon further investigation, we notice that in both groups the carrying ratio decreases, as can be seen in Figure 4.17. Two paired t-tests are performed to find whether these decreases are significant, where the test for the control group shows no significance. However, the experimental group does have a significant difference in comparing the first game (M = 1.2, SD = 0.4) with the second game (M = 0.8, SD = 0.5); t(26) = 3.54, p = .002.



(a) Comparing CR_1 and CR_2 respectively in the control group.

(b) Comparing CR_1 and CR_2 respectively in the experimental group.

Figure 4.17: A comparison of the differences in the participant's carrying boxes ratio (CR) per game in separate groups.

Response to help ratio

During the game, the robot calls the human for help with carrying a medium or heavy box. It is then for the human to decide how they respond to this. They could walk to the robot and carry the box together, or, in case of a medium box, decide to carry it alone, or even completely ignore the call for help. This is a big factor of indication of the participant's trustworthiness, strongly related to their benevolence. We use the following formula to make the responses of the participant

proportional:

$$RR_x = \frac{PH_x}{RC_x},$$

where PH is how many seconds it took on average for the participant to respond to the call for help from the
RC is the amount of times the robot called for help,
x is the number of the game

Here RR stands for the *response ratio*, calculated by how many seconds it takes for the participant to respond to the robot's call for help by walking to it and carrying the box together, timed from the moment the participant is available (e.g. if they are carrying a box alone, the timing starts as soon as that box is lowered). The average amount of seconds is calculated by this. Whenever the participant does not respond to the call for help, a thirty-second penalty is added. The overall response ratio is calculated by:

$$RR = RR_2 - RR_1$$

The result of this calculation is visualised in Figure 4.18, where a higher response ratio means that it took longer for the participant to respond. Two extreme outliers are removed from the data. An independent t-test is performed to compare these response ratios between the experimental group (M = -4.7, SD = 7.0) and control group (M = 15.4, SD = 38.8), showing a significant difference, t(28) = -2.64, p = .014.

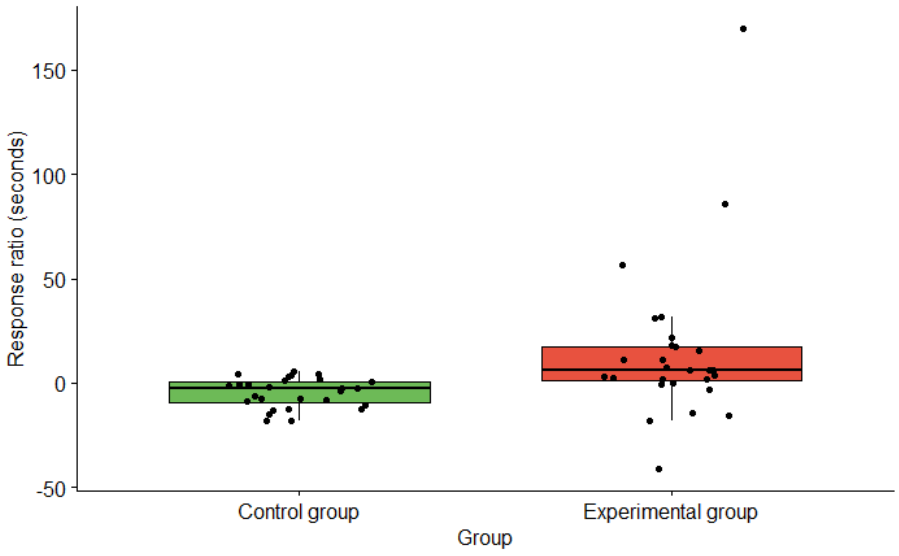
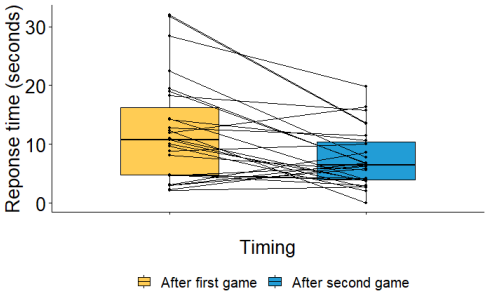


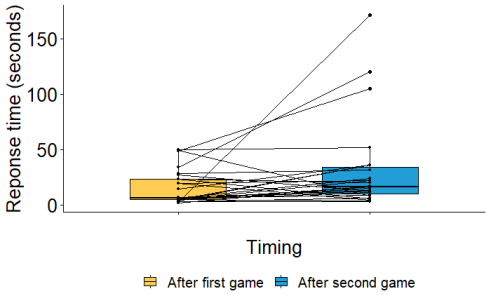
Figure 4.18: RR of the participants visualised in a box plot per group.

A paired t-tests shows a significant decrease in response rate for the control group when comparing the first game (M = 12.1, SD = 8.8) with the second game (M = 7.5, SD = 4.9); t(26) = 3.44, p = .002. Since the data from the experimental group is not normally distributed, a Wilcoxon signed-rank test is executed. This shows a significant increase from the first game (Md = 5.8, n

= 27) to the second game (Md = 16.0, n = 27), W = 85, p = .011. The data is visualised in Figure 4.19.



(a) Comparing RR_1 and RR_2 respectively in the control group.



(b) Comparing RR_1 and RR_2 respectively in the experimental group.

Figure 4.19: A comparison of the differences in the participant's responses to calls for help (RR) per game in separate groups.

Breaking boxes ratio

The game is built around the option to break boxes. This is designed so that the robot can clearly show that it is less trustworthy. With this, we expected that the participant would then also break boxes, skipping the heavy boxes, while still receiving extra points for the order. However, during the game it quickly becomes clear that the participants do not like to break boxes, even during the tutorial. Whenever a participant does break a box, it is in the first game, and merely because they forgot the rule of the safezone. The amount of broken boxes turns out to be extremely low (with a mean of around 0.2 boxes per game), such that without further examination we can say that there is no significant difference on this factor between the control group and the experimental group.

Call for help ratio

The last factor that can give an indication to the participant's trustworthiness is how many times they ask for the robot's help, indicating their level of cooperation. We again use a formula to make the calls for help proportional:

$$PCR = PC_2 - PC_1,$$

where PC_1 and PC_2 are the amount of times that the participant has called for help in first or second group respectively

Here PCR stands for *participant call for help ratio*. The data is visualised in Figure 4.20. To examine the change in asking for help depending on the two groups, an independent sample t-test is performed. The 27 participants in the control group (M = 0.3, SD = 3.5) compared to the 27 participants who experienced automation failure (M = -2.1, SD = 3.3) demonstrate a significant difference, t(52) = 2.55, p = .014.

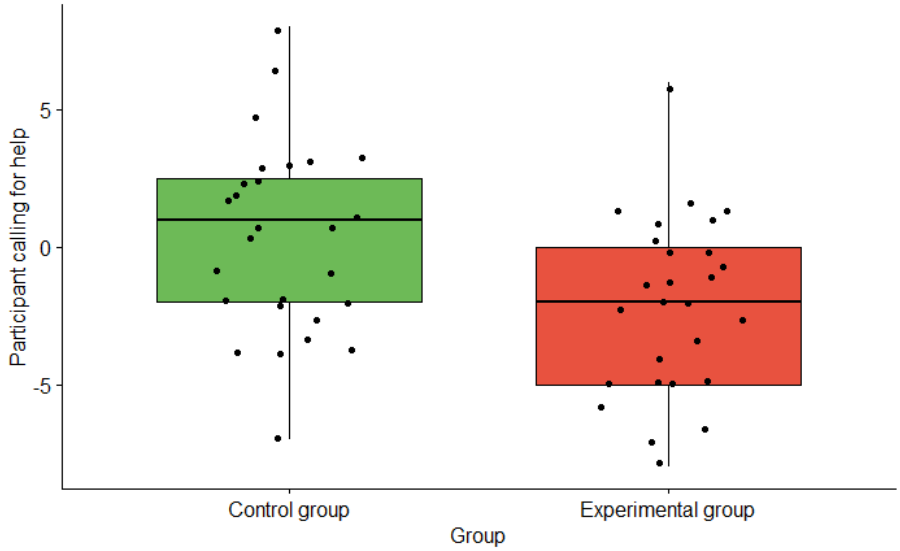
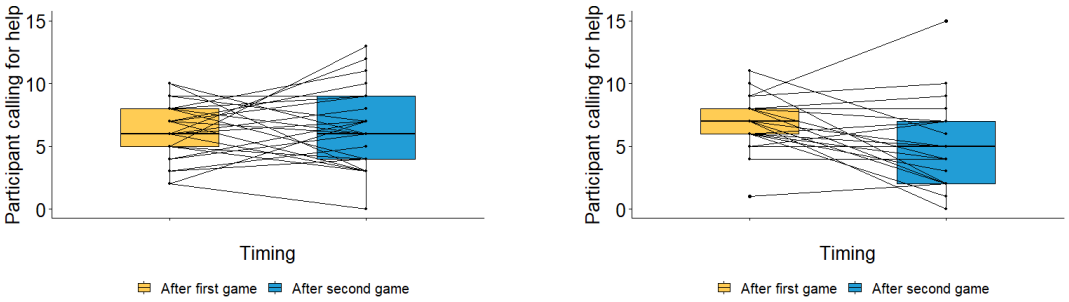


Figure 4.20: A box plot of PCR per group.

