Trust in human-robot interaction

The influence of transparency on trust repair with different strategies

Zhang, Ruohan

Award date:
2023

Eindhoven, 07-07-2023

Trust in human-robot interaction: The influence of transparency on trust repair with different strategies

by Ruohan Zhang

identity number 1670433

in partial fulfilment of the requirements for the degree of

Master of Science
in Human-Technology Interaction

Supervisors:
Dr. Chao Zhang
Margot Neggers

Abstract

This study investigates trust and trust repair in human-robot collaboration scenarios. Robots are playing an increasing role in industry, with more and more scenarios that require human-robot collaboration. However, the real working environment is often complex and changeable. When the environment changes, the robot may misidentify causing the task to fail. The failure of robots can lead to the loss of trust from human. Trust plays a vital role in human-robot interaction, and thus repairing trust when it is undermined has sparked the research interest. There is already research on trust repair between humans and robots, but since this is still a relatively new endeavor, the question of what the most efficient trust repair strategy is has not been fully studied. Compared to previous studies, this study focused on a specific scenario which is the food-processing industry. In addition, this study considered a combination of different trust repair strategies and test the effect of transparency. By an online experiment with 320 participants, the result highlighted the effectiveness of trust repair strategy combinations, and also found that transparency has little effect on trust repair. This study had guiding significance for promoting human-machine collaboration in food processing industry and other industries and provided new ideas for future human-robot interaction trust research.

Keywords: Industrial robot, Learning from demonstration, Human-robot interaction, Trust, Trust repair, Human-robot trust

Contents

1. Introduction

With the great achievement of machine learning research in recent years, advanced techniques have been increasingly developed in the domain of robotics and automation (Ravichandar et al., 2020). Robots have increasingly shown the potential to be applied into several industries to collaborate with humans (Ajoudani et al., 2018). For example, manufacturing is where industrial robots could be widely used (Chu et al., 2016; Maeda et al., 2017; Zhu & Hu, 2018). Robots can replace humans to complete some highly repetitive production line work, which can save high labor costs. The uses of industrial robots will also benefit the manufacturing efficiency improvement since it meets the multi-task transferability of production (Matheson et al., 2019). Another promising application area of industrial robots is healthcare. Machine learning algorithm has been shown to be effective in teaching medical robots a variety of specific movements while assisting with patient care or rehabilitation (Fong & Tavakoli, 2018; Lauretti et al., 2017; Ma et al., 2016; Meng et al., 2016; H. Wang et al., 2016). Robots also demonstrated the ability to assist children with intellectual disabilities (Najafi et al., 2017) and train people with cognitive disabilities (Moro et al., 2018).

A key determinant of the success of such human-robot interaction (HRI) applications is trust (Khavas, 2021). As an important component in human interaction, trust is considered to be a function of the objective reliability and performance of robots in HRI scenarios (Law & Scheutz, 2021). Compared to other automatic systems, robotic systems tend to be less reliable due to the complex and dynamic ways they are expected to perform (Baker et al., 2018). As a result, human's perceived trust on a robot can dramatically change (Salem et al., 2015b). In order to further promote the widespread of robot application in the industry, trust needs to be investigated in each specific industry scenario of human-robot interaction.

Food processing industry is another promising area where industrial robots have great

application prospects (Iqbal et al., 2017). This is thanks to the development of learning from demonstration (LfD) which is an emerging robotics technology (Argall et al., 2009). Unlike traditional programming methods, LfD is a robot training method in which robots acquire the ability to perform new tasks by imitating teachers (normally a human) (Chernova & Thomaz, 2014). It is not necessary for the food processing workers to have robot expertise or make extra effort to place the robot's movements, they can just work normally, and the robot will learn and perform different tasks through passive observation. However, due to the immaturity of technology, robots may make mistakes in their works if parameter changes in the environment. The imperfect performance of robots may lead to a loss of human trust in them, and the lack of trust will affect the quality of the collaboration (J. D. Lee & See, 2004). Hence, it is necessary to study when applying this emerging technology that is not yet fully mature, how should we deal with the situation of trust reduction in the face of robot instability.

Accordingly, the thesis will investigate trust and trust repair in human robot collaboration scenarios. The introduction first provides the discussion about trust definition and its measurement. After that, it gives literature reviews on trust in human-human interaction then transits to trust in human-robot interaction. To the end of introduction part, the research questions of the thesis will be provided as well as the hypotheses.

1.1 Trust and trust measurement

1.1.1 Trust definition

Before diving into the trust research, it is wise to firstly study how trust is defined and measured. The definition of trust is rather complex, and numbers of scholars have made a description to it in their own ways (Baker et al., 2018). According to Barber (1983), trust is an “expectation of technically competent performance”, which associates trust with competence. However, the most accepted definition of trust is made by Mayer et

al. (1995), who made a deeper interpretation of trust (Rousseau et al., 1998). In their paper, they indicated that trust is "the willingness of a party to be vulnerable to the outcomes of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer et al., 1995, p.712). As the model described, trust is considered to be a state which involves two party within a specific time point, rather than the general trait of an individual towards others that some other scholars believe (Rotter, 1967).

There are many more profound discussions about trust. For example, while some scholars believed that trust is just a simple antagonistic relationship between trust and distrust (Baker et al., 2018), which means trust is a single dimension, others have pointed out that trust is multidimensional (Lewicki et al., 1998), high trust and high distrust of something can coexist. As a primarily study, instead of going deeper into this, trust will be treated as a simple unidimensional state. That is to say, trust is treated as a value on a single segment. The endpoints of the segment are outright distrust and full trust, respectively. This study will focus on the violation and repair of trust, and the measurement of trust will be represented by a scalar.

1.1.2 Trust measurement

If we want to conduct experiment with trust change, understanding the approaches to measure trust is indispensable. Before we dive into trust measurement of Human-Robot interaction, it is also necessary to have a look at Human-Human interaction first. In examples of H-H trust research, there are two main approaches to measure trust, which are self-report measures and observational measurement respectively (Baker et al., 2018). Both the two methods have their advantages and limitations.

Self-report method has the most obvious advantage of its ease of implementation and its flexibility. Participants are asked to fill in questionnaires consists of several trust-

measure-related questions to report their perceived trust. The most common type of question is scales. Another advantage of this method is that it directly measures the participant, rather than through other intermediaries (WHEELESS & GROTZ, 1977). Also, the format of self-report can be used to follow up measurements at different stages of an experiment, such as at different time points (Couch et al., 2010). The drawback of this measurement method is that it relies on the subjective response of the participant, and it is impossible to obtain accurate measurement results assuming that the participant is unaware. From another perspective, however, trust itself is a subjective concept, which may be why many studies still use this measurement method (Baker et al., 2018).

The method of observing does not directly make the conversation with the participant, instead measures through other references. The common practice is to map trust with the behavior of the participant and express the trust by observing and quantifying the behaviors. In the experiment of Glaeser et al. (2000), they informed subjects that envelopes containing dollars were randomly dropped in several public areas and measured the subjects' trust in pedestrians who might take some money from the envelopes by calculating the subjects' valuation of the amount inside the envelopes. The advantage of observational methods is that it eliminates the influence of subject bias on the outcome because it measures objective data on participant behavior rather than subjective perceptions. However, the premise of this benefit is that participants are not aware of being measured, and behaviors that measure trust are also not always easy to find.

Despite the relatively small number of studies, the current H-R interaction trust experiments generally adopt the same measurement approaches as H-H empirical research (Baker et al., 2018). That is to say, the measurement are no more than subjective measurement (self-report format) and objective measurement (observation methods). Freedy et al. (2007) evaluated trust through the observation to human behaviors in a H-R collaboration scenario in their experiment. They hypothesized that trust is reflected in the expectation of willingness to hand over control to the robot and

measured it through observing the number of times a human operator handed over to the robot under the expected benefit target. This is a good attempt to establish objective measurements, but it cannot be a universal method because it only works for such specified interaction scenarios. Generality is also required in subjective methods, as the items in the measurement scale should not be scenario specific. The measurement scale developed by Charalambous et al. (2016) has the effectiveness in measuring trust in an industry setting of H-R interaction, but it is too context-limited. “Trust Perception Scale-HRI” provided by Schaefer (2016) could be applied to a wider range of scenarios, it is also widely recognized for its ability to identify the basic components of trust (Baker et al., 2018). However, due to its versatility, some items in the scale may need to be adjusted for specific scenarios.

1.2 Trust violation and repair in Human-Human Interaction

Trust violation is a common phenomenon in H-H interaction (Gillespie, 2017), and the trust changes in human relationships are very complex (Baker et al., 2018). Along with this, there are many behaviors that can lead to a loss of trust between humans, including mistakes and betrayals (Wildman, 2011), or incompetence (P. H. Kim et al., 2013a). The factors of these violations of trust can be summarized into two categories, which are competence and integrity (P. H. Kim et al., 2013b; Mayer et al., 1995). Competency-related violation behaviors relate to the person's own abilities, that is, the violation is due to a lack of capacity. Integrity-related violation behaviors have nothing to do with human ability, people have the ability to complete tasks but violate good faith, such as conventions. When a trust violation occurs, the party causing the loss of trust will often want to adopt strategies to repair the other's trust. For different types of trust violations, the most effective trust repair strategies are also different (P. H. Kim et al., 2009).

Corresponding to the two basic categories of trust violation mentioned above, there are also two basic trust repair strategies (Sebo et al., 2019). When a competency-related trust violation is committed, there is evidence shows that *apology* is the most effective

strategy (P. H. Kim et al., 2004). In the experiment conducted by Kim et al. (2004), participants will be asked to imagine themselves as hiring managers, and the results shows that they were more willing to trust candidates who apologized for their mistakes rather than denied them. In this context, the candidate's trust violation behaviors are related to competency.

In another case, where trust violations are integrity-related, *apology* may not be the most effective strategy, but *denial* (P. H. Kim et al., 2004). In the study of (Sigal et al., 1988), it is found that if politicians deny some of their personal and financial problems instead of acknowledging and apologizing, it will make it easier for them to get votes. This may be due to the greater weight of negative information that is included in the trustworthy assessment criteria on integrity-related trust issues and an apology confirms the existence of this negative information (Skowronski & Carlston, 1987). However, as (P. H. Kim et al., 2006) pointed out, the prerequisite for the success of denials is that the violator also provides a convincing additional explanation to get rid of suspicion. They also suggested that integrity-related in itself is a more serious trust violation, so probably all strategies are worth trying.

In addition to the remediation strategies corresponding to these two types of trust violations, other repair strategies are mentioned, such as *trustworthy action* (Schweitzer et al., 2006). After a loss of trust, a series of behaviors that do not violate trustworthiness persisting for a period of time may be enough to restore the trust. This approach requires that the two parties continue to interact after the trust violation occurs. Another strategy is *promise* (Schweitzer et al., 2006). Commitment to change in behavior will speed up the process of rebuilding trust, but this strategy will be less effective if deception occurs in previous interactions.

Determining the type of trust violation is critical to the choice of trust repair strategies. However, as Schweitzer et al. (2006) pointed out, a single repair strategy may not be sufficient in the process of repairing trust. Associating with daily-based human

interaction, it is inspired that human normally use the combination of strategies instead of one to repair trust violation. Rethinking about H-R interaction, we may also envisage the possibility of multiple strategy combinations.

1.3 Trust in Human-Robot Interaction

After the discussion about trust research in H-H interaction domain, we now come to the H-R interaction scenarios. The interaction process is not quite same in H-R since the success of interaction is affected by many factors (J. D. Lee & See, 2004). To investigate how trust change in human robot collaboration, the factors that may affect the interaction as well as the learning from demonstration method will be reviewed first, then the trust violation behaviors and corresponding repair strategies in H-R research will be summarized. Finally, the last section of this chapter will explain the effective trust measurement methods in H-R interaction experiment design.

1.3.1 Factors that affect trust in H-R Interaction

Trust perceived in H-R interaction can be influenced by factors from multiple perspectives. According to the meta-analysis by (Hancock et al., 2011), there are three main categories of factors that may have an impact on trust, which are human-related, robot-related, and environment-related factors accordingly. However, the study also concluded that among the three categories robot-related factors plays the most significant roles in the development of trust, while environmental factors are only moderately associated, and there are little evidence shows that human-related factors also have effects. The existing research of robot-related factors mainly fall on three aspects, which could be summarized as robot's appearance, reliability, and system transparency.

