



DEGREE PROJECT IN TECHNOLOGY,  
FIRST CYCLE, 15 CREDITS  
*STOCKHOLM, SWEDEN 2017*

# **A comparison of three robot recovery strategies to minimize the negative impact of failure in social HRI**

**SARA ENGELHARDT**

**EMMELI HANSSON**

# **A comparison of three robot recovery strategies to minimize the negative impact of failure in social HRI**

SARA ENGELHARDT, EMMELI HANSSON

Date: June 5, 2017

Supervisor: Iolanda Leite

Examiner: Örjan Ekeberg

Swedish title: En jämförelse mellan tre robotåterhämtningsstrategier för att minimera negativa effekter av misslyckanden i sociala människa-robot interaktioner

School of Computer Science, KTH



## Abstract

Failure happens in most social interactions, possibly even more so in interactions between a robot and a human. This paper investigates different failure recovery strategies that robots can employ to minimize the negative effect on people's perception of the robot. A between-subject Wizard-of-Oz experiment with 33 participants was conducted in a scenario where a robot and a human play a collaborative game. The interaction was mainly speech-based and controlled failures were introduced at specific moments. Three types of recovery strategies were investigated, one in each experimental condition: ignore (the robot ignores that a failure has occurred and moves on with the task), apology (the robot apologizes for failing and moves on) and problem-solving (the robot tries to solve the problem with the help of the human). Our results show that the apology strategy scored the lowest on measures such as likeability and perceived intelligence, and that the ignore strategy lead to better perceptions of perceived intelligence and animacy than the employed recovery strategies. In conclusion, problem-solving clearly minimized the negative effects of failure better than apology, but no recovery, the ignore condition, often scored at least as well as problem-solving.



## Sammanfattning

De flesta sociala interaktioner misslyckas ibland, kanske oftare för interaktioner mellan en robot och en människa. Denna rapport undersöker olika återhämtningsstrategier som robotar kan använda för att minimera de negativa effekterna på människors uppfattning av roboten. Ett Wizard-of-Oz-experiment med 33 deltagare utfördes där en robot och en människa samarbetade för att spela ett spel. Interaktionen var främst tal-baserad och kontrollerade misslyckanden introducerades vid givna tillfällen. Tre olika återhämtningsstrategier testades, på varsin grupp deltagare. Strategierna är: ignorera (roboten ignorerar att ett misslyckande skett och fortsätter med uppgiften), ursäkt (roboten ber om ursäkt för att den misslyckats och fortsätter sedan med uppgiften) och problemlösning (roboten försöker lösa problemet med hjälp av människan). Våra resultat visar att ursäktsstrategin fick lägst resultat på bland annat upplevd intelligens och sympati, och ignorerastrategin ledde till högre resultat på upplevd intelligens och animacitet än de använda återhämtningsstrategierna. Sammanfattningsvis så minskade problemlösning de negativa effekterna av misslyckanden betydligt bättre än ursäktsstrategin, men ingen strategi, ignorera, var ofta minst lika bra som problemlösning.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Purpose . . . . .	1
1.2	Problem Statement . . . . .	2
1.3	Scope . . . . .	2
1.4	Outline . . . . .	2
<b>2</b>	<b>Background</b>	<b>3</b>
2.1	Terminology . . . . .	3
2.2	Social Robot . . . . .	4
2.3	HRI Interaction . . . . .	4
2.4	Social failure in HRI . . . . .	5
2.5	Related work . . . . .	5
2.6	Strategies of recovery . . . . .	10
<b>3</b>	<b>Methods</b>	<b>12</b>
3.1	Wizard of Oz (WoZ) . . . . .	12
3.2	Experiment . . . . .	13
3.2.1	Protocols . . . . .	14
3.2.2	Nao . . . . .	14
3.2.3	Programming . . . . .	15
3.3	Survey . . . . .	17
3.4	Test Subjects . . . . .	18
3.5	WoZ Guidelines . . . . .	20
<b>4</b>	<b>Results</b>	<b>25</b>
4.1	Godspeed . . . . .	25
4.1.1	Likeability . . . . .	25
4.1.2	Perceived Intelligence . . . . .	26
4.1.3	Animacy . . . . .	27
4.2	RoSAS . . . . .	28

4.2.1	Competence . . . . .	28
4.2.2	Discomfort . . . . .	28
4.3	Experiment length . . . . .	29
4.4	WoZ guidelines . . . . .	29
<b>5</b>	<b>Discussion</b>	<b>30</b>
5.1	Compare strategies . . . . .	30
5.1.1	Influence of Homogeneous Participant Pool . . .	30
5.1.2	Influence of Experiment Design . . . . .	31
5.1.3	Sources of Errors . . . . .	32
5.1.4	Ethical Issues with the Method . . . . .	33
5.2	Limitations . . . . .	33
5.3	Future Research . . . . .	34
<b>6</b>	<b>Conclusion</b>	<b>35</b>
	<b>Bibliography</b>	<b>36</b>
<b>A</b>	<b>Experiment Description</b>	<b>38</b>
<b>B</b>	<b>Experiment Protocol</b>	<b>41</b>
B.1	Experiment Protocol Template: . . . . .	41
B.2	Fail-recovery: Ignore . . . . .	42
B.3	Fail-recovery: Apology . . . . .	43
B.4	Fail-recovery: Problem-solving . . . . .	43
<b>C</b>	<b>Questionnaire</b>	<b>44</b>

# Chapter 1

## Introduction

Robots and artificial intelligence are rapidly being integrated into our everyday life. Virtual assistants like Google Assistant, Siri (for Apple's products), Cortana (Windows' products) and Alexa (Amazon's) are just a few that are available in many homes and phones today. Robots are also used to help children with autism as well as help in the care of elderly. As they become more common and continue to be developed, we need to learn how to interact with them, or teach them how to interact with us, especially in the failure cases. But robots are machines, so how do you interact with them? Can you interact with them the way you interact with other people? Do people already have preconceived notions about how to interact with robots? When the Human Robot Interaction (HRI) fails, what do we do?

### 1.1 Purpose

The purpose of this report is to study robot failure, specifically, how people perceive robots depending on how the robot acts when it fails. There are many different strategies for the robot to follow. There are also many types of failures. We will focus on social failure and recovery, where the robot has a conversation with the human and needs to recover from a failure in that kind of social situation.

## 1.2 Problem Statement

The question we seek to answer is the following one: Which is the best strategy that robots can use to minimize negative impact of failure in social interactions with a human?

To study this we need to select what strategies we intend to study. We will then want to compare these strategies to be able to determine how the robot was perceived by people and the influence the conditions will have. We will test the following three strategies: apology, problem-solving and ignore. More information about the strategies can be found in section 2.6.

## 1.3 Scope

The focus we have chosen is to see what the robot can do to lessen the negative impact of failure on the interaction. This means that we will not look at the detection of social failure, but rather the appropriate reaction to when this happens. Consequently, we will devise situations where the robot will purposefully fail, and we will test different recovery strategies to see how well they work.

## 1.4 Outline

First we will look at earlier related work in section 2, the Background, to see what strategies have been tested for robot failure and what results they have generated. This will be followed by the Methods, in section 3, where we describe what we will do to answer our question, as well as how. Then the Results, section 4, generated by our experiments will be described. Following that, we will have the Discussion, section 4, where we discuss the results, the limitations we had and the future research we suggest.

# Chapter 2

## Background

### 2.1 Terminology

The following are a few key concepts. Godspeed, RoSAS and WoZ will be further explained in section 3, Methods.

**Between-subject experiment/study:** A study in which each participant only tests one condition.

**Human-Robot Interaction (HRI):** The field of interactions between humans and robots.

**Godspeed:** A questionnaire series to measure people's perception of robots.

**Robotics Social Attributes Scale (RoSAS):** Another questionnaire that measure people's perception of robots. Builds on the Godspeed Questionnaire Series.

**Wizard-of-Oz (WoZ):** An experiment technique where the robot's autonomy is simulated by a human, without the participants' knowledge.

## 2.2 Social Robot

There is more than one way to define a social robot according to Dautenhahn, and there can be different levels of social intelligence [4]. We are not interested in all of these ways, since they will not all be relevant to the kind of social robot that we intend to study. Furthermore, we will only simulate some of the social reactions and awareness of the robot, since we will not study the social skills of the robot itself, but rather the consequences of a certain part of those social skills. Therefore we will not have deep social intelligence in our robot, and the definition of a social robot from Dautenhahn that interests us is the socially situated robot. This kind of robot can see its environment and can react to it, as well as see the difference between objects and social agents [4]. These are the kind of social skills we want to give the impression that our robot has. Like the playmate for autistic children that Dautenhahn described [4], we will use a simple robot in a limited situation where we will make it appear social.

