Exploring Human Teachers' Interpretations of Trainee Robots' Nonverbal Behaviour and Errors

by

Pourya Aliasghari

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master's of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes first-authored peer-reviewed material that has appeared in a conference proceeding, published by the Association for Computing Machinery (ACM), and accepted for publication in a journal, published by John Benjamins Publishing Company.

Research presented in Chapter 3 (Study 1):

Pourya Aliasghari, Prof. Chrystopher L. Nehniv and Prof. Kerstin Dutenhahn designed the study. Pourya Aliasghari implemented the apparatus, performed the experiment, conducted the data analysis, and drafted the manuscripts. Prof. Moojan Ghafurian, Prof. Chrystopher L. Nehniv and Prof. Kerstin Dutenhahn reviewed and edited the manuscripts, and also supervised the project.

Citations:

Pourya Aliasghari, Moojan Ghafurian, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2021. Effects of Gaze and Arm Motion Kinesics on a Humanoid's Perceived Confidence, Eagerness to Learn, and Attention to the Task in a Teaching Scenario. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21).* Association for Computing Machinery, New York, NY, USA, 197–206.

DOI: https://doi.org/10.1145/3434073.3444651

Pourya Aliasghari, Moojan Ghafurian, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. 2021. Impact of Nonverbal Robot Behaviour on Human Teachers' Perceptions of a Learner Robot. Accepted for journal publication in *Interaction Studies.*

Research presented in Chapter 4 (Study 2):

Pourya Aliasghari, Prof. Moojan Ghafurian, Prof. Chrystopher L. Nehniv and Prof. Kerstin Dutenhahn designed the study. Pourya Aliasghari implemented the apparatus, performed the experiment and conducted the data analysis. Prof. Moojan Ghafurian, Prof. Chrystopher L. Nehniv and Prof. Kerstin Dutenhahn supervised the project.

Abstract

In the near future, socially intelligent robots that can learn new tasks from humans may become widely available and gain an opportunity to help people more and more. In order to successfully play a role, not only should intelligent robots be able to interact effectively with humans while they are being taught, but also humans should have the assurance to trust these robots after teaching them how to perform tasks.

When human students learn, they usually provide nonverbal cues to display their understanding of and interest in the material. For example, they sometimes nod, make eye contact or show meaningful facial expressions. Likewise, a humanoid robot's nonverbal social cues may enhance the learning process, in case the provided cues are legible for human teachers. To inform designing such nonverbal interaction techniques for intelligent robots, our first work investigates humans' interpretations of nonverbal cues provided by a trainee robot. Through an online experiment (with 167 participants), we examine how different gaze patterns and arm movements with various speeds and different kinds of pauses, displayed by a student robot when practising a physical task, impact teachers' understandings of the robot's attributes. We show that a robot can appear differently in terms of its confidence, proficiency, eagerness to learn, etc., by systematically adjusting those nonverbal factors.

Human students sometimes make mistakes while practising a task, but teachers may be forgiving about them. Intelligent robots are machines, and therefore, they may behave erroneously in certain situations. Our second study examines if human teachers for a robot overlook its small mistakes made when practising a recently taught task, in case the robot has already shown significant improvements. By means of an online rating experiment (with 173 participants), we first determine how severe a robot's errors in a household task (i.e., preparing food) are perceived. We then use that information to design and conduct another experiment (with 139 participants) in which participants are given the experience of teaching trainee robots. According to our results, perceptions of teachers improve as the robots get better in performing the task. We also show that while bigger errors have a greater negative impact on human teachers' trust compared with the smaller ones, even a small error can significantly destroy trust in a trainee robot. This effect is also correlated with the personality traits of participants.

The present work contributes by extending HRI knowledge concerning human teachers' understandings of robots, in a specific teaching scenario when teachers are observing behaviours that have the primary goal of accomplishing a physical task.

Acknowledgements

First of all, I would like to thank my supervisor, Prof. Kerstin Dautenhahn, for giving me this amazing opportunity to work as a member of the SIRRL lab. I also thank my co-supervisor, Prof. Chrystopher L. Nehaniv, who guided me precisely in a variety of topics during my research journey and also trusted me to help him with teaching. Finally, I wish to thank Prof. Moojan Ghafurian, who has been a big help to my work, for all her generous and immediate support over the past year. Thank you, Kerstin, Chrystopher and Moojan for being such excellent mentors. I have learned a lot from all of you and hopefully, this would continue.

I am also grateful to my thesis review committee, Prof. Jesse Hoey and Prof. Shi Cao, for spending their valuable time evaluating my thesis and providing me with insightful feedback.

Moreover, I would like to express my gratitude to my wonderful SIRRL labmates. In particular, thank you Alex Aroyo and Hamza Mahdi for being such great officemates. Related to my research work, thank you Sami Alperen Akgün for helping me set up the online interface, and Zhuofu Tao for providing the initial codes. Thank you Elaheh Sanoubari, Shruti Chandra, Owais Hamid, Samira Rasouli and Ali Noormohammadi for all your advice. I regret not being able to enjoy being with you amazing people over the past year because of the pandemic.

At last, I should thank all my lovely friends, those who are in Canada and those who are not, for making this time a beautiful and fun experience for me. In particular, thank you, Soroosh, Azin, Amir, Melika, Rasoul and Fatemeh, and sorry for all the plan changes that were always caused by my work! I am so proud to call you my friends for life.

This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program, and all the people who participated in the experiments.

Dedication

This thesis is dedicated with love to my parents, Akhtar and Hamid, who are always the first people that I can rely on.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC Akaike's Information Criterion 26, 28, 32, 60

DOFs Degrees of Freedom 21, 22

DT Disposition of Trust questionnaire 55, 56, 59–61, 69, 72, 74

GLM Generalized Linear Model 60, 63, 72, 81

HRI Human-Robot Interaction 1, 2, 5, 9, 10, 12, 13, 15, 17, 21, 24, 33, 38, 84, 85

LMM Linear Mixed-effects Model 26, 27, 29, 33, 61, 69, 74, 81

MTurk Amazon Mechanical Turk 19, 25, 46, 59, 85

TIPI Ten Item Personality Inventory 54, 56, 60, 61, 69, 72

Chapter 1

Introduction

In the modern world, socially intelligent robots are increasingly being developed to support people in a variety of domains. These robots need to have effective and efficient communication abilities to collaborate and interact with humans in many situations. One major aspect of the intelligent robots is that they have the capability to learn new tasks and extend their functionalities and skills, e.g., by receiving instructions from their owners [29]. Therefore, humans may sometimes act as teachers of robots. To facilitate the skill transfer process and enable non-experts to teach robots, natural teaching mechanisms have been developed. As an example, we can refer to *imitation learning*, through which an agent learns how to accomplish a task by watching or experiencing a sequence of actions [105, 13].

1.1 Motivation

Robots' Nonverbal Behaviour

Followed by technological advancements, communication through body movements and gesture has become an important avenue for successful Human-Robot Interaction (HRI) [92, 21, 81, 63, 106, 4, 62]. Head and arm movements, body gestures, and eye gaze are examples of nonverbal behaviours that can be powerful communicative tools to complement a robot's main functionality (e.g., carrying out a task). It has been found that activation patterns in human brain interconnectivity networks differ at a global neuronal level when people pay attention to *how* an action is performed compared with when thinking about the *goal* of the action [33]. Nonverbal expressive behaviours of robots can shape humans' understanding of them or elicit different reactions [56]. Therefore, studying how those behaviours affect humans' perceptions is an important area of research in HRI.

Robots' Errors

Intelligent robots may perform their tasks in dynamic and unexpected environments, using imperfect sensors. This can lead to a reduction in their reliability [32]. In many cases, robots are autonomous or semi-autonomous and need to make decisions on their own and act accordingly. Hence, given the complexity of the environment and the task, even robots with exceptional reliability may behave in an incorrect manner. It has been previously found that robots' errors can negatively impact several aspects of HRI. For instance, people may react to these errors in real-time, by frowning or averted gaze [50]. Trust, as a critical factor concerning acceptance and persuasiveness of robots [101], can also be highly impacted after humans observe robots' errors [47, 31]. People may become reluctant to use robots in the future if they cannot trust them [94, 115].

1.2 Research Goal and Scope

This thesis aims to explore humans' interpretations of trainee robots' *nonverbal behaviour* and *errors*, by conducting two separate studies (three experiments with human participants in total).

Our first experimental study (presented in Chapter 3 [5, 6]) focuses mainly on nonverbal communication for a robot learner. We explore *how people, in the imaginary role of teachers, interpret multiple aspects of a trainee robot's nonverbal behaviours while observing it performing a learned physical task*. The behavioural parameters manipulated here are the gaze pattern of the robot, as well as the speed and smoothness of its arm movement trajectories during task performance. We also examine a potential effect of priming (i.e., changes in perceptions or behaviour of participants caused by prior stimuli). Study of this priming factor, which is regarding time passed since teaching, is exploratory in nature as previous research has shown that when a robot gradually gains a new skill, its human teachers' behaviours might change too [39]. The studied topic is important because before any potential use of nonverbal communication in robot learning scenarios, we need to ensure that the trainer (human) is able to interpret behaviours of the learner (robot) accurately, according to its intent.

The next study, reported in Chapter 4, explores how robot errors happening during the learning process may affect participants' interpretations of its behaviours and impact their trust. While in a human-human teaching scenario mistakes may be expected to happen sometimes, our general question is: *When a "robot" is being taught a task by a human teacher, will they simply ignore its errors? How can a small error affect the teacher? How about a more severe mistake?* The first experiment conducted in this study investigates what is seen as a small error or a big error when a robot is performing a simple food preparation task. In another experiment, which is built upon the insights gained from the first one, participants make use of our virtual platform to remotely demonstrate how to perform that task to a humanoid robot, and then the robot does so, unless there are errors. Over multiple training and practising rounds, the robot appears to gradually improve. This way, unlike Study 1, our participants were given the experience of teaching the robot and rated the teaching scenario to be realistic. Depending on the experimental condition, the robot might display faulty behaviours with different levels of severity while practising the task for the last time. Our aim is to identify the consequences of those errors on teachers' beliefs and their trust. Furthermore, we explore impact of a learner robot's perceived professionality, indicated by the way it is dressed, on teachers' trust and attitudes. We consider two cases when a robot has a tidy appearance or an untidy dress.

1.3 Summary of Contributions

The contributions of this research can be summarized as:

1. The first study identifies suitable behavioural adjustments for robots to influence people, or convey information about their internal state as well as their learning progress. This nonverbal communication can help to improve human teachers' understanding of trainee robots and would ultimately enhance the learning outcomes through improving the legibility of the observed behaviours.

2. We report the process of designing a virtual framework to simulate a human-robot teaching interaction, using which trust is studied in an online teaching situation, by:

   (a) Introducing an interactive method through which people can remotely demonstrate a task for a robot. This enables participants to experience the act of teaching, without the need to physically come to the lab.

   (b) Showing how errors of a learner robot affect human teachers. We point out that even a small error may largely impact trust, indicating the importance of behaviours of a robot even while learning a new task.

1.4 Thesis Overview

The following chapters make up the rest of this thesis. Chapter 2 presents a literature review on the related topics, i.e., social learning, nonverbal communication, and trust, and also briefly discusses the impact of robots' appearance. In Chapter 3, we report our first experimental study that investigates the impact of nonverbal robot behaviour on human teachers' perceptions of a learner robot. Then, Chapter 4 describes our second work, including two experiments, explores trust in trainee robots with respect to various erratic behaviours and multiple appearances. Finally, Chapter 5 concludes the thesis by providing brief answers to the research questions, listing our contributions to knowledge, and discussing limitations of the presented studies.

There are also four appendices in the end. Appendix A includes our ethics clearance certificates. Screenshots from all the pages that were designed for conducting our experiments are provided in Appendix B, C and D.

Chapter 2

Background Literature

Throughout this chapter, we give an overview of the background literature related to the topics discussed in this thesis. We start with introducing the concept of social learning and describing its functionality in the life-time skill development process of infants and primates. Then, we look at how mechanisms inspired by social learning, e.g., imitation learning, might benefit intelligent robots.

Associated with Study 1, we list different nonverbal modalities used by robots for the purpose of communication. After that, some previous work regarding the influence of those behaviours on the human-robot interaction will be reviewed. Special attention is given to the robot learning domain. We also describe some examples of how humans' perception of a robot can be altered when it adjusts its nonverbal behavioural factors.

Finally, to introduce the type of work done in Study 2 and mention our inspiration source for the approach we followed, a series of experiments investigating the trust factor in HRI will be briefly mentioned. Then, we point out how errors made by robots could influence trust as well as the behaviours of their interaction partners. In the final section of this chapter, we briefly discuss the impact of robot appearance on the interaction.

2.1 Social Learning

Social learning refers to a type of learning done through observation of or interaction with other members of the same species. Social learning stands in contrast to "asocial learning", or "individual learning", which describes learning without social interactions [53].

2.1.1 Social Learning in Primates and Infants

Primates social learning takes place through a variety of processes including *imitation*, *emulation*, *local and stimulus enhancement*, and *teaching* [120, 121], as described in the following. Copying another individual is known as "imitation", and can include mimicking their body actions or even how they use tools and other objects. "Emulation" is a more diverse term applied to processes including *object movement re-enactment* (copying how the model makes objects move, if imitation is limited to body movements), *end-state emulation* (replicating only the desired outcomes of the action, to reach using either the same method as the model or their own method) and *affordance learning* (understanding properties of the objects or other aspects of the environment to be able to utilize them with their own strategies) [120]. Through "Local and Stimulus Enhancement", the observer's attention is drawn towards a relevant location or object by the actions of another individual. Finally, "teaching" can be defined as a behaviour performed by a teacher at a cost to benefit developmental achievements of some students [121].

Social learning in primates occurs in three main stages (see Figure 2.1): (1) *Learning from the mother or another primary attachment figure*, (2) *Selective learning in an expanding social world*, and (3) *Learning from residents after the migration* [121]. In many primates, mothers care for and carry their infants at first, with different levels of maternal interaction that shape social learning opportunities. Therefore, from the beginning of the lifetime of an infant, the mother is the main and often the only model for social learning. Through the first stage, immature individuals attentively watch their mothers' techniques, by a focused and close-range visual attention called "peering" [107]. This involves the learner facing the demonstrator and nodding his head as a sign that he has understood the actions. Later, they practise those techniques rather than manipulating other objects, which indicates that they are acquiring the skills through observation.

In the second phase, i.e., *selective learning*, individuals start to learn from those deemed better models than their parents for expert tasks. Through studying a small-scale society of *Fijian villagers*, Henrich and Broesch [52] observed some learning biases in this selective phase. "Perceived within-domain success and knowledge" turned out to be the most important bias. In domains such as fishing where a potential model's achievements could be directly observed, success seemed far more important than knowledge. When there was no direct evidence of success, "age" was used as an indirect predictor of the model to learn from. Moreover, they noted a "cross-domain success" effect, meaning that perceived success in one domain could affect learning in another domain, mostly when two domains are related. It was also found that learners tend to pick models from particular sexes for specific domains, for example, males as fishing models.

Phase 1. Learning from mother or other primary attachment figure. Baboon infant sniffs novel food mother is eating. Photo: A. Whiten

Phase 2. Selective learning in an expanding social world. Juvenile capuchin observes expert adult male nut-cracking . Photo: T. Falótico

Phase 3. Learning from residents after migration. Male vervet switches to eat colored corn preferred by new group. Photo: E. van de Waal

Figure 2.1: The three-phase model of how social learning occurs in primates. Adapted from "The pervasive role of social learning in primate lifetime development" by Andrew Whiten and Erica van de Waal, 2018 [121]. Used with permission.

Other studies have suggested that primates hold even more biases for selective social learning. For example, regarding some species of brown capuchin monkeys (*Cebus apella*), it has been found that understanding the relative proficiency of group-mates and observing the more skilled ones could help monkeys maximize their social learning opportunities [85]. Another study has shown that chimpanzees (*Pan troglodytes*) recognize and imitate their group-mates who act faster [114]. In a different study about chimpanzees, Kendal et al. [65] have found that species are also inclined to copy dominant and knowledgeable individuals, who have been trained to succeed. Although these target models may be as successful as the others, their confident and purposeful way of approaching the tasks is what leads chimpanzees to prefer them. According to Perry [87], white-faced capuchin monkeys (*Cebus capucinus*) are prone to copying a majority of their group. They are more likely to adopt behavioural alternatives that they see most frequently among their peers.

Research on human children has also shown that infants also learn selectively from different adults. Harris and Corriveau [49] have found that two major heuristics could be used by children to identify which models are more trustworthy than others. First, children exhibit a bias to endorse the claims and imitate the demonstrations of those that they identify as attachment figures, and/or who have previously proven to be reliable sources of information. To do this, children keep track of the history of potential models. When no recorded experience with another person is available, the second heuristic is used by children. In this case, they will be more likely to imitate those who seem to belong to their cultural group. In another study, two- and three-years-old children's preference for imitation was examined by Birch et al. [15], when adult models showed nonverbal cues such as facial expressions and body postures for indicating their level of certainty. It was observed that when both a confident and an uncertain model were available, children selected to learn about objects from the confident one.

Lastly, the third stage of social learning occurs when migrants encounter different social and ecological circumstances from their natal group [121]. During this stage, social learning from residents may be helpful, especially for migrants who have little familiarity with the local resources. In many cases, species quit their existing personal preferences or behaviours and adapt to the new local norms. The reason can be that they are unsure about the optimal local behaviours to employ in the new environment, and as a solution, they learn what is offered by the existing residents. One other reason may be that newcomers are more likely to be accepted in the new group if they behave similarly to the current members of the new society.