Compared to humans, the most intuitive difference when interacting with a robot can probably be the appearance (Rau et al., 2010). Also, the unique point of robotics compared to other forms of automatic might be its life-like characters (Prendinger &

Ishizuka, 2004). Robots are often designed to have different appearances in different application areas, such as service robots often have an anthropomorphic appearance (Riek et al., 2008), while industrial robots have a machinelike appearance (Ravichandar et al., 2020). Human's perception of robot's appearance look is even influenced by a variety of factors, including cultural background (Rau et al., 2010). In fact, the effect of robot appearance on human perception can be developed into a separate area of research. Notably, there is evidence that the physical form of robots does affect human perception of robot trustworthiness (Schaefer et al., 2012).

The second widely cited factor affecting trust in robot is the reliability. In other words, the ability of robot to complete assigned tasks results in the difference in level of trust (de Vries et al., 2003). Much research has shown that reliable automation leads to a greater degree of human trust (de Vries et al., 2003; J. Lee & Moray, 2007; Ross et al., 2008). Although some researchers argued that occasional errors do not significantly reduce human trust in a highly reliable automation if not successively failed (Parasuraman, 1997), there is sufficient evidence that robot errors do reduce human trust in the short term and have been applied to H-R interaction research as a condition for manipulating trust (Esterwood & Robert, 2021, 2022; Sebo et al., 2019). However, from another point of view, it is the incompetence of robots that creates new spaces for human assisting robot in completing tasks (Ullman et al., n.d.), which may eventually lead to the result of enhancing human trust in robots.

Another factor that cannot be ignored is transparency. In order to improve the collaboration between human and robot, humans need to understand the scope of their work partners' capabilities and when they need their own help, and then the robot's interpretability to their own behavior is particularly important (Wachter et al., 2017). The explainability, the ability to help humans understand the logic of their decision-making process and the principles of their functions, is described as the transparency of robots (Hoff & Bashir, 2015). Robot transparency has been shown to impact trust in a number of areas, including manufacturing (T. Kim & Hinds, 2006), autonomous driving

(Verberne et al., 2012), and the military (Boyce et al., 2015; Dzindolet et al., 2003). In the LfD process being discussed in this thesis, transparency is even more noteworthy since the human plays the role of teacher to help robot execute its tasks (Ravichandar et al., 2020). However, there is still not enough research on how transparency specifically affects trust, and whether it can affect trust in any type of human-robot interaction scenarios.

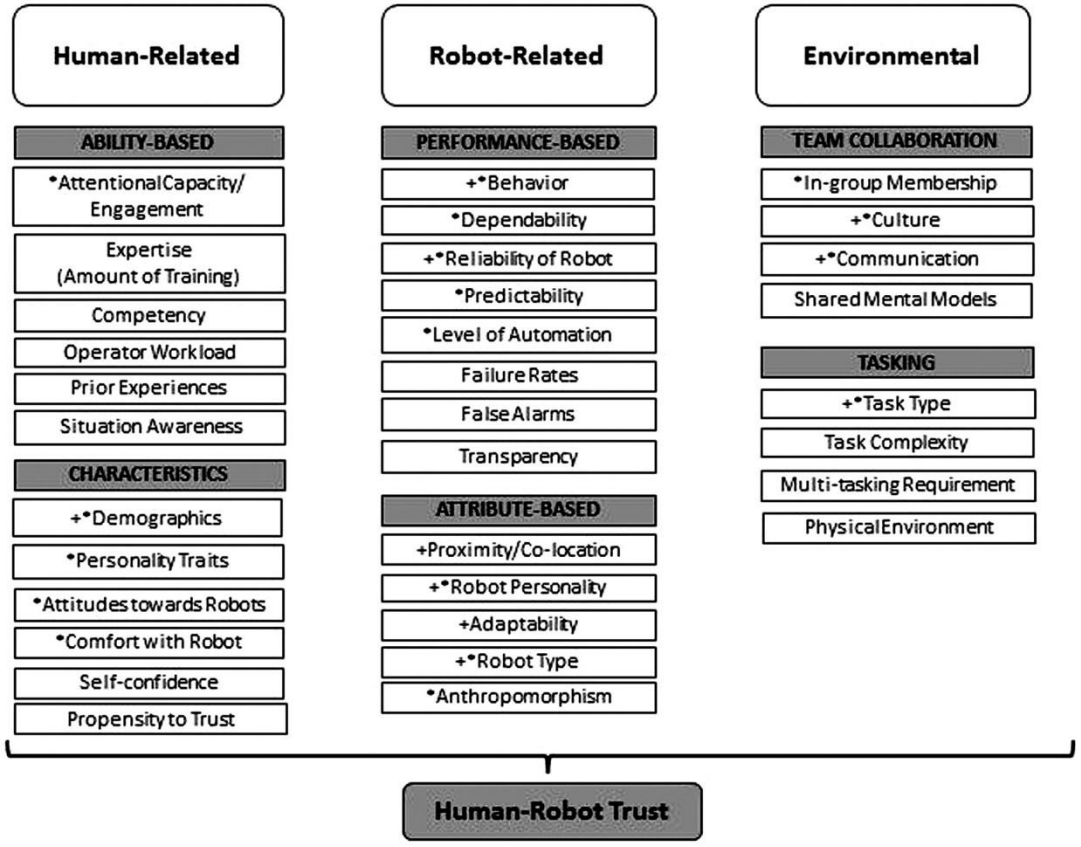

Figure 1. Factors of trust development in human-robot interaction (Hancock et al., 2011, p.521). These factors were identified a priori via literature review and subject matter expert guidance.

Since trust in H-R interaction is a relatively new field of study, there might be other factors that may have an influence and still not be discovered yet. In addition to the impact of these factors on human trust in robots, it is curious whether they will also affect the repair of trust after robots have committed trust-violating behaviors. To have

a clear view, the common types of trust-violating behaviors are needed to be studied as well as known effective trust repair strategies.

1.3.2 Trust repair strategies in H-R Interaction

Compared to H-H trust research, there are few studies that have explored trust violation and repair in H-R interaction. In order to identify different types of trust violation behaviors as well as corresponding repair strategies, Marinaccio et al. (2015) have developed a framework to summarize the findings of previous empirical studies. In the framework, trust violation behaviors are considered to be equivalent to robot errors and are divided into four types, which are *slips*, *lapses*, *mistakes*, and *violations* accordingly. *Slips* refers to errors of commission that an intend action was wrongly executed, and the most suitable repair strategy for this type of error is denials. *Lapses* is errors of omissions and apologies can be the best solution. Apologies also works for *mistakes* which stands for errors of planning or judgement. For *violation* errors, which means an intentional commission error, the best way to repair trust would also be denials. The difference between *slips* error and *lapses* is that the former is the wrong execution of an appropriate action, usually associated with attention deficits, while the latter is not performed at all, associated with memory deficits (Reason, 1990). And the difference between *mistakes* and *violations* lies in initiative.

It is not hard to recognize that the framework is based on the previous competence and integrity categorization provided by H-H empirical research (P. H. Kim et al., 2013a; Mayer et al., 1995). *Slips* and *violation* involve with integrity-based trust violation behaviors so denials would be the most effective repair strategy, while competency-related behaviors like *lapses* and *mistakes* need the strategy of apologies more. Even though the model is based on H-H research and there are few H-R research has verified this, there is still evidence that could be found that the model also works with H-R interaction (Sebo et al., 2019). The strategies of apologies and denials were proved to have effect on repairing trust when trust violation behaviors were committed by robots

in this study, and the suitable appliable cases are also corresponding with H-H situation.

The strategies of explanations and promises which have also been mentioned in H-H studies may also have effects on H-R trust repair (Esterwood & Robert, 2021, 2022; Robinette et al., 2015; N. Wang et al., 2018). However, there were no studies combined them with the competency-integrity framework. In the experiment of Robinette et al. (2015), a virtual 3D office environment was built, and participants were asked to follow the navigation robot to the exit when a simulated dangerous situation occurred. The robot made obvious detours during the lead, and then took different strategies to repair trust. The result of the experiment shows that promises is actually a more effective strategy in repairing trust compared to apologies and denials. Their study also investigated the best time for robots to adopt trust repair strategies, and it was found that robots should take actions before the trust decision is made by human for a larger efficacy, instead of right after the trust violation behavior. The effectiveness of promises was also verified by Esterwood & Robert (2021), where explanations was also found to have a large impact on trust repair when the robot is anthropomorphic. However, in other studies, explanations have been found to have no effectiveness to repair trust (Kox et al., 2021; M. K. Lee et al., 2010).

Through these literatures, it is discovered that the validity of apologies and denials and the conditions for their application are less controversial, in which apologies apply more to competency-related violations, and denials work better for integrity-related violations. But there does not seem to be a uniform conclusion on explanations and promises, neither enough evidence to show their effectiveness, nor to integrated them with the competency-integrity framework. Although these H-R trust repair studies have already given us some inspiration and can serve as a good starting point, there is still a lack of evidence on what is the most effective trust repair strategy, and there is no systematic summary of various situations and corresponding repair means. More empirical studies still need to be conducted in H-R settings.

1.4 Research gap and current study

There is some research that already explored the effectiveness of different trust repair strategies after trust violation in H-R interaction scenarios. The extent to which *apologies*, *denials*, *explanations*, and *promises* affect trust repair are compared (Esterwood & Robert, 2021, 2022), while *apologies* and *denials* are specifically compared based on the competency-based and integrity-based errors (Sebo et al., 2019). However, these studies only compared the performance of single strategy, and none of them have investigated the combination of strategies. Thinking back to H-H situations, people do not always use a single strategy when trying to save trust, but use a combination of them, and as Schweitzer et al. (2006) suggested, a combination of strategies tends to be more effective than a single strategy. In H-R contexts, the combination of *explanations* and *promises* has already been revealed to outperform the single *promises* strategy (N. Wang et al., 2018). However, this study only investigated the combination of *explanations* and *promises* and did not consider the other possible combinations. It will be interesting to compare and find the most effective combinations of repair strategies. In this way, we may further improve the solidity of H-R collaboration by applying the findings into robot design.

In addition to this, transparency has already been shown to increase trust in robots on trust (Boyce et al., 2015; Dzindolet et al., 2003; T. Kim & Hinds, 2006; Verberne et al., 2012). However, it is still not clear how it will have an influence on trust repair effectiveness. This is also an important area of study, especially in the context of LfD. Therefore, it is necessary to conduct a related study to reveal the importance of transparency in H-R interaction with robot LfD.

Finally, no such trust study has been conducted in the scenarios of LfD. The LfD approach has a unique approach to human-robot collaboration, with humans playing the role of teachers, unlike conventional programming methods for robots used in previous studies (Ravichandar et al., 2020). So there is reason to believe that there will

be a difference in the change in trust in LfD process as well. As an emerging technology, its full potential depends on whether applications can foster appropriate trust on these robots.

This study explores the possible combinations of known effective single trust repair strategies to find out if they can help rebuild trust or decrease trust loss after trust violation in H-R collaboration scenarios. Specifically, the study will compare the effectiveness of strategies combinations *apologies* & *explanations*, *apologies* & *promises*, and *explanations* & *promises*. According to the classification by Marinaccio et al. (2015), there only exists competency-based errors in the industry setting. Based on the evidence found by Sebo et al. (2019), *denials* are only effective after integrity-based violation behaviors. Due to this reason, *denials* will not be taken into account in this study. The scenario will be specified to a food processing industry setting where the LfD method is applied. As a preliminary study, the combination of all three single strategies as well as per single strategy will not be included into the comparison. Meanwhile, transparency as an important factor will also be induced into the study, and it will be manipulated as an experimental condition.

In this way, participants will observe different types of trust repair behaviors with whether a transparent setting or a nontransparent setting after the observation of robot trust violation behaviors. The main research question of the study can be formulated as follows:

Which combination of strategies will be the most effective to repair trust after trust violation in the robot LfD process?

The second research question cares about the effects of transparency. The research question is summarized as follows:

How will transparency of the robot learning process influence the effectiveness of

trust repair strategies?

Considering that transparency has already been shown to affect trust in H-R interaction (Boyce et al., 2015; Dzindolet et al., 2003; T. Kim & Hinds, 2006; Verberne et al., 2012), we have a good reason to infer that it will being a similar effect on trust repair efficiency. As such the hypothesis for this question will be:

Trust repair strategies with high transparency have overall higher effectiveness in repairing trust compared to low transparency.