Wizard-of-Oz (WoZ) studies are often used when studying what social skills robots need in the future to be autonomous and socially intelligent [15]. The WoZ approach simulates the robots autonomy in a social situation when the existing or available technology isn't enough and this is the approach we will use in our study also.

## 2.3 HRI Interaction

There are many ways to interact with robots, and they can be vastly different, from physical to social interactions. Therefore we need to define what interaction we intend to study. Since we want to study the impact of robot failure in interactions with humans, it is some variation of a social interaction that we will study. More specifically we will study human-centered HRI, since, based on the definition given by Dautenhahn, it is the interaction where the acceptance and comfort of the human in the interaction is the focus [4]. The consequences of robot failure in HRI are directly connected to how the human perceives the failure, that is, how the human feels about the robot following the failure, and whether the human feels comfortable with the robot or not afterwards.

Human-centered HRI is still too broad a concept to study, so we still

need to limit it further to have a feasible study. There are both non-verbal and verbal social interactions. Non-verbal interactions can be expressions and movements. We do not have the resources or time to define an experiment with which to study this. In a verbal interaction we can make the robot verbally apologize or ask for help, and the robot we intend to use, NAO, has the capacity to speak [11]. In a verbal interaction we can also study the consequences if the robot does not speak about or acknowledge its failure.

We also intend to have an interaction where the robot and the human interacting have a common goal and need to collaborate to reach this goal, to make the interaction more interesting and the failure in the social interaction more obvious.

## 2.4 Social failure in HRI

Interactions are not always successful. Humans have developed rules for how to interact with each other. These rules are called social norms. Sunstein defined social norms as “social attitudes of approval and disapproval, specifying what ought to be done and what ought not to be done” [13]. Because even interactions between humans fail at times, it is not surprising that Human-Robot interactions fail as well.

Giuliani et al found two types of failures in HRI: social norm violations and technical failures. A violation of a social norm is defined by Giuliani et al as “a deviation from the social script or the usage of the wrong social signals”. These failures were often due to planning failures; actions that are executed correctly, but inappropriate for the situation. An example of a planning failure would be the robot asking the user the same question several times even though an appropriate answer has been given. An example of use of inappropriate social signals is the robot not looking at the person it is talking to. Technical failures were often a result of execution failures, meaning an appropriate action was carried out, but done so incorrectly. This report will focus on social norm failures, limited to verbal failures. [8]

## 2.5 Related work

HRI is a big field of study. This section will introduce a few studies made with social HRI. Pictures of all the robots used can be found in



Table 2.1.

*Modeling robotic behaviour to mitigate malfunctions with the help of the user:* This study, by Bajones, Weiss, and Vincze [2], meant to offer some insight into recovery strategies for robot failure, more specifically, whom a robot should ask for help when it malfunctions. Using a Wizard-of-Oz experiment the researchers had 19 pairs of participants interact with a robot called HOBbit. [2]

The participants were separated by a screen and asked to build a lego model showed by the robot. Two conditions were tested. In the first, the participants had different roles; one builder and one director. In the second, the participants had the same role, but still had to collaborate to finish the task. During the second of three building tasks, the robot malfunctioned repeatedly. All of the malfunctions were navigational. When the robot malfunctioned it stated in a short manner what the problem was (for example “I’m stuck”), and how the participants could help it recover. [2]

The study showed that the person most likely to help the robot was the person who gave it its last command, followed by the person closest to it. In between each new lego model, the participants filled out a task contribution questionnaire, the perceived intelligence and likability scales from Godspeed, as well as three open-ended questions. The result of the questionnaires showed a tendency for a negative impact on perceived intelligence, likability and robot contribution when the robot malfunctioned. However the negative impact was small, which the researchers attributed to the robot’s recovery strategies, that made it able to fulfill its tasks in the end. The researchers also noted that while helping the robot made the task more engaging at first, the repeated demands for help soon became an annoyance for the participants. [2]

*How a robot should give advice:*

A study made by Torrey, Fussell, and Kiesler [14] aims to show that using hedges and discourse markers will help robots be perceived positively when offering advice. Giving advice is believed to threaten the autonomy of the person receiving the advice. Building on politeness theory, the researchers used informal speech and hedges to mitigate the “face-threatening” aspects of giving advice or orders. The experiment was divided into four communication conditions: discourse

markers, hedges, both and neither. The researchers hypothesized that both hedges and discourse markers would have a positive effect on the interaction, and that both of them combined would result in a stronger positive outcome. [14]

The 77 participants viewed four videos each of a person trying to bake cupcakes. Each time the baker had a helper, which was either a human or a robot. The robot was digitally spliced over the human in the videos, and no actual human-robot interaction took place during the experiment. Each participant saw all four communication conditions, where two had a human helper and two had a robot helper. After each video the participants were given statements about the interaction and asked to rate their agreement with them. The statements measured how considerate, controlling and likable the participants perceived the helper to be. [14]

All three factors were improved by both hedges and discourse markers. However, combining them had no additional effect. The use of discourse markers was more effective in reducing the perception of the helper as controlling for a robot helper than a human helper. This indicates that robots using politeness might have an even bigger positive effect on the interaction than humans doing the same. [14]

*Dynamic multi-party social interaction with a robot agent:*

In a study done in 2012 by Foster et al. [5], a robot (JAMES) is used as a bartender. Three scenarios were tested. In the first, the participant approaches the robot bartender alone. In the second, another person stands by the bar during the interaction, but does not attempt to interact with neither the bartender nor the participant. In the third scenario, the participant approaches the bartender together with another person. [5]

The study used both objective and subjective measures. The objective measures consisted of task success, dialogue quality and dialogue efficiency. The subjective measures consisted of ratings for each interaction on a scale from 1-10, and the five GODSPEED questionnaires. 31 people participated in the study, of which 22 were male. [5]

The objective measures showed that the robot was generally successful, but dialogue efficiency and quality could be improved. Dialogue efficiency and task success had the biggest impact on the subjective measures, which showed a generally positive result for the first two interactions, as well as perceived intelligence and likeability. [5]

*Comparing task-based and socially intelligent behaviour in a robot bartender:* This study from 2013, by Giuliani et al. [7], focuses on the effect of appropriate social behaviour in a human-robot interaction. The same team and the same robot were used as in the previous study. Using a between-participants design, two interaction styles were tested: a task based style and a more socially intelligent style. Half of the 40 participants interacted with each version. In the task based design the interaction was limited to the bartender asking customers for orders and serving the given drink orders to the customers. In the socially intelligent design, the robot has a more sophisticated behaviour, such as serving the drinks in the order that the orders were given, and acknowledging new customers with a nod, but finishing the current transaction before approaching the new customer. [7]

Again task success, dialogue quality and dialogue efficiency, was used for the objective measures. The subjective measures were collected with the GODSPEED questionnaire series this time as well. The participants filled out the questionnaires both before and after the interaction, to test for user expectations. [7]

The socially intelligent robot resulted in slightly smoother interactions. However, the difference between the two interaction styles did not affect the subjective ratings much. The overall length of the interaction positively affected the ratings. The results showed both a cultural difference, and a difference between genders. Females were served slightly slower and the interaction was longer. This difference is believed to partly be a result of the face recognition software mainly being trained on males. The cultural difference consisted of higher pre-test scores for participants that chose English over German as the interaction language. The researchers hypothesize that the difference is due to a cultural difference in attitudes. They also hypothesize that the participants that chose German were native German speakers, while the ones choosing English were mostly international students and not native English speakers. [7]

*Gracefully mitigating breakdowns in robotic services:*

Lee et al. [9] studied four different strategies for mitigating robot failures. One was an expectancy-setting strategy, where the participants were forewarned that the robot might experience difficulties. The other three were recovery strategies. The recovery strategies included in this

study were: apology, compensation and options. [9]

317 people participated in the online between-subject scenario survey. Each participant first viewed a short video of one of the two robots used in the experiment, the Snackbot robot, or the HERB robot. Then a questionnaire meant to measure the participants' evaluations of a service provider was filled out. This was followed by one of the 18 different scenarios of the human-robot interaction. Assignment to a scenario was randomized. [9]

The interaction consisted of a human asking the robot to get a drink, followed by the robot getting the right drink (two success control scenarios), or getting the wrong one (16 scenarios). The human notes that it is the wrong drink in all the fail scenarios, but the robot's reaction varies depending on recovery strategy. In the fail control scenario the robot's only response is "OK". In all the other scenarios the robot first explains that it didn't realize it had made a mistake, and then proceeds to the given strategy (apology, compensation, options). All scenarios but the success control were tested both with and without forewarning, where the participant's expectations are managed by the robot saying that it might have trouble with the task. Every scenario was tested with both robots. After the scenario, the participants were asked to fill out the questionnaire again. [9]