Further analysing this, two paired samples t-tests are performed. This data is visualised in Figure 4.21. There is no significant difference found for the control group. However, there is a significant difference in the amount of times the participant asks for help in the first game (M = 7.0, SD = 2.0) and the second game (M = 4.9, SD = 3.4) for the experimental group in terms of a decrease; t(26) = 3.31, p = .003.



(a) Comparing PC_1 and PC_2 respectively in the control group.

(b) Comparing PC_1 and PC_2 respectively in the experimental group.

Figure 4.21: A comparison of the differences in the participant's calls for help (PC) per game in separate groups.

Strategy

The end of each part of the questionnaire contains a question about the participant's *strategy*. They can tick off which strategy they were following, where multiple answers are possible. Figure 4.22 shows the answers after the first game. We observe that there is no notable difference when comparing the participants from the control group with those from the experimental group.

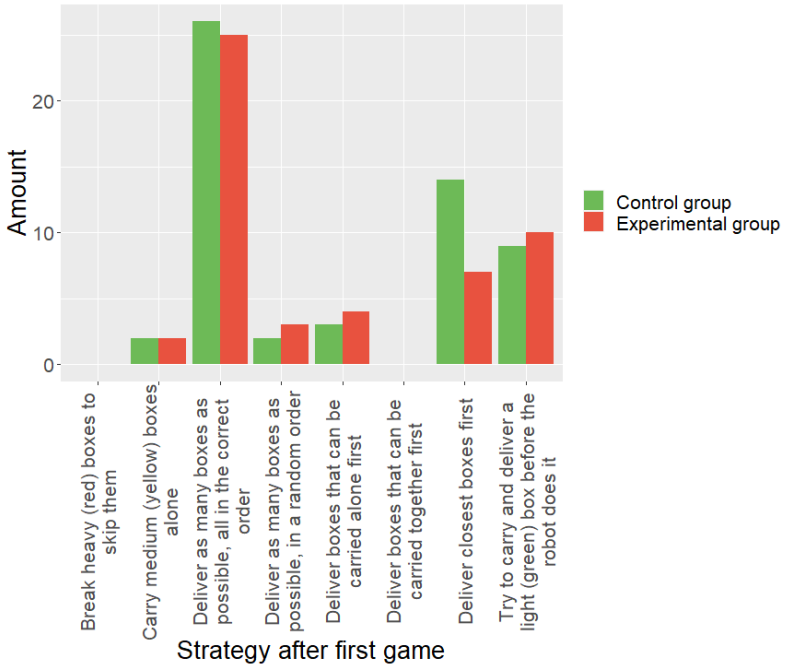


Figure 4.22: The distribution of chosen strategy in the first game per group.

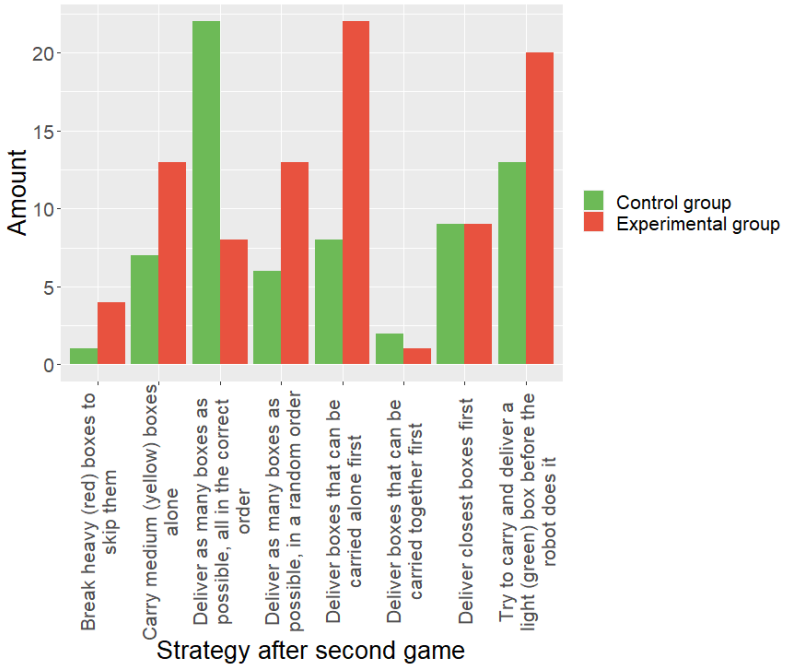


Figure 4.23: The distribution of chosen strategy in the second game per group.

Figure 4.23 displays the answers after the *second* game, where the participants from the experimental group experienced automation failure. We observe a change in strategy in both conditions. We again observe very few people in either group has a strategy that involve breaking boxes. However, their way of carrying and delivering boxes does change. In both groups there is an increase for carrying medium boxes alone, but we observe a much larger increase in the experimental group. Moreover, participants from the control group generally use the same strategy regarding the order of delivery, while participants from the experimental group change their strategy from delivery in the correct order to delivery in a random order. Another noticeable change is the increase of the amount of participants deciding to deliver boxes that can be carried alone first. In the control group, this is doubled, while in the experimental group the amount of people going for that strategy has become five times as much. With this delivery, there is an increase for delivering the closest boxes first for only the participants in the control group. Lastly, both groups show an increase for trying to carry a light box before the robot does it, but the increase in the experimental group was greater.

Ending the questionnaire, participants can indicate why they had changed their strategy. Most participants from the control group usually report that they had better knowledge of the game or the way the robot thinks, making this change in strategy a choice based on the score they want to obtain. Twenty-two participants from the experimental group report issues with the performance of the robot and their trust in the robot. Two state that they only changed their strategy because they were not able to get the high-score in the previous game, and three people did not answer the question.

4.4 Liking of robot

The questionnaire contains questions about the human liking the robot or not. From these questions, we can again receive a relative rating:

$$L_{rel} = L_2 - L_1,$$

where L_1 and L_2 are the liking of robot after the first or second group respectively

The data of this rating is visualised in Figure 4.24. We want to test whether there is a difference in this relative liking in the two groups. The data of the control group violates the assumption of normally distributed data for the independent t-test. We therefore conduct a Wilcoxon rank-sum test, which reveals a significant difference when comparing the human's relative liking of the robot in the control group (Md = 0.2, n = 27) with the experimental group (Md = -2.2, n = 27), W = 723, $p < .001$.

To compare the liking after the first game with the liking after the second game, we perform paired tests. Since the data of the control group does not have a normal distribution, we perform a Wilcoxon signed-rank test. This shows us a significant increase from the first game (Md = 4.0, n = 27) to the second game (Md = 4.4, n = 27), Z = 378, $p < .001$. This can be observed in Figure 4.25a.

Since the experimental group does have a normal distribution, we perform a paired samples t-test. The data is visualised in Figure 4.25a. This shows that between the first game (M = 4.42, SD = 0.59) and the second game (M = 2.27, SD = 0.84) there is a significant decrease; $t(26) = 11.9$, $p < .001$.

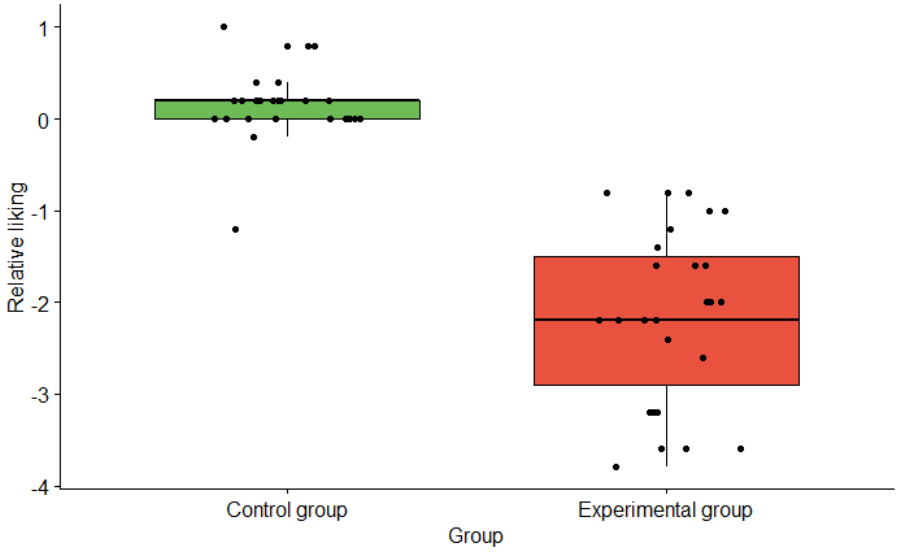
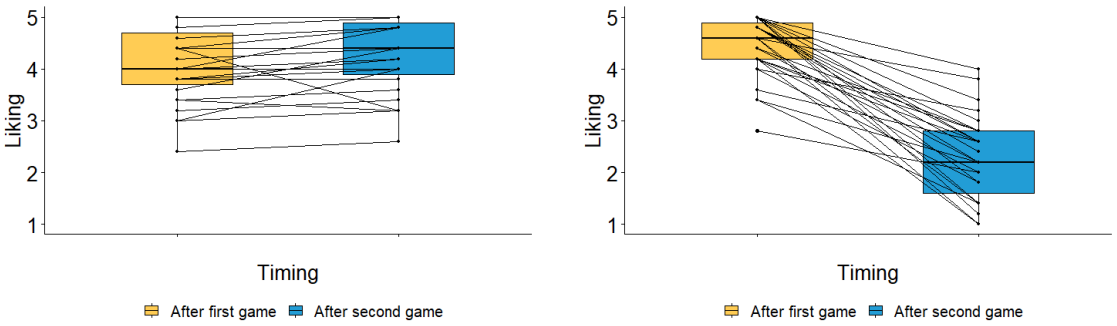


Figure 4.24: A box plot of L_{rel} per group.