Besides the two main research question, the study also will explore some other sub questions of interest. First sub question would be, is the effect of transparency on different trust repair strategies combinations the same? And since we are curious about whether the perceived trust is influenced by how well people know about robots, the second sub question is, is the effectiveness of the trust repair strategy affected by the robot expertise level of participants? Last but not the least, we plan to collect participants' subjective suggestions on the most acceptable means of trust repair, which will also inspire future research and promote human-robot collaboration.

2. Method

To test our hypothesis, an online experiment was conducted using pre-recorded videos to create a simulated human-robot collaboration scenario in a food processing factory.

2.1 Design

The experiment was a 2 (transparency: transparent vs. nontransparent) x 4 (trust repair strategies: non-strategy vs. apologies & explanations vs. apologies & promises vs. explanations & promises) between-subjects design. The independent variable were transparency and trust repair strategies. The dependent variable of the study is perceived trust on the robot. The transparency conditions were transparent and nontransparent, in which participants were introduced the working principle of the robot or not. For trust repair strategies conditions, it included a control group without any trust repair strategy, and the other three groups of conditions under which the robot will adopt combination strategies of apologies & explanations, apologies & promises, and explanations & promises to repair trust. Table 1 provides an overview of each condition that participants were assigned to and the experimental design of the study.

Table 1

Experimental design to study the effect of repair strategies and transparency on perceived trust

Transparency	**Trust repair strategies**			
	Non-strategy	Apologies& explanations	Apologies& promises	Explanations& promises
Transparent	1.1	1.2	1.3	1.4
Nontransparent	2.1	2.2	2.3	2.4

2.2 Participants

The required sample size of this study is calculated through R Superpower package. The desired alpha value of the study is $\alpha = .05$, and the target power is $1 - \beta = .90$. The smallest effect size of interest for this study was based on an assumption of detectable difference on the 7-point Likert scale that was used to measure perceived trust to the robot. The smallest interest effect size of the study was $.50$ perceived trust scale points between different strategies, and a pooled stand deviation of 1.50 was assumed. Under such design, a sample size of 40 participants per condition, which meant 320 in total, would achieve a minimal power of $1 - \beta = .96$ and $1 - \beta = .90$ respectively for the two assumptions of the study. Therefore, a sample size of 320 participants (40 per condition) was required for the study.

The experiment was completed by 320 participants (133 females, 186 males, and 1 third gender) from 30 countries, most of them (224) from Europe. Their ages ranged from 18 to 75 ($M = 34.3, SD = 11.8$). The sample consisted of 44 students, 42 part-time employees, 171 full-time employees, 7 employers, 30 unemployed, 12 retired, and 14 other occupation participants. All participants were recruited through the online research platform Prolific. Participants were required to be sufficiently skilled in English and should have the minimal approval rate of $.90$ provided by Prolific. Participants were paid £2 for their participation.

2.3 Experiment settings

2.3.1 Robot

The robot used in the experiment is developed by Wageningen University & Research (WUR). Figure 2 shows an image of the working robot. The robot has the appearance of a robotic arm, drove by motors. The torso of the 6-axis robotic arm is provided by Universal Robots which has a working radius of 850 mm. The gripper at the front end

of the robotic arm is the product RG2 Gripper produced by Onrobot. With these components, the robot can flexibly pick and move objects. To implement the LfD function, the robot is also equipped with a camera (Intel® RealSense™ D435) that can capture fast movements.

Figure 2. The robot is gripping a banana.

The fruit and human pose detection function of the robot is realized by Python, based on the “cascade maskRCNN” model (fruit recognition model) and “HRNet” (human pose estimation model). Human activity recognition function of the robot is based on temporal convolutional networks. These functions allow the robot to apply LfD method to imitate human behaviors. The robot could also be manipulated manually through Python scripts, which allows us to create different experimental scenarios by manipulating its actions.

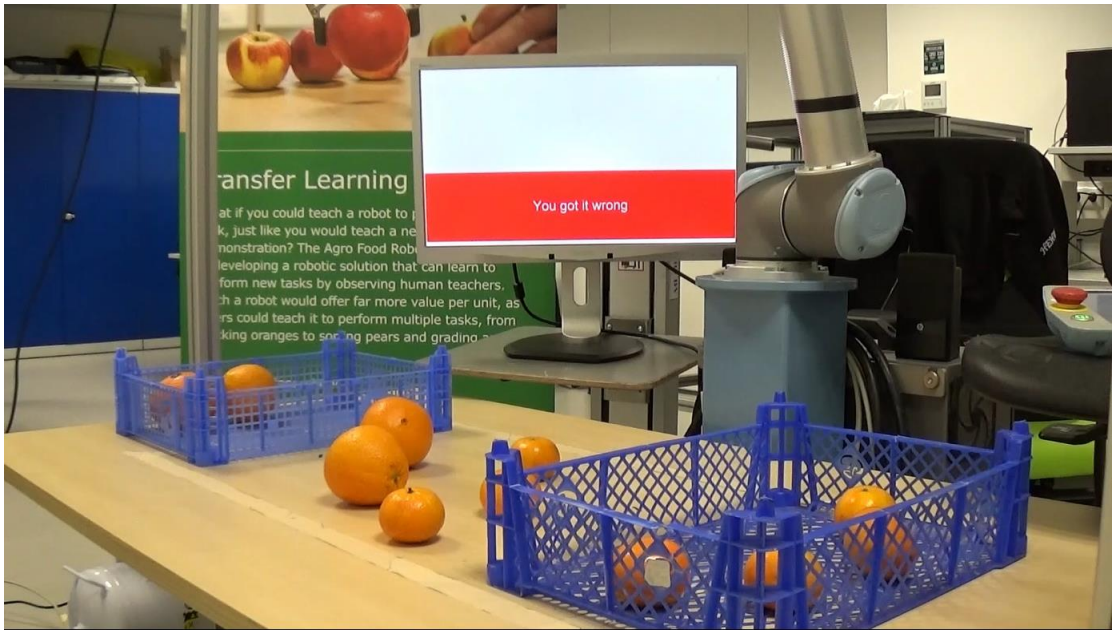

Figure 3. An experimental scene

The robot itself has no communication function, and in order to enable it to implement the trust repair strategies, a screen and speaker are installed next to it. Through a program written in Python, a bright red button will appear on the screen with the text "You got it wrong" on it. When the robot makes an error, the robot can be informed by clicking the button, so as to realize the dialogue between the human and the robot. The statement made by the robot will also be displayed on the screen, while the speaker will play the statement of the robot at the same time.

2.3.2 Stimuli

The eight experimental conditions were determined by a combination of transparency and trust repair strategies. Half of the participants were in the transparent group, watched a video showing the robot's learning process (figure 4), which included a human demonstration of two tasks (sorting oranges by their sizes and sorting bananas by their ripeness). The other half of participants, as nontransparent group, did not watch this video, but continued the rest of the experiment instead.

All participants were asked to watch the two robot normal working videos. In the first video, the robot placed large oranges in a box on one side of the table, while small oranges were placed in a box on the other side. The difference in size of oranges was very obvious and could be easily observed. The robot completed the task perfectly without any errors. In the next video, the robot also performed perfectly with no mistakes, placed the highly ripe bananas in the box on one side of the table and the low-ripe bananas on the other side. The ripeness of bananas could be easily observed by the difference in color (yellow and green).

*Figure 4***.** A screen in the transparency video

The video watched by all the participants also included the robot making an error. There were two versions of the video where the robot made a mistake, one that made a mistake in picking oranges and one that made a mistake when picking bananas. Participants watched one of them randomly. The robot put the fruits that should be distinguished into boxes on the same side, which was achieved by a Wizard-of-Oz method of remotely controlling the robot to make intentional mistakes by program to simulate possible errors in the real working situation of the robot. In these robot making mistakes

scenarios, the box on the side of the table was different from the normal operation situation, which simulated the trigger for the robot to make a mistake in a real environment, but participants were not required to notice this.

Right after the mistakes, a human worker clicked on the "You got it wrong" button and then the robot adopted trust repair strategies. In the four different conditions of trust repair strategies, the robot made various statements according to each condition. In the condition of Apologies & Explanations, the robot stated "*I am sorry I did not put the fruits into right boxes this time. I failed to do it correctly because the boxes are changed.*" In the Apologies & Promises condition, the robot stated "*I am sorry I did not put the fruits into right boxes this time. I will do better next time and get my job done*." In the Explanations & Promises condition, the robot stated "*I failed to do it correctly because the boxes are changed, I will do better next time and get my job done.*" These statements were displayed on the screen next to the robot and were presented to the participant by voice simultaneously.

All videos were recorded in the laboratory in Phenomea building on Wageningen University & Research campus. The shooting location was set as a fruit processing factory in the videos, however, not too much environment information was recorded. The main shot was a table with fruits and a robot fixed to one side of the table. The left and right sides of the robot were placed with the same or different boxes (corresponding to normal work and error situations).

2.4 Measurements

The main measurement object of the experiment was trust perceived by participants during the human-robot interaction process. This was done by self-report questionnaires based on the "Trust Perception Scale-HRI" developed by Schaefer (2016), since the scale was designed to measure perceived trust after observing robot's behaviors, and the usability of this scale is widely recognized ((Baker et al., 2018). The original scale

contains 40 items, since trust needs to be measured multiple times, overly long questionnaires may degrade the quality of responses. Based on this consideration, fifteen items on the scale were selected and adjusted into a 7-point scale to fit the experiment scenario, including three reversed coded items. Items like "*know the difference between friend and foe*" and "*warn people of potential risks of the environment*" were not selected since they were not appliable to the experiment scene. The trust measurement questionnaire was also subjected to a Cronbach's alpha scale reliability test to generate the perceived trust variable. The trust measurement questionnaire could be seen in table 2.

Each participant was asked to indicate the extent they want to continue hiring these robots by choosing answers on a 5-point scale from *never again* to *definitely*. The result could further evaluate participant's trust on the robot. Three open questions included "*To what extent will you continue hiring these robot workers and the reason*", "*do you think it is necessary for a robot to do something to repair your trust after making mistakes*", and "*what would you like the robot to do after making a mistake to restore your trust in it*" followed.

The demographic questions include age, gender, occupation, and self-reported expertise level of robot based on a 5-point scale where 1 corresponded to "*I have never touched anything robot-related*", 2 to "*My knowledge of robots only stays in non-scientific works*", 3 to "*I have limited knowledge and a few experiences with robot*", 4 to "*I have systematically learned about robots and have several working experiences*", and 5 to "*I am a robot expert and perhaps work on robot currently*", since it provided the information of the variety of the participants, also the experiment not only wants to find effects in students or people with high robot experience level, but also people in different groups.

Table 2

Trust measurement questionnaire

Please indicate to what extent you felt this robot will be …	
Number	Item
1	Dependable
2	Reliable
3	Unresponsive
4	Predictable
5	Act consistently
6	Have errors
7	Provide feedback
8	Meet the needs of the mission/task
9	Provide appropriate information
10	Communicate with people
11	Perform exactly as instructed
12	Follow directions
13	Incompetent
14	A good teammate
15	Perform a task better than a novice human user

2.5 Online survey setup

The online survey was built by software Qualtrics Core XM (Appendix X). The software enabled us to integrate videos into the survey and had the randomizer function that was convenient to assign participants into different experimental conditions. Two randomizers were used before and after the transparency video. Participants were first randomly assigned to transparent or nontransparent group. After this, including the

control group, taking into account the random assignment of the task of picking fruits in the experimental conditions, participants were assigned to one of eight conditions (one of four repair strategies, orange task or banana task). The randomizers could evenly assign participants to make sure each condition had same number participants.

The experiment interface was presented in a brief and intuitive design with clear feedback on participants' clicks. While the experiment is in progress, the participant's progress was displayed on the top of the interface with the format of a visualization progress bar. Once participated, the IP address of the participant was recorded. Despite the responsive design of the online survey, participants were expected to participate in the experiment on a computer or tablet whenever possible for the best experience.

In order to filter good quality responses, several means were used in the survey. First, in all the pages with videos displaying of the experiment, the *Next* button was not displayed when the time set to stay on the page reached the length of the video, preventing participants from skipping watching the video and going directly to the next step of the experiment. Second, in the pages of trust measurement questionnaires, the total time participants stayed on the page and the number of clicks will be recorded. If participants spent little time, we had reason to question the quality of the answers.