The robot failure decreased all ratings of the robot compared to the successful interaction, except how much they liked the robot. The forewarning strategy improved the evaluations of the robot, but did not improve the judgment of the service greatly. All of the recovery strategies increased the ratings of the politeness of the robot. To increase the perception of customer service satisfaction, compensation worked best. However, the other two strategies increased the perceived likelihood that the customer would return more. Overall, the apology strategy scored best, especially among people that scored high on relational orientation. On the contrary, people with low relational orientation or high utilitarian orientation liked compensation best, and actually preferred no recovery strategy over both apology and options. This suggests that recovery strategies need to be tailored to a person's orientation to services. Different scenarios might also benefit from different strategies depending on whether customer satisfaction or willingness to return is the most important factor. [9]

## 2.6 Strategies of recovery





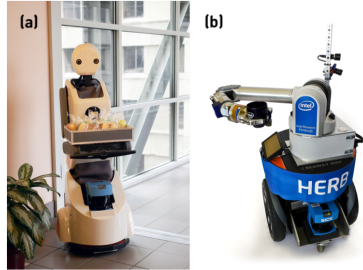
Lee et al. write about different recovery strategies for when the robot fails [9]. Among other strategies, they tried apologies and giving the human options for ways to help the robot correct its mistake. Both of these strategies yielded positive results, and increased the likelihood that participants believed the customer would want to use the service again [9]. They also had a control version where the robot ignores the failure [9]. Politeness was also found to be important in human-robot interactions by Torrey, Fussell, and Kiesler [14].

The three strategies of mitigating the negative impact of failure in social contexts that we want to study are loosely based on the strategies mentioned previously, and are as follows:

- Ignore: The robot ignores that it has failed, and just keeps going (the strategy to control against)
- Apology: The robot apologizes for its failure and then keeps going
- Problem-solving: The robot tries to solve its failure with the help of the human

In summary, our apology condition is just like the one Lee et al. write about, our problem-solving condition is based on their option condition, and our ignore conditions is just like their failure control condition.

Table 2.1: Previous work in the field and the robots used.

Previous work	Robot used
Markus Bajones, Astrid Weiss, and Markus Vincze. “Help, Anyone? A User Study For Modeling Robotic Behavior To Mitigate Malfunctions With The Help Of The User”. In: (2016) [2]	 [2]
Cristen Torrey, Susan Fussell, and Sara Kiesler. “How a robot should give advice”. In: <i>Proceedings of the 8th ACM/IEEE international conference on human-robot interaction</i> . HRI '13. IEEE Press, 2013, pp. 275–282. ISBN: 9781467330558 [14]	 [14]
Mary Ellen Foster et al. “Two people walk into a bar: dynamic multi-party social interaction with a robot agent”. eng. In: <i>Proceedings of the 14th ACM international conference on multimodal interaction</i> . ICMI '12. ACM, Oct. 2012, pp. 3–10. ISBN: 9781450314671 [5]	 [5]
Manuel Giuliani et al. “Comparing task-based and socially intelligent behaviour in a robot bartender”. In: <i>Proceedings of the 15th ACM on international conference on multimodal interaction</i> . ICMI '13. ACM, Dec. 2013, pp. 263–270. ISBN: 9781450321297 [7]	 [7]
Min Kyung Lee et al. “Gracefully mitigating breakdowns in robotic services”. In: <i>Proceedings of the 5th ACM/IEEE international conference on human-robot interaction</i> . HRI '10. IEEE Press, 2010, pp. 203–210. ISBN: 9781424448937 [9]	 [9]

# Chapter 3

## Methods

### 3.1 Wizard of Oz (WoZ)

As stated in the background we will use WoZ to simulate much of the social aspects of our robot.

Wizard of Oz is an experimental technique often used in HRI research. It involves a person (the wizard) controlling the robot remotely. The level of control versus the level of autonomy of the robot can vary from experiment to experiment, and the purpose of using Wizard of Oz is to simulate how an HRI interaction could look like in the future [10].

However there are a few concerns with the implications of using HRI, that are lifted by L.D. Riek [10]. The main concerns center around the fact that a human is in charge of the behaviour, while the participants are lead to believe that it is the robot reacting on its own. This can lead to ethical issues because the participants are tricked, as well as difficulties for real robots to live up to the expectations set by these studies. To minimize the negative effects, Riek suggests limiting the wizard's freedom in reacting to the participant, for example by using specific scenarios. It might also help to reveal the setup to the participants after the experiment to minimize the misconceptions afterwards. [10]

To minimize these concerns and have a better WoZ method, Riek gives guidelines to follow [10]. They can be found in Table 3.2, in section 3.5. Every question in the guidelines might not be relevant or important to the experiment in question, but it is important to at least have reflected upon them [10]. We intend to follow these guidelines

to lessen the concerns previously mentioned and to have a WoZ study that is the best that it can be.

## 3.2 Experiment

The experiments will be testing social interaction failure in conversation. The failure is the robot interpreting human words wrong, through a game with cards where the robot needs help. Twelve cards will be placed on the table, face down, in spots labeled from A to L. The robot holds one card, specifically a queen of hearts in our experiment. Because of language barrier, where it might not come naturally to all participants to name the symbol on the card, we ask the test persons to only say the number on the card.

The goal of the game will be to find the other queens. However, the robot can't turn the cards himself and needs the humans help to turn the cards and say what card is where, in order to find the hidden queens. The failure will consist of the robot hearing the wrong card. To make it clear that the robot fails, the robot always repeats the card it just heard and asks the participant to confirm or deny. When the participant answers "no" to this question, different recovery strategies will be used depending on the experimental conditions. Each recovery strategy will have its own protocol of how to handle and recover from the failure (more details in section 3.2.1).

Foster et al. [5] and Bajones, Weiss, and Vincze [2] showed that task success can have a big impact on the perception of the robot. So, to prevent task success from affecting the results, the outcome of the game will always be the same; all queens are found, and the main task is successful. Also, to simplify the experiment and the results, and make sure the task is successful, the robot will only hear incorrectly in the case when the card is not a queen. Each participant will experience three failures, three queens found and heard correctly and three other cards that were heard correctly. More details about the experiment, such as the exact order and frequency of the fails, can be found in the Appendix A.

Additionally, the experiments for the three strategies need to be approximately the same length. This is based on the study by Guiliani et al., that found that the length of the interaction affected the ratings [7]. Further, the robot's speech cannot be too repetitive and run the



risk of boring the user, or making them uninterested and uninvested in the experiment. This may affect the results.

### 3.2.1 Protocols

Each strategy has its own way of handling the failure. In ignore, the robot always says "OK", and then keeps going. In apology, the robot apologizes (for example, by saying "I'm sorry, sometimes I don't interpret speech correctly") for its failure and then moves on. In problem-solving, the robot tries to solve its failure with the help of the human by asking him/her to repeat the card. After hearing the card one more time, the robot acknowledges that it understood which card the participant is referring to.

The exact lines for the robot to say are specified in protocols, one for each condition. These, as well as a description of the experiment can be found in the Appendix A and B.

### 3.2.2 Nao

NAO is a humanoid robot created by Aldebaran Robotics (now owned by Softbank robotics). It is 0.58 meters tall. NAO's abilities include speaking, recognizing speech, making gestures such as wave, and recognizing shapes and objects [11].



Figure 3.1: The Nao Robot

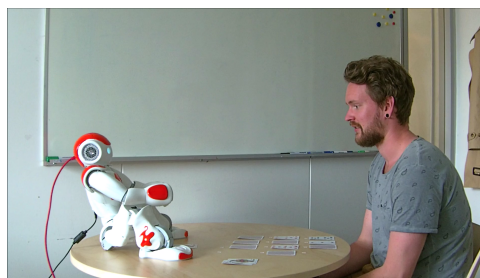


Figure 3.2: The experiment setup

The Nao is a common robot used in HRI studies, but the Nao's

speech recognition isn't always good enough and it is common to use WoZ when it is used for HRI studies [15].

### 3.2.3 Programming

For the experiment to be of the type Wizard of Oz, the robot's actions will be controlled by a so-called wizard. The level of control and the level of autonomy of the robot in a Wizard of Oz experiment can vary [10]. In our case we, as wizards, need to control when the Nao robot says what, because the Nao robot will not have the autonomy of recognizing speech and reacting to it. Also we have an exact protocol of what the robot will say in what order for each scenario. We want to simulate a social failure, and it needs to be the same failure for every test-subject. So, we don't want the robot to actually fail, since then we wouldn't have a controlled failure with a controlled failure recovery. The wizard enables us to give a controlled failure scenario.

To do this, we need the wizard to remotely tell the robot to say certain things, in real time. To communicate with the Nao robot in real time, we need an interface from which we can control it. The Nao robot has an API for Python which can be found in the Nao documentation.