(a) Comparing L_1 and L_2 respectively in the control group.

(b) Comparing L_1 and L_2 respectively in the experimental group.

Figure 4.25: A comparison of the differences in the human's liking of the robot (L) per game in separate groups.

4.4.1 Correlation between liking and trust

As we now have observed both the human's trust in the robot and the human's liking of the robot, we want to examine whether these factors are in some way related to each other. Figure 4.26 shows a plot of the two against each other, where Local Polynomial Regression Fitting tries to find a smooth curve. When observing this plots, we see that there might be a linear correlation between the two factors, irrespective of the groups to which the participants were assigned. We compute a Pearson correlation coefficient to assess the linear relationship between relative liking and relative trust. There was a positive correlation between the two variables, $r(52) = .88$, $p < .001$. A visualisation of this result is shown in Figure 4.27.

We observe that this plot also shows two clusters, one for each group in our experiment.

Examining this, we execute a K-means cluster analysis with an accuracy of 94%, of which the result is shown in Figure 4.29.

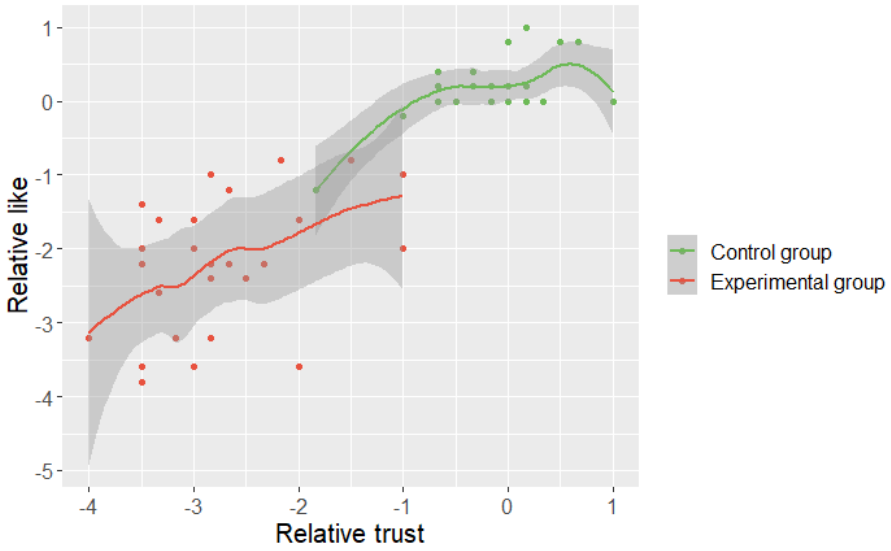


Figure 4.26: A plot of the collected relative liking in combination with the relative trust per group, where Local Polynomial Regression Fitting creates a smooth curve.

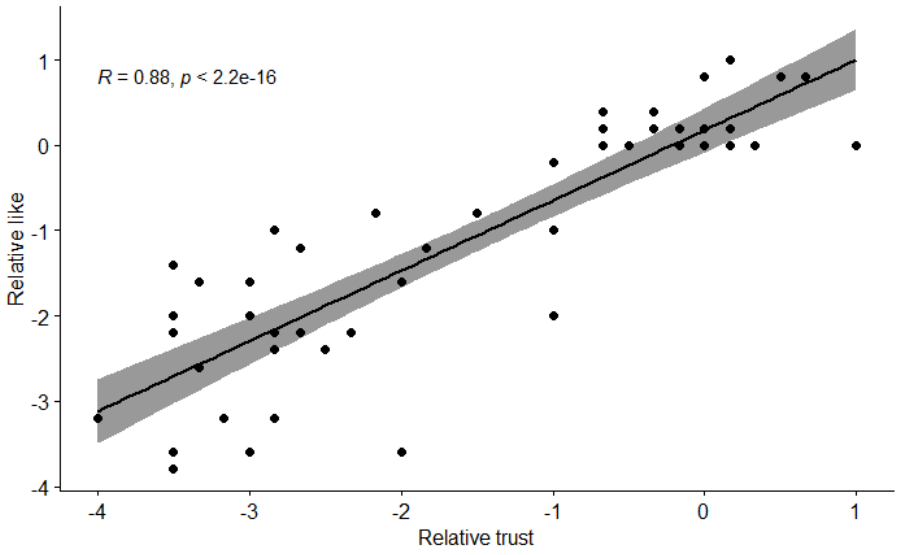


Figure 4.27: Pearson correlation between relative liking and relative trust.

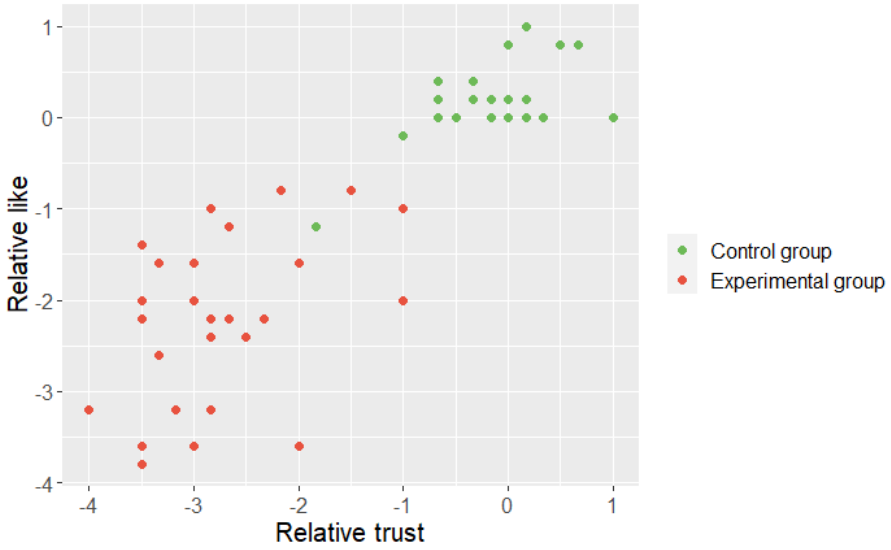


Figure 4.28: A plot of the collected relative liking in combination with the relative trust per group.

Figure 4.29: K-means cluster analysis of the relative trust with relative liking.

4.4.2 Correlation between liking and trustworthiness

We examine whether there is also a correlation between the human's liking of the robot and the human's trustworthiness. Figure 4.32 shows a plot of the two against each other, where Local Polynomial Regression Fitting tries to find a smooth curve. We are interested in whether there is a linear relationship between relative trustworthiness and relative liking. Here we found a positive correlation as well, $r(52) = .46$, $p < .001$.

We observe that this plot also shows two clusters, one for each group in our experiment. Examining this, we execute a K-means cluster analysis, receiving an accuracy of 93%, of which the result is shown in Figure 4.33.

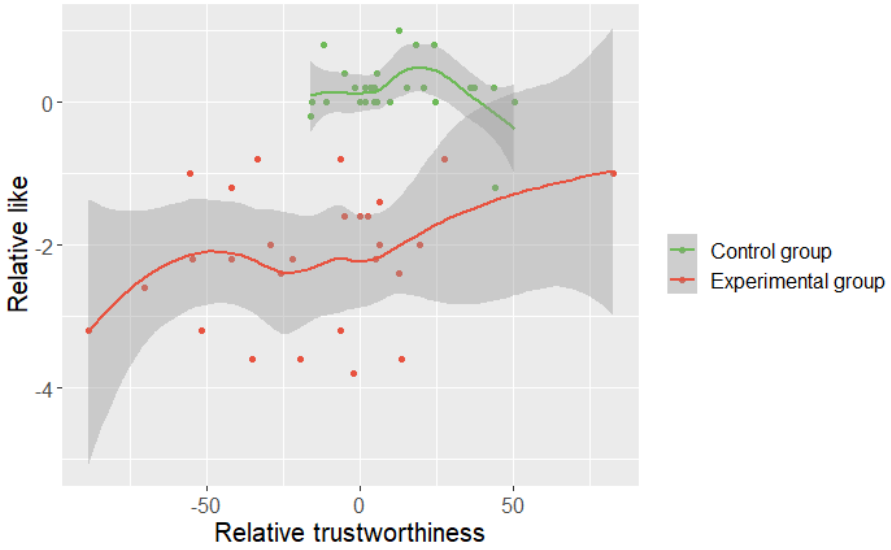


Figure 4.30: A plot of the collected relative liking in combination with the relative trustworthiness per group, where Local Polynomial Regression Fitting creates the smooth curve.