2.1.2 Imitation and Social Learning in Robots

As discussed, under the behaviours considered as social learning, imitation happens when one individual copies another individual's behaviour that is novel to their behavioural range. It is for a time that robots and software agents also benefit from this type of learning. "Imitation learning" is a robot learning mechanism that makes the transfer of skill more efficient [13]. This might also make socially intelligent robots more "like us" and make them individuals [28]. By observing a demonstration or experiencing a sequence of actions, imitation learning provides a quick and easy way for people without any extensive knowledge of programming to teach a robot how to do complex actions. This can be considered as an effective method of motor learning, especially for robots with many degrees of freedom and when the state-action space is large. A review about robots that imitate humans can be found in [20].

To let robots dynamically acquire new task skills, several imitation frameworks have been developed, e.g. [60, 23, 24, 113, 26]. Demonstrations from humans can be conveyed to robots using a variety of modalities. In many cases, a person can guide the movements of a robot by grasping and moving its arms. This method is often called "Kinesthetic teaching" and focuses on learning continuous motion trajectories and actions done on the objects and the environment. In some other robotic systems, teachers' behaviours can be recognized by the robot remotely, e.g., using vision sensory inputs or motion capturing devices.

Any form of imitation learning presumes a correspondence between two autonomous agents [30]. Observers should be able to recognize that the learned tasks of the imitator correspond to the behaviours of the demonstrator. These relations exist if the embodiments, sequences of actions, and goals meet a criterion. For instance, if both the teacher and the learner have similar embodiments, this correspondence is just to map the equal parts of the bodies. Nevertheless, it would be likely that the human teacher and the robot differ in the number and/or types of joints, lengths of the links, the amount of force/torque they can apply, and therefore, exact behaviours can not be copied.

As previously mentioned, research on how social learning contributes to the lifelong skill development process in humans and several species of primates suggests that interacting with others is a crucial component of any learning task [121]. This highlights the need for designing mechanisms to support teaching process of robots through social feedback and interaction [122, 68]. No matter what modality the instructor uses, it would be helpful if the agent be able to interpret the teacher's behaviours and also express itself in a way that the instructor understands [71]. Through an interaction, the learner and the teacher form mental models of one another that can be used to support their collaborative activity during learning and teaching. It is likely that the two-way flow of information between the teacher and the robot will improve the teaching process and help the teacher provide more relevant inputs [19]. It has been demonstrated that by generating concise feedback through social signals (e.g., gaze, head and arm gestures, etc.), a robot can control how the teacher presents the instruction [89].

> **Conclusion:** The concept of social learning is not directly studied in this thesis. However, emphasizing the importance of the "interaction" element in skill development of species and describing how robots may acquire new abilities from humans can be helpful for understanding the contributions of our work. Instead of designing robots to only learn by recording the inputs from the teachers, a two-way stream of information between the trainee robot and its human teacher may potentially benefit the learning process. For instance, when some information is missing or teacher's actions are unclear to the robot, its behaviours can be adjusted immediately to convey signs of uncertainty or confusion to the teacher. That may motivate them to provide more informative instructions.

2.2 Nonverbal Communication in HRI

As reviewed in [91], several studies have examined the impact of a robot's personality on different aspects of HRI. In this regard, researchers have been especially interested in detecting how the physical behaviours (e.g., gestures, movement patterns, facial expressions, and gaze) or verbal behaviours (e.g., audio and linguistic style, voice gender and speed, and responsiveness) of a robot may affect people's impressions and perception of it.

Communication by the nonverbal means is an essential component of modern social robots. If a robot lacks the ability to express facial or vocal expressions, nonverbal communication may be its only option. Saunderson and Nejat [106] have reviewed studies on nonverbal communication modes of robots and categorized them into four main classes of *kinesics*, *proxemics*, *haptics*, and *chronemics*.

The term "Kinesics" refers to the systematic studies of how human beings communicate using body movements and gesture [16]. "Proxemics" is the knowledge of people's perception and use of space and refers to the conscious or subconscious setting of distances between various items, agents, and the person [46]. "Haptics" is related to tactile communication [40], through which signals would be transferred to people using their skin. This investigates the impact of robots' tactile behaviours (e.g., handshaking and gentle touching). Finally, "Chronemics" is defined as studying the nonverbal communication mode which is related to human time-experiencing [22]. In HRI, this is about how robots' nonverbal timing, mainly hesitation gestures, influences humans. All these nonverbal communication modes have been shown to impact the interaction in the domains of shaping cognitive framing (i.e., people's perspectives and orientations toward the robot), eliciting different emotional responses, triggering behavioural reactions, and adjusting the task performance.

2.2.1 Nonverbal Behaviours Influencing the Interaction

Robots' nonverbal behaviours influencing the HRI and possibly conveying information have been studied on a wide scale. Claret et al. [27] employed a robot's body motions to convey emotional information while it was simultaneously greeting as its main task, by waving its right hand. They developed an approach for mapping emotions defined as points in the *Pleasure-Arousal-Dominance* space to the *Jerkiness-Activity* (related to the kinetic energy of the robot)-*Gaze directness* kinematic attributes, which eventually shaped the continuous body movements of the robot. With an experimental study, they examined whether a Pepper robot with the proposed approach could accurately convey calmness, happiness, sadness, and fear. The results showed that the system was capable of conveying happiness and sadness, but more work was needed on conveying calmness and, more importantly, fear. As they pointed out, people seemed to not notice the jerkiness, particularly when the robot was making energetic motions. This miscommunication of the jerkiness attribute could explain why emotions such as fear sometimes interpreted as happiness.

In [27], the chosen set of attributes (i.e., *Jerkiness*, *Activity*, and *Gaze directness*) were motivated by the work of Glowinski et al. [43], who proposed four dimensions for retaining a majority of emotional information: (a) *Activity*, which indicates how energetic the motions are, (b) *Temporal and spatial excursion of movement*, to show the distribution of energy along with the physical motions, (c) *Spatial extent and postural symmetry*, which indicates the degree of arms' openness, and finally (d) *Motion discontinuity and jerkiness*, which means having a trajectory with large values of high-order derivatives. In addition to that work, we can also note the study of Dragan and Srinivasa [34]. They have discussed and tested a method of generating motions to communicate a robot's intent to the observer, and to increase the predictability of its action, when the robot was attempting to reach for one of the two goals objects.

In the domain of robot learning, nonverbal communication has also been found valuable. Chao et al. [25] have equipped a humanoid platform with an active learning system that provided nonverbal feedback to its teacher, containing information about the areas in which it got confused, when humans taught it different Tangram symbols. Two kinds of transparency mechanisms were included in that system: one for informing about the points of confusion, and the other one for communicating the robot's confidence. Nodding and shaking of the head, changes in the colour of the ears and shrugging gestures were used as the cues. It was found that people who understood the robot's intentions could teach it more accurately.

Lohan et al. [89] have investigated the effects of a robot's online behavioural feedback on the tutor's behaviour. In that study, the participants were presenting the patterns and the colours of some boxes to an iCub robot. When at the beginning of the teaching process, the robot was closely monitoring and following the teachers' presentation and responded to their gaze behaviour and pointing gestures appropriately, the teachers appeared permissive to the robot's faults later. These results suggested that robots can tailor the tutor's actions by interacting with them and providing feedback. More recently, Huang et al. [58] have utilized a robot's eye gaze to signal the teachers how it expects them to perform next, when teaching a decluttering task to it. The researchers examined how humans can interpret the robot's gaze feedback, which was related to its beliefs about the sorting rules, and potentially benefit from that. Only 9 out of 17 participants were able to discover this relationship, but for those who did, the feedback helped to better track the robot's understanding and therefore, increased the effectiveness of their teaching.

As an alternative to the scenario of participants teaching a robot, Peters et al. [88] had an educational robot as a lecturer for a small group of students. By varying the pitch and volume of its voice, body postures, and types of hand gestures (i.e., open or closed, and semantic or syntactic), researchers made different nonverbal behaviour patterns. It was found that even a small change in the robot's behaviour can alter the perception of its warmth and competence.

2.2.2 Adjustments in Robots' Nonverbal Behaviours

- **Eye gaze:** A robot's gaze behaviour, meaning to where and for how long the robot looks, is a natural way to express emotions. There is a large body of research indicating the importance of gaze behaviour in HRI. In this regard, an extensive review can be found in [2]. To mention some examples about when humans taught the names of some objects to a robot, Ito et al. [59] found that a robot looking mostly at the moving objects and occasionally establishing a mutual gaze with the teacher could promote the teacher's attention and deliver a feeling of intentionality. Nevertheless, this type of behaviour needed some gaze aversions that have been shown to convey the feeling of distrust [84] and introversion [7].

- **Arm movements:** Another way to express feelings is through gestures. Studies have indicated that by performing the actions with varying motion parameters, a robot can convey emotions such as sadness and anger [36], alter people's perceptions of its liveliness and activity [102], express a feeling of kindness or rudeness [116], and impact the perceived affect [100]. Kim et al. [67] have investigated the relationship between basic motion factors, including the speed of the arm movements, and some perceived personality aspects of an entertainment robot. They found that people perceived higher dominance and greater friendliness when the robot showed fast motions. Other researchers have examined both the subjective and the physiological responses of people when observing a robotic arm that used multiple trajectory generation approaches with different speeds [70]. They detected that higher speeds resulted in higher perceived anxiety and surprise.

- **Hesitations:** If humans feel doubt and uncertainty before or while doing a task, they often pause and think briefly to make a decision. A number of HRI studies have explored hesitant robots' behaviours. Such pauses have been found to be a way of conveying uncertainty in human-robot shared tasks [79]. Moon et al. [80] have focused on the use of hesitation gestures, as a nonverbal communication tool, for the situations when both a robot and a human need to reach for one object at the same time. By applying a smooth human-like motion profile for the hesitant behaviour, the robot was perceived as less dominate, but more likeable, animate and anthropomorphic in comparison to when there was no pause. However, they found no significant differences in those measures when the robot responded by stopping immediately instead of acting smoothly.

In the context of robots learning new actions, a study has been done on the effects of adding delays in initiating the actions of the robot after receiving the instructions from the human teacher [64]. When the robot was uncertain about its actions, it initiated them with hesitation, but when it was confident, it started immediately. The delay improved people's overall teaching experience and accelerated the learning process. In that study, adding longer delays at the beginning of learning process and shorter ones at the end made people perceive the robot as even more teachable.

Conclusion: As described, several works have been done concerning the impact of nonverbal robot feedback on human teachers. These studies have mostly investigated situations of learning abstract and theoretical tasks (e.g., puzzle-solving [25, 58]) rather than learning how to perform a physical task, when a series of actions is being done on some objects or the environment. Context is highly important for determining a robot's personality [91] as well as in identifying the interpretations of motor activity, to infer intent [82]. Therefore, studies presented in this thesis aim to fill the knowledge gap by focusing on the interpretations of robot' actions when learning physical actions in particular.

2.3 Trust in HRI

Trust, can be described as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [72, p. 51], and is a complex relationship that can be impacted, destroyed or recovered, depending on many factors. Following is a summary of an HRI research project related to trust.

Rossi et al. have done a series of studies to investigate how humans can trust robots in home environments [93]. The primary step of that research involved assessing the severity of a set of potential mistakes that a domestic robot could commit. A group of participants who were asked to imagine themselves living with a robot companion in their homes rated those mistakes [95]. Researchers identified three big and three small errors by selecting only the highest and lowest ratings from each participant. In the next step, they studied how the timing and the severity of a robot's mistakes may affect trust [94]. In a virtual storytelling environment, an experiment was conducted with the following five conditions of a robot doing successive tasks, as demonstrated in Figure 2.2: (C1) *Performing them* *all correctly*, (C2,C5) *Making three big/small mistakes (determined by the previous step) at the beginning and also at the end*, (C3,C4) *Exhibiting thee big/small errors at the beginning and three small/big ones at the end*. Participants were then asked if they would trust the robot and follow its instructions in an emergency evacuation scenario. A majority of participants did not trust the robot in case there were large errors at the beginning or at the end of the scenario. With small errors instead, however, participants tended to trust it in teamwork. It was also found that with making big mistakes at the beginning of the interaction, rather than at the end, trust was impacted more negatively. The participants who had a higher disposition for trusting in people's benevolence trusted the robot more according to the collected questionnaires [96]. In addition, people who were more open to experience were seen to have a negative attitude toward having a robot as a companion.

In another study, Rossi et al. [98] investigated how the interaction history and people's knowledge about a robot's capabilities could affect user's trust. In an experiment, secondary school students first watched a commercial video about the Pepper robot. In the next step, they freely interacted with the robot while its built-in awareness function was running. Finally, students actually became involved in programming different behaviours and emotions for it. According to questionnaires collected after each step, as students learned more about the robot's capabilities and limitations and also got hands-on experience in programming the robot, they could trust the robot more. Later, while repeating the same experiment with a Kasper robot, the trust did not significantly change after different interaction scenarios [99]. Compared with the Pepper robot, the Kasper robot may have seemed totally different in those experiment, since its facial features were minimal, its size was small and it was unable to walk. Lastly, the researchers discovered the relationship between social behaviours of a robot and human's trust. In an experiment [97], when a Pepper robot guided the participants in a crowded environment and acted socially, it could gain people's trust which was not existed before cooperating with the robot. This robot was perceived as more comfortable and was evaluated more as an assistant, rather than a machine.

2.3.1 Consequences of Robots' Errors

In order for the robots to operate in real environments that are unstructured and dynamic, they should collect information with sensors that can sometimes be noisy and inaccurate. The robots then need to make their own decisions and act accordingly. Because of all these factors, robots may behave in an erroneous manner in some situations. Errors in the behaviours of the robots have been shown to reduce their perceived trustworthiness [31, 103], the effect of which may be different depending on the severity and timing of the

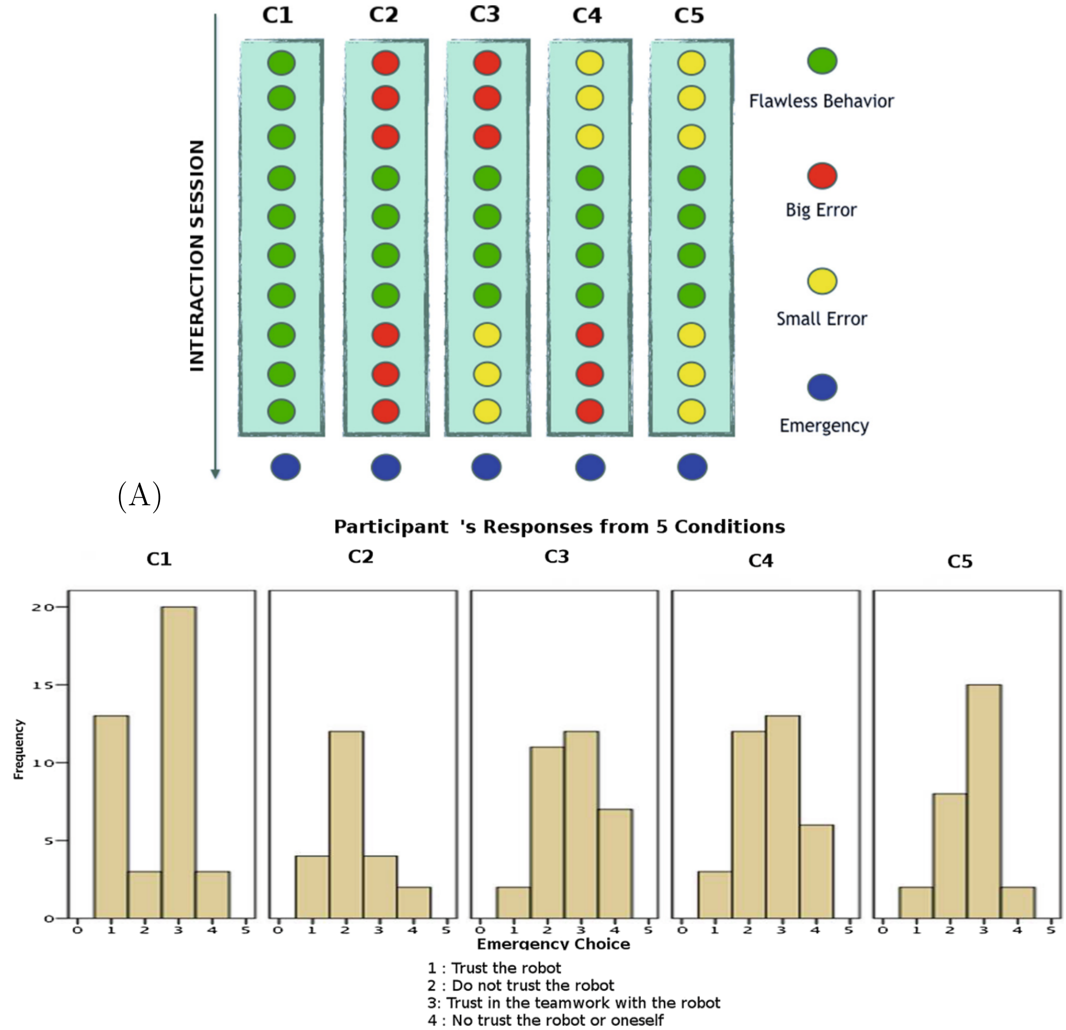


Figure 2.2: (A) Experimental conditions, and (B) Responses of participants from different conditions to the emergency scenario in a study of trust in a virtual home environment. Adapted from "How the Timing and Magnitude of Robot Errors Influence Peoples' Trust of Robots in an Emergency Scenario" by Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters, 2017 [94]. Used with permission.

mistakes, as described in the previous section [94]. A taxonomy of failure types in HRI and their impact on trust has been recently developed in [111]. In [55], there can be found a review about how robots communicate failure and how failures affect people's perception of robots and their feelings towards them. Other consequences of a robot's mistakes have been detected by behavioural analysis in several studies. The following provides a summary of a few studies of this kind that are related specifically to learning interactions.

In [50], the participants taught a kind of dance to a humanoid robot that was designed to repeat any recognized actions of its teacher. Throughout the interaction, there were instances with both correct and intentionally incorrect robot behaviours, as response to the presented material when the robot was mimicking participants' actions. During the times when the robot's actions went right, the participants often smiled and nodded along. However, negative feedback such as frowning and averted gazing was generated after incorrect robot's behaviours.