The program itself<sup>1</sup> is separated into three parts. The first part is the Command Line Interface (CLI) file. The code here is inspired by the example code found in the python documentation [6]. This is where the commands that are written in the terminal by the wizard are handled.

The second part of the program is the scenario file, where all the lines of the protocol for each scenario are stored in three different arrays and the WhatToSay file, where the choice of the scenario and index of the array is handled. Depending on which scenario the wizard started, the appropriate array will be used.

The third part of the program is the Nao-interface part, the part of the program that is communicating with the Nao robot. Here, the commands from the CLI are "translated" to the actual action to be done, and a method that sends an action to the Nao robot is called. This method gets a line from the scenario file and sends that line to the robot to say. A counter is used to keep track of which line is the current one, which increases every time the wizard uses the command to say the next line.

---

<sup>1</sup>The code can be found at: <https://gits-15.sys.kth.se/emmelih/NaoCLI-KEX17>

What the wizard can do through the CLI is shown in Table 3.1.

Table 3.1: Commands that the wizard can do in the CLI.

Command	What it does	Purpose
start [scenario number]	Launches the first thing the robot says and has an argument that is the scenario number to be started (1, 2 or 3)	Start the experiment.
next	Makes the robot say the next line, by increasing the line counter by one, and getting the line at that index in the array.	Make the conversation go forward, simulating the robot having heard what the user says and responding to it.
rep	Makes the robot repeat the same line again.	If the user doesn't hear it the first time (user error).
prev	Makes the robot say the line before.	If the user hadn't heard it, or had misunderstood, or if the wizard made a mistake.
Jump	Increments the line counter by one, but without making the robot say the next line.	To jump ahead in the protocol, if user answers yes instead of no.
to [number]	Jumps to the number entered after "to"	If we need to jump ahead more than one line. Check printed protocol for what line number to enter.
panic "text"	An option for the robot to say something to solve an unexpected problem, but that doesn't interfere with the recovery strategy of the particular scenario being run. The line said will be the text entered as argument by the wizard.	To use in unexpected scenarios, if the user asks some unexpected question, if the wizard doesn't hear or understand what the user says, or makes some other mistake..

y or n	Makes the robot say yes or no.	To use when user asks unexpected questions that can be answered by yes or no.
bye	This ends the interaction by closing the CLI.	Ends the experiment and stops the timer.

There is also a counter, called “unexpected”, that counts the number of times the wizard needs to deviate from the protocol, so that we can see if that has an effect on the experiment results. Furthermore, we have a timer which starts when starting a scenario and ends when bye is typed in the CLI. With this information we can see how the time the participants spend doing the experiment varies and if it has an effect on the results.

Additionally, we had the time to add movements to the Nao robot to make the interaction seem more interesting. We added face tracking with the help of the example code from the Nao documentation, so that the Nao robot follows the face of the participant it is talking to [1]. We also added gestures while it is speaking, to make it seem more alive. This is done with a mode, called contextual, where the robot decides autonomously what gestures to do while speaking, depending on what it says. So these two things are not controlled by the wizard. Rather, they are autonomous actions of the robot.

### 3.3 Survey

To evaluate the different conditions we used a survey. The survey consists of four parts: some general questions about the participant, three of the Godspeed questionnaires, two parts of the RoSAS questionnaire, and finally a few open ended questions. The questionnaires were filled out on a computer, and the open ended questions were asked and recorded by the researchers.

The general questions concern age, gender, occupation and rating experience with programming and experience with robots on a five-point scale.

The Godspeed Questionnaire Series is an established way to measure people’s perception of robots. There are five questionnaires in total: Anthropomorphism, Animacy, Likeability, Perceived Intelligence

and Perceived Safety. We chose to only use likeability, animacy and perceived intelligence since they were the only ones that measured items relevant to this study. All Godspeed questionnaires consist of five-point semantic differentials, that is, scales from one word to its antonym, for example 1 represents dislike, and 5 represents like. The scales from the different questionnaires appeared in a mixed order.

The Robotics Social Attributes Scale (RoSAS) builds upon the Godspeed Questionnaire Series, but seeks to improve the cohesiveness of the measurements [3]. It consists of three parts, meant to measure: warmth, competence and discomfort. The questionnaire uses a 9-point likert scale with unidimensional items, for example 1 represents definitely not associated with interactive, 9 represents definitely associated with interactive. We chose to exclude the warmth-part because it seemed less relevant to this study. The scales from the different parts appeared in a mixed order in this part as well.

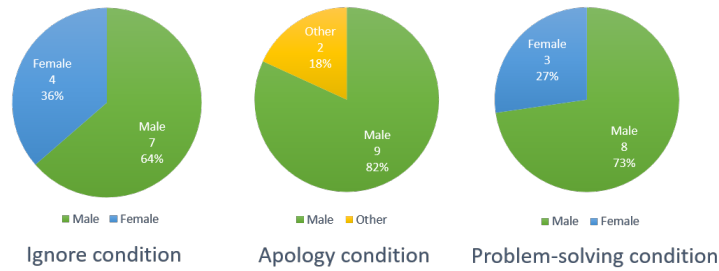
The open-ended questions are meant to check whether the participants realized that the robot failed and that it was a wizard-of-Oz-experiment, add additional information about their experience, and give some insight into their expectations for the robot. The questions were the following:

- Did you notice something wrong?
- How did you experience the robot?
- What is your opinion about the robot interaction as a whole?
- Did the robot live up to your expectations?
- Did you notice that we were controlling what the robot says?

### 3.4 Test Subjects

We had 33 participants in total and 11 participants per condition. The participants were drafted from the researchers' personal networks within the Bachelor's and Master's in Computer Science programs at KTH and the participants were all at least in their second year. There is a majority of male students studying this program at KTH, which is mirrored in the gender distribution of our participants.

Figure 3.3: Gender Distribution for the Different Conditions.



Additionally, since the number of participants for each group is rather small, the uneven distribution is more obvious and might have bigger effects.

The fact that they are all students from the computer science program means that they are not a representative group. It also means that they probably have a higher programming expertise than an average person, as well as a difference in perception of a robot such as Nao. People with programming background tend to have a different outlook on them compared to a naïv user Stadler, Weiss, and Tschelegi [12]. So, all participants have a programming background, but, they are in different years in the computer science program, and might have different programming expertise. The distribution of the programming expertise in the scenario is seen in diagram 2, which is information collected from the survey, where we asked the participants to rate their programming experience on a scale from 1 to 5, where 1 is “none”, and 5 is “experienced”. Overall the programming experience is high, see Figure 3.4.

We also asked the participants to rate their experience with robots from 1 to 5, where 1 is “none” and 5 is “extensive”. The distribution between scenarios of that expertise is seen in Figure 3.5. Overall, the experience with robots is rather low.

The median age of the participants was 22, with a maximum of 27 and minimum of 20. So our age group is also young. The age distribution between the different scenarios can be seen in Figure 3.6.

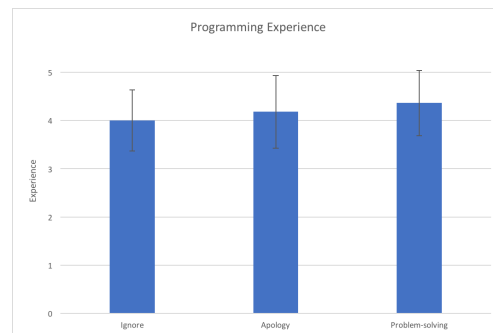


Figure 3.4: Programming Experience.

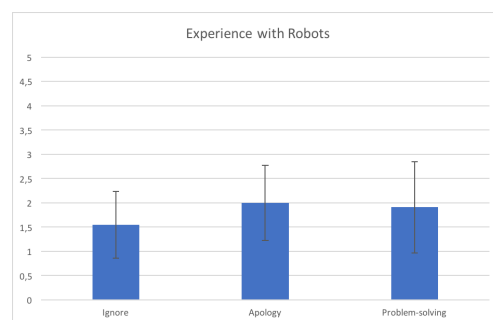


Figure 3.5: Experience with Robots.

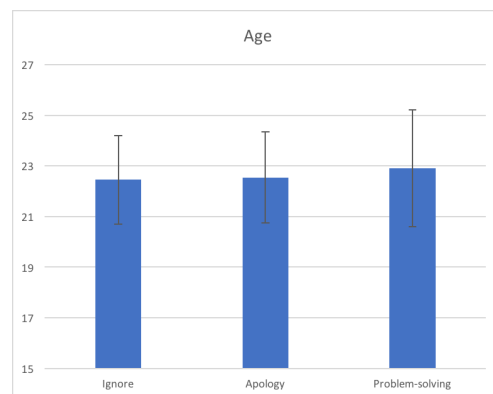


Figure 3.6: Age.

### 3.5 WoZ Guidelines

The WoZ guidelines referenced in 3.1 and our answers to each one can be found in Table 3.2.