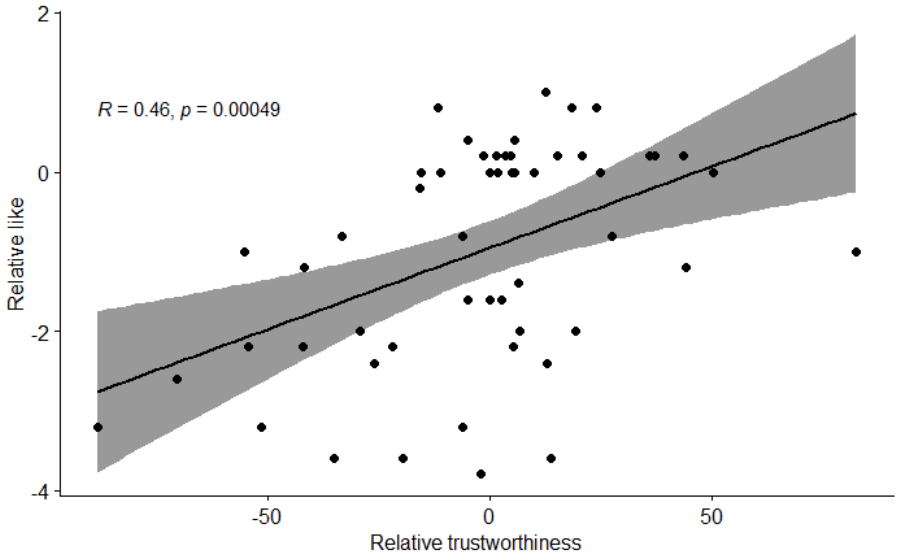


Figure 4.31: Pearson correlation between relative liking and relative trustworthiness.

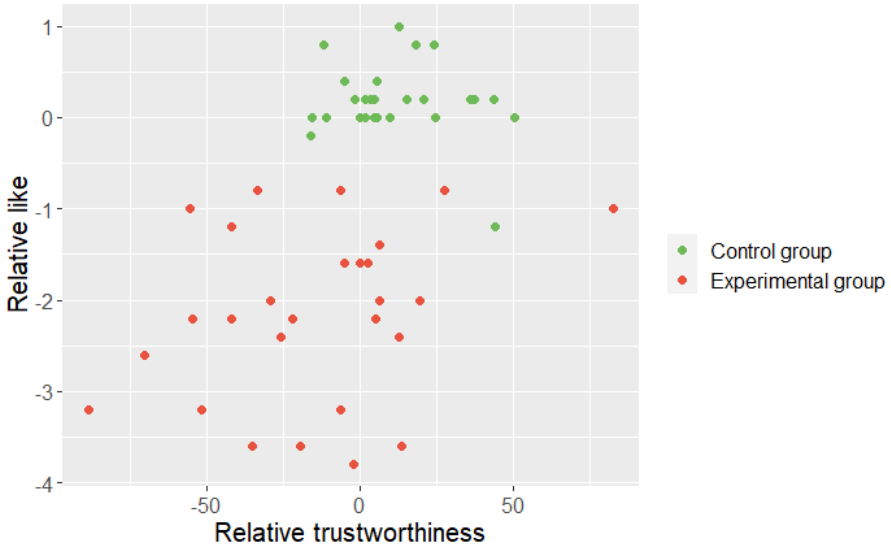


Figure 4.32: A plot of the collected relative liking in combination with the relative trustworthiness per group.

Figure 4.33: K-means cluster analysis of the relative trustworthiness with relative liking.

4.5 Clustering with trust, liking and trustworthiness

It is interesting to see the clustered plots of paired factors, as given in Figures 4.15, 4.29 and 4.33. This shows a clear division between the control group and experimental group, as the clusters as given by the K-means algorithm are close to the truth. We plot the normalised values of the relative trust, relative trustworthiness and relative liking in a three-dimensional figure, coloured

by the groups from the experiment. This is given in Figure 4.34. Similarly, we plot the results of the K-means cluster analysis, shown in Figure 4.35. Comparing the truth with the found clusters shows a clear segregation that is 93% accurate.

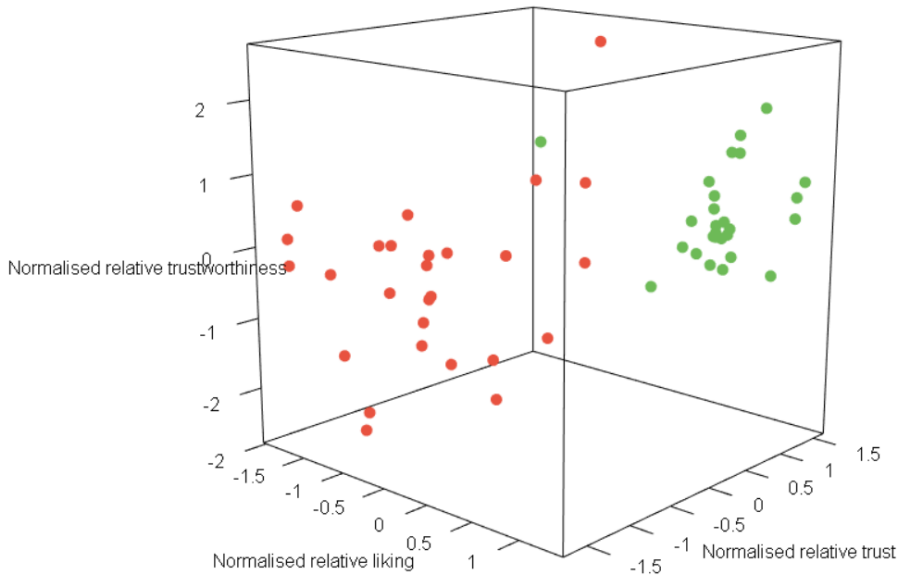


Figure 4.34: A three-dimensional scatter plot of the normalised relative trust, relative trustworthiness and relative liking, divided by the groups from the experiment.

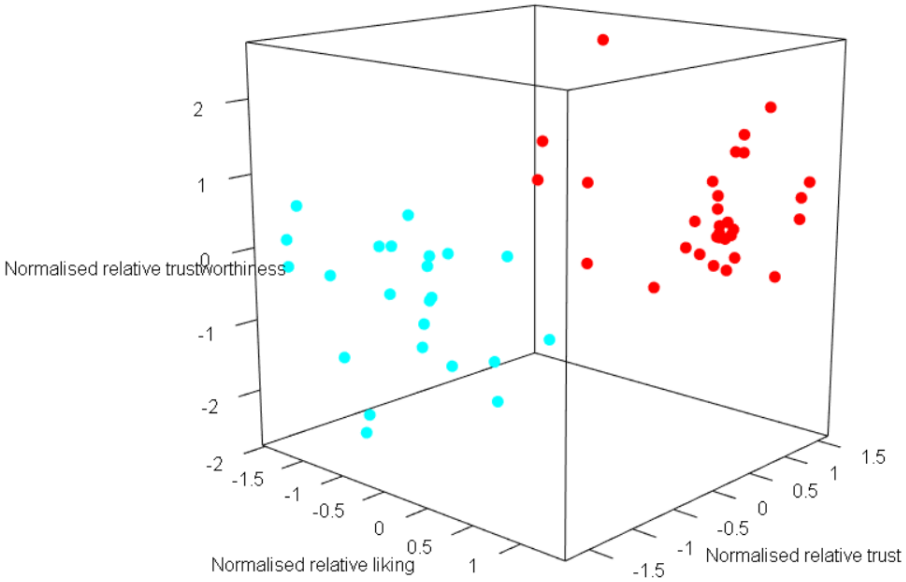


Figure 4.35: K-means three-dimensional cluster analysis of the relative trust, relative trustworthiness and relative liking.

4.6 Summary

In this chapter, we talked about all the factors that were measured during the experiment. A test was performed for every factor to compare the two groups, trying to find a significant difference between them. If this was found, we would further investigate the change between the games for each group individually. Table 4.1 shows an overview of these test results. A row containing only stripes has shown no significant difference between the two groups (e.g. for the breaking boxes ratio). If only the part of a row of a specific group shows stripes, there was no significant change between games (e.g. for the case of trust).