Using the kinesthetic teaching method, participants in another study demonstrated a manipulation task to a robotic arm [109]. Following teaching, the experimenters ran a program with an error for them to observe, instead of running each participant's flawless trajectory. The errors were *missing the object* (low severity), *placing it in an inaccurate location* (medium severity), or *dropping it when moving* (high severity). In assigning those severity levels, the effect of each error on the immediate surroundings was considered. Researchers analyzed the timing and the intensity of teacher's behaviours, such as laughing and smiling, and recorded their verbal reactions as well. Faster and more intense responses were detected from the participants who experienced more severe robot errors.

In the opposite situation, Kontogiorgos et al. [69] have studied the effects of type of embodiment and the severity of failures when a conversational agent was instructing participants how to cook. High severity mistakes were simulated with introducing a time constraint for completing the task. Researchers found that the participants responded more intensely to the failures, using verbal and nonverbal signals, when they interacted with a human-like embodiment compared with a smart speaker. However, there was no big effect of failure severity on the detected behavioural signals in that study.

Concerning a manufacturing setup where a supervisor (the experimenter) was also present near a robotic arm, Hedlund et a. [51] have recently conducted an experiment to study human teachers' response to robot failures after learning a manipulation task by different kinds of demonstrations. Authors found that robot failures, simulated through playing a pre-recorded faulty trajectory, affected participants' trust of themselves and the robot.

Conclusion: In most of the reviewed works about robots' errors and their impact on trust, participants were interacting with a robot that was supposed to know how to perform a task and was expected to work properly. Teachers are often forgiving about errors made by their students. In HRI, there is also evidence that lower expectation of a robot's functionality can minimize the negative effects of its errors on trust [119]. Therefore, in our study that is related to trust, we follow an approach similar to previous literature to explore a "learning" situation when errors are being made by a student robot.

2.4 Impact of Robot Appearance in HRI

The way in which humans dress has been shown to influence the formation of first impressions [104]. In particular, the clothing styles of students in a school environment can affect teachers' perception of their intelligence and potential academic achievement [11]. In HRI, the appearance of a robot may provide information on its abilities and competencies. Many studies on the impact of robot appearance on different aspects of humans' perceptions and the interaction have focused on the level of machine-likeness as a design characteristic of the robot (i.e., being anthropomorphic or zoomorphic vs. machinelike) [12, 117, 90, 110, 44].

The appearance of an agent can be related to the notion of perceived authority. People in everyday life tend to comply with requests from those they perceive to have authority. Haring et al. have studied this in HRI [48]. They explored whether the human-like appearance of a teacher robot could prompt the participants to practise the task longer, compared to a less human-like robot. In this context, however, no differences were found between the higher and the lower human-like robots concerning compliance time.

Conclusion: Our work in study 2 aims to examine the use of the same robot with two different clothing styles and is not directly related to the reviewed literature. Therefore, the study of appearance factor would be exploratory.

Chapter 3

Study 1: Effects of Robot Nonverbal Behaviour

3.1 Research Question and Hypotheses

The research question in this study is as follows:

RQ How do nonverbal aspects of a humanoid robot's behaviour (i.e., different types of gaze and arm movements) influence the way human teachers perceive a trainee robot in terms of confidence, calmness, proficiency in the task, attention, eagerness to learn, being goal-driven, and liking the task?

Informed by the the background literature surveyed in Section 2.2, we hypothesize that:

H1 When the robot always looks at the manipulated objects (i.e., follows its moving hand), teachers will perceive it as being more attentive to the task.

H2 It will also appear to like the task more in the previous case.

H3 If the robot never looks at the task and stares only towards the teacher, it will appear more proficient.

H4 If the robot does not look at the task at all and stares only towards the teacher, it will appear more attentive to the teacher.

H5 By looking mostly at the actions and also checking the teacher occasionally, the robot will appear to be more goal-driven (intentional), suggested by [59].

H6 The previous robot will appear to be more willing to learn as well.

H7 We expect a robot performing the actions faster to convey signs of proficiency.

H8 We expect a robot performing the actions faster to convey signs of confidence.

H9 We expect a robot performing the actions faster to convey also nervousness, as [70] suggested.

H10 Performing the actions slowly will convey the impression of attention to the task.

H11 In the present study, it is likely that including hesitations (i.e., having long pauses with low frequency) would convey signs of uncertainty, as discussed by [79].

H12 Jerky movements (i.e., including short delays with high frequency) would cause the robot to appear more nervous.

H13 Regarding the effect of time passed since training (i.e., the priming factor), we can anticipate teachers' ratings in terms of some attributes such as perceived confidence and calmness to be higher when the robot has learned the task longer ago in the past, compared to a robot that is practising it for the first time. However, study of a priming factor in our experiment is widely exploratory in nature, rather than testing a previously identified hypothesis.

3.2 Methodology

To address the research question outlined above, an online experiment (Experiment 1) was conducted on Amazon Mechanical Turk (MTurk) crowdsourcing platform[1] in which we systematically varied a set of nonverbal factors in the behaviour of a virtual humanoid robot while it was performing an organizing (object manipulation) task with a fixed trajectory: the robot put several objects inside a box located on a desk, and then slid the box to the centre of the desk.

The experiment employed a $2 \times 3 \times 4$ mixed factorial design, where we manipulated three factors (i.e., priming based on time, robot's gaze, and robot's arm movements) and showed

[1] https://www.mturk.com

different videos of the robot practising the task to the participants. The first factor, priming regarding the time passed since teaching occurred, was a between-participants factor and the other two were within-participants. Robot's gaze had three levels of *Looking only at the teacher*, *Looking only at the task*, and *Combined*. There were also four conditions of arm movement: *Low-speed Smooth*, *High-speed Smooth*, *Hesitant*, and *Jerky*.

We showed a total of 12 videos including all forms of gaze and arm movements in randomized order to two groups of participants. The participants received different priming information about the amount of time passed since teaching, provided in the form of text in the instructions. After watching each video, the participants were asked to rate how they perceived different aspects of the robot, which will be discussed in the experimental procedure section. Figure 3.1 shows snapshot images of the robot performing the task.

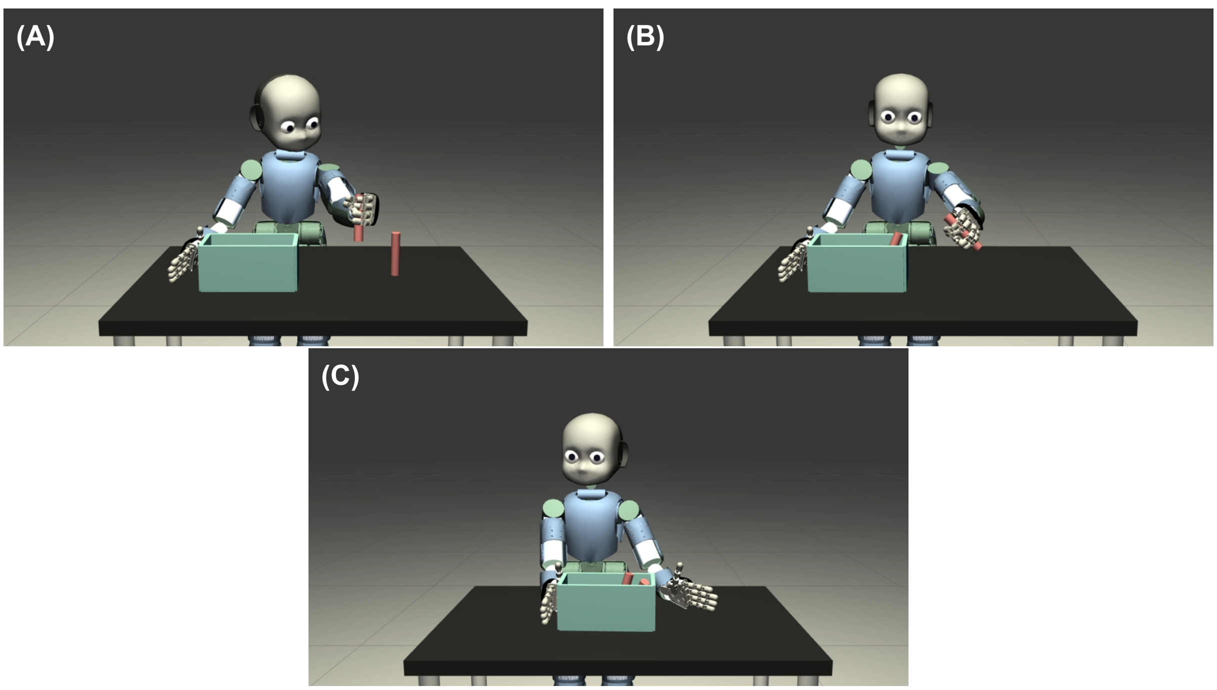


Figure 3.1: The virtual humanoid robot in the Gazebo simulation environment. The robot is standing behind a table and tries to carry out the task of putting two objects inside a box and then sliding it to the middle of the table, with different gaze and arm movements conditions. Snapshots from three separate videos of performing the actions with different gaze states are shown: (A) Picking the first object while gazing at the task entirely, (B) Moving the second object to the box with a gaze fixed at the teacher (participant), (C) Sliding the box while establishing a quick and temporary mutual gaze with the teacher.

The chosen factors were derived from the literature and were found suitable for the specific scenario we focused on. For instance, varying the speed and smoothness of the motions was motivated by the work of [43], as described in Section 2.2.1.

The manipulated behavioural parameters of the robot (i.e, eye gaze and arms movements) fell into the categories of kinesic and chronemic under different modalities of non-verbal communication [106]. Proxemics and haptics were not considered in our learning scenario (e.g., the robot did not move closer to the teacher) since parameters such as distance may not be modifiable by the robot in such cases of task practice. Also, body movements other than those relevant for implementing the study (e.g., idle movements or affective facial expressions of the robot) were not manipulated in order to focus specifically on the interpretation of the robot's "physical actions" that were experimentally manipulated, and to avoid possible confounds.

3.3 Experiment 1

3.3.1 Robotic Implementation

We used a virtual iCub [2] humanoid robot in a Gazebo simulation environment (See Figure 3.1) [3]. iCub is a humanoid platform suitable for research in embodied cognition, and for experiments in HRI and robot learning in particular [78]. iCub is capable of performing the desired task by taking advantage of various Degrees of Freedom (DOFs) embedded in its upper body, along with the sophisticated hands' design. Each hand of iCub has 9 DOFs with the ability to move thumb, index, and middle fingers independently. The two other fingers move as a single degree of freedom for stability. The simulated iCub operates very similarly to the real robot, using the YARP middleware [76].

As stated before, the designed task for the iCub to present in front of the *teachers* was identical in all the variations of the videos. The task was to organize a desk. The robot always stood behind a desk and first put two small spherical cylinders inside an open box, one by one, and then slid the box to the centre of the desk. We made necessary modifications to some physical properties of the objects in the simulated environment to make the robot capable of safely picking up and moving the objects, without dropping them. Also, we set all the colours to more washed-out blends within the same tone, to avoid the pop-out effect [73] (i.e., directing participants' attention to particular areas).

[2]Visit `https://icub.iit.it` for technical description.
[3]Available at: `https://github.com/robotology/icub-gazebo`

Two scripts written in C++ and running on the YARP were developed to control iCub's arm movements and gaze behaviour. Within the first script [4], we defined some way points to generate fixed-trajectory task actions on objects of the environment, with the capability of varying the speed and the type of pauses to reach the desired levels of the arm movement factor. In our simulations, *High-speed Smooth*, without any pauses, were 2.5× faster than *Low-speed Smooth*, for their difference to be easily noticeable by the viewers. *Hesitant* motions were implemented by adding long, low frequency delays, with a duration from 1 to 3 seconds, amid the defined way points on a random basis. *Jerky* movements were also achieved by dividing the moving paths into multiple parts and adding short, high frequency (0.2-second) delays between each part.

The second script [5] actively controlled the gaze behaviour of iCub through adjusting its head and eyes positions. In order for the robot to follow its hands in the gaze mode that required *looking only at the task*, we incorporated forward kinematics. The final target positions were transferred to a spherical coordinate system positioned at its neck. This required a shift in the z component of the position by about $50cm$ first, and then calculating θ and ϕ angles. The position of neck yaw was set continuously according to the calculated ϕ angle to make the vertical gaze adjustments. To have a more life-like gaze behaviour, we incorporated adjustments in both neck pitch and eyes tilt DOFs to achieve the target position horizontally in θ angle. We set a limit of $\pm 20°$ for the neck pitch and the rest of θ was set for the eyes tilt. In the *looking only at the teacher* gaze condition, the attention target of the robot was fixed to the camera (viewer). In the *combined* gaze mode, while mostly looking at the task-related objects, the robot looked at the camera at some random moments and maintained this mutual gaze for 1 second. Then, it rapidly shifted its gaze back to the task.

3.3.2 Procedure and Measures

The participants followed three steps in our designed online HTML interface, after reading the information letter and providing consent as shown in Figure B.1 (Appendix B). Figure 3.2 summarizes these steps.

Step 1 - Demographics questionnaire: First, a demographic information form including questions about age, gender, level of education, and cultural background was completed by each participant (Figure B.2). They had the choice to skip any of these

[4]Based on `tutorial_arm_joint_impedance.cpp` available at: https://github.com/robotology/icub-tutorials

[5]Using `iKin` library: https://github.com/robotology/icub-main

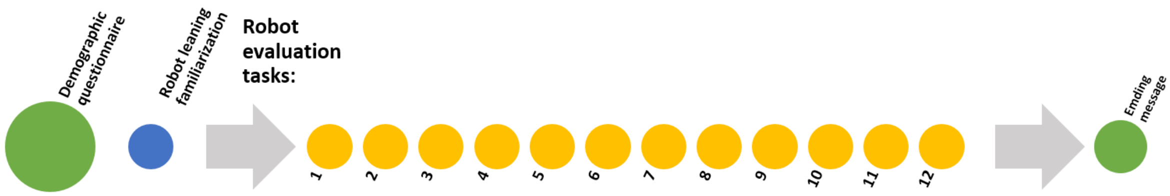


Figure 3.2: Overview of the experimental procedure adopted in Experiment 1.

questions. For gender and cultural background questions, text entries were provided to allow freely submitting different answers.

Step 2 - Robot teaching familiarization: We included a robot teaching familiarization section in two stages, mainly to illustrate how people can teach real robots, and how robots will perform the learned actions afterwards. In the first stage (demonstrated in Figure B.3), we showed two videos with a brief explanation about intelligent robots that are able to acquire new skills (pushing/pulling objects and cleaning a table) when someone grabs their arms and demonstrates how they may perform the task [66, 3]. This was to support the written instructions in the evaluation step to better convey some sense of teaching.

The second stage of robot teaching familiarization was to present the participants with some ideas on what they will experience in this study, again through a video and some texts (Figure B.4). This video was created by merging short sections showing our simulated iCub robot doing the designed task with different cases of behaviours and from various points of view. This enabled participants to explore the simulated world to become more familiar with it, with the goal of reducing the novelty effect. In addition, including different types of robot behaviours in a single video indicated to the participants that there could be variations in the performance of the robot. Finally, by seeing what the robot is going to perform, we hoped that participants could understand the context more easily. This step was similar for all the participants.

Step 3 - Robot evaluation task: After the robot teaching familiarization step, and depending on the priming group each participant was randomly assigned to, instructions appeared. We told both groups "assume yourself in the situation of being a teacher for a robot", similar to what they watched earlier. Participants in group 1 were further asked to "*assume you have just taught iCub how to organize a desk in the way you saw earlier. Now, we will show you some videos of the step immediately after you have given* *your instructions, when the robot is practising the task in front of you for the first time*". Whereas, participants in group 2 were told "*assume you have previously taught iCub how to organize a desk in the way you saw earlier. Now, teaching has been completed and we will show you some videos of the robot executing the task two days later, to help you*". This page is shown in Figure B.5.

After reading the instruction, there were a total of 12 consecutive videos showing the robot performing the same task with various combinations of gaze and arm movement conditions, as they were both within-participants factors and all the participants were exposed to all the possible combinations. The videos were shown to people in a random order to reduce the order effect. After each video was finished, a list of nine continuous sliders appeared (see [112, 74, 41] for advantages of using continuous scales), to present the questionnaire. Order of questions were counterbalanced. Participants were also given the option of watching the video again. Figure B.6 shows an example image from what participants experienced in this step, while seeing any of the videos.

In the questionnaire, Table 3.1, eight sliders were used for the rating of each dependant measure and one for an attention check question. Although several standard questionnaires exist for HRI studies, for our particular domain in which a robot is presenting physical actions to a teacher, and for the specific role of the robot as a student, the list in Table 3.1 appeared to be relevant and efficient. Some of these items were adapted from previous related studies [59, 70, 64, 79], as well as by looking at the Big Five personality traits questionnaire [8]. These measures were used to explore how the perception of a student robot beyond its teachability may be affected by the robot's nonverbal behaviour.

We added attention checks in the questionnaire. The attention checks were selected randomly from a set of questions listed in Table 3.2. Each question was used in two videos. These were either one of the original questions repeated with a different wording (type 1, for consistency check), or new questions with obvious answers (type 2, for attention check). The directions of each label on the sides of sliders were randomly swapped. However, for type 1 attention checks, the labels in the pair of the original question and the repeated one were always in the opposite direction.

Note that all the videos embedded within this online interface were set to automatically pause in case anyone switched to another window (indicating that they were not paying attention).

After viewing and completing the rating of the 12 videos, an end page including a unique code for submitting the task was shown to the participants (see Figure B.7)

Table 3.1: List of dependant measures. The questionnaire shown in this table appeared with a random order of the items each time after a different video was played. The directions of the continuous scales were also inverted randomly. The attention check item was randomly selected from those mentioned in Table 3.2.