2.6 Procedure

Figure 5 shows the overview of the experiment process. First, participants were welcomed to participate in the experiment, they were also asked to fill in their Prolific ID. As the only personal information collected, it enabled us to filter good quality responses. As the experiment started, participants were asked to imagine themselves as managers of a food processing factory and inspect the robots working. Then, participants in transparent group were given a video of robot training, while participants in nontransparent group did not watch this video. Then the video began by showing participants two scenarios where the robot is working, and participants were asked to

fill out a trust measurement questionnaire after each. Then a video of the robot's failure was played. At this time, according to different experimental conditions, the robot made different statements, and the statements were transmitted to the participants (through video footage) in the form of text and audio at the same time. Each statement was followed by another trust measurement questionnaire. Next, participants were asked to answer some general open questions about the human-robot interaction scenarios. In the end of the experiment, some demographic information was collected. The experiment lasted approximately 10 minutes, in the last page of the survey, a completion code was provided, and the participants could copy and paste the code on Prolific website to inform the platform that they had completed the study, in this way they were sent to the waiting approval list and their responses could be checked by the experimenter. Once approved on the website, the platform automatically sent participants the compensation of £2.

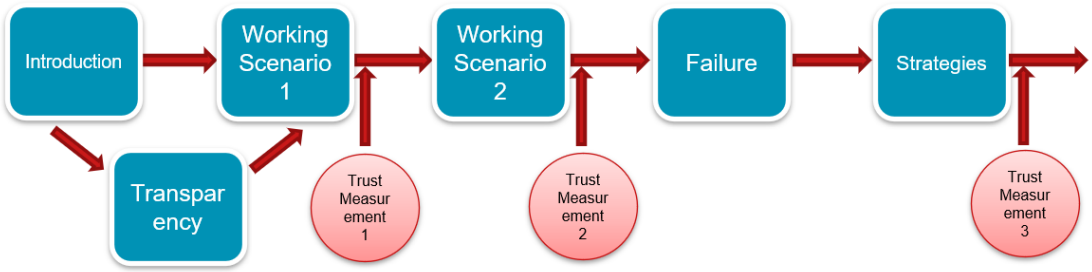

Figure 5. The procedure of the experiment

2.7 Statistical analysis

All the experiment data were exported from Qualtrics Core XM to an excel file as well as a STATA readable dataset. Then statistical analyses were done by software STATA 17. First, Cronbach's alpha test was done to generate the perceived trust variables. After this, the data were checked if there existed missing values or outliers using the general indicator that whether the absolute z-score is higher than 3.00 or not. The normality of the data per condition was checked by Shapiro-Wilk tests as well as skewness and

kurtosis tests. Paired *t*-tests were conducted to check the manipulation to the experiment. For research question exploration and hypothesis testing, instead of directly using trust change as the dependent variable in the statistical test, a more suitable approach according to Lüdtke & Robitzsch (2020) was used. A factorial analysis of covariance (ANCOVA) was conducted with the perceived trust in measurement 3 as the dependent variable, trust in measurement 2 as a covariance, and transparency and group as independent variables. This was followed by several contrast analysis to look the results into details. The visualization of the result was implemented by ggplot library in R.

What's more, the responses to general open questions were analyzed by the method of thematic analysis (the likelihood of continue hiring the robot was summarized in excel and tested the correlation between the last perceived trust). The analysis was conducted in the online software tool Miro. First, the initial codes were generated based on the answers of participants on each question. Next, the codes were collected and analyzed to form different themes under each question. The themes were then reviewed again to find connection between our main research question and hypothesis.

3. Result

3.1 Data quality check

In total, the study collected all the 320 responses from 320 participants. Ideally, there should be equal participants under all eight experimental conditions. However, the randomizer feature of Qualtrics Core XM was applied to all people who enter the online questionnaire, including some who fail to complete all of the experiments, and their presence resulted in the final unevenly distributed responses. Table 3 shows the detail of number of participants per condition.

Table 3

Number pf participants per condition

Transparency	**Trust repair strategies**				Total
	Non-strategy	Apologies & explanations	Apologies &promises	Explanations &promises	
Transparent	41	39	43	39	162
Nontransparent	42	41	38	37	158
Total	83	80	81	76	320

3.2 Descriptive statistics

To have a brief overview at all the variables, a descriptive summary was done (Table 4). The mean of age of participants is 34, while the self-reported expertise on robot ranged from 1 to 4. The perceived trust ranged from 2.87 to 7 with the mean of 4.77 in the first measurement. For the second measurement, it ranged from 1.67 to 7 and the mean was 4.64. The highest value for perceived trust in the last measurement was 6.33,

and the lowest was 1.4 while the mean was 3.70.

3.2.1 Perceived trust change process

First, a Cronbach's alpha reliability test was conducted to the items of the trust measurement questionnaire to measure the internal consistency between the items. Since the trust was measured three times, the response per item was obtained three times despite of the scale used. The test resulted in an alpha value of $\alpha = .82$, $\alpha = .85$, $\alpha = .88$ respectively, which is relatively high. The alpha value could be further increase by selectively removing several items from the battery, however, since the alpha was already quite high, it was not necessary.

Table 4

Descriptive analysis of all variables

Variable	Obs	Mean	SD	Min	Max
Gender	320	.59	.50	0	2
Age	320	34.33	11.81	18	75
Expertise	320	2.47	.81	1	4
Trust measurement 1	320	4.77	.77	2.87	7
Trust measurement 2	320	4.64	.87	1.67	7
Trust measurement 3	320	3.70	1.04	1.4	6.33

Note. For variable Gender, Female=0, Male=1, Third gender=2. The variables Trust measurement 1/2/3 were generated by the item battery using means of Cronbach's alpha.

The responses on all fifteen items in three measurements are calculated separately, and

the result can be found in Appendix B. Items 3, 6, 13 were reversely coded. For item scale calculation and the generation of new variable, these items were also reversed in advance. Based on the high internal consistency between the items, the variables *trust measurement 1/2/3* were generated from the mean of the fifteen items per observation (Table 4).

To better observe the process of trust change, a line chart was plotted based on the average of the three measurements of all participants and the corresponding 95% confidence intervals (figure 6). It is easy to see from the graph that trust gradually decreases in the process of human robot interaction, especially after observing robot mistake (between trust measurement 2 and 3). This indicates our manipulation of trust loss is successful. The result also received statistical support by paired *t*-tests with a $t(319) = 3.63, p < .001$ between trust measurement 1 and 2, and a $t(319) = 15.96, p < .001$ between trust measurement 2 and 3.

From the distribution of fifteen items in the three measurement it can be observed that not all the items in the scale change in the same way. it can be observed that not all the items in the scale change in the same way. Therefore, another line chart (figure 7) was made to visualize the change process of the score of each item. From the graph it is noticed that most of items follow the same trend as overall trust change process, however, two items *Communicate with people* and *Provide feedback* as well as three reversely coded items *have errors*, *incompetent*, and *unresponsive* show a clear upward trend conversely. Among them, the upward trend in *have errors* was most pronounced, while the change in *unresponsive* was very gentle. In addition, item *Provide appropriate information* remains almost same in the second and third measurement. The results indicates that the trust repair strategies adopted by the robot indeed influenced participants' judgments about whether the robot was competent and whether it provided feedback.

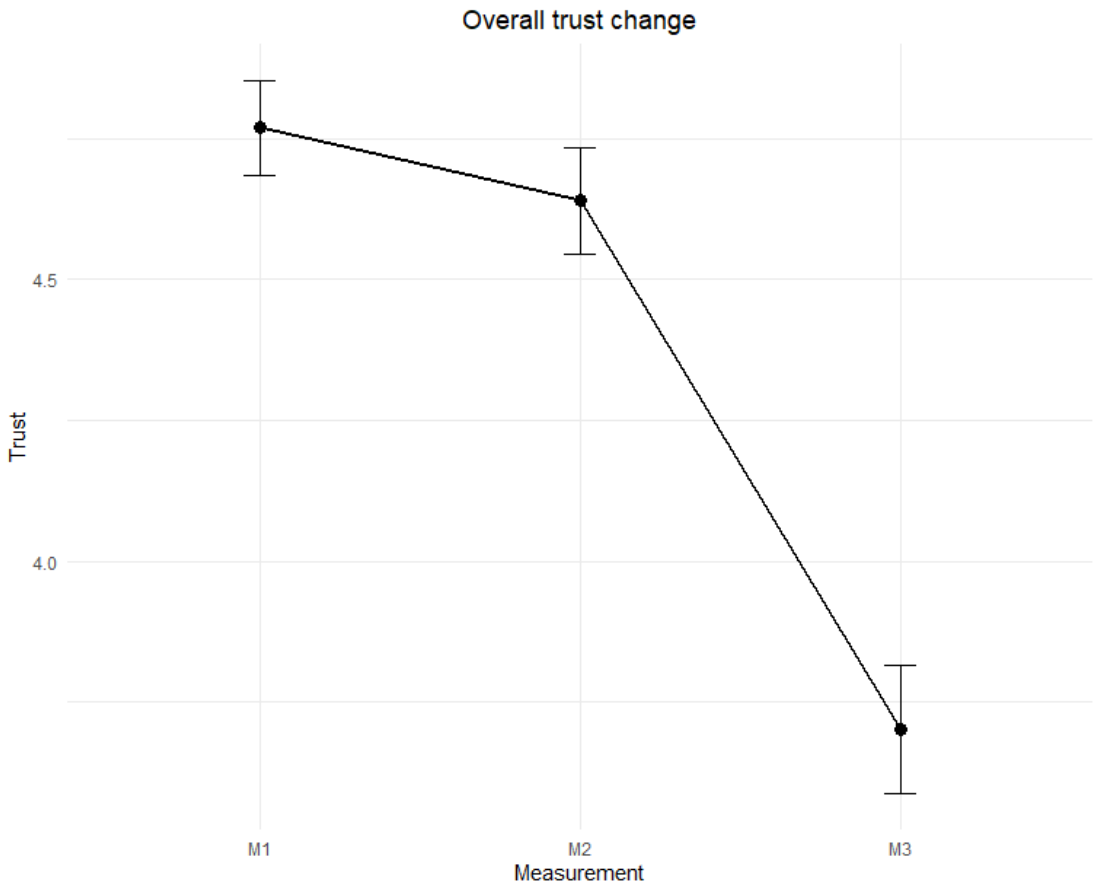

Figure 6. Overall trust change process of all participants

3.3 Overall trust change process

In order to further test our hypothesis, the variables trust measurement 1/2/3 as the main dependent variable needs to be statistical analyzed and compared between different conditions. Hence, before the statistical test, normality of the distribution of the variables should be checked. The trust variables were divided into eight conditions by the independent variable transparency and trust repair strategies, as such, normality of the trust variables should be checked for each of the conditions. A Shapiro-Wilk test as well as a skewness and kurtosis test were conducted, and three trust measurements in all the eight conditions had the result of $p > .05$ for both the two normality tests, only except for two conditions in which trust measurement 2 did not pass. Therefore, we assumed normality for these two conditions as well. In this way, we assumed that trust measurements variables are normally distributed for all the experiment conditions.

To obtain statistical support for the perceived trust change in the human robot interaction process, paired *t*-tests was conducted between trust measurement 1 and trust measurement 2, and a significant result was obtained ($t(319) = 3.63, p < .001$). This implies that even without a mistake by robot, participants had a perceived trust loss in the process of watching two robot working videos. Another *t*-test was run between trust measurement 2 and trust measurement 3, the result was also significant with a $t(319) = 15.96, p < .001$. Therefore, the process of trust change could be described as a relatively small decrease at first, and a large decrease after an error was observed.