Group		**Control**		**Experimental**	
		change	*p*	*change*	*p*
Subjective trustworthiness		increase	.004	decrease	.033
Objective trustworthiness	*Carrying ratio*	decrease	NS	decrease	.002
	Breaking boxes ratio	-	-	-	-
	Response to help time	decrease[1]	.002	increase[2]	.011
	Call for help ratio	-	-	decrease	.003
Trust		-	-	decrease	<.001
Liking		increase	<.001	decrease	<.001

Table 4.1: A summary of the test results for all tested factors.

[1]Decrease in time, which means an increase in trustworthiness according to the definition in this study.
[2]Increase in time, which means a decrease in trustworthiness according to the definition in this study.

5

Discussion

In this chapter, we discuss the results of the research conducted in this thesis. We do this by revisiting the research question and interpreting the results of the previous chapter. This chapter ends with our limitations and future work.

5.1 Results

The interpretation of our results will be elaborated using the model presented in Figure 5.1. This interpretation revolves around the main research question:

What is the effect of automation failure on the human's trustworthiness in human-automation teamwork?

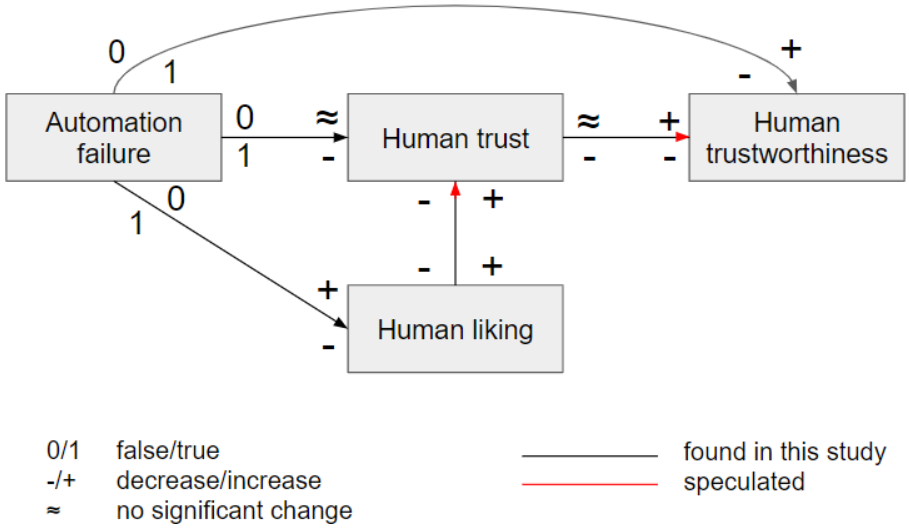


Figure 5.1: The effect of automation failure on the human's trust, trustworthiness, and liking as found in this study.

5.1.1 Trustworthiness

Section 4.3 has shown us that there is a significant difference between the *subjective trustworthiness* of the participants from the control group compared to the participants from the experimental group. When looking at these groups individually, we observe a significant increase in the control group, while there is a significant decrease in the experimental group. This confirms our hypothesis, stating that automation failure has a negative effect on the human's trustworthiness.

When observing the subjective trustworthiness as measured *per question* in Figure 4.3.1, the experimental group shows a decrease in only question four, about the response to help, which is consistent with our objective findings. Although we cannot conclude anything from the questions separately, as we view the questions as an indication of the trustworthiness as a whole, we can still speculate on the results. For example, it is interesting to see an increase of competence in the control group, while there is no significant increase in the experimental group. The participants always executed the game for a second time, which would logically increase their competence, as practice usually does. However, it seems like the participants from the experimental group could not make the same improvement as those from the control group, merely because of the automation failure. We need to keep in mind that this is subjective, which could mean that their competence could still have increased, but they did not feel like it did because of, for example, a decrease in score. It would be interesting to further observe which of the factors of trustworthiness are increasing or decreasing with or without automation failure.

We also examined the *objective trustworthiness* that was defined and observed in the experiment. This consists of the carrying ratio, breaking boxes ratio, response to help ratio, and call for help ratio, and is motivated by their strategy. We observe significant differences between the control group and the experimental group on the factors of carrying ratio, response to help ratio, and the call for help ratio. Within these ratios, overall we observed an increase of trustworthiness in the control group, and a decrease of trustworthiness in the experimental group, as was summarised in Table 4.1. We mostly observe an increase of trustworthiness by the decrease of response time in the control group, while the experimental group shows a decrease of trustworthiness on almost all factors. Only the breaking boxes ratio did not show any significant results, as participants only broke boxes on accident. These results again support our hypothesis.

When comparing the *strategy* of the first game with the second game per group, as displayed in Figure 4.22 and Figure 4.23 respectively, we observe no significant differences in the first game, indicating a good balance between the groups. The second game does show a great difference. We observe a great preference for carrying boxes alone in the experimental group, where the order is mostly discarded. Although we see a slight increase in preference for carrying boxes alone in the control group as well, it is still chosen by only a third of the participants, while in the experimental group it is chosen by almost all. Moreover, in the control group almost all participants would still follow the correct order of the boxes, while in the experimental group more than half of them switched strategies to a random order. Lastly, we see an increase in participants who tried to carry a box before the robot is able to in both groups. For the control group this is an increase of half of the previous amount, while in the experimental group the previous amount is doubled. We think this overall increase is due to the speed of the robot, which can be seen as quite slow for the average gamer. These changes and differences between the two groups confirms the decrease in trustworthiness mentioned before.

Our research question was about finding an effect of automation failure on the human's trustworthiness. The results show a decrease in both subjective and objective trustworthiness whenever there is automation failure (experimental group). We confirm our hypothesis, stating that automation failure has indeed a negative effect on the human's trustworthiness in this study. This is indicated in figure Figure 5.1 by the direct arrow between automation failure and the human trustworthiness.

5.1.2 Trust

The results show a decrease of human trust in the robot whenever there was automation failure. When there was no automation failure, there was no significant difference between the two games. Correia et al. (2016) suggests that the development of trust may need longer interactions, which may

be why we do not see a significant increase in trust in the group with no automation failure. From the decrease in trust in the experimental group, we determine a relationship between automation failure and human trust, as indicated by the arrow in Figure 5.1.

Figure 4.11 shows a two-dimensional plot of the relative trustworthiness with the relative trust. When trying to predict the interaction effect between the two factors, we see linear lines. Further investigating those lines, the Pearson correlation test from Figure 4.13 shows just a moderate correlation.

It is interesting to see how the points from each group from the experiment are close together, showing a division between the groups. When using a K-means cluster analysis, as visualised in Figure 4.15, the result is close to the truth. As this algorithm has found almost exactly the same clusters as we have intended, we know that there must be a difference between the effect of the lack of automation failure compared to the presence of automation failure on the human's trust and trustworthiness. We observe that the participants from the control group are all close to having no change in trust and trustworthiness, while most participants from the experimental group have both a negative relative trustworthiness as negative relative trust. This shows how automation failure causes both trust and trustworthiness to reduce.

There seems to be a correlation between the relative trust and the relative trustworthiness. This leads us to the arrow between these two factors in Figure 5.1. This study cannot conclude the direction of this arrow, but we hypothesise that it is directed from human trust to human trustworthiness because of prior research (Tullberg, 2008; Falcone & Castelfranchi, 2004).

5.1.3 Liking

Liking is a factor that we did not anticipate when starting this study. It was merely included to make use of the full questionnaire from Merritt (2011). The author sees liking as an affect-related attitudinal construct, defined as "the degree to which the user feels positively toward the automated system" (Merritt & Ilgen, 2008, p. 358). Other studies refer to liking as satisfaction (Wang et al., 2011; Donmez et al., 2006) or positive affect (Kim et al., 2020) associated with a system.

The results have shown that liking decreases when automation failure occurs. Additionally, it increases when there is no automation failure, as in the control group. This results into the direct line from automation failure to human liking, as given in Figure 5.1.

Correlation with trust

We tried to find a correlation of liking with trust. Figure 4.26 shows an almost linear line, suggesting that the relative trust and liking are highly correlated. A high (.88) Pearson correlation coefficient confirms this, of which the data is shown in Figure 4.27. This is also confirmed by prior research that found a positive correlation between likeability and trust (Kim et al., 2020; Donmez et al., 2006, e.g.). This leads us to the arrow between human trust and human liking, as in Figure 5.1. The direction of this arrow cannot be derived from this study. However, Nicholson et al. (2001) found that liking has a direct significant positive impact on trust. This study has been done in human-human relationships regarding buyer's trust in the sales representative. Since this study involves different agents and a different context than what we studied in our thesis, the direction of the arrow is merely a speculation.

The plot also shows two clear clusters per group. The control group again seems to have no negative changes in trust and liking, hovering around one and zero on both factors. The experimental group, on the other hand, shows a decrease for every participant on both factors. This makes the distinction clear. A K-means cluster analysis, as visualised in Figure 4.29, shows

clusters that are close to the truth, indicating that our idea of two segregated clusters is correct, and that there is indeed an effect of automation failure, as an algorithm can find the same distinction as we designed there to be. All participants from the experimental group show a decrease in both trust and liking, while most of the participants from the control group show either no change or an increase.