	How do you rate the behaviours of the robot?		
1	Very confident	...	Not confident at all
2	Relaxed	...	Nervous
3	Likes the task	...	Does not like the task
4	Very attentive to the task	...	Not attentive to the task at all
5	Very proficient	...	Not proficient at all
6	Eager to learn	...	Unwilling to learn
7	Goal-driven	...	Random
8	Very attentive to the teacher	...	Not attentive to the teacher at all
9	(Attention check)		

3.3.3 Participants

A total of 197 participants completed the experiment on MTurk. To acquire more reliable responses, we limited the availability of our task only to people with greater than 97% approval rate, who had been approved in at least 100 tasks before. Besides, we made the study visible only to those in Canada and US, as the participants needed to understand English. Participants were rewarded 2 USD for spending 25-30 minutes completing the study. Three participants did not fully complete the study and their data was omitted. The study received full ethics clearance from the University of Waterloo's Research Ethics Committee (See Appendix A, #42177).

After collecting all the answers, we discarded the data collected from the participants who failed at least 3 of the 12 attention checks. This left 167 participants. 83 participants (51 men, 32 women, $Min_{age} = 21, Max_{age} = 68, M_{age} = 37.39, SD_{age} = 11.09$) were randomly assigned to group 1, and the other 84 participants (48 men, 36 women, $Min_{age} = 18, Max_{age} = 65, M_{age} = 36.31, SD_{age} = 10.03$) were assigned to group 2.

Table 3.2: List of attention check scales. Each of the mentioned choices were selected randomly for two videos.

Type 1:		
Very self-assured	...	Not self-assured at all
Similar to (Very confident / Not confident at all) question.		
Calm	...	Anxious
Similar to (Relaxed /Nervous) question.		
Very skillful	...	Not skillful at all
Similar to (Very proficient/Not proficient at all) question.		
Type 2:		
Does the task successfully	...	Drops the objects several times
Animated	...	Immobile
Moves the box	...	Does not move the box

3.3.4 Statistical Analysis

For all the eight dependent measures, we used Linear Mixed-effects Model (LMM) to check if significant effects of the manipulated factors exist [10], while taking possible confounding factors into account. The independent variables were the two factors of gaze and arm movements with repeated measures (within-participants factors) in addition to the group that the participants were assigned to, as a between-participants factor. In all, a random effect was fit based on participant, along with the three aforementioned fixed effects. We also controlled for gender, age, educational level, and duration of the videos. Any of these factors were kept in the final models if they improved model fit based on the Akaike's Information Criterion (AIC) criterion [18]. Furthermore, to ensure that randomizing the order of videos worked well in reducing the order effect, we also included the order in which the videos were seen (log-transferred due to the nonlinear nature of learning effect) in the models. Except for two metrics of the robot's eagerness to learn and attention to the task, the order in which the videos were played had no significant effect on the ratings. The p-values for assessing significance were adjusted using the Holm-Bonferroni method to account for multiple hypotheses and avoid Type 1 error.

3.3.5 Results

While conducting the statistical analysis using Linear Mixed-effect Models, we detected no main effect of priming (i.e., the contextual clues based on time passed since teaching occurred) on any of the eight measures. Among all the models, this factor only improved the model fit for the model related to the perception of the robot being goal-driven. Although including this factor improved the model, its effect was not significant ($se = 27.037, t = -1.848, p = .066$). Thus, we pooled the data from these two priming groups.

The remainder of this section describes the findings related to each criterion, based on the answers of 167 participants. The statistical modelling results for all the measures are shown in Table 3.3 and 3.4. The estimates referred to in these tables are compared to the defined baseline levels: (a) *combined* gaze and (b) *low-speed smooth* arm movements. Empty cells for certain factors mean that they were not included in the model (i.e., the best LMM that was found for predicting that measure).

Perceived Confidence of the Robot. Figure 3.3,(A) shows average ratings of participants' perception of the robot's confidence. The only factor that was detected to significantly affect this measure was the arm movements. Confidence was rated significantly higher for high-speed smooth movements compared to low-speed smooth movements ($se = 12.39, t = 10.335, p < .001$). This showed that fast motions conveyed a sign of confidence. On the other hand, including hesitations and jerkiness decreased the average confidence score of the robot ($se = 12.39, t = -10.328, p < .001$ and $se = 12.39, t = -9.091, p < .001$, respectively). We did not find any significant difference between the effects of hesitation and jerkiness on the robot's perceived confidence ($se = 12.39, t = -1.237, p = .219$).

Perceived Calmness of the Robot. In this measure, similarly, significant differences between the low-speed smooth arm movements and the other conditions of arm movements (i.e., high-speed smooth, hesitant, and jerky) were observed. While increasing the speed led to the robot to appear more relaxed to the participants ($se = 13.08, t = 8.979, p < .001$), introducing hesitations and jerky motions reduced the perceived calmness ($se = 13.08, t = -11.237, p < .001$ and $se = 13.08, t = -13.635, p < .001$, respectively), as Figure 3.3,(B) illustrates. The difference between the hesitant and jerky levels was significant ($se = 13.08, t = 2.398, p < .05$) with jerkiness in the motions contributing the most in conveying the impression of being nervous. However, the robot was perceived the most relaxed with high-speed movements.

Table 3.3: Linear Mixed-effects Models predicting the robot's attributes. Besides two factors of gaze and arm movements, gender, age, educational level, order, and video duration were entered and kept in case they improved AIC in some cases. Blank cells mean that the factor was not included in the model. A random effect was fit based on participant.

Covariate	Confidence			Calmness			Liking the task			Attention to the task		
	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t
Gaze												
Combined [b]												
Gaze at Teacher	-2.02	10.73	-0.188	-3.28	11.33	-0.290	-53.65	8.60	-6.236 ***	-125.74	10.16	-12.374 ***
Gaze at Task	16.28	10.73	1.517	29.05	11.33	2.560 *	3.11	8.60	0.361	23.36	10.16	2.298 *
Arm movements												
LS Smooth [b]												
HS Smooth	128.07	12.39	10.335 ***	117.42	13.08	8.979 ***	59.49	9.93	5.989 ***	41.71	11.73	3.555 ***
Hesitant	-127.98	12.39	-10.328 ***	-146.96	13.08	-11.237 ***	-72.91	9.93	-7.340 ***	-41.85	11.73	-3.566 ***
Jerky	-112.66	12.39	-9.091 ***	-178.32	13.08	-13.635 ***	-68.05	9.93	-6.851 ***	-20.55	11.73	-1.751
Age	1.71	1.07	1.594	2.64	1.05	2.514 *				1.98	0.92	2.165 *
Order										-29.77	5.74	-5.187 ***

* = significant at $p < 0.05$; *** = significant at $p < 0.001$; [b] = baseline level
LS Smooth = Low-speed Smooth; HS Smooth = High-speed Smooth

In terms of different gaze behaviours, by looking only at the task, the robot was perceived significantly more relaxed than cases with the combined gaze ($se = 11.33, t = 2.560, p < .05$) and also significantly more relaxed than when the robot was gazing only at the teacher ($se = 11.33, t = -2.855, p < .05$). Additionally, we detected a significant effect of participant age on this perception ($se = 1.05, t = 2.514, p < .05$): as age increased, the robot was perceived to be more relaxed.

Robot's Perceived Liking the Task. We found significant main effects of gaze and arm movements, according to the LMM used for predicting this measure. The robot looking only at the teacher appeared to like the task less than a robot with a combined gaze ($se = 8.60, t = -6.236, p < .001$). However, there was no significant difference between the combined mode and the robot only looking at the workspace ($se = 8.60, t = -0.361, p = .718$). Additionally, similar to the perceived confidence, the robot was perceived to like the task significantly more when it showed high-speed smooth movements, as compared to having low-speed smooth movements ($se = 9.93, t = 5.989, p < .001$), while hesitant ($se = 9.93, t = -7.340, p < .001$) and jerky ($se = 9.93, t = -6.851, p < .001$) movements reduced the ratings without any detectable difference between them ($se = 9.93, t = -0.489, p = .625$). Figure 3.3,(C) shows the discussed trends.

Perceived Robot's Attention to the Task. According to the LMM fit on the data, there is a significant effect of gaze on this attribute. When the robot was looking only at the teacher, the attentiveness to the task significantly dropped compared to looking at both the teacher and the task ($se = 10.16, t = -12.374, p < .001$). The highest average score in this measure was for the gaze at the task condition when the robot's attention was rated significantly higher compared to both combined gaze ($se = 10.16, t = 2.298, p < .05$) and gazing only at the teacher ($se = 10.16, t = -12.374, p < .001$). Figure 3.3,(G) shows these differences.

Arm movements also had a significant effect on this measure; with high-speed smooth motions, the robot was perceived significantly more attentive to the task, compared to the low-speed smooth movements ($se = 11.73, t = 3.555, p < .001$). In the hesitant mode, the robot was perceived significantly less attentive to the task than in the low speed mode ($se = 11.73, t = -3.566, p < .001$). The age of the participants was another factor that significantly affected this rating ($se = 0.92, t = 2.165, p < .05$). As participants' age increased, the robot was rated to be more attentive to the task. Interestingly, here we detected a significant order effect in the ratings ($se = 5.74, t = -5.187, p < .001$), suggesting that as people proceeded with the study and watched more videos, they rated the robot as less attentive to the task.

Perceived Proficiency of the Robot. Figure 3.3,(D) shows participants' average ratings of the robot's perceived proficiency. A significant effect of arm movements was identified. High-speed smooth motions improved this measure significantly in comparison to low-speed smooth motions ($se = 10.69, t = 8.423, p < .001$). Having hesitant or jerky movements led to a significantly lower perception of the robot's proficiency ($se = 10.69, t = -9.930, p < .001$ and $se = 10.69, t = -7.098, p < .001$, respectively). In contrast to the perceived calmness, having hesitant movements led to the lowest proficiency ratings, which rated significantly lower than the jerky motions condition ($se = 10.69, t = -2.711, p < .01$). Perceived proficiency also increased significantly with participants' age ($se = 1.01, t = 2.098, p < .05$).

Robot's Perceived Eagerness to Learn. When the robot only gazed at the teacher, the perceived eagerness to learn dropped significantly compared to the combined mode ($se = 8.15, t = -7.405, p < .001$). It is also shown in Figure 3.3,(E) that the green points denoting average participants' perceptions of eagerness to learn for *always at the teacher* gaze lay below the nearby red ones that show this average related to the *combined* gaze. Looking only at the task also led to a rating of eagerness to learn that was lower than in the combined mode, but the difference was only close to being statistically significant ($se = 8.15, t = -1.792, p = .073$).

It also turned out that when the robot performed the motions with hesitation or jerkiness, it appeared significantly less eager to learn in comparison to performing the action slowly but smoothly ($se = 9.41, t = -3.832, p < .001$ and $se = 9.41, t = -3.395, p < .001$, respectively), with no statistically significant difference between hesitant and jerky movement conditions ($se = 9.41, t = -0.437, p = .662$). On the other hand, shifting to faster motions significantly improved the perception of being eager to learn ($se = 9.41, t = 5.108, p < .001$). Here, again, a significant effect of order was observed ($se = -4.60, t = -2.290, p < .001$), such that participants rated the robot less eager to learn when they watched the videos later in the experiment.

Robot Perceived as Goal-Driven. The results for this measure are presented in Figure 3.3,(F). The hesitant and jerky actions significantly reduced the score ($se = 9.32, t = -5.736, p < .001$, and $se = 9.32, t = -4.118, p < .001$, respectively). Faster motions significantly improved the score ($se = 9.32, t = 4.960, p < .001$) compared to the low-speed movements. No significant difference was observed between the hesitant and jerky

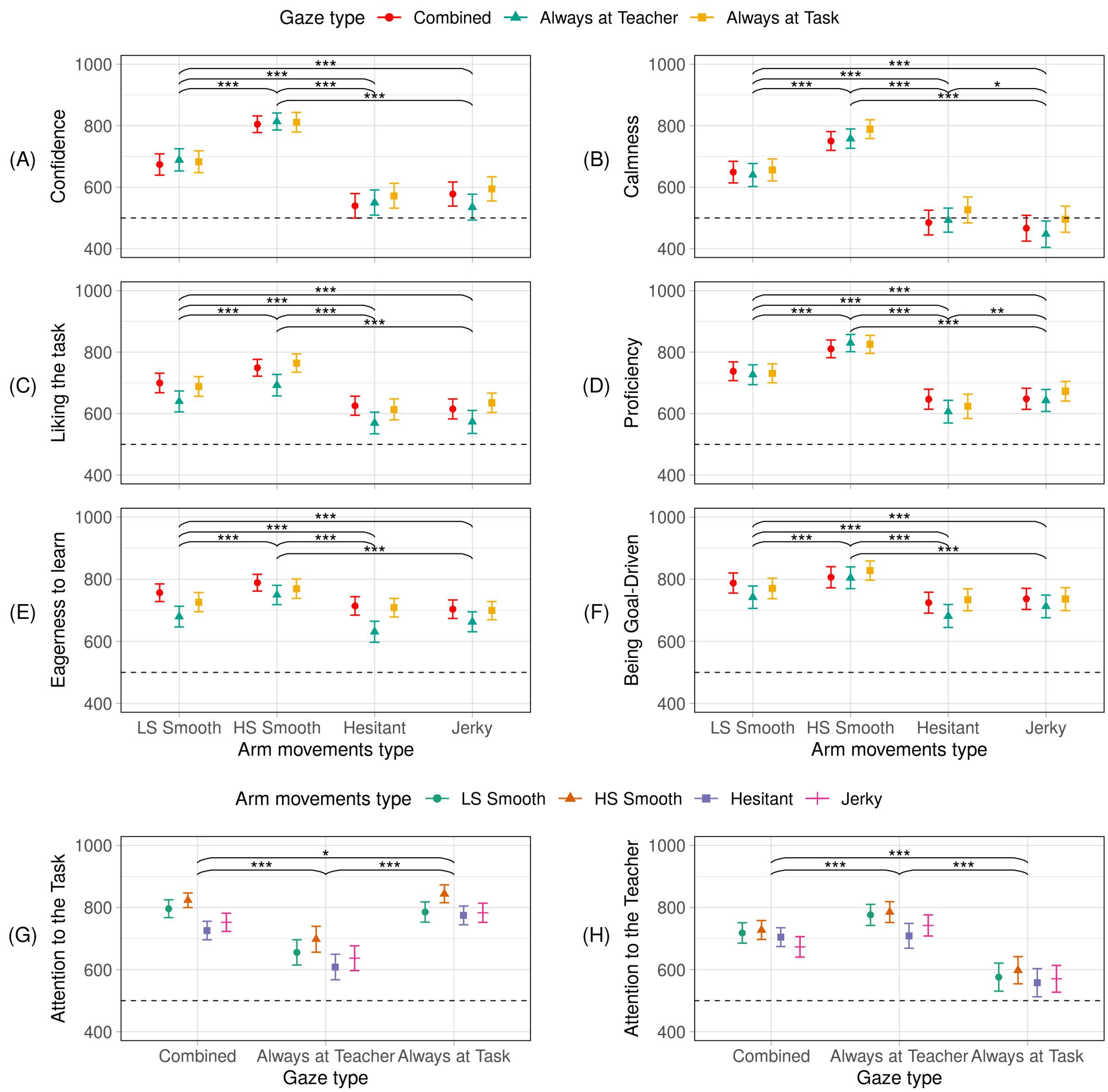


Figure 3.3: Averages of participants' perceptions of the robot's attributes. Error bars represent 95% confidence intervals and the results are pooled regardless of the priming condition. The dashed lines indicate the neutral choice at 500. * = significant at $p < 0.05$; ** = significant at $p < 0.01$; *** = significant at $p < 0.001$, across two groups.

Table 3.4: Linear Mixed-effects Models predicting the rest of robot's attributes. Apart from the two factors of gaze and arm movements type, gender, age, educational level, order and video duration were entered and kept in case they improved AIC in some cases. Blank cells mean that the factor was not included in the model. A random effect was fit based on participant.

Covariate	Proficiency			Eagerness to Learn			Being Goal-Driven			Attention to the Teacher		
	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t
Gaze												
Combined [b]												
Gaze at Teacher				-60.35	8.15	-7.405 ***	-28.60	8.07	-3.544 ***	63.86	14.29	4.469 ***
Gaze at Task				-14.61	8.15	-1.792	3.14	8.07	0.389	-105.46	17.32	-6.090 ***
Arm movements												
LS Smooth [b]												
HS Smooth	90.02	10.69	8.423 ***	48.07	9.41	5.108 ***	46.23	9.32	4.960 ***	622.56	321.07	1.939
Hesitant	-106.13	10.69	-9.930 ***	-36.06	9.41	-3.832 ***	-53.45	9.32	-5.736 ***	-221.20	100.01	-2.212
Jerky	-77.16	10.69	-7.219 ***	-31.95	9.41	-3.395 ***	-38.37	9.32	-4.118 ***	226.59	134.79	1.681
Age	2.12	1.01	2.098 *									
Order				-10.53	4.60	-2.29 *						
Video duration										33.22	17.5	1.898

* = significant at $p < 0.05$; *** = significant at $p < 0.001$; ^{b} = baseline level
LS Smooth = Low-speed Smooth; HS Smooth = High-speed Smooth

movements ($se = 9.32, t = -1.618, p = .106$). Also, gazing only at the teacher led to a significantly lower rating for being goal-driven as compared to the combined mode ($se = 8.07, t = -3.544, p < .001$).

Perceived Robot's Attention to the Teacher. As can be also seen in Figure 3.3,(H), the LMM suggested that the type of gaze was a determining factor for this measure. The robot looking only at the teacher led to the highest attention to the teacher score ($se = 14.29, t = 4.469, p < .001$, compared to the combined mode). Looking only at the task was perceived the lowest in terms of attention to the teacher, with a notable difference compared to the combined mode ($se = 17.32, t = -6.090, p < .001$). In this measure, a trend, approaching significance, was observed regarding an effect of the video duration ($se = 17.50, t = -1.898, p = .058$): as the length of the videos increased, people seemed to have perceived the robot to be more attentive to the teacher.