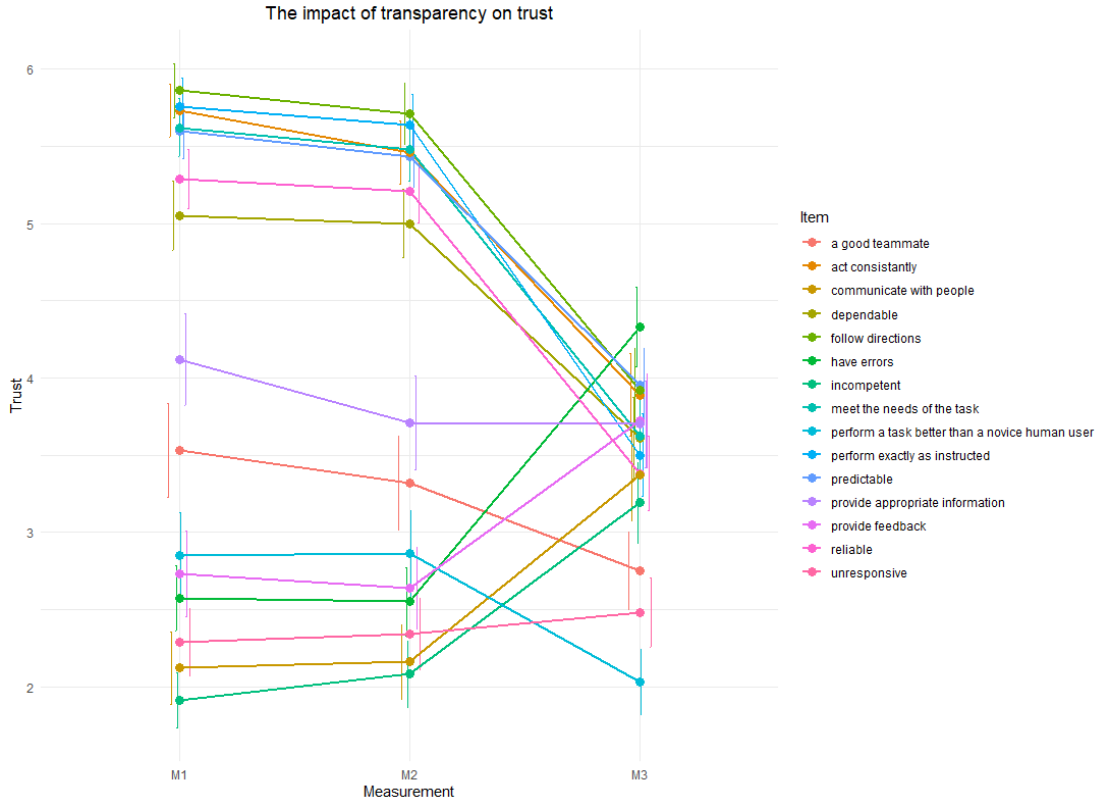

Figure 7. Trust change process of each item

3.4 Trust repair strategies

To answer the first research question that which combination strategies has the largest effect on repairing trust among all the strategies after trust violation behavior, two line graphs (figure 8) were made to show the trust change process under four repair strategies combinations, including the control group, in the condition of transparent or

nontransparent. As can be seen from the graph, the three combinations of strategies have a similar slope after robot's error in nontransparent group, clearly higher than control group. However, groups apologies & promises and explanations & promises have a different upward trend before mistake observed. In transparent group, all the four groups follow the similar trend of trust change. The control group with non-strategy has a clear lower trust level compared to other groups, while groups apologies & explanations and explanations & promises are above other groups. The graph indicates that trust repair strategies have effects on reducing trust loss after trust violation behaviors. Since the normality assumption was already met for all eight conditions, further statistical analysis could be conducted.

A factorial ANCOVA was conducted with trust measurement 3 as the dependent variable, trust measurement 2 as a covariance, and transparency and group as independent variables. The model run by ANCOVA was significant with an $F(8, 311) = 13.04, p < 0.001$, and the adjusted $R^2 = 0.23$. The combination of trust repair strategies also shows a significant effect on trust repair ($F(3.311) = 12.17, p < 0.001, \eta_p^2 = 0.11$. The effect of transparency and the interaction effect will be discussed in the following sections.

To answer RQ1, a contrast analysis was further conducted between pairs of different groups. Table 5 shows the results of the pairwise comparisons. There exists a significant difference on trust repair effectiveness between non-strategy group and apologies & explanations group ($p < .001$). Compared with apologies & promises group, both apologies & explanations group and explanations & promises group to have a significantly larger effectiveness in repairing trust ($p = .045, p = .037$). However, no significant difference was found in between the effectiveness of these two groups ($p = .91$). Also, in addition to comparison with control group, the two significant differences between strategies had a contrast of $C = .29, C = .31$ scale points respectively on the 7-point scale, which were smaller than the minimum effect size of interest. The differences were considered to be too small to be practically interesting, but they indeed

showed that the combination of apologies and explanations as well as the combination of explanations and promises had larger effectiveness in trust repair.

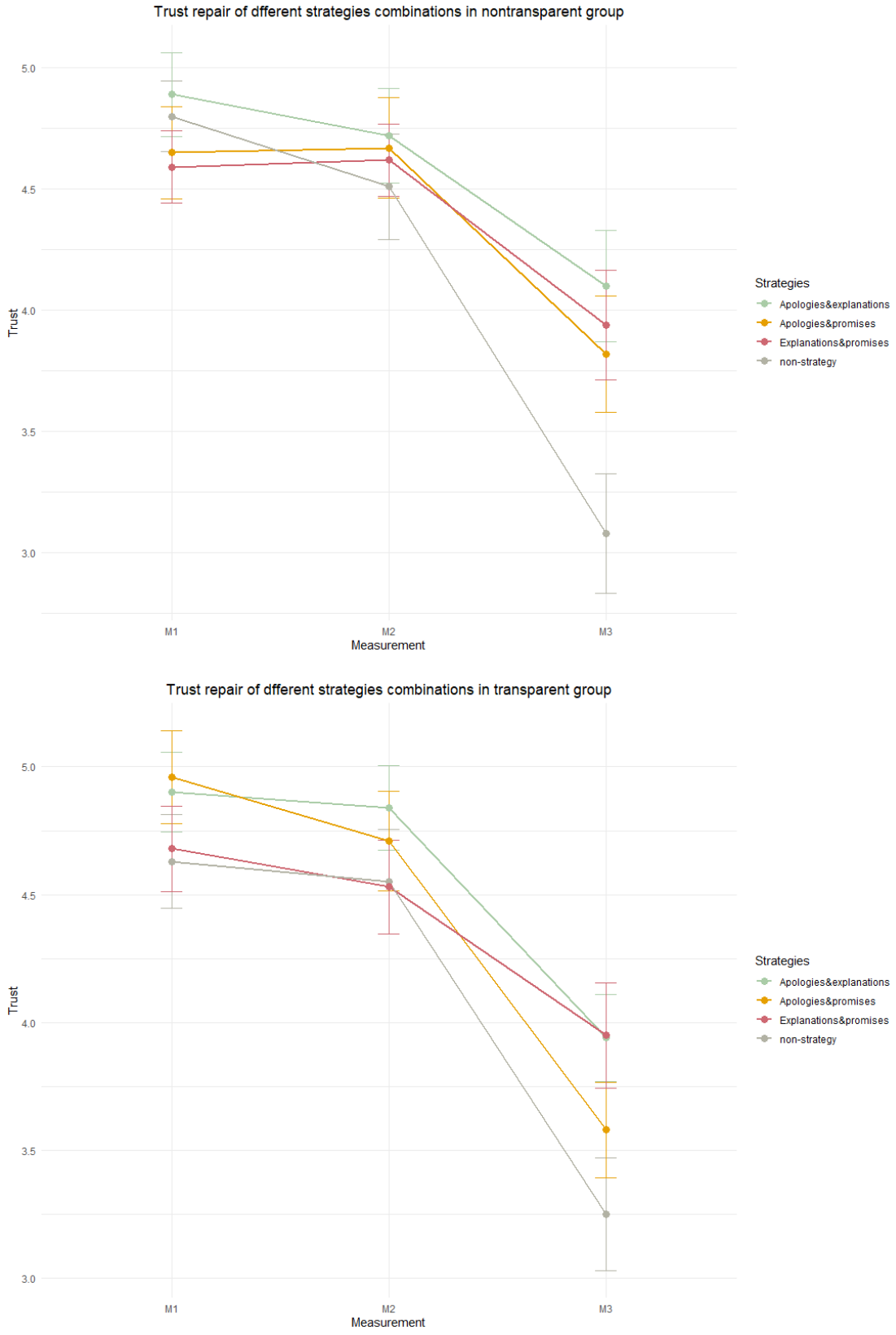

Figure 8. Trust change process under each combination of trust repair strategies

The results show the combination of apologies and explanations, and the combination of explanations and promises are the most effective trust repair strategies after robot's trust violation behavior. However, since a significant difference in effectiveness between these two strategies was not found, it cannot be stated that the combination of explanations and promises is the most effective trust repair strategy. Therefore, the first hypothesis was rejected. It at least can be verified that explanations & promises is a more effective strategy combination compared to apologies & promises as well as non-strategy.

Table 5

The result of contrast analysis in ANCOVA

Comparison	**Contrast**	**F**	**p**	**95% CI**
Non-strategy-Apologies & Explanations	-.74	26.94	<.001	[-1.03, -.46]
Apologies & Explanations-Apologies & Promises	.29	4.05	.045	[.01, .57]
Apologies & Promises-Explanations & Promises	-.31	4.41	.037	[-.59, -.02]
Apologies & Explanations-Explanations & Promises	-.02	0.01	.91	[-.30, .27]

3.5 Transparency and trust repair strategies

To test the hypothesis that trust repair strategies with high transparency have overall higher effectiveness in repairing trust compared to low transparency, the line graphs (figure 9) were firstly observed to compare the trust change process in transparent group and nontransparent group, with different repair strategies. The difference between the two groups was not clear judged from the graph, a look into the details of statistical analysis was needed.

The result of the factorial ANCOVA did not show a significant effect of transparency on perceived trust change ($F(1, 311) = .46, p = .496, \eta_p^2 = 0.002$). Given the small effect size and the non-significant result, there is no evidence shows that transparency had a direct effect on trust repair effectiveness. Therefore, the hypothesis was rejected. Through the observation of the graph, the result of first trust measurement seemed to be different between two groups. Hence, a *t*-test was run to further explore the possible influence by transparency on perceived trust. However, the result also did not find a significant difference between the two groups on trust measurement 1 with a $t(317.9) = -.70, p = .49$. The result indicates that transparency does not affect the effectiveness of trust repair after robot's trust violation behavior.

To further explore the influence of transparency on robot trust repair, the interaction effect of transparency and repair strategies might exist was examined. With the factorial ANCOVA, a non-significant result was obtained ($F(3, 311) = .96, p = .41, \eta_p^2 = 0.009$). As such, there is also no interaction effect found between the two independent variables.

3.6 Other exploratory research

3.6.1 Robot expertise influence on trust

To investigate the exploratory research question that whether the effectiveness of trust repair strategy affected by the robot expertise level of participants or not, the new factor *expertise* was added to the previous factorial ANCOVA model. The expertise of participants on robot was measured by means of self-report, with the minimum result of 1 and maximum of 4. No participants identified themselves as robot experts (a value of 5). In the model, it can only be observed a marginally significant effect of expertise on trust repair effectiveness ($F(3, 308) = 2.30, p = 0.078, \eta_p^2 = .02$). Hence, we could not conclude that expertise is an important factor that might affect trust repair.

3.6.2 Subjective reviews analysis

Open questions were set in the experiment to find out the attitude of participants toward hiring these industry robots as well as their subjective feeling and expectation to these robots. Review of the responses to these questions might give us more inspiration in the design of human-robot collaboration scenarios. The first question given to participants was the likelihood that they would continue to hire these robots as factory managers. Histogram in figure 10 shows the distribution of the number of responses. It can be observed that the responses are approximately normally distributed, while more participants hold a relatively negative attitude towards the robot workers. It is worth mentioning that a correlation test was done to the likelihood of continue hiring robots and the last trust measurement, with a significant result ($r = .43, p < .001$) it shows that participants' attitude is in line with their perceived trust on the robot.