Correlation with trustworthiness

Figure 4.32 shows a plot of the relative liking with the relative trustworthiness. Again analysing the linearity with a Pearson correlation test, we find only moderate results. As this plot also seems to show two clusters, although less obvious, we perform a K-means cluster analysis. This shows two clusters that are quite similar to the truth. However, the plots do not seem to show things as clear as the previous plots on the other factors. Where the previously examined correlations show a decrease on both factors for the experimental group, here it is all a bit more divided. We see a clear separation of the two groups in terms of liking, but none of that in terms of trustworthiness. We therefore speculate that there is no direct correlation between relative liking and relative trustworthiness, as indicated by the lack of arrow in Figure 5.1. This conclusion is supported by Kim et al. (2020), as they found no direct effect of liking on trustworthiness.

5.1.4 Clustering with trust, liking and trustworthiness

We discussed the correlations that relative liking, relative trust and relative trustworthiness have pairwise, showing two-dimensional figures. As we are interested to see how these three would relate to each other, a three-dimensional plot was given in Figure 4.34. This plot clearly shows a division between the two groups. When applying a K-means clustering analysis, of which the results are shown in Figure 4.35, we obtain a similar figure, confirming the segregation. The data points from the participants in the experimental group lie in the negative space for all three factors, visualising the effect of automation failure, thus confirming the negative effect of automation failure on all three factors as given in our model.

5.1.5 General discussion

We see an interesting linear correlation between the trust in and liking of the automation. As we also see a correlation between trust and trustworthiness, we want to highlight the speculated importance of likeability in automations. The likeability of an automation could indirectly influence the trustworthiness of the human team member, perhaps improving their teamwork.

In general, the results show a decrease in trustworthiness, trust, and liking of the human whenever there is automation failure. Clustering the data with a highly accurate K-means analysis visualises the segregation of the two groups, confirming the negative effect of automation failure. This also either confirms the mistake that Salem et al. (2015) made in reporting the results, or shows that the effect highly depends on the type of automation failure.

Seeing the effect of automation failure in our experiment, we could use this to foresee undesirable consequences and improve human-automation teamwork. If we view this from a human perspective, we could know that the human tends to be less trustworthy towards the automation, deteriorating the teamwork. As we now have seen that there is a large effect on human-automation teamwork, we see the importance of having automations that show the least amount of failures.

Viewing this from the automation's perspective, it could be programmed to know when it fails. It would then know that the human is less trustworthy, and could thus anticipate on it. This could be done by either using a repair strategy, or by keeping in mind that the collaboration with

their human team member will be less effective. It could then, for example, change its way of communication or how much trust it puts in its team member.

5.2 Limitations

In the course of this research, we stumbled upon a few limitations. For example, the ability of the participant could have been observed more closely, providing us with another indicator of their objective trustworthiness. We kept track of the scores and whether the participant was carrying the box alone or together, but by making some kind of division for the team score to individual scores, we are still not anticipating the effect of the automation failure, or fully grasping the participant's ability. For example, if we would give individual scores to the agents by observation (shared when they worked together, or individual points when one worked alone) and the participant would decide to work alone, they could potentially score more points in the second game than in the first game because the points are not shared, while they are not necessarily more capable than in the first game. This needs to be thought through, creating a solution for this experiment or one that involves a different type of experiment.

Moreover, it sometimes became clear that the participant did not understand every rule of the game. This did not happen often enough to discard the work, and it was not always the same rule that was forgotten (e.g. some participants forgot that a box would break, some forgot the effect of a broken box, some did not understand the rules of delivering in a certain order in combination with breaking boxes). Since they would understand after the first game, this could have affected the participant's behaviour and thus the objective results from the second game. This could have been avoided by a longer tutorial, where they could participate in the game more independently. We expect that they would stumble upon their misinterpretations of the rules during this independent game, while not yet establishing an opinion about the robot, since it can be left out for this part. Another solution would have been to do a knowledge check on the rules. This would show their knowledge on the aspects of the experiment that could not have been observed by the instructor (e.g. the instructor might think that the participant knows the rule about the order of the boxes by their behaviour, but that is just a coincidence).

Lastly, as presented in Figure 4.3.2, what we measured as an increase in objective trustworthiness, could just be a choice of efficiency. For example, in both groups participants decided to carry light boxes first, and trying to get to them before the robot does. Participants from the control group reported that they did this to get a higher score. This is understandable when we consider that most participants were quicker than the robot. Participants from the experimental group reported that they decided to do this because they did not trust the robot to safely deliver it. Although the reasoning makes the division clear, such a division would be clearer in a group where a change in strategy for efficiency would lead to other participant behaviour than a change in strategy because of a decrease in trust. This should have been considered in the design of the experiment.

5.3 Future work

In the future of this research, it would be interesting to see the causality between trustworthiness, trust, and liking, as we can now only hypothesise. For example, we can raise the question whether the trustworthiness decreases because of the decrease in trust, or because of a decrease in liking. We do not know which of these factors affect which.

Moreover, we do not know whether all components of trustworthiness decrease. For example, it is possible that the participant's ability increases, while their benevolence and integrity

decreases. Our analysis displayed that only question four showed an individual significant decrease (see Figure 4.3.1). Although we cannot conclude anything within this research, since we measured trustworthiness as a whole, this could be interesting to further investigate. Knowing this, we could not only improve human-automation teamwork, but also use this information for the better of the participant (e.g. intentional automation failure to increase ability).

Lastly, we are curious to see which types of automation failure (e.g. false alarms compared to misses) has a larger effect on the human's trustworthiness. This could involve other contexts, for example a more serious context like a self-driving car. It is possible that people would have a steeper decrease of trust, liking and trustworthiness when a failure could do more harm to them, as would be logical when we look back at Figure 2.2: activity context and perceived risk and reward matter. This was a specific context with no real risks, and rewards only in score, which would not influence anything outside the experiment. Knowing the degree of effect of such failures and other contexts does not only extend our knowledge of trust in human-automation teamwork, but could let us anticipate on the effects if necessary, or improve a study for repair strategies.

6

Conclusion

In this thesis, we researched the effect of automation failure on the human's trustworthiness in a human-automation collaborative setting. Starting our research, we studied existing work on this topic. To answer our research question, we have executed an experiment to gather the necessary data (chapter 3). The data we gathered was on the participant's trust in the automation, liking of the automation, and their own trustworthiness. The experiment was simulated in a two-dimensional grid-world where the human collaborates with an automation, a visualised robot, to deliver boxes to the correct location. We gathered data during this experiment by subjective measurements (questionnaires) and objective measurements (logging the behaviour).

While the experiment could be improved to gather more insight, interpreting the results of the analysed collected data (chapter 4) showed interesting outcomes, visualised in a model.In our experiment, automation failure shows a reduction on the human's subjective liking, trust, and trustworthiness. Moreover, objective measures show that their trustworthiness is reduced in terms of executing more sub-tasks alone, helping the robot less often. We therefore confirmed our hypothesis, providing an answer to our research question (chapter 5).

The results in our research show that automation failure negatively affects the human's trustworthiness (both subjectively and objectively), and raises the question whether all factors of trustworthiness are affected, and whether all types of automation failures have this effect. This research shows relevant findings of previous research, closing a gap between human-human research and human-automation non-collaborative research, contributing to a better understanding of the nature and dynamics of trust in human-automation teams, and the possibility to foresee undesirable consequences and improve human-automation teamwork.

References

Adams, B. D., Bruyn, L. E., Houde, S., & Angelopoulos, P. (2003). Trust in automated systems. *Ministry of National Defence*.

Aggarwal, L. P. (2019, 11). Data augmentation in dermatology image recognition using machine learning. *Skin Research and Technology*, *25*(6), 815–820. doi: 10.1111/srt.12726

Beardsworth, T., & Kumar, N. (2019, 5). *Trades by Robot cost Hong Kong businessman $20mn, who does he sue?* Retrieved from `https://www.hindustantimes.com/world-news/trades-by-robot-cost-hong-kong-businessman-20mn-who-does-he-sue/story-GhkyAIvxsHGCBPklEeX9IM.html`

Blitch, J. (1996). *Artificial intelligence technologies for robot assisted urban search and rescue* (Tech. Rep.).

Bostrom, N. (2006). *AI set to exceed human brain power.* Retrieved from `http://edition.cnn.com/2006/TECH/science/07/24/ai.bostrom/`

Botvinick, M. M., & Rosen, Z. B. (2009, 9). Anticipation of cognitive demand during decision-making. *Psychological Research*, *73*(6), 835–842. doi: 10.1007/s00426-008-0197-8

Bradshaw, J. M., Ch Meyer, J.-J., van den Bosch, K., Harbers, M., Johnson, M., Feltovich, P., & Meyer, J.-J. (2011, 8). Explanation in Human-Agent Teamwork. In *International workshop on coordination, organizations, institutions, and norms in agent systems* (pp. 21–37). Berlin: Springer.