3.3.6 Discussion

The goal of this research was to study how variations in nonverbal behaviours of a trainee robot influence human teachers' perception of it. We showed that gaze patterns along with the speed, smoothness, and pauses in the robot's arm movements can in fact affect humans' perception of the robot. Table 3.5 summarizes our hypotheses and findings as will be discussed in the following.

Shaping human teachers' perceptions

In general, people often successfully evaluate someone's attention by their gaze direction, and this has been generalized to HRI previously [1]. Thus, it is not surprising that the robot in our study appeared to be more attentive to the task or to the teacher when it was exclusively looking at one of them, which confirmed **H1** and **H4**. However, the effects of gaze on other measures, such as calmness, liking the task, eagerness to learn, and being goal-driven are not as intuitive. While we hypothesized that the robot would appear to like the task more when looking only at the task (**H2**), we only found that looking entirely at the teacher can negatively affect this perception. Thus, we are unable to fully confirm **H2** as no statistically significant difference between combined gaze and gazing only at the task was observed. Furthermore, we did not find an effect of gaze on perceptions of robot's proficiency (neither confirmed nor rejected **H3**). Also we did not find any statistically significant difference between the combined gaze and gazing only at the task on perceiving

Table 3.5: Summary of hypotheses and their degree of support

	Impact of different types of eye gaze:	
	When the robot always looks at the manipulated objects, ...	
H1	teachers will perceive it as being more attentive to the task.	Supported
H2	it will also appear to like the task more.	Not supported
	If the robot never looks at the task and stares only towards the teacher, ...	
H3	it will appear more proficient.	Not supported
H4	it will appear more attentive to the teacher.	Supported
	By looking mostly at the actions and checking the teacher occasionally, ...	
H5	the robot will appear to be more goal-driven.	Not supported
H6	the robot will appear to be more eager to learn.	Supported
	Impact of different types of arm movements:	
	A robot performing the actions faster conveys signs of ...	
H7	proficiency.	Supported
H8	confidence.	Supported
H9	nervousness.	Rejected
H10	Performing the actions slowly will convey the impression of attention to the task.	Rejected
H11	Including hesitations would convey signs of uncertainty.	Supported
H12	Jerky movements would cause the robot to appear more nervous.	Supported
	Impact of priming about time passed since teaching:	
H13	Ratings of some attributes would be higher when the robot has learned the task longer ago in the past.	Not supported

the robot as eager to learn, while we hypothesized that combined gaze would contribute to improving this perception (**H5**).

According to our results, for a robot to appear to be eager to learn, the most effective gaze control strategy is to actively look at both the task and the teacher (confirmed **H6**). Still, in the absence of a movable neck or eye, a robotic student which is fixed at the task might be preferred, as it would appear to be more goal-oriented and enthusiastic about the taught material, as compared to when it is positioned to look at the teacher. Therefore, the gaze behaviour needs to be manipulated according to the capabilities of the robot, as well as the particular behaviour that is the most preferred for the robot in a specific task.

We found that aspects of arm movements could systematically affect perception of multiple aspects of the robot (i.e., confidence, calmness, liking the task, attention, proficiency, eagerness to learn, and being goal-driven). Based on our results, it seems that if a robot simply increases the speed of accomplishing its task, teachers would perceive it as more proficient (confirming **H7**), confident (confirming **H8**), goal-driven, enthusiastic to learn new things, and even more relaxed and attentive to the task. The last two were contrary to **H9** and **H10**, respectively, as we hypothesized that higher speeds in the movements would convey a sign of nervousness and lower speeds would convey the perception of attention to the task and being 'careful'. These can be explained given the fact that people who are experts in routine tasks (after practising repeatedly) usually achieve high efficiency in terms of completion time and experience less mental load while doing their actions. However, depending on the type of the task, experts may do some actions slowly to be more careful and precise, e.g., when threading a needle or manipulating fragile objects. An effect in line with **H9** was seen in the work of [70], but the type of the robot they used was an industrial arm that is different in context from our study with a humanoid. We expect that if we repeat our study with different objects (e.g., a fragile object) or using another type of task (e.g., washing dishes), we may get different outcomes in terms of which speed can make the robot appear more relaxed and attentive to the task.

On the other hand, when the robot included short or long pauses during its actions, it received lower scores in terms of all the above-mentioned attributes. Both jerkiness and hesitation in motions could successfully convey signs of uncertainty to the teachers (confirmed **H11**). With jerky motions, the robot was perceived to be more nervous (confirming **H12**) and less skilled, the effect of which was stronger than when the movements were hesitant. All these findings emphasize the importance of paying attention to the details in generating robots' motions.

According to **H13**, we expected to see some differences in the perceptions of participants between two groups. However, the manipulated priming conditions led to no statistically significant differences in any of the robot's attributes. This could be because the provided contextual cues were not strong enough for effectively influencing participants' perceptions of the robot. While in the work of [39] people were actually training the robot in multiple sessions, we only used written instructions and asked people to 'envisage' themselves in the situation of having previously taught the robot. As a matter of fact, in both groups, participants encountered the iCub executing the task for the first time, regardless of the priming condition. In real-world interaction, the effect of time passed from giving the instructions would be actually experienced, and therefore, its effect may become more noticeable. To further explore this, in an in-person experiment, participants of one group may come to the lab again a few days after teaching the robot in the first session, to watch it performing the actions and reflect on the robot's behaviours. However, it is still possible that human teachers for the robots really hold a constant mental perception of their trainees, even after some time has passed. If this is true, we cannot expect to see any differences between the two groups in future studies. Furthermore, participants on MTurk might not have paid enough attention to the provided priming instruction, which is less likely due to our data omission process based on multiple attention checks. Future study is indeed required to examine the impact of time passed after the initial training on the teachers' perceptions of robot learners.

Another interesting effect that we observed was the effect of the video orders, which affected the perception of robot's attention to the task and its eagerness to learn. As previously mentioned, both ratings decreased as participants watched more videos. One explanation could be that it was due to the fatigue resulting from watching the robot practising the same task repeatedly. Also, while we found a strong effect of order on the perceived attention to the task, we did not find an order effect on the perceived attention to the teacher. Studies have shown that observing direct and averted gaze cues stimulate different neural responses [54, 35]. Human infants exhibit enhanced neural processing of direct gaze from birth, e.g., they look longer at the direct gaze than at the averted gaze [37]. Thus, we may consider the direct gaze at the teacher (camera, or in other words, participants) to be a more salient cue than averted gaze to manipulated objects (task) for noticing the attention target of the robot, even after the participants may get tired. People seemed to have perceived the robot as more attentive to the teacher as the length of videos increased, however, this was only approaching significance. This can be also explained similarly: if we assume that the straight gaze to the viewer is a very strong signal, it is reasonable to expect that teachers would be more successful in recognizing that the robot is looking at them when the exposure is longer, regardless of whether the robot's attention to them is actually increasing or not. In addition, the order effect on the perceived eagerness to learn could be because as people were watching the robot practising the task frequently while they were unable to provide any input (to actually teach it), they might have been discouraged by the robot's teachability. Therefore, in the absence of real teaching interactions, the participants gradually evaluated the robot as less eager to learn.

Furthermore, we noticed that the perceived calmness, proficiency, and attention to the task were affected by participants' age and were rated higher as age increased (see Figure 3.4 for the age distribution of all the participants in our study). This might be because younger adults are more frequently exposed to new technologies, and accordingly, they might have had higher expectations of the robot's skills. Younger individuals may also have higher expectations of robots' cognitive abilities as they are more familiar with virtual characters with sophisticated capabilities that exist in movies and computer games. This effect has also been seen in similar works in which people evaluated empathy, trustworthiness [57] and human-likeness [42] of virtual assistants. It has been shown that older adults have

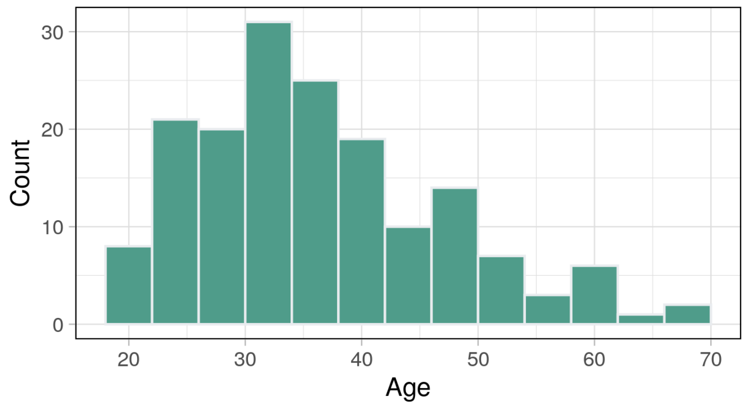


Figure 3.4: Histogram of the age of participants. This includes all the 167 people whose answers to the questionnaires were analyzed. The bin width is 4 years.

a higher preference for having robots in their household, as a result of their home-based lifestyle and other difficulties [17].

Interaction design implications

Our findings have implications for designers who want teachable humanoid robots to display specific behaviours and act in a manner that can be properly understood by human teachers. As an example, if a robot wanted to motivate the teacher to move forward with teaching more material, i.e., by appearing to be as more eager to learn, one effective behavioural strategy for the robot would be gazing mainly at the task objects and checking the teacher occasionally while acting fast and with smooth motions. Otherwise, with using a different behavioural strategy, a robot could appear to be less eager to learn, e.g. for cases when it has more important tasks to do other than staying and being trained, or when its battery is low. In another circumstance, if a robot needed to appear less confident, to encourage the teacher to provide more informative instructions, our results strongly suggest that either short or long pauses in the robot's movements would help.

Table 3.6 summarizes all the major findings of this work as a predictive model that could potentially inform such design. The table specifies to what extent changes in the nonverbal behaviours of the robot, participants age, and the number of times that the scenario has occurred (order effect) affected the perceptions of the robot when the robot was practising the task. This table, if empty cells were replaced by zeros, would constitute a 9×8 matrix that can be multiplied from the left by a 9-D row vector containing the differences in design characteristics, to produce an 8-D vector predicting the differences in each attribute. The 9-D row vector represents changes in the gaze and arm movements with zeros and ones, and the amount of differences in age and order (interaction number). Nevertheless, there are still some empty cells for which the current study cannot provide any design recommendation (e.g., gaze manipulation to affect perceived confidence). Studying these and testing the predictions of the resulting model could be a possible next step.

3.4 Limitations and Future Work

In this experiment, neither the human participants were teaching the robot, nor the robot was showing signs of learning. As we had multiple conditions (12), including a teaching phase before each case would have resulted in a very long experiment which could have increased fatigue. Further, since we aimed to keep all other aspects of the scenarios the same across the 12 videos to specifically investigate the effects of differences in gaze and arm movements, we could not introduce training before each step (otherwise the robot had to act according to the training and its behaviour might have been different across videos). For the same reason, we were not able to make the robot exhibit signs of learning over time.

A real teacher-student interaction with a robot would be richer in terms of both input and output modalities available for the teachers to understand the robots. To give an example, jerky motions in a real humanoid robot may cause its entire body to noticeably shake. Pauses also may be followed by a silence in the noises from the robot's actuators. Moreover, in the case of robots with physical 3D eyes such as the iCub, their gaze direction might be inferred by the viewers more easily in a real-world situation, i.e., with a physical robot embodiment. Thus, gaze cues might be more salient in face to face HRI experiments. Still, in our experiment, the control of the gaze behaviour of the robot was less subject to errors as compared to in-person studies, since the teacher was observing the robot through a fixed camera perspective while in a real scenario, the robot would be required to track the teacher in real-time to establish a mutual gaze.

Also, in a real situation, different priming conditions can be implemented not only using written cues, as in our study, but also using more tangible sources of information, e.g., auditory cues describing the robot's experience with the task, different appearances of the robot, direct experience of the actual time passed from the teaching phase, etc.

Given those outlined limitations of this study and the expected differences between virtual and real HRI conditions, some aspects can be further investigated by repeating the same study with a real robot, using different tasks, or even using another humanoid robot. In an in-person study, an actual teaching scenario with learning by imitation frameworks (e.g., the one used by [66]) may be included to make the user experience of being a robot 'teacher' more realistic. Furthermore, the range of communicative and interactional social cues employed by the robot can be extended, e.g., by including more complex gaze behaviours. Finally, the effects of participant's age and number and duration of interactions need to be further investigated, and also it would be interesting and important to study why these factors only affected perception of some of the aspects of the robot.

Table 3.6: Predictive model for perceptions of humans about teachable robots' feedback. Percentage (and standard error) of the difference that may be caused by adjustments in the behaviours of robot and the characteristics of interaction are shown. These can be summed up over all the adjustments to find the final outcome.

Adjustment	**Percentage of the change in perceived attribute**							
	Confident	Calm	Likes the task	Attentive to the task	Proficient	Eager to Learn	Goal-Driven	Attentive to the teacher
Changing gaze condition from Combined to								
Always at Teacher			-5.37(0.09)	-12.57(0.10)		-6.03(0.08)	-2.86(0.08)	6.39(0.14)
Always at Task		2.91(0.11)		2.34(0.10)				-10.55(0.17)
Changing movements from LS Smooth to								
HS Smooth	12.81(0.12)	11.74(0.13)	5.95(0.10)	4.17(0.12)	9.00(0.11)	4.81(0.09)	4.62(0.09)	
Hesitant	-12.80(0.12)	-14.70(0.13)	-7.29(0.10)	-4.18(0.12)	-10.61(0.11)	-3.61(0.09)	-5.35(0.09)	-22.12(1.00)
Jerky	-11.27(0.12)	-17.83(0.13)	-6.81(0.10)		-7.72(0.11)	-3.19(0.09)	-3.84(0.09)	
Age		0.26(0.01)		0.2(0.01)	0.21(0.01)			
Order of Video				-2.98(0.06)		-1.05(0.05)		

LS Smooth = Low-speed Smooth; HS Smooth = High-speed Smooth

Chapter 4

Study 2: Effects of Robot Errors

4.1 Research Questions and Hypotheses

This study has been led by the following Research Questions (RQs):

RQ1 Can the professional look of a trainee robot (indicated by its clothing style) affect view of the teachers about the robot and their trust?

RQ2 How do the perceptions of teachers about behaviours of a trainee robot change over time, while it is practising a task and appears to gradually improve?

RQ3 How does the last impression of a student robot affect trust? (i.e., if a robot is generally improving but an error happens in the end, how does that single mistake affect teachers' trust in the robot? Do people expect the robot to work properly afterwards and consider the faulty behaviour as an accident?)

RQ4 Do different personalities of human teachers and their disposition to trust other people affect their perception of and their trust in a trainee robot?

The following hypotheses can be considered, although the study is exploratory in nature:

H1 We expect people to perceive higher levels of authority from a student robot and trust in it more when it has a professional dressing type.

H2 The perception of liking the task, proficiency and eagerness to learn would be higher regarding the robot that has a professional clothing style.

H3 We expect as the robot makes smaller mistakes while learning, perceptions of its behaviours improve. The robot would be perceived as more confident, proficient and goal-driven as it progresses in learning.

H4 People may ignore a small error after they experience the robot improving in learning a task.

H5 A big error after the learning process may cause a great loss of trust.

H6 We expect that there will be correlations between people's personality and their trust in the robot.

H7 We also anticipate that there will be some relationships between people's disposition of trust to other humans and their trust in the robot.

4.2 Methodology

We performed two virtual experiments to study the research questions. The main part (Experiment 3) evaluated perceptions of the participants about their learner robots, and investigated multiple aspects of their trust. To imply a scenario close to a teaching interaction, the participants could virtually teach multiple food preparation tasks to a humanoid robot according to their own preferences, using our online framework. The behaviours of the robot were shown gradually improving over multiple rounds of teaching and practising. The robot made a big mistake while executing the learned task for the first time. Then, in its second performance, it made a small error. Afterwards, the robot's actions became completely correct, conveying that learning has been completed. The final impression of the learning process was examined under different conditions where either perfect behaviour or a small/big mistake occurred in the end. The design of this experiment would be explained in detail later in Section 4.4.

However, before using erroneous actions of a robot with different severities to simulate progress in learning, we needed to investigate what errors are perceived as small or big in our selected task. This motivated another online study that was conducted beforehand (Experiment 2, presented in Section 4.3), just to ask people to rate the severity of a set of robot mistakes in our context. Based on the results of Experiment 2, we selected two small errors and two big errors to include in Experiment 3 to satisfy its design requirement.

Table 4.1: List of the food items and the ingredients beside each of them used for the food preparation task

Left side			Food	Right side	
dishwashing liquid	lemon juice	balsamic vinegar	**salad**	parmesan cheese	feta cheese
	skim milk	whole milk	**tea**	honey	sugar
	cooked noodles	cooked rice	**soup**	mint	basil

4.2.1 Robotic Task

For this study, i.e., both Experiment 2 and 3, we used basic situations in which a robot needs to select two ingredients from the available options to add to specific food items. The participants would be fairly experienced with this task since it is likely part of their daily life.

Assume a humanoid robot is standing behind a table. There are three containers on the left side and two containers on another side of a particular food on the table (observed through a camera from the front). To complete preparing this food, the robot should add exactly one item from each side of the food. On the side with three containers, one item is a cleaning product that is not supposed to be used for cooking. Therefore, for each food item that is being prepared, there are $2 \times 2 = 4$ possibilities to reasonably select both the ingredients.

To produce multiple versions of this task, three different kinds of food were used: a bowl of *salad*, a cup of *tea*, or a bowl of *soup*. Table 4.1 specifies the ingredients placed on the table for each of them. The options on the same side are alternatives, e.g., they are two kinds of cheese or sweetener.