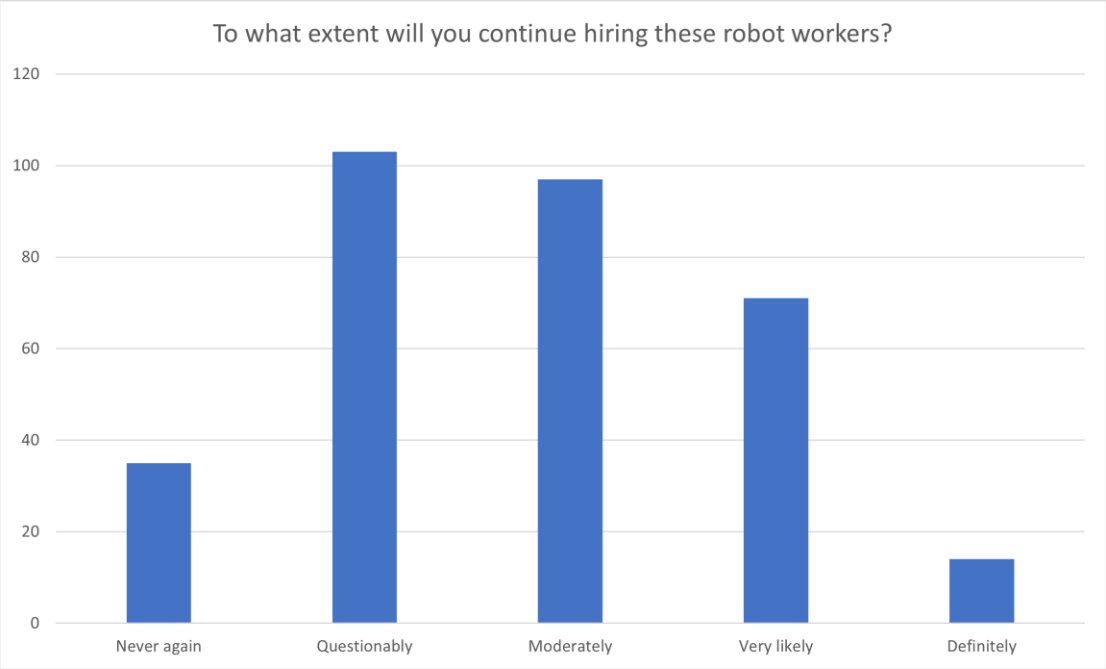

Figure 9. Distribution of responses in likelihood of continue hiring the robot

Afterward, participants were asked for the reasons of their responses and whether they felt necessary to have the repair strategies or not. They were also asked for their own opinions to the effective trust repair methods. From their responses, the themes were structured according to the three subjective open questions.

For the first theme *Reason to/to not continue hiring robot workers*, the most repetitive answer was that robots were considered too slow to outperform humans for the same job. The answers included “Slower than a human”, “Time wasting”, and “the robot is extremely slow” and so on. Besides, some participants were reluctant to continue choosing robots because of ethical concerns and worries about robots replacing human jobs. Some stated, “On the base that people need to work, more machine labor less jobs and more poverty”. In addition, some participants believe that robots need a lot of training time to avoid errors, which increases the cost of use. They mentioned, “Robots will be expensive in the beginning and require maintenance and checking errors which might costs more than hiring humans”. Very few participants felt they could not trust the robot for no reason, because they thought robots are just machines and cannot be

treated as humans.

For the second theme *Necessity of the trust repair strategy*, there were obviously more participants who thought it was unnecessary for the robot to adopt a trust repair strategy after making a mistake. Their main reason was robot itself had no responsibility (“The programmer would be the one who needs to improve”). For the ones who thought it was necessary, they believed that the robot need to prove that they can handle the job. Both participants with the two attitudes agreed that they need to know what lead to such errors so that the robot could be improved (“Yes, I think they should explain the reason for their mistake so it can be easily fixed”, "No, I think feedback on what led to the wrong decision is needed”).

The third theme is *Expected robot’s behaviors after mistakes*. Participants were asked to share their own ideas about the best way to restore their trust on the robot. Most of the answers are that the most effective way for robots is to complete the task accurately, without resorting to a repair strategy after a mistake. Some also suggested that the robot could be asked to perform the same task again to prove its ability (“Perform the task correctly for a number of times again to show that it was still competent at doing the job”). A quite large portion of the participants pointed out that the most important thing for them was to know why the robot made mistakes, so explanation was what they most wanted (“Explain its mistake to me”). Very few participants reckoned whatever the robot do would not make them trust it.

4. Discussion

Participants were asked to participate in an online experiment which involved several robot working videos in the context of fruit processing factory. Participants observed robot's failure in performing a specific task, which could be recognized as a trust violation behavior. The robot adopted different repair strategies after the error. The perceived trust was measured three times in the experiment to check the trust change process of each participant. The effectiveness of different combinations of trust repair strategies was investigated, as well as the influence of transparency on these human-robot collaboration scenario.

4.1 Effective combination of trust repair strategies

To answer the first research question, a line graph was made to illustrate the trust change process between different repair strategies, and a factorial ANCOVA as well as contrast tests were done to test the significance. In line with studies by Esterwood & Robert (2021), Robinette et al. (2015), and N. Wang et al. (2018), explanation and promises were considered the most effective repair strategies. The results of statistical analysis did find a significant difference in the effectiveness between combination apologies & explanations and combination apologies & promises, and a significant difference between combination explanations & promises and apologies & promises. However, the differences were smaller than the defined minimum effect size of interest, so the result was not considered to be practically meaningful. Also, a significant difference between the two effective combinations apologies & explanations and explanations & promises was not found.

According to Wang et al. (2018), the combination of explanations and promises outperforms single promises strategy. The result of our experiment did not verify the effectiveness of promises since apologies & explanations has a similar effect with explanations & promises. However, the effectiveness of explanations was further

validated, since the two combinations involved with explanations both performed better in repairing trust, which confirmed the studies by Esterwood & Robert (2021). The difference from this research is that the validity of explanations was found to apply not only to anthropomorphic robots, but also to non-anthropomorphic industrial robots. Based on the framework developed by Marinaccio et al. (2015), the type of error robot made in this experiment belonged to *mistake* since it made wrong judgement. Accordingly, apologies should also be effective since robot's failure could be recognized as a competency-based trust violation behavior. However, the effectiveness of apologies was not verified by the experiment since it did not outperform both explanations and promises. In the subjective reviews from the participants, apology was also not mentioned to be a strategy that is expected to happen after trust violation. Explanations, but was mentioned many times in the open questions by the participants because they wanted to know the reason why the robot made a mistake.

We could infer from the result that the reason apologies were not effective could be explained by the research result of Sebo et al. (2019). It was pointed out in their study that one of highly influential factors on trust repair is a reciprocal relationship between human and robot. Once a reciprocal promise is made by human to the robot, that is, when humans assume that robots will abide by reciprocal agreements with themselves, their perceived trust on the robot after trust violation and apology strategy will be increased compared to no agreements made. In the online experiment setting, even though participants were asked to imagine themselves as factory managers, it was hard for them to convince themselves that the robot would directly affect their benefits. Hence, the perceived trust after the violation behavior could not be repair effectively.

To explain the low effectiveness of promise strategy, research by (Robinette et al., 2015) provided a reasonable answer. The timing for robot trust repair was carefully discussed and it was found that promises would only have effect on repairing trust if it is adopted before the next decision to trust by human, instead of taking the strategy immediately after the trust violation. In our experiment, the process stopped right after the robot's

error, with no further robot working scenarios, and the trust repair strategies were adopted right after the trust violation. Given this condition, the promise strategy did not make the difference it was supposed to do.

4.2 The impact of transparency on trust repair

To answer the second research question that whether transparency will have an effect on trust repair effectiveness, similar analysis as the repair strategies study was conducted. The result of perceived trust neither showed a significant difference in different transparency group, nor found interaction effect of transparency with repair strategies. The result did not confirm our hypothesis that robot with higher transparency will be more effective in repairing trust. The hypothesis was made based on the consideration that transparent condition allowed participants to have preliminary knowledge about the possible errors might occur, and the explanations provided by the robot after trust violation corresponded to this knowledge. The result implies that transparency will not have an effect on trust repair.

Since transparency was found to have significant positive effect on perceived trust in several research (Boyce et al., 2015; T. Kim & Hinds, 2006; Verberne et al., 2012), this study came to a different conclusion. However, it exists the possibility that our manipulation on transparency was not successful or not convincing enough. To distinguish between different transparency, the only manipulation is providing a demonstration video of robot working principle. It is likely that the video did not catch too much attention from participants and there was also not a transparency perception check in the experiment. Hence, it is hard to tell if the transparency level of the two groups was unequal.

Another conjecture that may provide an explanation is that transparency in the system might lead to the perception of participants on some other defects in the robot. In the research by J. Wang et al. (2021), transparency was discovered not to be the higher the

better, too much transparency can even affect user pleasure. Several subjective reviews by participants mentioned the concern about high time cost on training the robot, which also indicated that the training process shown in the transparency video sometimes received negative impression. Therefore, with such bias on robot cognition might interfere with the impact of transparency on perceived trust itself.

4.3 Implications for future human-robot collaboration design

Instead of comparing single trust strategies, the study explored the effectiveness of different combinations of strategy in repairing trust in human-robot interaction process for the first time. The results show that the combination of strategies can indeed effectively repair the loss of human trust in robots after observing mistake. Although experiments did not find a strategy combination that was significantly superior compared to others, it found the importance to adopt explanations in the application of repair strategy combination. Therefore, in future robot work scenarios that require human supervision, industrial robot developers can add functions that provide feedback to the robot and allow the robot to explain the cause of its errors.

For the first time, a trust study in robot LfD application was conducted. Specifically, the study investigated whether the transparency of LfD process is an important factor in trust repair. The results show that transparency in the human demonstration process did not increase trust in robots. Hence, for the development of future industrial robots with LfD techniques, transparency should not be the main consideration, but more attention should be paid to the accuracy and efficiency of the robot.

In the study we also observed a significant loss in trust even the robot was normally working. This shows that in human-robot collaboration in industrial environments, it is not only the robot's mistakes that cause the loss of human trust, but also the robot's performance and the efficiency it demonstrates. As shown in the subjective questions section of the experiment, people show expectations for robotics techniques and

question the efficiency of robots. This makes us realize that in the context of completing industrial tasks, the functionality of robots is always the most important condition for promoting the widespread application of human-robot collaboration in the future.

The finding of the study is not only relevant with food processing industry, but also other industries that require relatively simple and highly repeatable human execution. For instance, it has implications for manufacturing and health care which are both promising application area for LfD methods (Fong & Tavakoli, 2018; Zhu & Hu, 2018).

4.4 Implications for trust measurement

The trust measurement method used in the study was self-report questionnaires based on the general HRI trust measurement scale developed by Schaefer (2016). The results of the study results suggested that this commonly used trust measure scale reflected different aspects of trust, since the items within the scale showed different temporal developments in the trust change process. Therefore, the factor structure of the scale could be further analyzed in the future, so that the scale could be developed to measure different components of trust. Besides, the method of objective measurement (behavioral measure) of trust could be adopted in the future study to compare its validity with the self-report measurement.

4.5 Limitations and future research

The first limitation of this study is as the previous section mentioned, the manipulation of transparency might not be successful. Future research is suggested to adopt more convincing methods to manipulate such variables or add a manipulation check in the experiment. According to Vigni et al. (2022), transparency of a non-humanoid robot could be expressed in a multi-model form which could be one option.

The second limitation is that this study only compared the difference between different combinations of trust repair strategy, instead of adding single strategies into the

comparison. Considering the feasibility of the experiment, no more conditions were added. In the future research, it is expected that the comparison between single strategy and combination of strategy could be made, to further investigate if combination truly improved the effectiveness in repairing trust after trust violation behaviors.

What's more, as Esterwood & Robert (2022) pointed out, the robot might not only have tasks performed wrong once in the real working context. However, the robot working scenarios in this experiment ended directly after the only trust violation behavior by robot. Since some participants suggested in the subjective questions that the performance of the robot after first trust violation was expected to be observed, future study could extend the experiment process to have more working as well as error scenarios of robot corresponding with trust measurements. The idea is also in line with the study of Robinette et al. (2015), and also provide the opportunity to find effects on timing to trust repair.

Another limitation of this study is the experiment was conducted online. Compared with physically interacting with robots, watching videos of human-robot collaboration scenarios will not enable participants to have a real immersive experience. Despite of the challenge to gather a large number of participants into the laboratory, it is recommended to conduct experiments in the real world setting as much as possible if the conditions are met in order to achieve the largest effect.

Last but not the least, it is encouraged to add more factors that might affect trust repair into the experiment in future research. In this preliminary research, only transparency was discussed with the trust repair strategies in repairing trust. Robot-related factors such as appearance, human-related factors such as expertise level on robotics, as well as environment-related factors such as noise, all of these could be the object of future human-robot interaction trust research.