Cahour, B., & Forzy, J. F. (2009, 11). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, *47*(9), 1260–1270. doi: 10.1016/j.ssci.2009.03.015

Centeio Jorge, C., Mehrotra, S., Jonker, C. M., & Tielman, M. L. (2021). Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams. In *Proceedings of the international workshop in agent societies.* Retrieved from `http://ceur-ws.org`

Chen, J. Y., & Barnes, M. J. (2014, 2). Human - Agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, *44*(1), 13–29. doi: 10.1109/THMS.2013.2293535

Correia, F., Alves-Oliveira, P., Maia, N., Ribeiro, T., Petisca, S., Melo, F. S., & Paiva, A. (2016). Just follow the suit! Trust in human-robot interactions during card game playing. In *Robot and human interactive communication (ro-man), 2016 25th ieee international symposium on* (pp. 507–512).

Dagli, M. (2018). *Designing for Trust Exploring Trust and Collaboration in Conversational Agents for E-commerce* (Unpublished doctoral dissertation). School of Design, Carnegie Mellon University.

Dastin, J. (2018, 10). *Amazon scraps secret AI recruiting tool that showed bias against women.* San Francisco. Retrieved from `https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G`

Deutschi, M. (1960). The Effect of Motivational Orientation upon Trust and Suspicion. *Human Relations*, *13*(2), 123–139. doi: 10.1177/001872676001300202

de Visser, E. J., Peeters, M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020, 5). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. doi: 10.1007/s12369-019-00596-x

Donmez, B., Boyle, L. N., Lee, J. D., & McGehee, D. V. (2006). Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(6), 387–398. doi: 10.1016/j.trf.2006.02.001

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94.

Falcone, R., & Castelfranchi, C. (2004). Trust dynamics: how trust is influenced by direct experiences and by trust itself. In *Proceedings of the third international joint conference on autonomous agents and multiagent systems, 2004. aamas 2004.* (pp. 740–747).

Fan, X., McNeese, M., Sun, B., Hanratty, T., Allender, L., & Yen, J. (2010, 3). Human-agent collaboration for time-stressed multicontext decision making. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *40*(2), 306–320. doi: 10.1109/TSMCA.2009.2035302

Frieainan, B. (1995). "It's the Computer's Fault" -Reasoning About Computers as Moral Agents. In *Conference companion on human factors in computing systems* (pp. 226–227).

Groom, V., & Nass, C. (2007, 10). Can robots be teammates? Benchmarks in human-robot teams. *Interaction Studies*, *8*, 483–500.

Guznov, S., Lyons, J., Nelson, A., & Woolley, M. (2016). The effects of automation error types on operators' trust and reliance. In *International conference on virtual, augmented and mixed reality* (Vol. 9740, pp. 116–124). Springer. doi: 10.1007/978-3-319-39907-2{_}11

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011, 10). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. doi: 10.1177/0018720811417254

Hern, A. (2018, 1). *Google's solution to accidental algorithmic racism: ban gorillas.* Retrieved from `https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people`

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. *arXiv preprint arXiv:1812.04608*.

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an emperically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, *4*(1), 53–71.

Johnson, M., & Bradshaw, J. M. (2021). The role of interdependence in trust. In *Trust in human-robot interaction* (pp. 379–403). Elsevier. doi: 10.1016/b978-0-12-819472-0.00016-2

Kahneman, D., & Klein, G. (2009, 9). Conditions for Intuitive Expertise: A Failure to Disagree. *American Psychologist*, *64*(6), 515–526. doi: 10.1037/a0016755

Kearns, E. (2016, 3). Talos. In *Oxford research encyclopedia of classics.* Oxford University Press. doi: 10.1093/acrefore/9780199381135.013.6212

Kim, W., Kim, N., Lyons, J. B., & Nam, C. S. (2020, 5). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied Ergonomics*, *85*. doi: 10.1016/j.apergo.2020.103056

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten Challenges for Making Automation a "Team Player" in Joint Human-Agent Activity. *IEEE Intelligent Systems*, *19*(6), 91–95.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021, 10). *Measurement of Trust in Automation: A Narrative Review and Reference Guide* (Vol. 12). Frontiers Media S.A. doi: 10.3389/fpsyg.2021.604977

Laurent, K. S., Mandal, A., Khalili, W., Beaubrun, K., Mccray, S., Khalili, S., & Mandal, P. K. (2019). *Current and Emerging Applications of Innovative Artificial Intelligence in Modern Medicine and Technology* (Vol. 3; Tech. Rep. No. 1). Retrieved from `https://www.researchgate.net/publication/342449458`

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. doi: 10.1080/00140139208967392

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80.

Levin, S. (2016, 9). *A beauty contest was judged by AI and the robots didn't like dark skin.* San Francisco. Retrieved from `https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people`

Lewandowsky, S., Mundy, M., & Tan, G. P. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, *6*(2), 104–123. doi: 10.1037/1076-898X.6.2.104

Licklider, J. C. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. doi: 10.1109/THFE2.1960.4503259

Lohr, S. (2021, 7). *What Ever Happened to IBM's Watson?* Retrieved from `https://www.nytimes.com/2021/07/16/technology/what-happened-ibm-watson.html`

Madhavan, P., & Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: is trust in humans comparable to trust in machines? In *Proceedings of the human factors and ergonomics society annual meeting* (pp. 581–585).

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. doi: 10.1080/14639220500337708

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human factors*, *48*, 241–256.

Madsen, M., & Gregor, S. (2000). Measuring Human-Computer Trust. In *11th australasian conference on information systems* (pp. 6–8). Citeseer.

Matyszczyk, C. (2016, 11). *Fatty the robot smashes glass, injures visitor.* Retrieved from `https://www.cnet.com/culture/fatty-the-robot-smashes-glass-injures-visitor/`

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *Academy of management review*, *20*(3), 709–734. Retrieved from `https://www.jstor.org/stable/258792?seq=1&cid=pdf-`

McCausland, P. (2019, 11). *Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk.* Retrieved from `https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281`

Merritt, S. M. (2011, 8). Affective processes in human-automation interactions. *Human Factors*, *53*(4), 356–370. doi: 10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., Lachapell, J., & Lee, D. (2013, 6). I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors*, *55*(3), 520–534. doi: 10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008, 4). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. doi: 10.1518/001872008X288574

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive Automation, Trust, and Self-Confidence in Fault Management of Time-Critical Tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44–58. doi: 10.1037/0278-7393.6.1.44

Morgan, T. (1992). Competence and responsibility in intelligent systems. *Artificial Intelligence Review*, *6*, 217–226.

Nicholson, C. Y., Compeau, L. D., & Sethi, R. (2001). *The Role of Interpersonal Liking in Building Trust in Long-Term Channel Relationships* (Tech. Rep.).

Olsen, D. R., & Goodrich, M. A. (2003). Metrics for Evaluating Human-Robot Interactions. In *Proceedings of permis.*

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, *30*(3), 286–297. doi: 10.1109/3468.844354

Robinette, P., Howard, A. M., & Wagner, A. R. (2017, 8). Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations. *IEEE Transactions on Human-Machine Systems*, *47*(4), 425–436. doi: 10.1109/THMS.2017.2648849

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015, 3). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Acm/ieee international conference on human-robot interaction* (Vol. 2015-March, pp. 141–148). IEEE Computer Society. doi: 10.1145/2696454.2696497

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-Induced "Complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, *3*(2), 111–122. doi: 10.1207/s15327108ijap0302{_}2

Tullberg, J. (2008, 10). Trust-The importance of trustfulness versus trustworthiness. *Journal of Socio-Economics*, *37*(5), 2059–2071. doi: 10.1016/j.socec.2007.10.004

Victor, D. (2016, 3). *Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.* Retrieved from `https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html`

Wagner, A. R., Robinette, P., & Howard, A. (2018, 11). Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems*, *8*(4). doi: 10.1145/3152890

Wang, L., Jamieson, G. A., & Hollands, J. G. (2011). The effects of design features on users' trust in and reliance on a combat identification system. In *Proceedings of the human factors and ergonomics society* (pp. 375–379). doi: 10.1177/1071181311551077

Webber, S. S. (2008, 12). Development of cognitive and affective trust in teams: A longitudinal study. *Small Group Research*, *39*(6), 746–769. doi: 10.1177/1046496408323569

Wei, L. S., Gan, Q., & Ji, T. (2018). Skin Disease Recognition Method Based on Image Color and Texture Features. *Computational and Mathematical Methods in Medicine*, *2018*. doi: 10.1155/2018/8145713

Wood, G. (2003). *Living Dolls: A Magical History of the Quest for Mechanical Life.* Faber & Faber.