Table 3.2: WoZ Guidelines.

Robot:	
How many robots were used?	1
What kind(s) of robot(s)? (e.g, humanoid, zoomorphic, mechanical, android?)	Humanoid, Nao
What level(s) of autonomy? (i.e., which components of the robot(s) were autonomous and which were controlled by the wizard?)	Speech is controlled by the wizard. Movement (of head and arms) is autonomous.
What were the robot's capabilities?	Talking, recognizing speech, holding a conversation, following a person's face, moving its arms.
What hypotheses did the researcher have for the robot?	none
User:	
How many users participated in total, and per experimental trial?	33 in total, one person per experiment, 11 per condition tested.
What were the user demographics, sampling procedure, etc.?	The users were all computer science students at KTH, from the researchers' personal network. They are in year 2 and higher in the Computer Science Program.
What instructions were provided to the user?	From robot protocol: "Hi! I am Nao and I have a queen of hearts card. I need your help to find the other three queens, because I can't lift these cards myself and I can't read very well. But I can hear you. Would you please look at card K and tell me what card it is? (You don't need to say the symbol or color, only the number)" And then, after a short exchange: "You can leave the card there, facing up or facing down, as you like."
What behavioral hypothesis does the researcher have about the user?	We have expected answers from the user in each part of the protocol.



Was the simulation convincing to the user?	See section 4.4
What expectations did the user have about the robot, before and after the experiment?	We tried to tell as little as possible about the experiment when recruiting the participants, in order not to bias them in any way. The information given was simply that they would interact with a robot and that we were testing HRI. See section 4.4 for their expectations in answer to the open ended questions.
Wizard:	
How many wizards were used?	1
What were the wizard demographics? (e.g., the researcher, lab mates, naïve?)	The researchers, who wrote the interface used (us)
Did the wizard know the behavioural hypothesis of the experiment?	Yes (since they are us)
What were the wizard recognition variables and how were they controlled for?	We have the expected input from the protocols, as well as the possible unexpected situations stated in Table 3.3.
How did the experimenter control for wizard error (deliberate and accidental)?	We had possible wizard error and solutions, found in Table 3.3.
How much and what sort of training did wizards receive prior to starting the experiment?	The wizards are the researchers who developed the experiment and the interface used, so they knew what to expect and exactly how it worked. There was also a few (4) test runs of the experiment on other users.
General:	
Where did the experiment take place?	In a meeting room. One meeting room for the majority (32) of the experiments, and another similar meeting room for a minority (1) of the experiments.

What were the environmental constants and how were they controlled?	The door is closed during the experiment, and the experiment is done the same way every time.
What scenarios did the researchers employ?	The three failure recovery conditions, see section 3.2.1 Protocols and section 2.6 Strategies.
Was this experiment part of an iterative design process?	No (we will not develop a robot capable of this failure recovery as a part of this project)
Does this paper discuss the limitations of WoZ?	yes, see section 4.4, for the answers to the open ended questions concerning this, and section 5.1.4.

Table 3.3, lists the possible mistakes or unexpected situations that can be made by either the wizard or the user, as well as the action the wizard should to take to resolve the situation. The term “unexpected” is simply a reference to the fact that it is a deviation from the protocol.

Table 3.3: Possible Wizard and User mistakes.

Mistake	By who?	Solution	Comment
User doesn't hear what robot says	User	repeat	
User misunderstands what robot says	User	repeat, if the mistake matters	
Wizard doesn't hear what user says	Wizard	repeat, or panic 'can you repeat that?' if it doesn't disrupt the scenario.	At wizards discretion, if it is important to the experiment to have heard and understood what user said.
User didn't hear what robot said, but wizard already typed next	User and Wizard	prev	

Wizard accidentally makes robot say next line before it is supposed to	Wizard	prev, maybe preceded by panic 'I am getting ahead of myself'	Might disrupt scenario or give away wizard of oz.
Wizard accidentally makes robot repeat a line.	Wizard	-	Might influence users opinion of robot.
Wizard starts wrong scenario.	Wizard	If it is discovered at the beginning, the scenario can be changed by writing the right start command.	It is however preferable to do the wrong scenario instead, because the switch might make it weird, if one of the scenarios have a line that is further along at that index.
User says yes when no is expected.	User	jump	Depending on what scenario is run, make sure the correct number of lines are skipped.
User asks an unexpected question that can be answered by yes and no.	User	y or n	
User asks a question that can be answered by a simple line.	User	panic "If you like" or "correct" for example	At wizard's discretion. Probably try to avoid, if not necessary. Might take time to write message.

# Chapter 4

## Results

### 4.1 Godspeed

#### 4.1.1 Likeability

The third questionnaire in the Godspeed series is meant to measure likeability. The questionnaire includes five scales: dislike to like, unfriendly to friendly, unpleasant to pleasant, unkind to kind, and finally, awful to nice. Generally the participants in all three conditions rated the robot quite high on likeability. Participants in the apology-condition rated the robot slightly lower ( $M = 4.3, SD = 0.4$ ), than ignore ( $M = 4.5, SD = 0.7$ ) and problem-solving ( $M = 4.5, SD = 0.2$ ). However, the variations in opinions of the robot varied more for the ignore-condition than the other two conditions, and very little for the problem-solving-condition.

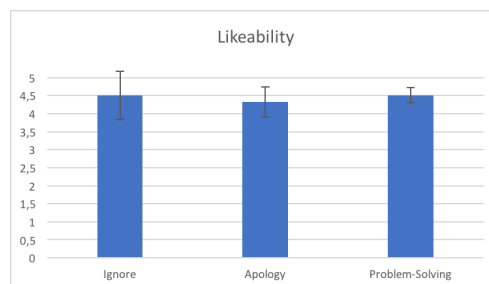


Figure 4.1: Likeability.

On the specific dislike to like-scale the apology-condition scored lowest, with the ignore-condition slightly higher, and the problem-solving-condition highest (diagram 4.2). The robot scored very high on the unfriendly to friendly-scale. The averages for all three conditions were between 4,8 and 5,0. The scores followed the same pattern as the general likeability rating, with the apology-condition slightly lower than the other two.

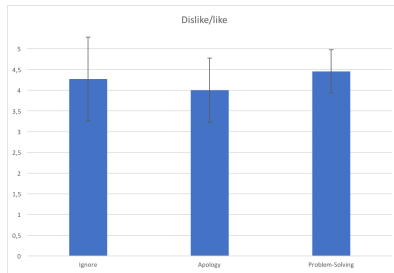


Figure 4.2: Like-scale.

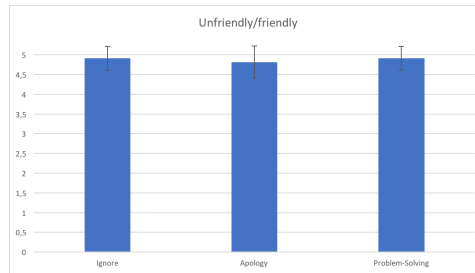


Figure 4.3: Friendly-scale.

### 4.1.2 Perceived Intelligence

The fourth questionnaire in the Godspeed series measures perceived intelligence. This questionnaire consists of: incompetent to competent, ignorant to knowledgeable, unintelligent to intelligent, irresponsible to responsible, and foolish to sensible. In general, the robot was perceived as fairly competent and intelligent, with average scores mostly between 3 and 4. The ignore condition scored highest ( $M = 3.7$ ,  $SD = 0.6$ ), followed by problem-solving ( $M = 3.4$ ,  $SD = 0.4$ ), and apology ( $M = 3.1$ ,  $SD = 0.6$ ).

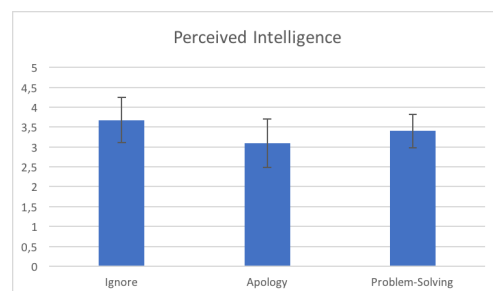


Figure 4.4: Perceived Intelligence.

The pattern holds for the unintelligent to intelligent-scale. On the incompetent to competent-scale, the problem-solving condition scored slightly higher, making it equal to the ignore-condition.

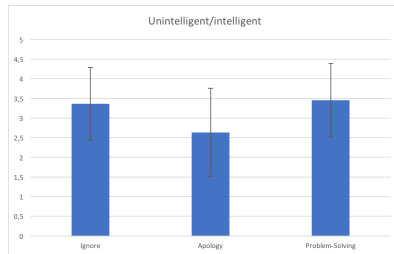


Figure 4.5: Intelligent-scale.