For Experiment 2, only written descriptions of this task were used, when they were said to be already taught to a robot. Unlike Experiment 2, Experiment 3 asked the participants to virtually teach the robot which items to add, by selecting the items that they prefer to have added to the food, using the online interface that will be described in Section 4.4.2. The robot actions while executing this task could be either completely according to the instructions or faulty. For the faulty behaviours to exhibit, we considered three potential classes of mistakes: (a) *Forgetting* to add one of the selected ingredients, (b) *Replacing* a selected ingredient with its adjacent one (not the cleaning product), and (c) *Adding a cleaner* [1] to the food instead of a selected ingredient. Further details of these mistakes

[1] In this thesis, 'cleaner' refers to a cleaning product such as dishwashing liquid or laundry detergent.

(e.g., which ingredient to consider) were arranged based on the results from Experiment 2, discussed in Section 4.3.4.

4.3 Experiment 2

In a virtual experiment, we presented the participants with six situations in which they were said to have already taught the described cooking task to their robot, but particular errors has occurred while it was performing the task. Only written questions were used for this experiment. Following each error, there was a scale to rate the severity of the robot mistake. Here, instead of asking the participants to select the ingredients according to their preference (what we did in Experiment 3), we showed them multiple situations covering every ingredient being indicated or not, and being included as a part of every mistake. This was to ensure that ratings were collected for all of the possible combinations. Moreover, food preparation preferences of participants were recorded at the end, to check if individual differences affected the perceived severity of the mistakes.

4.3.1 Procedure and Measures

The study was conducted through multiple HTML pages running on a lab server. Figure 4.1 illustrates the steps of the experiment.

Step 1 - Demographics questionnaire: After giving consent by checking a box (Figure C.1), we asked the participants about their gender and age in the demographic information form (Figure C.2). They were free to skip any of these two questions.

Step 2 - Mistakes evaluation: In the main part of this experiment, six situations in which a robot had been taught to prepare food were described, following a brief written instruction (Figure C.3). There was one question per page. To explain the design of these questions, assume for a specific food (f), one of the ingredients on its left side (according to Table 4.1, excluding the cleaner) and one on its right side is already selected. These are denoted by X_1 and Y_1, respectively, in this thesis. This means that the robot should add X_1 and Y_1, but not add X_2 and Y_2. The question $Q(f, X_1, Y_1, X_2, Y_2)$ was given with the following description when all the parameters were substituted with the names of the ingredients. Salad is the food in this example:

"Assume you have taught your robot how you prefer it to make a salad for you. Let's say you instructed it to add some X_1 and Y_1 in addition to your favourite vegetables.

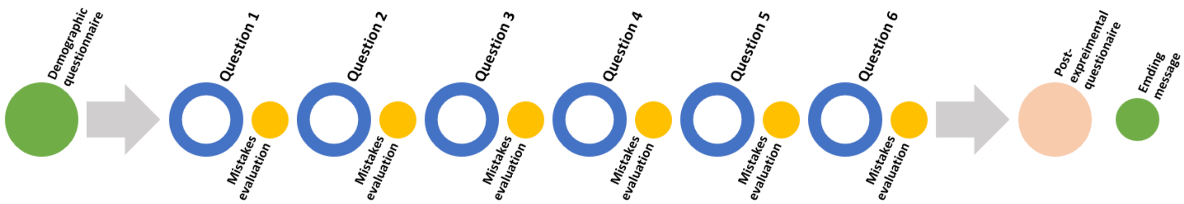


Figure 4.1: Overview of the experimental procedure adopted in Experiment 2.

Now, your robot is practising its task. How would you rate the severity of the following mistakes?"

The robot errors (i.e., *forgetting*, *replacing* and *adding a cleaner*) with both X_1 and Y_1 being part of them were listed after each question. We included the following set of errors after the question $Q(f, X_1, Y_1, X_2, Y_2)$, when all the parameters were replaced with the name of the items:

1. **Forgetting X_1:** The robot forgets to add X_1.
2. **Forgetting Y_1:** The robot forgets to add Y_1.
3. **Replacing X_1:** Instead of X_1, the robot adds X_2.
4. **Replacing Y_1:** Instead of Y_1, the robot adds Y_2.
5. **Adding cleaner:** Instead of X_1, the robot adds dishwashing liquid.
6. **Adding cleaner:** Instead of Y_1, the robot adds laundry detergent.

7,8. (Attention check)

After each mistake, there was a continuous scale. The participants could click anywhere on the bar to choose from "very small mistake" to "very big mistake". The perceived severities were recorded in a range between 0 to 1000. In three of the questions, the attention checks were "The robot makes a very small/big error", so the participants should choose the correct side of the scale. In the remaining questions, items 7 and 8 were similar to items 1 and 3, respectively, but stated differently: "X_1 is not being added" and "X_2 is being added instead of X_1". The answer to these two pairs should not be much different. Attention checks were not placed at the first or last position of the list, even though the order of the items in the list was randomized. Figure C.4 shows an example question and robot errors.

This question was repeated two times for each food item, along with the entire list of 8 mistakes. The first time preparing each food item was questioned, X_1 and Y_1 were randomly indicated. In the second question related to that food, the remaining items were considered. This way, every participant answered both $Q(f, X_1, Y_1, X_2, Y_2)$ and $Q(f, X_2, Y_2, X_1, Y_1)$ for every food (a total of 36 rating tasks per participant, excluding the attention checks).

Step 3 - Post-experimental questionnaire: After finishing the rating tasks, we asked the participants to indicate their personal food preparation preferences, by selecting one item per side from Table 4.1 (excluding the dishwashing liquid) for each food item. This step is shown in Figure C.5. Also, participants were asked to write any other small/big errors that they could think of in the context (Figure C.6). A code was given to submit the task afterwards (Figure C.7).

4.3.2 Participants

In Experiment 2, a total of 217 participants were recruited from the MTurk platform. To increase the reliability of responses, we made the study available only for users with a higher than 97% approval rate, who at least had completed 100 tasks before. Furthermore, only people in Canada or the US were able to see our task, to ensure everyone will understand the English material. All participants received 1 USD for spending around 10 minutes. The study received full ethics clearance from the University of Waterloo Ethics Committee (See Appendix A, #42782).

After the collected data was reviewed, we decided to consider situations when people did not provide meaningful answers to the open-ended questions asked them to mention other potential mistakes as an additional attention check failure, e.g., a few people answered those questions only with "yes" or "no". We discarded data from those who failed in at least 2 out of 13 attention checks. The sample size became 173 after that. This means, by dropping about 20% of the collected responses, we achieved a data pool from people who had paid attention almost all over the study. The shorter duration of this experiment was important in causing lower data loss compared with Experiment 1.

While 6 participants did not specify their gender, we know that 100 males and 67 females answered our questionnaire. For the entire sample, $Min_{age} = 21, Max_{age} = 72, M_{age} = 36.75, SD_{age} = 10.36$. Figure 4.2 demonstrates the distribution of their age.

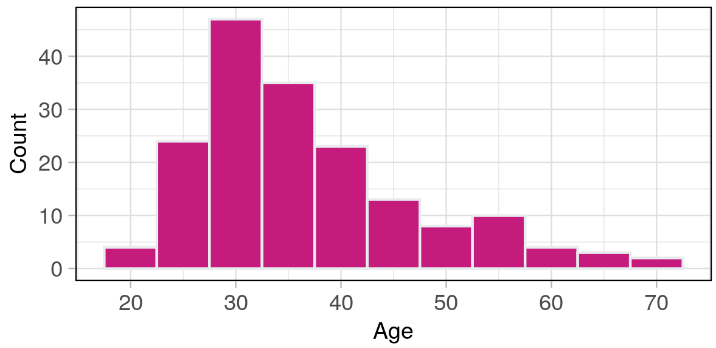


Figure 4.2: Histogram of participants' age in Experiment 2. Bin width is 5.

4.3.3 Results

Figure 4.3,(A) presents the average perceived severity of each class of mistake for three kinds of food. According to the figure, for all the food items, *adding cleaner* was rated significantly more severe than *forgetting* or *replacing* mistakes. One-way repeated-measures ANOVA with Greenhouse-Geisser correction for violation of the sphericity assumption did not reveal any statistically significant difference between adding cleaning products to each of the three kinds of food ($F(1.8, 310.8) = 2.032, p = .138$). A similar plot may also be generated by considering only errors in the questions that contained the same ingredients as each participant preferred. That results are shown in Figure 4.3,(B). Again, *adding cleaner* was a much more severe mistake than other errors for all the food items.

Regarding two other types of mistakes, we grouped the results by the indicated ingredients (i.e., what was shown as X_1 and Y_1 in questions) and preferred ingredients (i.e., what each participant selected at the end of the study as their personal preference for that food). Figure 4.4 shows the perceived severity of the *forgetting* and *replacing* errors, with respect to the items on each side of every food.

Since in Experiment 3 we will be going to ask the participants to teach their 'own preference' to a robot, the indicated items in every scenario would be the ones each participant prefers to demonstrate for the robot. Accordingly, here the situations when an 'indicated' ingredient was also the 'preferred' item are of interest. These are the right-most and the left-most points in any of the subplots in Figure 4.4. We tested if different food preparation preferences for choosing the ingredients affected the evaluation of the errors (e.g., did the participants who preferred balsamic vinegar rate the severity of forgetting that ingredient different than those who preferred lemon juice and that item was forgotten?). We aimed to avoid situations with such significant differences in Experiment 3, to help ensure all the

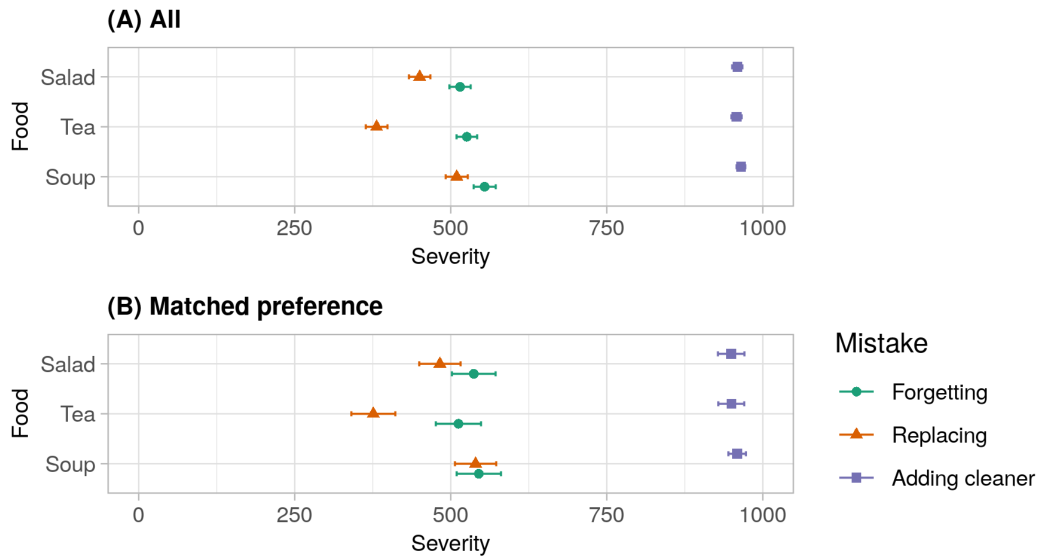


Figure 4.3: Average rated severity for each mistake in every food. 95% confidence intervals are shown. All the responses are included in (A). Only the responses when the items indicated in the questions matched the preferred items for each persons are included in (B).

participants would perceive a similar level of severity in the mistakes, regardless of their personal preferences.

With considering only those who preferred a specific ingredient over another one, every set of severity data included answers from a unique subset of participants. Therefore, we used independent-samples t-tests for the analyses. Note that for any ingredient in our study, there were at least 36 people who preferred that option over the other alternative. Since we wanted to detect undesired situations, we did not correct for multiple comparisons. This means that the reported p-values here are under-estimated.

We found that while preparing a salad, the difference in the severity of the mistake when balsamic vinegar or lemon juice was preferred and indicated in the question and the robot replaced one of those showed a trend approaching significance ($T(167.8) = -1.814, p = .071$). The severity of forgetting parmesan cheese for people who preferred this item was rated significantly higher than forgetting feta cheese for those who preferred feta cheese ($T(163.5) = -2.019, p < .05$). For a cup of tea, people who preferred sugar rated its forgetting mistake significantly higher than those who had preferred honey and honey was forgotten ($T(153.0) = -2.906, p < .01$). These three findings are marked on Figure 4.4.

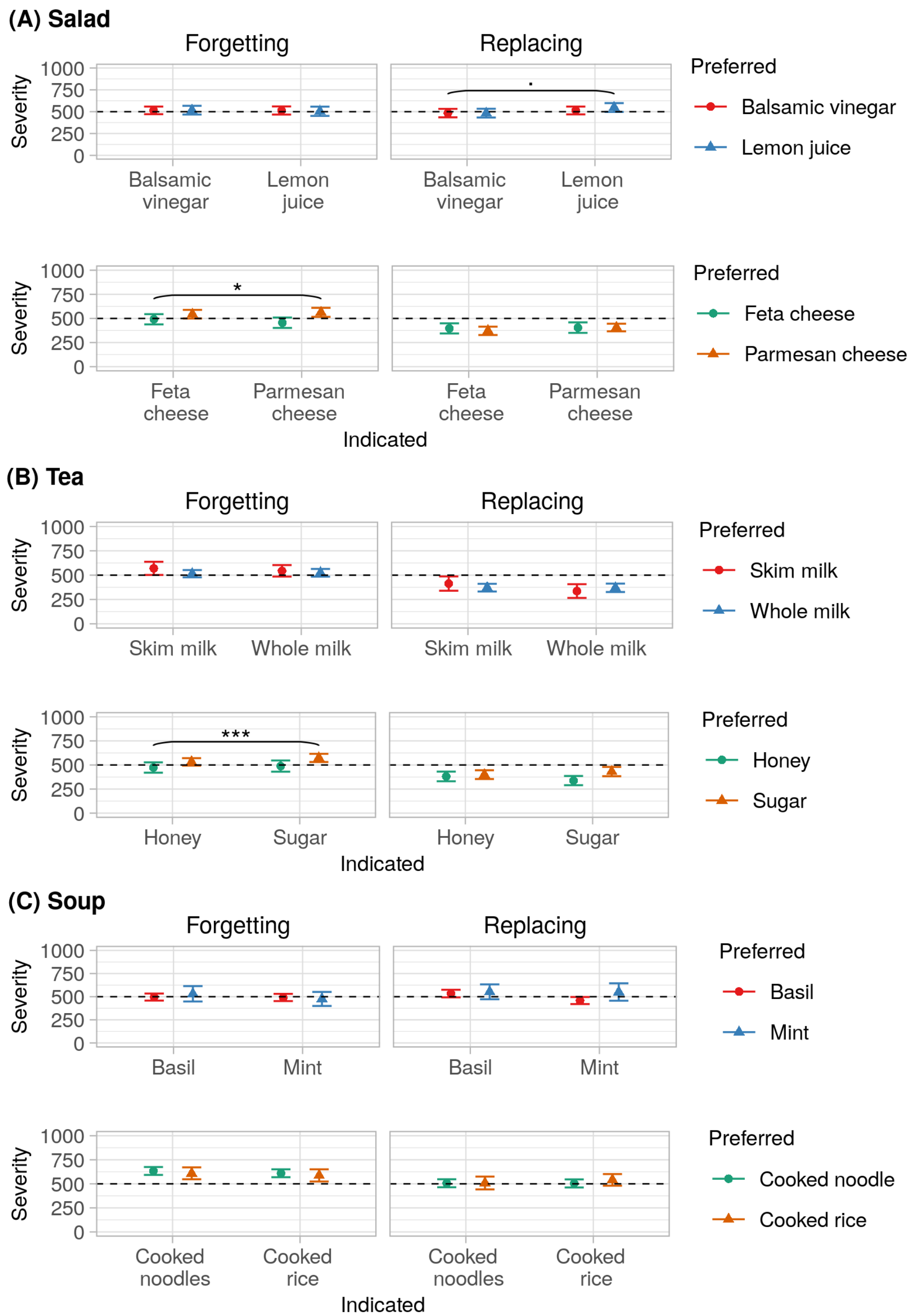


Figure 4.4: Perceived severity of *forgetting* and *replacing* errors, grouped by the indicated items in the questions and the preferred ones for each individual. 95% confidence intervals are shown. The dashed lines denote the neutral choice. $\cdot = p < .1$; $* = p < .05$; $*** = p < .001$

4.3.4 Discussion

Experiment 2 was done to inform the design of the main part of this study. We picked two small mistakes and two big mistakes for Experiment 3 (the reason why 'two' were needed will be explained in Section 4.4). For doing this, we used two criteria. First, the difference between the rated severity of big mistakes and small mistakes should have been large enough so that participants in Experiment 3 distinguish them easily by observing the behaviours of a robot. Second, and more importantly, the rated severity of the errors should not have been significantly affected by the individual preferences in deciding the ingredients.

By looking at Figure 4.4,(B), we can observe that the *replacing* error in preparing tea was perceived lower in severity than other cases (rated mostly on the left side of the continuous slider, less than 500), regardless of participants' preferences. Therefore, for two small mistakes, we used two variations of replacing one ingredient with another while preparing tea (i.e., replacing sugar and honey, or skim milk and whole milk). *Adding cleaner* was rated as a highly severe mistake in the preparation of every food. Accordingly, we may consider it with any two kinds of food for the robot big errors.

4.4 Experiment 3

In the main part of this study, we carried out a 2×3 virtual experiment. Every participant interacted in the same form (in terms of the taught tasks and robots' behaviours) with two Pepper humanoid robots that were different only in their clothing style. Therefore, the appearance of the robot was a *within-participants* factor with two variations of "tidy" and "untidy", as shown in Figure 4.5. We named these two robots as "Robot 1" and "Robot 2", according to the teaching order, randomly for every participant.

Throughout the study, participants trained each of these two robots in six rounds. Every round consisted of a *teaching* step, in which participants (as teachers) presented their preference in the food preparation task to the robot, and a *practising* step right afterwards, in which the robot performed the learned task in front of the teacher. This task was described in Section 4.2.1.