Wright, J. L., Chen, J. Y., Barnes, M. J., & Hancock, P. A. (2016). The effect of agent reasoning transparency on automation bias: An analysis of response performance. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 9740, pp. 465–477). Springer Verlag. doi: 10.1007/978-3-319-39907-2{_}45

Yadron, D., & Tynan, D. (2016, 7). *Tesla driver dies in first fatal crash while using autopilot mode.* San Francisco. Retrieved from `https://www.theguardian.com/technology/2016/jun/30/tesla-autopilot-death-self-driving-car-elon-musk`

A

Questionnaires

A.1 Informed consent

You are being invited to participate in a research study called Moving Out. This study is part of a master's final project of the student Nikki Bouman from the TU Delft, supervised by Prof. dr. Catholijn M. Jonker, Dr. Myrthe L. Tielman and Carolina Jorge (PhD).

The purpose of this study is to explore the dynamics of trust relationship in human-AI teamwork. You will be filling in questionnaires and playing in a small online game in which you will be moving boxes with the help of a robot. This simulation will be executed on a laptop, and you will only have to use the keyboard for interaction. It will take approximately 30 minutes to complete. The data will be used for analysation.

To the best of our ability your answers in this study will remain confidential. We will minimize any risks by anonymizing the data, such that only your gender, age, gaming experience, and answers to non-personal questions in the questionnaire is being stored.

We will archive your anonymized data at 4TU.ResearchData for at least 10 years, so it can be used for future research and learning. This data will be publicly available for non-commercial use only.

Your participation in this study is entirely voluntary and you can withdraw up until five minutes after the end of the experiment. After the end of your participation in the experiment, your data cannot be removed. You are free to omit any questions.

1. I have read and understood the study information, or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.

   ❍ Yes

2. I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study up until five minutes after the end of the experiment, without having to give a reason.

   ❍ Yes

3. I understand that taking part in the study involves completing a survey questionnaire and playing in a simulation in which I will be moving boxes with a robot.

   ❍ Yes

4. I understand that the study will end after I have completed the last question in the complete questionnaire.

   ❍ Yes

5. I understand that taking part in the study involves collecting your age range, gender, and gaming experience, with the potential risk of my identity being revealed. I understand that

the data will be anonymized and the data will be stored securely to minimise the threat of a data breach and protect my identity in the event of such a breach.

- ❍ Yes

6. I understand that personal information collected about me that can identify me will not be shared beyond the study team.

   - ❍ Yes

7. I understand that after the research study the de-identified information I provide will be used for analysis, of which the results will be published in a master thesis.

   - ❍ Yes

8. I give permission for the de-identified data (age range, gender, gaming experience and answers to the survey) that I provide to be archived in 4TU.ResearchData so it can be used for future research and learning.

   - ❍ Yes

9. I understand that access to the repository of this study is public and available for non-commercial use only.

   - ❍ Yes

A.2 Pre-test

10. What is your age?

    - ❍ 18-29
    - ❍ 30-39
    - ❍ 40-49
    - ❍ 50-59
    - ❍ 60-69
    - ❍ 70+

11. What is your gender?

    - ❍ Male
    - ❍ Female
    - ❍ Other / Prefer not to say

12. How often do you play videogames?

    - ❍ Never (or almost never)
    - ❍ A few times a month
    - ❍ A few times a week
    - ❍ Daily

13. Please ask the instructor for your number

❍ 1

❍ 2

14. Automations are defined as any sensing, detection, information-processing, decision-making or control action that could be performed by humans but is actually performed by a machine. Examples are Google Assistant, and self-driving cars.

    Answer the following questions regarding your current behaviors, not what you want your behaviors to be.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I usually trust automations until there is a reason not to	❍	❍	❍	❍	❍
For the most part, I distrust automations	❍	❍	❍	❍	❍
In general, I would rely on an automation to assist me	❍	❍	❍	❍	❍
My tendency to trust automations is high	❍	❍	❍	❍	❍
It is easy for me to trust automations to do their job	❍	❍	❍	❍	❍
I am likely to trust an automation even when I have little knowledge about it	❍	❍	❍	❍	❍

A.3 Mid-test

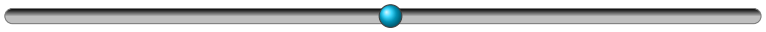
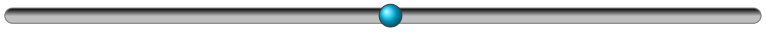
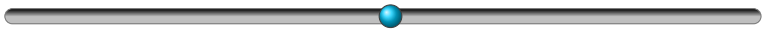
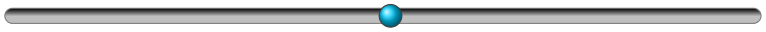
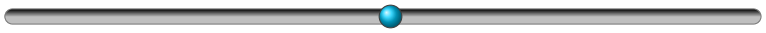
15. What did you think about the robot?

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I believe the robot is a competent performer	❍	❍	❍	❍	❍
I trust the robot	❍	❍	❍	❍	❍
I have confidence in the actions of the robot	❍	❍	❍	❍	❍
I can depend on the robot	❍	❍	❍	❍	❍
I can rely on the robot to behave in consistent ways	❍	❍	❍	❍	❍
I can rely on the robot to do its best every time it has to move a box	❍	❍	❍	❍	❍

16. What did you think about the robot?

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I like working with the robot	❍	❍	❍	❍	❍
I wish the robot wasn't around	❍	❍	❍	❍	❍
I dislike the robot	❍	❍	❍	❍	❍
I'm glad I have the option to use the robot	❍	❍	❍	❍	❍
Overall, I feel positive toward the robot	❍	❍	❍	❍	❍

17. What did you think about yourself during the task? (the first simulation)

	Strongly disagree -100 — Neither agree nor disagree 0 — Strongly agree 100
I was a very competent performer	
The robot was able to trust me completely to collaborate with it	
The robot was able to have complete confidence in my actions	
The robot was able to depend on me entirely to help it when it asked for help	
I behaved in very consistent ways	
The robot was able to rely on me completely to do my best every time I had to move a box	

18. What was your strategy for the task? (multiple answers possible)

- ❒ Deliver as many boxes as possible, all in the correct order
- ❒ Deliver as many boxes as possible, in a random order
- ❒ Deliver boxes that can be carried alone first
- ❒ Deliver boxes that can be carried together first
- ❒ Deliver closest boxes first
- ❒ Try to carry and deliver a light (green) box before the robot does it
- ❒ Carry medium (yellow) boxes alone
- ❒ Break heavy (red) boxes to skip them
- ❒ Other: _____

A.4 Post-test

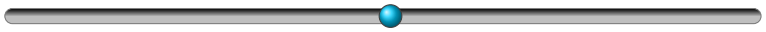
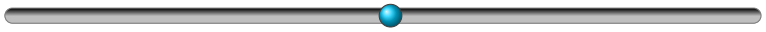
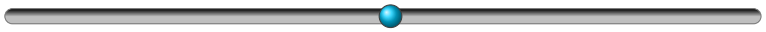
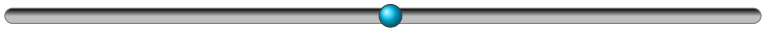
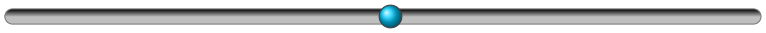
19. What did you think about the robot?

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I believe the robot is a competent performer	❍	❍	❍	❍	❍
I trust the robot	❍	❍	❍	❍	❍
I have confidence in the actions of the robot	❍	❍	❍	❍	❍
I can depend on the robot	❍	❍	❍	❍	❍
I can rely on the robot to behave in consistent ways	❍	❍	❍	❍	❍
I can rely on the robot to do its best every time it has to move a box	❍	❍	❍	❍	❍

20. What did you think about the robot?

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
I like working with the robot	❍	❍	❍	❍	❍
I wish the robot wasn't around	❍	❍	❍	❍	❍
I dislike the robot	❍	❍	❍	❍	❍
I'm glad I have the option to use the robot	❍	❍	❍	❍	❍
Overall, I feel positive toward the robot	❍	❍	❍	❍	❍

21. What did you think about yourself during the task? (the **second** simulation)

	Strongly disagree -100 — Neither agree nor disagree 0 — Strongly agree 100
I was a very competent performer	
The robot was able to trust me completely to collaborate with it	
The robot was able to have complete confidence in my actions	
The robot was able to depend on me entirely to help it when it asked for help	
I behaved in very consistent ways	
The robot was able to rely on me completely to do my best every time I had to move a box	

22. What was your strategy for the task? (multiple answers possible)

    - ❒ Deliver as many boxes as possible, all in the correct order
    - ❒ Deliver as many boxes as possible, in a random order
    - ❒ Deliver boxes that can be carried alone first
    - ❒ Deliver boxes that can be carried together first
    - ❒ Deliver closest boxes first
    - ❒ Try to carry and deliver a light (green) box before the robot does it
    - ❒ Carry medium (yellow) boxes alone
    - ❒ Break heavy (red) boxes to skip them
    - ❒ Other: _____

23. If your strategy has changed from the previous part of the experiment, could you please explain why?

    _____