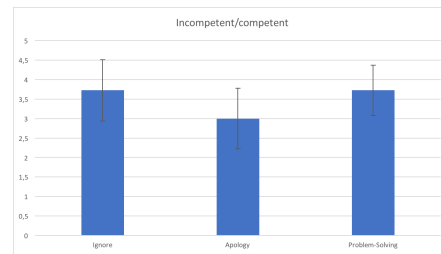


Figure 4.6: Competent-scale.

### 4.1.3 Animacy

The second questionnaire in the Godspeed series is meant to measure animacy. The questionnaire consists of six scales. They are: inert to interactive, artificial to lifelike, apathetic to responsive, stagnant to lively, dead to alive, and mechanical to organic. Overall, the robot was seen as fairly animated, with scores slightly above the middle of the scale. The ignore condition generated a slightly higher rating ( $M = 3.4$ ,  $SD = 0.7$ ) than apology ( $M = 3.2$ ,  $SD = 0.6$ ) and problem-solving ( $M = 3.2$ ,  $SD = 0.4$ ). The same pattern appears to apply to responsiveness as well. However, on interactiveness the problem-solving-condition scores highest, with the apology-condition generating the lowest score.

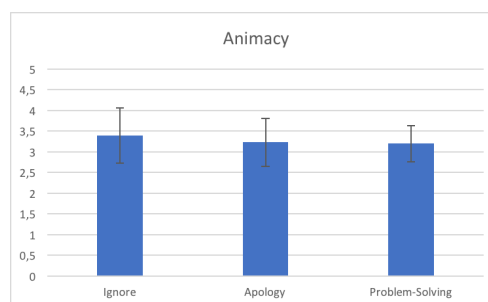


Figure 4.7: Animacy.

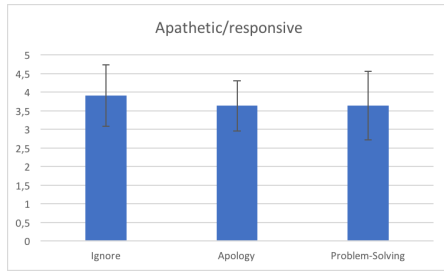


Figure 4.8: Responsive-scale.

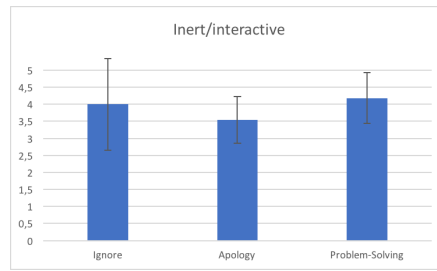


Figure 4.9: Interactive-scale.

## 4.2 RoSAS

### 4.2.1 Competence

The RoSAS Competence-score is based on six ratings. They are the following: capable, responsive, reliable, interactive, competent, knowledgeable. Generally, the apology condition ( $M = 5.3, SD = 1.3$ ) rated the robot lower on perceived competence, with the ignore condition ( $M = 6.2, SD = 0.9$ ) scoring slightly higher than problem-solving ( $M = 6.0, SD = 1.2$ ).

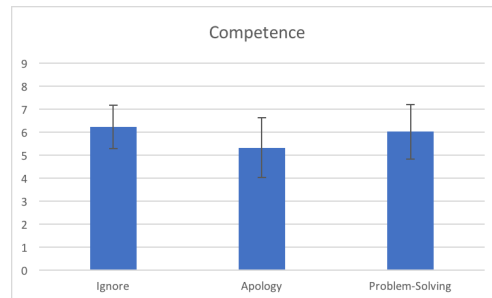


Figure 4.10: Competence.

### 4.2.2 Discomfort

The RoSAS discomfort score is based on the six scales: aggressive, awkward, scary, awful, dangerous, and strange. The robot scored low on discomfort, with the averages for all three conditions below 3. Ignore ( $M = 3.0, SD = 1.5$ ) scored slightly higher than apology ( $M = 2.7, SD = 0.9$ ) and problem solving ( $M = 2.6, SD = 0.9$ ).

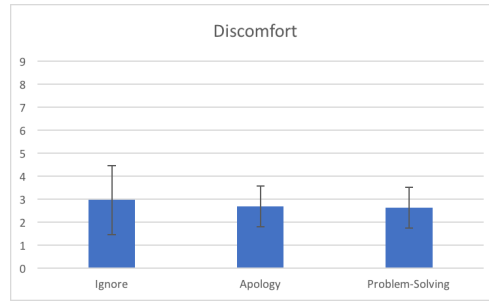


Figure 4.11: Discomfort.

### 4.3 Experiment length

The average length of the interaction with the robot, across all conditions, was about 3.0 minutes, with the average for the ignore condition slightly shorter ( $M = 2 : 49, SD = 15sec$ ) and for problem-solving slightly longer ( $M = 3 : 13, SD = 32sec$ ). The apology condition had the same average as the overall time ( $M = 3 : 01, SD = 17sec$ ).

### 4.4 WoZ guidelines

We had two questions in our open ended questions that are intended for the WoZ guidelines, the one where we ask about the participants' expectations from the robot and experiment and the one where we ask if the participant realized we were controlling the robot. Most participants didn't have any expectations or very low expectations, since we hadn't told them anything about the experiment beforehand. Some, who had seen the robot before had higher expectations, but overall the participants expectations were met or exceeded. Also, only a few participants realized that we were controlling the robot, so we can conclude that the simulation of the robot's autonomy was convincing.



# Chapter 5

## Discussion

### 5.1 Compare strategies

On average the robot was perceived as fairly competent and intelligent, with average scores around the middle of the scales on both Godspeed and RoSAS for all conditions. Additionally the robot was well liked across all conditions and the scores on discomfort were low. While the problem-solving scenario had the highest scores overall, the apology-scenario scored lowest on all intelligence and competence scores. The ignore-condition was meant to be less responsive and not try to recover when it fails. In the study by Lee et al., a similar behaviour resulted in a lower score for competence than behavior similar to both the problem-solving condition and the apology-condition [9]. Their study also showed that the apology-strategy produced the highest competence-rating [9]. On the contrary, our results showed the complete opposite, and the ignore condition scoring higher than expected. Unexpected participants behaviour might have decreased the perception of the robot's intelligence and competence for both the problem-solving-condition and the apology-condition, while at the same time increasing the ratings for the ignore-condition.

#### 5.1.1 Influence of Homogeneous Participant Pool

First, all participants have studied at least a year of computer science at a technical university and this education focuses heavily on problem-solving. This could be one reason why the participants in general preferred the problem-solving strategy over the apology-strategy.

Secondly, because the education includes a lot of problem-solving, many of the participants might have a similar attitude towards and interpretation of this behaviour. In the problem-solving condition of the tests, the standard deviations were, in general, lower than for the other two conditions, which suggests that the participants in that group agreed more in their perception of the robot. So, the focus on problem-solving in this condition combined with the similar attitude towards it could be one reason why they agreed more in their ratings than the participants in the other conditions.

Third, because all of the participants are young, have chosen to study computer science and engineering, and chose to participate in this study without any compensation, they might have a more positive attitude towards technology and robots than people in general have. This could be part of the explanation for why the robot was rated so high on friendliness and likeability. 29 of the 33 participants gave the robot the highest score on friendliness.

Fourth, the low scores on discomfort might be due to the participants' backgrounds as well. An interest in technology, experience with programming and some experience with robots could all reduce the feeling of discomfort in the presence of a robot. Furthermore, the fact that the wizards, which are classmates of the participants, were in the room might have reduced the discomfort-level further. Additionally, the robot's appearance, which several participants found cute, and that the experiment was a pleasant game can also have had positive effects on discomfort and likeability.

### 5.1.2 Influence of Experiment Design

The Animacy scores and length of the experiments were similar across all conditions, which is in line with what is expected since the interactions were made to be as similar as possible in terms of the amount the robot speaks and moves.

However, unexpected behaviour in the interaction between the robot and the participants might have affected the results, in particular many participants seemed to instinctively repeat the card that the robot heard incorrectly, a behaviour that did not comply with the script for the interaction. This means that, for the problem-solving condition, the participants in many cases had already repeated the card once when the robot asked them to repeat it again, introducing an additional plan-

ning failure. This might have influenced the participants perception of the robot and its capability negatively especially for interactiveness, responsiveness and overall intelligence and competence.

Furthermore, the open-ended questions showed that some of the participants seemed to mistake the robot's "ok" in the ignore-condition as acknowledgement of them repeating the card, rather than an acknowledgment that the robot heard incorrectly but wouldn't do anything to correct it. In this case, the participants would perceive the robot as having solved the problem, while for the apology condition it is obvious that the robot didn't solve the issue. This misconception might be why the apology-condition seems to be overall less popular, even compared to the ignore condition.