5. Conclusion

With the development of robot technology, more and more robots are applied in the industry, bringing more human-robot collaboration scenarios. Faced with the inevitable trust loss problem in these scenarios, this study aims to find the most effective trust repair strategies that robots can take after violating trust, and the impact of system transparency on trust changes. In this online experiment of 320 participants in fruit processing factory setting, different combinations of trust repair strategies and different level of transparency were manipulated, and the process of participants' trust changes was tracked. The results show that the most effective trust repair strategies are Apologies & Explanations and Explanations & Promises. Contrary to expectations, transparency does not show the ability to influence changes in trust. This study is the first to explore the effectiveness of different strategy combinations, combined with the application of LfD method, which brings new inspiration for promoting human-robot collaboration and provides new ideas for human-robot interaction trust research.

References:

Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schäffer, A., Kosuge, K., & Khatib, O. (2018). Progress and prospects of the human–robot collaboration. *Autonomous Robots*, *42*(5), 957–975. https://doi.org/10.1007/S10514-017-9677-2/FIGURES/5

Argall, B. D., Chernova, S., Veloso, M., & Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, *57*(5), 469–483. https://doi.org/10.1016/J.ROBOT.2008.10.024

Bairam, U., Bogaardt, M. J., Burg, S. van der, Chauhan, A., Elburg, E. van, Kool, J., Lokhorst, C., Hemming, J., Mencarelli, A., Nieuwenhuizen, A. T., Ouweltjes, W., & Riviere, I. J. la. (2020). *Autonomous collaborative robots for agri-food processes*. https://research.wur.nl/en/publications/autonomous-collaborative-robots-for-agri-food-processes

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems*, *8*(4). https://doi.org/10.1145/3181671

Barber, Bernard. (1983). *The logic and limits of trust*. 189.

Boyce, M. W., Chen, J. Y. C., Selkowitz, A. R., & Lakhmani, S. G. (2015). Effects of Agent Transparency on Operator Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, *02-05-March-2015*, 179–180. https://doi.org/10.1145/2701973.2702059

Calinon, S., Guenter, F., & Billard, A. (2007). On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *37*(2), 286–298. https://doi.org/10.1109/TSMCB.2006.886952

Charalambous, G., Fletcher, S., & Webb, P. (2016). The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration. *International Journal of*

Social Robotics, *8*(2), 193–209. https://doi.org/10.1007/S12369-015-0333-8/TABLES/3

Chernova, S., & Thomaz, A. L. (2014). Robot Learning from Human Teachers. *Robot Learning from Human Teachers*. https://doi.org/10.1007/978-3-031-01570-0

Chu, V., Fitzgerald, T., & Thomaz, A. L. (2016). Learning object affordances by leveraging the combination of human-guidance and self-exploration. *ACM/IEEE International Conference on Human-Robot Interaction*, *2016-April*, 221–228. https://doi.org/10.1109/HRI.2016.7451755

Couch, L. L., Adams, J. M., & Jones, W. H. (2010). The Assessment of Trust Orientation. *Http://Dx.Doi.Org/10.1207/S15327752jpa6702_7*, *67*(2), 305–323. https://doi.org/10.1207/S15327752JPA6702_7

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Esterwood, C., & Robert, L. P. (2021). Do you still trust me? Human-robot trust repair strategies. *2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021*, 183–188. https://doi.org/10.1109/RO-MAN50785.2021.9515365

Esterwood, C., & Robert, L. P. (2022). Having the Right Attitude: How Attitude Impacts Trust Repair in Human - Robot Interaction. *ACM/IEEE International Conference on Human-Robot Interaction*, *2022-March*, 332–341. https://doi.org/10.1109/HRI53351.2022.9889535

Fong, J., & Tavakoli, M. (2018). Kinesthetic teaching of a therapist's behavior to a rehabilitation robot. *2018 International Symposium on Medical Robotics, ISMR 2018*, *2018-January*, 1–6. https://doi.org/10.1109/ISMR.2018.8333285

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007). Measurement of trust in human-robot collaboration. *Proceedings of the 2007 International Symposium on Collaborative Technologies and Systems, CTS*, 106–114. https://doi.org/10.1109/CTS.2007.4621745

Gillespie, N. (2017). Trust dynamics and repair: An interview with Roy Lewicki. *Http://Dx.Doi.Org/10.1080/21515581.2017.1373022*, *7*(2), 204–219. https://doi.org/10.1080/21515581.2017.1373022

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring Trust. *The Quarterly Journal of Economics*, *115*(3), 811–846. https://doi.org/10.1162/003355300554926

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

Havoutis, I., & Calinon, S. (2019). Learning from demonstration for semi-autonomous teleoperation. *Autonomous Robots*, *43*(3), 713–726. https://doi.org/10.1007/S10514-018-9745-2/FIGURES/12

Hoff, K. A., & Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Iqbal, J., Khan, Z. H., & Khalid, A. (2017). Prospects of robotics in food industry. *Food Science and Technology*, *37*(2), 159–165. https://doi.org/10.1590/1678-457X.14616

Kaiser, J., Melbaum, S., Tieck, J. C. V., Roennau, A., Butz, M. V., & Dillmann, R. (2018). Learning to Reproduce Visually Similar Movements by Minimizing Event-Based Prediction Error. *Proceedings of the IEEE RAS and EMBS International Conference on Biomedical Robotics and Biomechatronics*, *2018-August*, 260–267. https://doi.org/10.1109/BIOROB.2018.8487959

Khavas, Z. R. (2021). *A Review on Trust in Human-Robot Interaction*. https://arxiv.org/abs/2105.10045v1

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013a). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14. https://doi.org/10.1016/J.OBHDP.2012.08.004

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013b). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14. https://doi.org/10.1016/J.OBHDP.2012.08.004

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The Repair of Trust: A Dynamic Bilateral Perspective and Multilevel Conceptualization. *Https://Doi.Org/10.5465/Amr.2009.40631887*, *34*(3), 401–422. https://doi.org/10.5465/AMR.2009.40631887

Kim, P. H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. external attributions for the repair of trust after a competence- vs. integrity-based trust violation. *Organizational Behavior and Human Decision Processes*, *99*(1), 49–65. https://doi.org/10.1016/J.OBHDP.2005.07.002

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- versus Integrity-Based Trust Violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication*, 80–85. https://doi.org/10.1109/ROMAN.2006.314398

Kox, E. S., Kerstholt, J. H., Hueting, T. F., & de Vries, P. W. (2021). Trust repair in human-agent teams: the effectiveness of explanations and expressing regret. *Autonomous Agents and Multi-Agent Systems*, *35*(2), 1–20. https://doi.org/10.1007/S10458-021-09515-9/TABLES/4

Lauretti, C., Cordella, F., Guglielmelli, E., & Zollo, L. (2017). Learning by Demonstration for Planning Activities of Daily Living in Rehabilitation and Assistive Robotics. *IEEE Robotics and Automation Letters*, *2*(3), 1375–1382.

https://doi.org/10.1109/LRA.2017.2669369

Law, T., & Scheutz, M. (2021). Trust: Recent concepts and evaluations in human-robot interaction. *Trust in Human-Robot Interaction*, 27–57. https://doi.org/10.1016/B978-0-12-819472-0.00002-2

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Https://Doi.Org/10.1518/Hfes.46.1.50_30392*, *46*(1), 50–80. https://doi.org/10.1518/HFES.46.1.50_30392

Lee, J., & Moray, N. (2007). Trust, control strategies and allocation of function in human-machine systems. *Http://Dx.Doi.Org/10.1080/00140139208967392*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). *Gracefully mitigating breakdowns in robotic services*. 203–210. https://doi.org/10.1109/HRI.2010.5453195

Lewicki, R. J., McAllister, D. J., & Bies, R. I. (1998). Trust And Distrust: New Relationships and Realities. *Https://Doi.Org/10.5465/Amr.1998.926620*, *23*(3), 438–458. https://doi.org/10.5465/AMR.1998.926620

Lüdtke, O., & Robitzsch, A. (2020). *ANCOVA versus Change Score for the Analysis of Nonexperimental Two-Wave Data: A Structural Modeling Perspective*. https://doi.org/10.31234/OSF.IO/5ZDME

Ma, Z., Ben-Tzvi, P., & Danoff, J. (2016). Hand Rehabilitation Learning System With an Exoskeleton Robotic Glove. *IEEE Transactions on Neural Systems and Rehabilitation Engineering : A Publication of the IEEE Engineering in Medicine and Biology Society*, *24*(12), 1323–1332. https://doi.org/10.1109/TNSRE.2015.2501748

Maeda, G. J., Neumann, G., Ewerton, M., Lioutikov, R., Kroemer, O., & Peters, J. (2017). Probabilistic movement primitives for coordination of multiple human–robot collaborative tasks. *Autonomous Robots*, *41*(3), 593–612. https://doi.org/10.1007/S10514-016-9556-2/FIGURES/21

Mandlekar, A., Zhu, Y., Garg, A., Booher, J., Spero, M., Tung, A., Gao, J., Emmons, J., Gupta, A., Orbay, E., Savarese, S., & Fei-Fei, L. (2018). *RoboTurk: A*

Crowdsourcing Platform for Robotic Skill Learning through Imitation. https://arxiv.org/abs/1811.02790v1

Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015). *A Framework for Rebuilding Trust in Social Automation Across Health-Care Domains*. https://doi.org/10.1177/2327857915041036

Matheson, E., Minto, R., Zampieri, E. G. G., Faccio, M., & Rosati, G. (2019). Human–Robot Collaboration in Manufacturing Applications: A Review. *Robotics 2019, Vol. 8, Page 100*, *8*(4), 100. https://doi.org/10.3390/ROBOTICS8040100

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Https://Doi.Org/10.5465/Amr.1995.9508080335*, *20*(3), 709–734. https://doi.org/10.5465/AMR.1995.9508080335

Meng, Y., Munroe, C., Wu, Y. N., & Begum, M. (2016). A learning from demonstration framework to promote home-based neuromotor rehabilitation. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, 1126–1131. https://doi.org/10.1109/ROMAN.2016.7745249

Moro, C., Nejat, G., & Mihailidis, A. (2018). Learning and Personalizing Socially Assistive Robot Behaviors to Aid with Activities of Daily Living. *ACM Transactions on Human-Robot Interaction (THRI)*, *7*(2). https://doi.org/10.1145/3277903

Mueller, C., Venicx, J., & Hayes, B. (2018). Robust Robot Learning from Demonstration and Skill Repair Using Conceptual Constraints. *IEEE International Conference on Intelligent Robots and Systems*, 6029–6036. https://doi.org/10.1109/IROS.2018.8594133

Najafi, M., Sharifi, M., Adams, K., & Tavakoli, M. (2017). Robotic assistance for children with cerebral palsy based on learning from tele-cooperative demonstration. *International Journal of Intelligent Robotics and Applications*, *1*(1), 43–54. https://doi.org/10.1007/S41315-016-0006-2

Parasuraman, R. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *HUMAN FACTORS*, *39*(2), 230–253.