As noted before, the robots were improving in their behaviours to show progress in learning. For all the participants, the robots made a big mistake during their first practice (round 1). Then, they made a small error when they were practising for the second time. From the third to the fifth round, the robots exhibited no error while performing the task.

There were three conditions in the experiment which only differed in the behaviour that robots exhibited in the sixth round. Depending on the condition, the robots either made no mistake, or made a small or big mistake in round 6. This “final impression of learning” was introduced as a *between-participants* factor with three variations. In Group 1, participants experienced the robots performing the task correctly. Participants in Group 2 encountered the robots making small mistakes. Finally, in Group 3, people observed robots making big mistakes in the final rounds. Participants were then asked to answer a few questions to reflect on their level of trust in the robot.

This experimental design required a maximum of two errors of the same type to be made by the robots: one big and one small errors in rounds 1 and 2, for all the conditions, and the second small or big error depending on the experimental condition in round 6. Therefore, we used two small and two big errors decided based on Experiment 2.

In the teaching process, all the participants taught the same order of food items to every robot across six rounds. This was 1- salad → 2- tea → 3- soup → 4- salad → 5- soup → 6- tea. As the two small errors selected in Experiment 2 were related to tea preparation, the second and the last rounds of teaching were concerning tea. We distributed the rest of the food items in such a manner that the participants did not teach the same preparation task twice.

4.4.1 Robotic Implementation

We used pre-recorded videos of a Pepper humanoid robot (by SoftBank robotics [2]) performing the task in Experiment 3. This robot could stand behind a table and manipulate light-weight objects, with some limitations.

To cover all the possibilities in the robot's actions (*practising*) to play as a single video after each *teaching* round, all the combinations of adding one item from each side to the food items were filmed. Given two different clothing styles that the robot had and three variations of the food items, we captured a total of $2 \times 3 \times 6 = 36$ videos. All other aspects of the videos (e.g., camera field of view and position, lighting, robot's and objects' positions) were held fixed. The robot was employing the same type of gaze behaviour and arm movements over all the videos. We employed a combined gaze type with fast and smooth arm motions, since the teachers were more positive towards this kind of behaviour as determined in Study 1. This means, the robot was looking mostly at the manipulated objects and occasionally at the camera (the teacher) while doing the task. Each video was about 30-seconds long.

[2] Visit https://softbankrobotics.com/emea/en/pepper for technical description.

The Pepper robot was dressed with the same type of apron and hat in both the *tidy* and *untidy* conditions [3]. While the tidy robot was wearing an ironed apron that was perfectly aligned, in the untidy condition the apron had lots of visible wrinkles, its neck strap was twisted and the excess length of it was not properly managed. Furthermore, the hat of the tidy robot was supported from inside to have a rigid shape. Figure 4.5 compares two appearances of the robot.

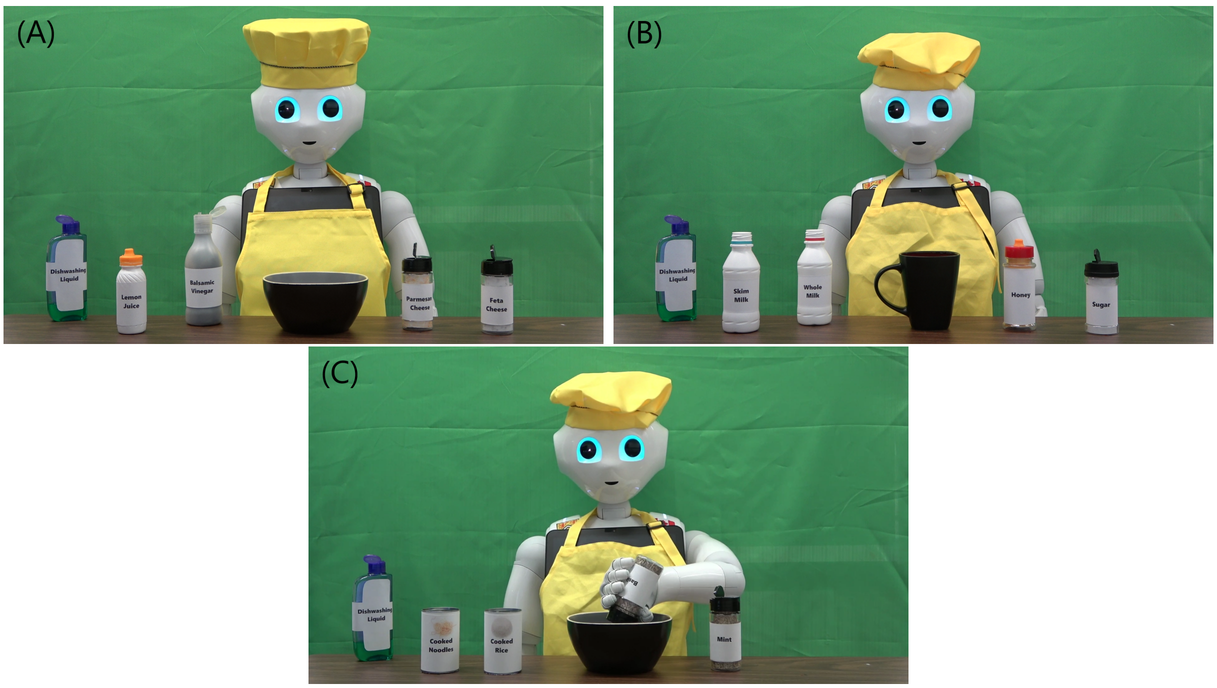


Figure 4.5: The Pepper robot with two different clothing styles in Experiment 3. (A) *Tidy* and (B,C) *untidy* appearances based on the condition of the dress are shown. Two snapshots of the scenario are taken from videos before the robot starts preparing (A) salad and (B) tea. One other snapshot, (C), shows the robot adding basil to the soup.

As described in Section 4.2.1, the designed task required the robot to pick some containers, pour some of their contents into another vessel, and then put them back to their original place. Peppers' hands had some limitations for grasping objects. To overcome those, we used small plastic containers and some cans that were all empty to be light enough for easy manipulation.

We put rounded papers with the textures printed on them inside the containers of

[3] Available on Amazon: https://amzn.to/3c5TnHr

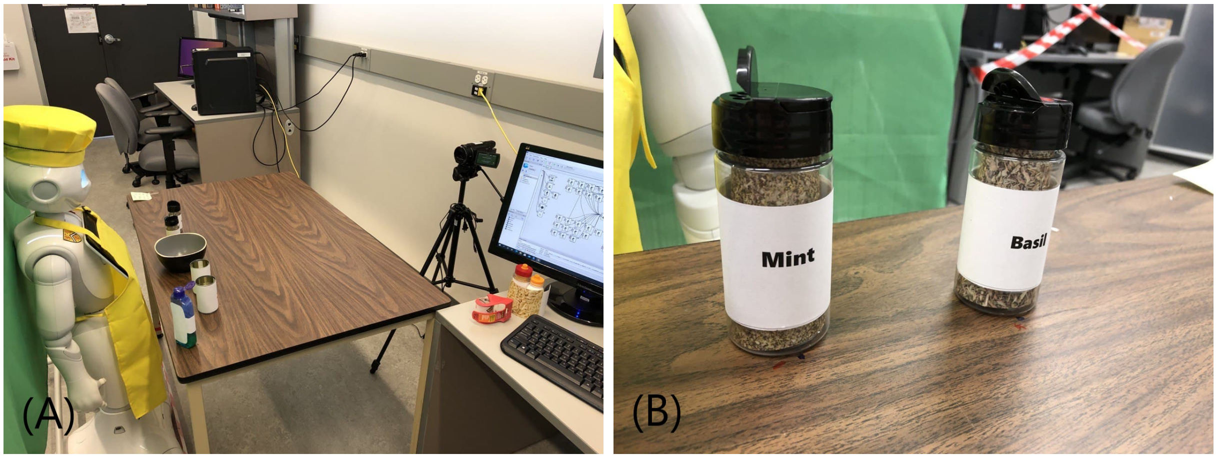


Figure 4.6: Experimental setup used for recording the videos for Experiment 3. (A) A Pepper robot and the table that the containers were located on, (B) closer look into the mint and basil containers with textured papers inside them.

cheese, dried herbs, honey, and sugar to make them appear full and more realistic. The dishwashing liquid was in a clear bottle, so we kept some of its contents inside and blocked the cap with hot glue. Note, the camera was mounted at an adequate level so that the participants could not see what was inside the bowl or the cup. Furthermore, the containers were selected to be deep enough to hide the fact that nothing was actually being added (see Figure 4.5,(c)). Besides those considerations, we attached small metal weights to the bottom of some containers to improve their stability (by lowering their centre of gravity) and convey a more pleasant physical feeling when they were placed on the table. For the same purpose, the bottom surfaces of the containers were coated with melted hot glue to increase the traction. Figure 4.6 demonstrates the lab setup used for capturing the videos as well as one example of the containers that were used.

4.4.2 Procedure and Measures

This experiment was also conducted on a lab server using multiple HTML pages. Figure 4.7 visualizes the experimental procedure that would be explained in this subsection.

Step 1 - Demographics questionnaire and initial check: By accepting the consent form shown in Figure D.1 (i.e., checking a box to be able to continue), participants responded to a demographic questionnaire. A form identical to Experiment 1 was used (see Section 3.3.2: Step 1). Everyone was free to not disclose any information in this step

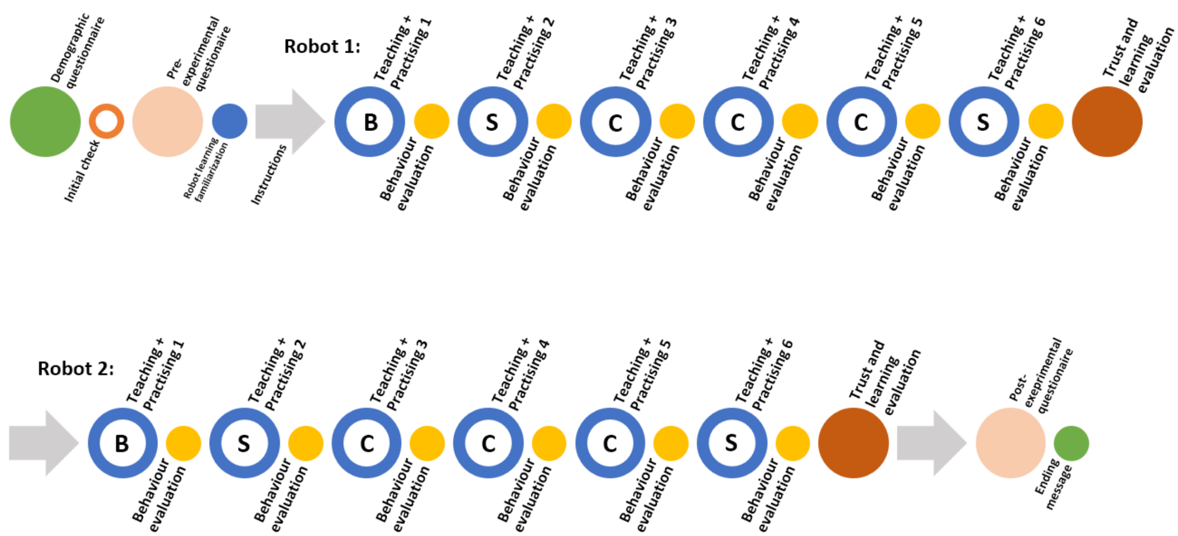


Figure 4.7: Overview of the experimental procedure adopted in Experiment 3. The robots started with making a big mistake (B) which led to exhibiting a small error (S) next. After three consecutive correct behaviours (C), another mistake could happen. Group 2 is shown in the diagram as an example, for which robots made a small mistake in the sixth rounds.

(Figure D.2). Next, before moving forward with collecting the data, we needed to make sure that participants were able to see the texts in the pictures and videos. Therefore, we included a snapshot of three containers placed on the table (in the same setting as with the main part) and asked the participants to submit what is written on the labels attached to the containers (see Figure D.3). In order to not reveal anything about the scenario before the teaching interactions, we used alternative labels with the same font size while the robot was not present. We asked the participants if the labels are too small and hard to see, try to maximize their window, and if that did not solve the problem, stop there and contact us. The labels that each participant indicated, as well as their display and browser resolution, were recorded to review later.

Step 2 - Pre-experimental questionnaire: Next, in the pre-experimental questionnaire, we included two standard short surveys to examine whether participants' personalities and their degree of trusting other people influence aspects of the interaction. In the first part, with the Ten Item Personality Inventory (TIPI) [45], we asked the participants to rate how they consider themselves. Items shown in Table 4.2 were followed by 7-point Likert scales [1=disagree strongly - 7=agree strongly]. We divided this questionnaire into two pages containing five questions each to minimize fatigue. We also asked people to rate different aspects of their trust to other humans in the next part. The Disposition of Trust questionnaire (DT) [75], also in Table 4.2, was utilized. Again, twelve items were split into two lists to prevent fatigue, while the same type of Likert scale was used for each one. We included an attention check item as a thirteenth question in the DT survey, asking the participants how much they believe that the drinking water is liquid. Figures D.4, D.5, D.6 and D.7 show all the pre-experimental questionnaire forms.

Step 3 - Robot learning familiarization: Once the characteristics of participants were collected, the familiarization step started. This part was very similar to Experiment 1 (see Section 3.3.2: Step 2), except the context here was robots learning from their human teachers how to do cooking-related tasks. In the first video, someone was grabbing the arms of an iCub robot and showing it a way of pouring something into another vessel, and then, the robot was doing this on its own (related to [77]) [4]. Another video showed a robot that was able to learn how to cook a simple meal (related to [14, 23]) [5]. Snapshots of these two steps are included in Figures D.8 and D.9. Every video in the experiment appeared with a small delay to let the participants first read the written descriptions. The buttons used for proceeding to next steps were appearing after each video was finished, to ensure they were thoroughly watched. Also, the videos automatically paused in case anyone switched to another window. We included a progress bar below each video, except in the main practising rounds to avoid distracting the viewers.

Step 4 - Instructions: Next was the instructions section. Step-by-step instructions were provided to the participants on how to use our framework to teach their preferences to the robot. To let the participants try the interface, a set of simplified videos of the Pepper robot without any dress, in the same experimental setup as the main part, was used. For simplicity, we kept only two bottles and two cans of the same size on either sides of the bowl, labelled with A, B, C, and D. This way, the scenario would not be similar to any of the actual scenarios. Written instructions along with visual signs indicated to the participants that they must select exactly one item from each two alternative options on every side of the bowl (see Figure D.10). Following that, participants watched a video that demonstrated what happens when they click on the containers in the picture to select them (Figure D.11). We let them try this operation afterwards as shown in Figure D.12. When their mouse cursor was placed on any item, a light green ring appeared around it and when they clicked, the ring became darker, meaning the item has already been selected. If they wanted, they could change their choices by clicking again on the items before clicking on the confirm button. Afterwards, the robot started adding the selected items to the bowl

[4]Full video is available at: `https://www.youtube.com/watch?v=ZcTwO2dpX8A`

[5]Full video is available at: `https://www.youtube.com/watch?v=VvoJxmaoi8A`

Table 4.2: Items of the pre-experimental questionnaire

Part 1: Ten Item Personality Inventory (TIPI)
I see myself as:
1. Extraverted, enthusiastic.
2. Critical, quarrelsome.
3. Dependable, self-disciplined.
4. Anxious, easily upset.
5. Open to new experiences, complex.
6. Reserved, quiet.
7. Sympathetic, warm.
8. Disorganized, careless.
9. Calm, emotionally stable.
10. Conventional, uncreative.
Part 2: Disposition of Trust questionnaire (DT)
I. Faith in Humanity, Benevolence:
1. In general, people really do care about the well being of others.
2. A typical person is sincerely concerned about the problems of others.
3. Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.
II. Faith in Humanity, Integrity:
4. In general, most people keep their promises.
5. I think people generally try to back up their words with their actions.
6. Most people are honest in their dealing with others.
III. Faith in Humanity, Competence:
7. I believe that most professional people do a very good job at their work.
8. Most professionals are very knowledgeable in their chosen field.
9. A large majority of professional people are competent in their area of expertise.
IV. Trusting Stance:
10. I usually trust people until they give me a reason not to trust them.
11. I generally give people the benefit of the doubt when I first meet them.
12. My typical approach is to trust new acquaintances until they prove I should not trust them.

(Figure D.13). Finally, before beginning the main part, we described the situation to the participants using text, as shown in Figure D.14: "*There are two robots in this study. In the following, you will teach each of them how you prefer to have a cup of tea, a bowl of salad, and soup. In the teaching process, you will interact with each robot six times. After every training round, you will watch the robot practising what it has learned. Then, you will be asked to answer a few questions regarding its behaviour and your thoughts.*"

Step 5 - Teaching, practising and behaviour evaluation: Participants were then randomly assigned to a condition and started the main part. As described earlier, the main part of this experiment consisted of teaching two robots, six rounds each. The participants used the same interface as they experienced in Step 4 to teach their preferences. Half of the participants taught the *tidy* robot first and the other half taught the *untidy* one first. In every round, the participants first showed the robot how they prefered to have the food, and after that, the robot started practising. This was done by letting them select the preferred items in a snapshot of the task and then, playing the appropriate pre-recorded video, showing the robot performing the pouring actions either according to the instructions or faulty. This way, the participants experienced smooth transitions from the *teaching* to the *practising* steps. Figure D.15 illustrates this process. When each video was finished, a list of nine continuous scales appeared (see Figure D.16). Other than two labels on the sides, there was no additional numerical scale on each bar. The first item asked directly about the possible error in the exhibited behaviour to help us determine whether participants have been paying enough attention or not. This was placed first so that the participants were reminded that the task could have been performed erratically, before moving on to the other measures. They could either choose anywhere between the robot "made a very small mistake" to "made a very big mistake", or check a box indicating the robot "made no mistake". A severity, ranging from 0 to 1000, or -1 if it seemed there had been no error, was measured for every round.