The ignore-condition was meant to be less responsive and interactive than the other two, by not responding to the fact that the participant says it misheard them. Nevertheless, the participants in that condition gave the robot higher ratings on responsiveness than the participants in both of the other conditions. They also rated it higher on interactiveness than the participants in the apology-condition. However, if the participants repeated themselves before the robot said "ok", this could be interpreted as the robot understanding them, which would explain the higher scores. Additionally, when the participants in the other two conditions also repeated themselves, it would seem like the robot ignored this, which could make it seem much less responsive. The fact that the robot asks them to repeat, and confirms what they say, could result in a higher rating for interactiveness, and yet not improve the responsiveness-score. Furthermore, Bajones, Weiss, and Vincze noted in their study that repetitive requests for help became an annoyance to the participants [2]. That could be a reason for the problem-solving condition not having better scores, compared to the ignore condition.

### 5.1.3 Sources of Errors

A first aspect of the experiment that introduces uncertainty in the accuracy of the results is the wording of the instructions for the part of the survey based on the RoSAS questions. We discovered during the experiments that the wording was unclear and that some of the participants answered the questions in the sections competence and discomfort for all robots, not only the Nao and the interaction that we had

them do. This could account for the standard deviation of these parts being high, meaning that the maximum and minimum of the answers of the participants were very different for some of the questions.

The second occurrence that could influence the results is the wizard and participant error. That is, the wizard did make errors sometimes, as well as the participants going off script in ways we had not made provisions for. The most common participant mistake was not quite hearing what card the robot asked them to look at. This resulted in confusion for some participants when they realized they had heard incorrectly, or when they heard another letter when the robot asked what number the card had. Other participants didn't seem to realize the mistake. Also, the different approaches the wizard could have made when the participants went off script were a little cumbersome, and during the experiments the wizard had to be quick when deciding what to do, mostly resulting in keeping to the script as much as possible. Hopefully, these mistakes were more or less evenly distributed across the three conditions, however the mistakes were not all taken into account by the unexpected counter (see section 3.2.3), and therefore this could not be checked accurately.

#### **5.1.4 Ethical Issues with the Method**

Furthermore, there are a few ethical downsides to using the WoZ method, since it entails in practice that the researchers are deceiving the participants into believing that the robot is acting on its own, as mentioned in section 3.1 in the Method. However, WoZ is an established way of conducting this type of studies. Also, when asked about it, the participants seemed to understand the necessity of this approach and not have any aversions about it. This might also be influenced by the fact that they are all university students in Computer Science, and have another understanding of this type of research compared to a group more representative of the general population.

## **5.2 Limitations**

The biggest limitation we had was the size of our participant group. We had 33 participants in total, and 11 per condition. This was a good number within the time we had and with the resources we had. However, it is a small number and it means that every mistake or misunder-

standing by the wizard and/or participant has a greater impact on the results and we could not afford to remove any experiments because of this either.

Furthermore, our participant group was very homogeneous, since all participants are between 20 and 27 and have studied at least one year of computer science at a technical university. They were also all drafted from the researchers' personal networks. This will have influenced the results. However, the homogeneousness of the group entails that it is easier to draw conclusions about it, even if they are not applicable on the larger society.

### 5.3 Future Research

The study conducted for this paper, as well as the results resulting from it are rather restrictive, so there is a lot that can be done for future research. A few suggestions that we have thought about are:

- To have an iterative process for the experiment, as suggested in the WoZ guidelines, to perfect the experiment as well as to minimize the impact of wizard and participant error.
- To have more participants, and maybe a group of participants that are more representative of society.
- To remove the wizards from the room, if possible, to see if this has an influence on the results.
- To have a control condition where the robot doesn't fail, to see how much the failure in itself influences the participants perception of the robot.
- To have a condition that combines apology and problem-solving, which we would have liked to study but we had to limit ourselves to three strategies.
- To improve the ignore condition, to see if experiment design, with "ok", had as big of an impact as we think
- To have a more responsive robot, meaning, if the user already starts to repeat, the robot doesn't ask to repeat again, to see if that had as much of an impact as we believe.

## Chapter 6

### Conclusion

Based on previous research, we expected that the apology strategy would be most successful, followed by problem-solving, and then ignore. However, our results show that the apology condition resulted in less positive perceptions of the robot in most of our measures of interest, and that the ignore condition did as well as, or better than, problem-solving in many cases. We believe one reason why the problem-solving condition surpassed the apology condition is that the homogeneous participant pool had a problem-solving background. Secondly, unexpected participant behaviour led to the ignore conditions seemingly solving the problem, and seem better than it is. Task success has been found to be an important factor in the perception of a robot, and the fact that the apology condition seemed to be the only one that did not fully succeed at the task could explain why it was the least popular condition. To answer our research question, of the two recovery strategies, problem-solving minimized the negative effects of failure most, but no recovery, the ignore condition, often scored at least as well as problem-solving.

# Bibliography

- [1] Aldebran. *Aldebran Documentation Tutorial or samples*. URL: <http://doc.aldebaran.com/2-1/naoqi/trackers/trackers-sample.html> (visited on 04/23/2017).
- [2] Markus Bajones, Astrid Weiss, and Markus Vincze. "Help, Anyone? A User Study For Modeling Robotic Behavior To Mitigate Malfunctions With The Help Of The User". In: (2016).
- [3] Colleen Carpinella et al. "The Robotic Social Attributes Scale (RoSAS): Development and Validation". In: *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. HRI '17. ACM, Mar. 2017, pp. 254–262. ISBN: 9781450343367.
- [4] Kerstin Dautenhahn. "Socially Intelligent Robots: Dimensions of Human-Robot Interaction". In: 362.1480 (2007), pp. 679–704. ISSN: 09628436.
- [5] Mary Ellen Foster et al. "Two people walk into a bar: dynamic multi-party social interaction with a robot agent". eng. In: *Proceedings of the 14th ACM international conference on multimodal interaction*. ICMI '12. ACM, Oct. 2012, pp. 3–10. ISBN: 9781450314671.
- [6] Python Software Foundation. 24.2. *cmd-Support for line-oriented command interpreters*. 2017. URL: <https://docs.python.org/3/library/cmd.html> (visited on 04/04/2017).
- [7] Manuel Giuliani et al. "Comparing task-based and socially intelligent behaviour in a robot bartender". In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ICMI '13. ACM, Dec. 2013, pp. 263–270. ISBN: 9781450321297.
- [8] Manuel Giuliani et al. "Systematic analysis of video data from different human-robot interaction studies: a categorization of social signals during error situations". In: *Frontiers in psychology* 6 (2015). ISSN: 1664-1078.

- [9] Min Kyung Lee et al. "Gracefully mitigating breakdowns in robotic services". In: *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction*. HRI '10. IEEE Press, 2010, pp. 203–210. ISBN: 9781424448937.
- [10] Laurel D Riek. "Wizard of oz studies in hri: a systematic review and new reporting guidelines". In: *Journal of Human-Robot Interaction* 1.1 (2012).
- [11] Softbank robotics. *Find out more about Nao*. URL: <https://www.alde.softbankrobotics.com/en/cool-robots/nao/find-out-more-about-nao> (visited on 03/29/2017).
- [12] Susanne Stadler, Astrid Weiss, and Manfred Tscheligi. "I Trained this robot: The impact of pre-experience and execution behavior on robot teachers". In: IEEE, Aug. 2014, pp. 1030–1036. ISBN: 978-1-4799-6763-6.
- [13] Cass R. Sunstein. "Social Norms and Social Roles". In: *Columbia Law Review* 96.4 (1996), pp. 903–968. ISSN: 00101958.
- [14] Cristen Torrey, Susan Fussell, and Sara Kiesler. "How a robot should give advice". In: *Proceedings of the 8th ACM/IEEE international conference on human-robot interaction*. HRI '13. IEEE Press, 2013, pp. 275–282. ISBN: 9781467330558.
- [15] Astrid Weiss and Christoph Bartneck. "Meta analysis of the usage of the Godspeed Questionnaire Series". In: IEEE, Aug. 2015, pp. 381–388. ISBN: 978-1-4673-6704-2.



# Appendix A

## Experiment Description

### What failure are the experiments testing?

Social interaction failure, conversation, where failure is the robot interpreting human words wrong.

### How?

A game with cards where the robot needs help.