Peters, R. A., Campbell, C. L., Bluethmann, W. J., & Huber, E. (2003). Robonaut task

learning through teleoperation. *Proceedings - IEEE International Conference on Robotics and Automation*, *2*, 2806–2811. https://doi.org/10.1109/ROBOT.2003.1242017

Prendinger, H., & Ishizuka, M. (2004). *Introducing the Cast for Social Computing: Life-Like Characters*. 3–16. https://doi.org/10.1007/978-3-662-08373-4_1

Rau, P. L. P., Li, Y., & Li, D. (2010). A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, *2*(2), 175–186. https://doi.org/10.1007/S12369-010-0056-9/METRICS

Ravichandar, H., Polydoros, A. S., Chernova, S., & Billard, A. (2020). Annual Review of Control, Robotics, and Autonomous Systems Recent Advances in Robot Learning from Demonstration. *Annu. Rev. Control Robot. Auton. Syst. 2020*, *2020*, 297–330. https://doi.org/10.1146/annurev-control-100819

Reason, J. (1990). *Human Error*. https://doi.org/10.1017/CBO9781139062367

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2008). How anthropomorphism affects empathy toward robots. *Proceedings of the 4th ACM/IEEE International Conference on Human-Robot Interaction, HRI'09*, 245–246. https://doi.org/10.1145/1514095.1514158

Robinette, P., Howard, A. M., & Wagner, A. R. (2015). Timing is key for robot trust repair. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9388 LNCS*, 574–583. https://doi.org/10.1007/978-3-319-25554-5_57/COVER

Ross, J. M., Szalma, J. L., Hancock, P. A., Barnett, J. S., & Taylor, G. (2008). *The Effect of Automation Reliability on User Automation Trust and Reliance in a Search-and-Rescue Scenario*. https://doi.org/10.1518/107118108X353444

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of Personality*, *35*(4), 651–665. https://doi.org/10.1111/J.1467-6494.1967.TB01454.X

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not So Different After All: A Cross-Discipline View Of Trust. *Https://Doi.Org/10.5465/Amr.1998.926617*, *23*(3), 393–404.

https://doi.org/10.5465/AMR.1998.926617

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015a). Towards safe and trustworthy social robots: Ethical challenges and practical issues. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9388 LNCS*, 584–593. https://doi.org/10.1007/978-3-319-25554-5_58/COVER

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015b). Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. *ACM/IEEE International Conference on Human-Robot Interaction*, *2015-March*, 141–148. https://doi.org/10.1145/2696454.2696497

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "trust perception scale-HRI." In *Robust Intelligence and Trust in Autonomous Systems* (pp. 191–218). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_10

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012). *Classification of Robot Form: Factors Predicting Perceived Trustworthiness*. https://doi.org/10.1177/1071181312561308

Schulman, J., Ho, J., Lee, A., Awwal, I., Bradlow, H., & Abbeel, P. (2013). Finding Locally Optimal, Collision-Free Trajectories with Sequential Convex Optimization. *Robotics: Science and Systems IX*, 1–10. https://doi.org/10.15607/RSS.2013.IX.031

Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19. https://doi.org/10.1016/J.OBHDP.2006.05.005

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). "I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair. *ACM/IEEE International Conference on Human-Robot Interaction*, *2019-March*, 57–65. https://doi.org/10.1109/HRI.2019.8673169

Sigal, J., Hsu, L., Foodim, S., & Betman, J. (1988). Factors Affecting Perceptions of

Political Candidates Accused of Sexual and Financial Misconduct. *Political Psychology*, *9*(2), 273. https://doi.org/10.2307/3790956

Skowronski, J. J., & Carlston, D. E. (1987). Social Judgment and Social Memory: The Role of Cue Diagnosticity in Negativity, Positivity, and Extremity Biases. *Journal of Personality and Social Psychology*, *52*(4), 689–699. https://doi.org/10.1037/0022-3514.52.4.689

Ullman, D., Malle Cognitive, B. F., & Sciences, P. (n.d.). Human-Robot Trust: Just a Button Press Away. *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3029798

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in smart systems: sharing driving goals and giving information to increase trustworthiness and acceptability of smart systems in cars. *Human Factors*, *54*(5), 799–810. https://doi.org/10.1177/0018720812443825

Vigni, F., Rossi, A., Miccio, L., & Rossi, S. (2022). On the Emotional Transparency of a Non-humanoid Social Robot. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13817 LNAI*, 290–299. https://doi.org/10.1007/978-3-031-24667-8_26/FIGURES/4

Vogt, D., Stepputtis, S., Grehl, S., Jung, B., & Ben Amor, H. (2017). A system for learning continuous human-robot interactions from human-human demonstrations. *Proceedings - IEEE International Conference on Robotics and Automation*, 2882–2889. https://doi.org/10.1109/ICRA.2017.7989334

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Transparent, explainable, and accountable AI for robotics. *Science Robotics*, *2*(6). https://doi.org/10.1126/SCIROBOTICS.AAN6080

Wang, H., Chen, J., Lau, H. Y. K., & Ren, H. (2016). Motion Planning Based on Learning from Demonstration for Multiple-Segment Flexible Soft Robots Actuated by Electroactive Polymers. *IEEE Robotics and Automation Letters*, *1*(1), 391–398. https://doi.org/10.1109/LRA.2016.2521384

Wang, J., Liu, Y., Yue, T., Wang, C., Mao, J., Wang, Y., & You, F. (2021). Robot Transparency and Anthropomorphic Attribute Effects on Human–Robot Interactions. *Sensors 2021, Vol. 21, Page 5722*, *21*(17), 5722. https://doi.org/10.3390/S21175722

Wang, N., Pynadath, D. V., Rovira, E., Barnes, M. J., & Hill, S. G. (2018). Is it my looks? Or something i said? The impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10809 LNCS*, 56–69. https://doi.org/10.1007/978-3-319-78978-1_5/FIGURES/3

WHEELESS, L. R., & GROTZ, J. (1977). THE MEASUREMENT OF TRUST AND ITS RELATIONSHIP TO SELF-DISCLOSURE. *Human Communication Research*, *3*(3), 250–257. https://doi.org/10.1111/J.1468-2958.1977.TB00523.X

Wildman, J. (2011). Cultural Differences In Forgiveness Fatalism, Trust Violations, And Trust Repair Efforts In Interpersonal Collaboration. *Electronic Theses and Dissertations*. https://stars.library.ucf.edu/etd/1728

Zhu, Z., & Hu, H. (2018). Robot Learning from Demonstration in Robotic Assembly: A Survey. *Robotics 2018, Vol. 7, Page 17*, *7*(2), 17. https://doi.org/10.3390/ROBOTICS7020017

Appendix A Informed consent form

Information form for participants

This document gives you information about the study “Trust repair in an HRI scenario of a food processing factory”. Before the study begins, it is important that you learn about the procedure followed in this study and that you give your informed consent for voluntary participation. Please read this document carefully.

Aim and benefit of the study

The aim of this survey is to investigate which combination of trust repair strategies of the robot is the most effective after trust violation in learning from demonstration process. The results of this survey will give useful insights into rebuilding trust in human-robot collaboration scenarios.

This study is performed by R. Zhang, students under supervision of Chao Zhang and Margot Neggers of the Human-Technology Interaction group.

Procedure

You do this experiment using your web browser on your own laptop. The online experiment is performed by software Limesurvey. During the experiment, it is important to remain seated and minimize distractions from your surroundings. You will only start the experiment if you consent. Once you start, you will firstly be asked to imagine yourself as a manager of a food processing factory and inspect the robots working. Then, depending on the experimental conditions, you will be assigned to one of them and watch the corresponding videos. After viewing each video, you will be asked to answer a series of questions about your trust perception. Then, you will be asked to answer some general questions about the experiment. Finally, you will be asked to provide some demographic information and your Prolific ID as the only personal data.

Risks

The study does not involve any risks, detrimental side effects, or cause discomfort.

Duration

The experiment will take approximately 10 minutes, including reading instructions, watching videos, and filling out questions.

Participants

You were selected because you are registered as a participant in the participant database of the online platform Prolific and you are aged above 18 with fluent English proficiency.

Voluntary

Your participation is completely voluntary. You can refuse to participate without giving any reasons and you can stop your participation at any time during the study. You can also withdraw your permission to use your data up to 24 hours after they were collected. None of this will have any negative consequences for you whatsoever.

Compensation

You will be paid £ 2 for your participation.

Confidentiality and use, storage, and sharing of data.

All research conducted at the Human-Technology Interaction Group adheres to the Code of Ethics of the NIP (Nederlands Instituut voor Psychologen – Dutch Institute for Psychologists), and this study has been approved by the Ethical Review Board of the department.

In this study demographic data (age, gender, occupation, expertise level of robot) and experimental data (responses to questionnaires) will be recorded, analyzed, and stored. Besides, your Prolific account ID will be collected to enable filtering good quality responses. This information will be stored in an encrypted environment and will be deleted after payment. The goal of collecting, analyzing, and storing this data is to answer the research question and publish the results in the scientific literature. To protect your privacy, all data that can be used to personally identify you will be stored on an encrypted server of the Human-Technology Interaction group for at least 10 years that is only accessible by selected HTI staff members. No information that can be used to personally identify you will be shared with others.

The data collected in this study might also be of relevance for future research projects within the Human Technology Interaction group as well as for other researchers. The aim of those studies might be unrelated to the goals of this study. The collected data will therefore also be made available to the general public in an online data repository. The coded data collected in this study and that will be released to the public will (to the best of our knowledge and ability) not contain information that can identify you. It will include all answers you provide during the study, including demographic variables (e.g., age and gender) if you choose to provide these during the study.

At the bottom of this consent form, you can indicate whether or not you agree with the use of your data for future research within the Human Technology Interaction and the distribution of your data by means of a secured online data repository with open access for the general public. You are not obliged to let us use and share your data. However, you must give your consent to share your data in this way in order to participate in this study. If you do not give your consent, you cannot participate in this study.

No video or audio recordings are made that could identify you.

Further information

If you want more information about this study, the study design, or the results, you can contact Ruohan Zhang (r.zhang@student.tue.nl).

If you have any complaints about this study, please contact the supervisor, Chao Zhang (C.Zhang.5@tue.nl). You can report irregularities related to scientific integrity to confidential advisors of the TU/e.

Informed consent form

Responses to a working industrial robot

- I have read and understood the information of the corresponding information form for participants.

- I have been given the opportunity to ask questions. My questions are sufficiently answered, and I had sufficient time to decide whether I participate.

- I know that my participation is completely voluntary. I know that I can refuse to participate and that I can stop my participation at any time during the study, without giving any reasons. I know that I can withdraw permission to use my data up to 24 hours after the data have been recorded.

- I agree to voluntarily participate in this study carried out by the research group Human Technology Interaction of the Eindhoven University of Technology.

- I know that no information that can be used to personally identify me or my responses in this study will be shared with anyone outside of the research team.

- □ I want and provide consent to participate in this study.

- □ I give permission to make my anonymized recorded data available to others in a public online data repository, and allow others to use this data for future research projects unrelated to this study.

Appendix B Distribution of items in three trust measurement

Table 6

The distribution of all the fifteen items in trust measurement 1

Please indicate to what extent you felt this robot will be …				
Number	Item	Mean	SD	Sign
1	Dependable	5.05	1.44	+
2	Reliable	5.29	1.24	+
3	Unresponsive	2.29	1.41	-
4	Predictable	5.6	1.14	+
5	Act consistently	5.73	1.10	+
6	Have errors	2.57	1.37	-
7	Provide feedback	2.73	1.80	+
8	Meet the needs of the mission/task	5.62	1.22	+
9	Provide appropriate information	4.12	1.91	+
10	Communicate with people	2.12	1.51	+
11	Perform exactly as instructed	5.76	1.18	+
12	Follow directions	5.86	1.14	+
13	Incompetent	1.91	1.16	-
14	A good teammate	3.53	1.96	+
15	Perform a task better than a novice human user	2.85	1.80	+

Table 7

The distribution of all the fifteen items in trust measurement 2

Please indicate to what extent you felt this robot will be …				
Number	Item	Mean	SD	Sign
1	Dependable	5.00	1.44	+
2	Reliable	5.21	1.31	+
3	Unresponsive	2.34	1.49	-
4	Predictable	5.43	1.26	+
5	Act consistently	5.46	1.34	+
6	Have errors	2.55	1.42	-
7	Provide feedback	2.64	1.71	+
8	Meet the needs of the mission/task	5.48	1.31	+
9	Provide appropriate information	3.71	1.97	+
10	Communicate with people	2.16	1.55	+
11	Perform exactly as instructed	5.64	1.26	+
12	Follow directions	5.71	1.27	+
13	Incompetent	2.08	1.38	-
14	A good teammate	3.32	1.96	+
15	Perform a task better than a novice human user	2.86	1.80	+

Table 8

The distribution of all the fifteen items in trust measurement 3

Please indicate to what extent you felt this robot will be …				
Number	Item	Mean	SD	Sign
1	Dependable	3.61	1.69	+
2	Reliable	3.38	1.56	+
3	Unresponsive	2.48	1.44	-
4	Predictable	3.95	1.57	+
5	Act consistently	3.89	1.71	+
6	Have errors	4.33	1.65	-
7	Provide feedback	3.72	1.97	+
8	Meet the needs of the mission/task	3.62	1.63	+
9	Provide appropriate information	3.7	1.81	+
10	Communicate with people	3.37	1.89	+
11	Perform exactly as instructed	3.50	1.73	+
12	Follow directions	3.92	1.74	+
13	Incompetent	3.19	1.69	-
14	A good teammate	2.75	1.63	+
15	Perform a task better than a novice human user	2.03	1.35	+