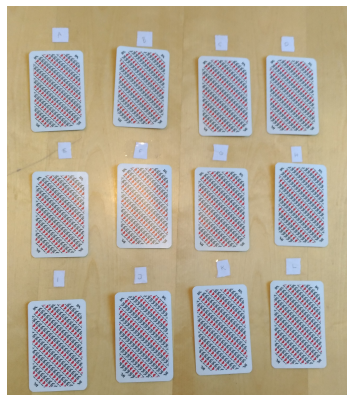


Figure A.1: Card Setup Picture

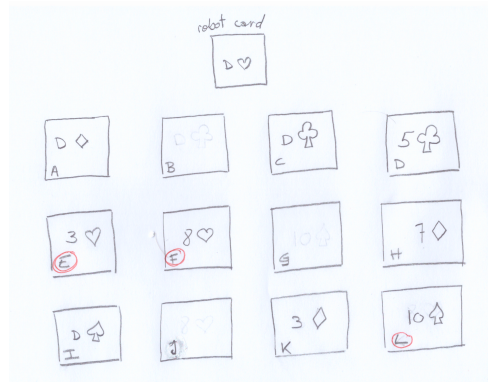


Figure A.2: Card Setup Sketch

### Premiss:

12 cards on the table, face down, placed in spots labeled from A to L. The robot holds one card, a queen of hearts for example.

Because of language barrier, where it might not come naturally to all participants to name the symbol on the card, we ask the test persons to just say the number on the card.

**Robots goal:**

To find the other three queens.

**Robots restriction:**

The robot can't turn the cards himself, he needs the humans help to turn the cards and say what card is where, in order to find the hidden queens.

**Dialog:**

"Hi! I am Nao and I have a queen of hearts card. I need your help to find the other three queens, because I can't lift these cards myself.

- Would you please look at card X and tell me what card it is?
- [test subject input]
- So X is a Y card?
- yes/no
- (if the card was a queen) Great! Can you please put that card aside?"

**Failure:**

Hearing the wrong card.

**Each recovery strategy in practice:**

**Ignore:** doesn't acknowledge anything, just goes on to asking for the next card.

"So is X a Y card

- no
- Would you please look at card Z and tell me what card it is?"

**Apology:** apologizes for not having heard correctly and the goes on to asking for next card.

"So is X a Y card?

- no
- I'm sorry, sometimes I don't interpret human speech correctly.

- Would you please look at card Z and tell me what card it is?"

**Problem-solving:** acknowledges the failure by asking human to repeat themselves.

"So is X a Y card?

- no

- Oh, can you please repeat what card you said?

- [user input]

- So it is a S card?

- yes"

**Outcome:**

All queens will be found. When they are, the robot says: "We found all the queens! Thank you for your help!"

**Order of cards and frequency of failure:**

K - hears correctly

A - queen of diamonds

F - fails to hear correctly

H - hears correctly

I - queen of spades

L - fails to hear correctly

E - fails to hear correctly

D - hears correctly

C - queen of clubs

# Appendix B

## Experiment Protocol

### B.1 Experiment Protocol Template:

“Hi! I am Nao and I have a queen of hearts card. I need your help to find the other three queens, because I can’t lift these cards myself and I can’t read very well. But I can hear you.

- Would you please look at card K and tell me what card it is? (You don’t need to say the symbol or color, only the number)
- It’s a 3 [expected subject input]
- So K is a 3? [K - hears correctly]
- yes [expected subject input]
- You can leave the card there, facing up or facing down, as you like.
- Could you check card A now and tell me what card it is?
- it’s a queen [expected subject input]
- So A is a queen? [A - queen of diamonds]
- yes [expected subject input]
- Great! Can you please put that card aside?
- [expected subject action]
- Now flip card F. Could you please tell me what rank it is?
- It’s an 8 [expected subject input]
- Did you say that F is a knight? [F - fails to hear correctly]
- no [expected subject input]

FAIL RECOVERY 1

- Would you please look at card H. Which rank is it?
- It’s a 7 [expected subject input]
- H is a 7? [H - hears correctly]
- yes [expected subject input]

- Now look at card I. What card is it?
- It's a queen [expected subject input]
- Did you say we found our third queen? [I - queen of diamonds]
- yes [expected subject input]
- Great! Can you please put it with the other queen?
- [expected subject action]
- Would you please look at card L and tell me what card it is?
- It's a 10 [expected subject input]
- So L is a 7? [L - fails to hear correctly]
- no [expected subject input]

#### FAIL RECOVERY 2

- Could you now please look at card E and tell me what it is?
- It's a 3 [expected subject input]
- E is a king? [E - fails to hear correctly]
- no [expected subject input]

#### FAIL RECOVERY 3

- Would you please tell me what rank D has?
- It's a 5 [expected subject input]
- Did you say that D is a 5? [D - hears correctly]
- yes [expected subject input]
- Now, please look at card C and tell me what card it is.
- It's a queen [expected subject input]
- So C is a queen? [C - queen of diamonds]
- yes [expected subject input]
- Great! We found all the queens! Thank you for your help!

## B.2 Fail-recovery: Ignore

#### FAIL RECOVERY 1

Okay.

#### FAIL RECOVERY 2

Okay.

#### FAIL RECOVERY 3

Okay.

### B.3 Fail-recovery: Apology

FAIL RECOVERY 1

- I'm sorry, sometimes I don't interpret human speech correctly.

FAIL RECOVERY 2

- I apologize, it seems that my hearing sensors aren't working very well

FAIL RECOVERY 3

- I'm so sorry, it appears my hearing failed again.

### B.4 Fail-recovery: Problem-solving

FAIL RECOVERY 1

- Oh! Can you please repeat what you said?
- It's an 8 [expected subject input]
- Is F an 8 then?
- yes [expected subject input]

FAIL RECOVERY 2

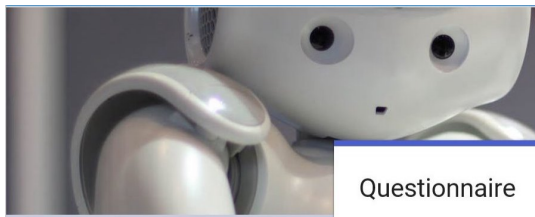
- Oh! Can you say that again?
- It's a 10 [expected subject input]
- Is L a 10 then?
- yes [expected subject input]

FAIL RECOVERY 3

- Oh! Can you please repeat it?
- It's a 3 [expected subject input]
- Is E a 3 then?
- yes [expected subject input]

# Appendix C

## Questionnaire



**Questionnaire**

Please answer these questions the best you can! There are three pages.

\*Required

**Age \***

Your answer \_\_\_\_\_

**Gender \***

☐ Male

☐ Female

☐ Other

☐ Don't want to specify

**Programming experience \***

	1	2	3	4	5	
None	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Experienced

**Experience with robots \***


	1	2	3	4	5	
None	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extensive

**NEXT**

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Additional Terms

Google Forms



## Questionnaire

\*Required

Please rate your impression of the robot on these scales:

1 \*

Dislike

1

2

3

4

5

Like

2 \*

Incompetent

1

2

3

4

5

Competent

3 \*

Inert

1

2

3

4

5

Interactive

4 \*

Unfriendly

1

2

3

4

5

Friendly

5 \*

Ignorant

1

2

3

4

5

Knowledge - able

6 \*

Unpleasant

1

2

3

4

5

Pleasant

7 \*

Artificial

1

2

3

4

5

Lifelike

8 \*

Unintelligent

1

2

3

4

5

Intelligent

Likeability

Perceived Intelligence

Animacy

Likeability

Perceived Intelligence

Likeability

Animacy

Perceived Intelligence



Likeability

Animacy

Animacy

Perceived Intelligence

Likeability

Animacy

Perceived Intelligence

Animacy

9 \*

Unkind

1

2

3

4

5

Kind

10

Apathetic

1

2

3

4

5

Responsive

11 \*

Stagnant

1

2

3

4

5

Lively

12 \*

Irresponsible

1

2

3

4

5

Responsible

13 \*

Awful

1

2

3

4

5

Nice

14 \*

Dead

1

2

3

4

5

Alive

15 \*

Foolish

1

2

3

4

5

Sensible

16 \*

Mechanical

1

2

3

4

5

Organic

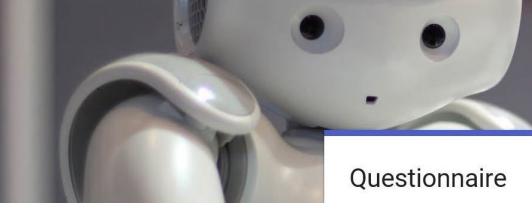
BACK

NEXT

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Additional Terms

Google Forms



## Questionnaire

*\*Required*

Using the scale provided, how closely are the words below associated with the category robots?  
1 = definitely not associated to 9 = definitely associated.

\*

	1	2	3	4	5	6	7	8	9
Capable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Aggressive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Responsive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Reliable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Awkward	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Interactive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Scary	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Competent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Awful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dangerous	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Knowledge-able	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Strange	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

BACK SUBMIT

Never submit passwords through Google Forms.

This content is neither created nor endorsed by Google. Report Abuse - Terms of Service - Additional Terms

Google Forms

**Competence**

**Discomfort**

**Competence**

**Competence**

**Discomfort**

**Competence**

**Discomfort**

**Competence**

**Discomfort**

**Competence**

**Discomfort**