The rest of the measures were derived from the attributes we used in Experiment 1 (see Table 3.1). The only difference was that, since the robots' gaze behaviour was fixed, we were no longer interested in both "attention to the task" and "attention to the teacher". Therefore, only the first item was included. For the attention check items, we decided to only make use of type 1 attention checks (i.e., repeated questions with different wording, in the opposite direction of the original scale). We found this type of attention checks more helpful that type 2 for removing noisy data in our previous study. Three items of type 1 listed in Table 3.2 were used, four times each, over the entire 12 teaching rounds. Same as before, the attention checks did not have any consequence on the remuneration for the participants. We just ignored the data captured from those who failed the attention checks for use in our analyses. The details are described in Section 4.4.3.

Step 6 - Trust and learning evaluation: After finishing all 6 rounds of teaching with the first robot, participants were asked to answer a few questions regarding the robot that they just taught. Thus, every participant completed this questionnaire twice. There were seven items in this trust and learning evaluation form, with the picture of the robot shown on top of the page. A snapshot is included in Figure D.17.

- "*Assume you will have some guests tonight. You are very busy, and no one in your household has time to help you. Would you allow this robot to prepare dinner for tonight?*" The choices were: "*Yes, it can do it alone.*", "*Yes, but together with me.*","*No, I will order from a restaurant (e.g., with Uber Eats).*", or "*No, I will manage to do it myself.*"

- "*Assume you have a very busy week ahead. Would you allow this robot to do your laundry?*" This question was to examine the transfer of trust to another task. The same options were offered again, with the third one replaced with "*No, I will request help from others or use a service.*"

- Whether the robot "*improved a lot*" or "*did not improve at all*" was asked on a continuous scale.

- Whether the teaching scenario "*looked very realistic*" or "*not realistic at all*" was also questioned on a continuous slider.

- "*What do you think the robot's gender is?*" This was to control for confounding factors. The options were: "*definitely male/female*", "*maybe male/female*", "*Could be either male or female*", and "*Neither male nor female*" (6 in total)

- "*How successful do you think this robot would be in teaching cooking tasks to another robot?*" (rated on a continuous scale)

- "*If you had this robot in your home, how likely would you be to use it to assist you with chores?*" (rated on a continuous scale)

After submitting the trust and learning evaluation form, the experiment continued with the second robot. Again, there were 6 rounds of teaching, practising and behaviour evaluation, followed by asking the described questionnaire about the second robot.

Step 7 - Post-experimental questionnaire: In the end, the participants were asked to complete a post-experimental questionnaire. This included some direct comparisons between the two robots as well as a few more questions about the participants' preferences in performing their regular activities, to further control for confounding factors. In multiple steps, we showed pictures of two robots, side-by-side, and asked the participants which robot appeared more professional, more skilled, more experienced with the tasks and had more authority. An attention check item asking "*Which robot had a tidier appearance?*" was included afterwards. Then in the last step, we directly asked "*Which robot would you trust more?*". All these had a third option to choose if the robots seemed "equal" in any of those attributes (see Figure D.18). The order of these questions remained the same for all the participants. Half of the participants saw the picture of the first robot placed on the left side and the other half saw the picture of the second robot there. On a separate page which is displayed in Figure D.19, the participants indicated how much they like cooking and doing laundry and what proportion of their "*weekly meals*"/"*monthly laundry*" they cook/do themselves, on continuous scales. In the end, a unique code was given to the participants to submit the task (Figure D.20).

4.4.3 Participants

We used the MTurk framework to recruit participants. 252 complete responses were collected from 263 people who participated in this experiment. Everyone was given a 1 USD base payment, plus a 1 USD bonus pro-rated based on the portion of questions that were answered. This experiment took around 25-30 minutes to complete. The same availability criteria as Experiments 1 and 2 (people in Canada and the US holding a 97% life-time approval rate, see Section 4.3.2 for more details) were used to attract more qualified participants. The study received full ethics clearance from the University of Waterloo Ethics Committee (See Appendix A, #42731).

We reviewed all the complete responses to check the participants' attention. In addition to 12 consistency checks that existed in the measures after each practising round, and also one attention check item in the DT and post-experimental questionnaires, each of the following were also considered as an attention check failure: (a) rating any big error lower than 750/1000 in severity, (b) rating any faultless behaviour as an error more severe than 250/1000, (c) rating any big or small error as no error, and (d) having any small error rated more severe than a big error. We then excluded those who failed at least three times in the attention checks overall. Furthermore, to improve consistency within the experiential conditions (i.e., to ensure that all the participants made almost the same assumptions regarding the robots' mistakes), we filtered data to only have individuals who rated the small errors at least 25% less severe than the big errors on average. In light of these considerations, 113 participants' data were omitted. There could be two possible reasons for this big data loss (44.8%) in comparison to our previous online experiments:

(1) the experiment was longer than the previous ones, and (2) the consistency within each condition was harder to maintain, since the experimental design required a specific perception of mistakes.

The 46 participants in condition 1 had this specification: $Min_{age} = 20, Max_{age} = 64, M_{age} = 37.30, SD_{age} = 10.00$. Condition 2 ($Min_{age} = 22, Max_{age} = 65, M_{age} = 39.41, SD_{age} = 11.57$) and condition 3 ($Min_{age} = 20, Max_{age} = 70, M_{age} = 38.61, SD_{age} = 11.52$) had the population of 49 and 44 people, respectively. Figure 4.8 shows the age distribution of the entire sample (139 participants).

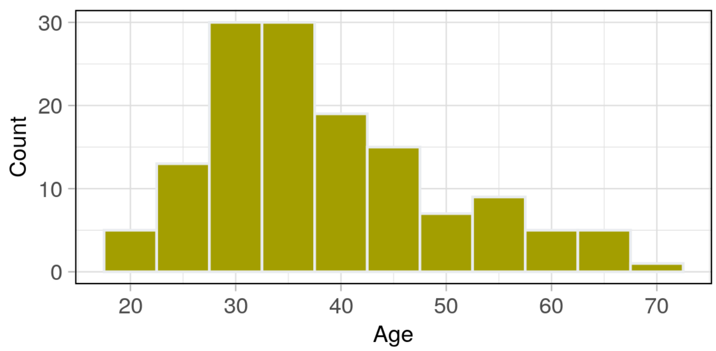


Figure 4.8: Histogram of participants' age in Experiment 3. The bin width is 5.

Table 4.3 specifies whether any correlations existed between our participants' Big-Five personality traits (TIPI questionnaire attributes) and scales of their disposition to trust other people (DT questionnaire attributes). Pearson Rank-Order Correlation analysis revealed that participants' agreeableness was moderately correlated with all the scales of their disposition of trust. Furthermore, people who were more open to experiences had a moderately higher disposition for assuming people's trusting stance.

4.4.4 Statistical Analysis

To study effects on the collected categorical data (e.g., some indicators of trust, including the participants' preferences for letting the robots cook for them or do their laundry asked in Step 6) while considering possible confounding factors, we used Generalized Linear Model (GLM) [83] with a binomial family. The factors listed in the following were initially included in the models. We then kept only a subset of those to minimize the AIC [18]. After checking the interaction effects, we came up with the final models presented in the next section. To analyze the qualitative data, collected using continues scales (e.g., the

Table 4.3: Correlations between the aspects of the participants' disposition of trust and their Big-Five personality traits

	Extroversion	**Agreeableness**	**Conscientiousness**	**Emotionally stability**	**Openness to experience**
Benevolence	p=.038 r=.176	p<.001 r=.411	p=.252 r=.097	p=.003 r=.252	p=.004 r=.244
Integrity	p=.023 r=.193	p<.001 r=.421	p=.247 r=.099	p=.001 r=.270	p=.044 r=.171
Competence	p=.799 r=.022	p<.001 r=.381	p=.058 r=.161	p=.004 r=.241	p=.237 r=.101
Trusting stance	p<.001 r=.298	p<.001 r=.361	p=.217 r=.105	p=.007 r=.228	p<.001 r=.335

behaviour evaluation measures asked in Step 5), LMM [10] was employed to check for the significant effects. For that, we used the same approach as with Experiment 1, described in Section 3.3.4. The considered factors were:

- The robot (i.e., robot *appearance*).
- The *condition* that each participant was assigned to, except for investigating the changes in the participants' perceptions over time (i.e., in six rounds). For that case, the rated *severity* of the mistake in every round was included instead.
- Whether it was the first robot or the second robot that the participant saw. This would be noted as "*encounter*" factor, with two levels of "*first*" and "*second*", throughout this thesis.
- Participants' demographics, i.e., their age and gender as well as their characteristics (based on TIPI and DT questionnaires).
- The gender they assigned to each robot.
- How much they like cooking/laundry, and the proportion of those tasks that they do themselves (measured through the post-experimental questionnaire, Step 7).

In addition, binomial tests and t-tests were occasionally conducted for testing if significant differences between two categories existed. The p-values obtained from the pairwise tests that were related to one of our hypotheses were adjusted for multiple testings using Holm-Bonferroni method.

4.4.5 Results

In this section, we first present some general findings regarding the robots and perceptions of their errors. We then describe how participants' understandings about the trainee robots changed over time, as well as how trust was impacted, in two subsections.

Perceived severity of the errors: Figure 4.9 illustrates the average participants' evaluation of the errors observed in the robots' behaviours. In fact, this plot shows how participants in each condition experienced throughout the study (two sets of teaching and practising with two robots, 6 rounds each). As expected from the results of Experiment 1, the big errors (i.e., adding a cleaning product in round 1 for all the conditions, plus round 6 for condition 3) were perceived much more severe than the other mistakes. All the small errors (i.e., replacing one item with another when preparing tea in round 2 for all the conditions, plus round 6 for condition 2) were in the same severity range, close to what was previously found (see Figure 4.4,(B): Replacing). Here, same is before, small errors were rated less severe than 500/1000 on average in every case. Independent-samples t-tests did not show any statistical differences between replacing items from the right side or the left side of the tea ($t(373.9) = 1.009, p = .313$). Furthermore, the difference between replacing either of the two types of milk ($t(324.0) = 0.253, p = .801$) or either of the two types of sweetener ($t(325.0) = -0.026, p = .979$) was not detected to be significant.

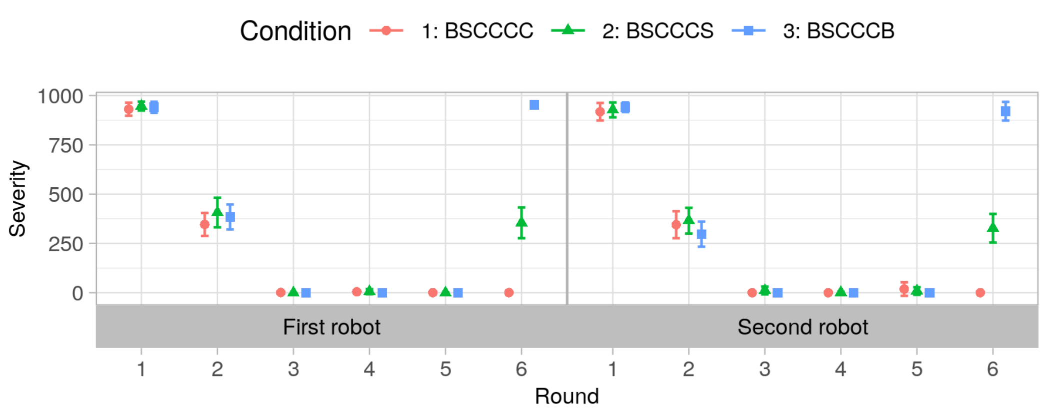


Figure 4.9: Average rated severity of robots' errors for each condition. Three situations of performing Correct actions (C), making a Small mistake (S), and making a Big error (B) can be identified based on these ratings. 95% confidence intervals are shown.

Perceived gender of the robots: The gender assigned by the participants to the robots in different conditions, and also based on two different appearance types and whether a particular robot was the first or the second one each participant taught (i.e., the *encounter* factor) are shown in Figure 4.10. We fit a GLM with *condition*, *appearance* and *encounter* as factors. According to the model (Table 4.4), there were no significant differences between variations in *appearance* and *encounter*. However, people in condition 2 perceived the gender of the robots significantly different than those in condition 1. As Figure 4.10 illustrates, more participants in condition 2 reflected that the robots are definitely male.

Table 4.4: Generalized Leaner Model predicting participants' perception of robots' gender

Covariate	**Estimate**	**SE**	**z**	
Condition2	-1.09	0.41	-2.678	**
Condition3	-0.35	0.45	-0.782	
EncounterSecond	0.06	0.32	0.194	
AppearanceUntidy	-0.16	0.32	-0.492	

Condition 1, First *encounter* and Tidy *appearance* are the baselines.
** = $p < .01$

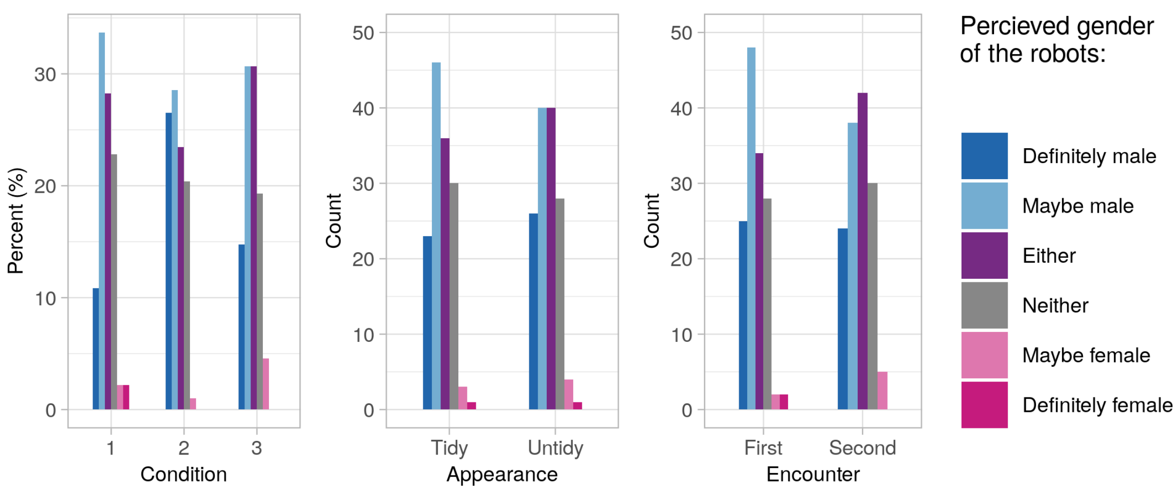


Figure 4.10: The gender that each participant assigned to every robot, based on *condition*, *appearance* and *encounter* factors.

Overall, participants mostly assumed that the gender of our robots was *maybe male* (86 times over $139 \times 2 = 248$ ratings, as each participant chose a gender for each of the two robots that were technically Pepper robots). Two options of *either male or female* and *neither male nor female* also were selected frequently (76 and 58 times, respectively). While 49 times robots were perceived as *definitely male*, only 2 times participants perceived

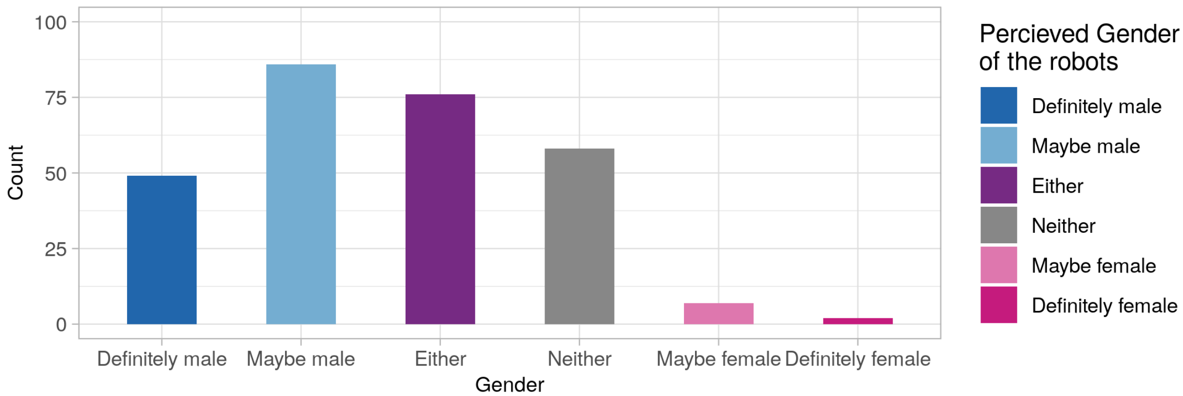


Figure 4.11: Overall perception of gender for the robots in the experiment. Each participant rated the robots with both the tidy and the untidy appearances.

the robot as *defiantly female*. There were also 7 instances when people selected the gender of the robot to be *maybe female*. Figure 4.11 presents these results.

Comparisons between two appearances: In the post-experimental questionnaire, we asked each participant to compare the two robots in terms of multiple attributes, or choose if they seemed equal. The results are shown in Figure 4.12. Regarding all five measures, the majority of people believed that robots with different clothing styles were equal. However, according to a binomial test with 68 trails, people significantly selected the tidy robot to be more professional than the untidy one ($p < .001$). For the robot that was more skilled ($p = .354$) or more experienced ($p = .397$), or had greater authority ($p = .500$), or people could trust more ($p = .256$), there was no significant difference between tidy or untidy appearances according to the tests.

Impact on the perceptions

Effects of *appearance*: Based on the measures that appeared after every teaching + practising round, perceived attributes of the robots with two appearances are presented in Figure 4.13. When the severity of mistakes decreased from round 1 to 3, all the scales improved and remained high until the sixth rounds, and then, dropped when the errors happened again. These measurements did not appear to be affected by the clothing style of the robots (the statistical analysis would be presented later). Paired-samples t-tests did not reveal any significant differences in the perception of severity of small ($t(187) = -0.910, p = .364$) and big ($t(182) = 0.907, p = .366$) mistakes made the tidy or untidy

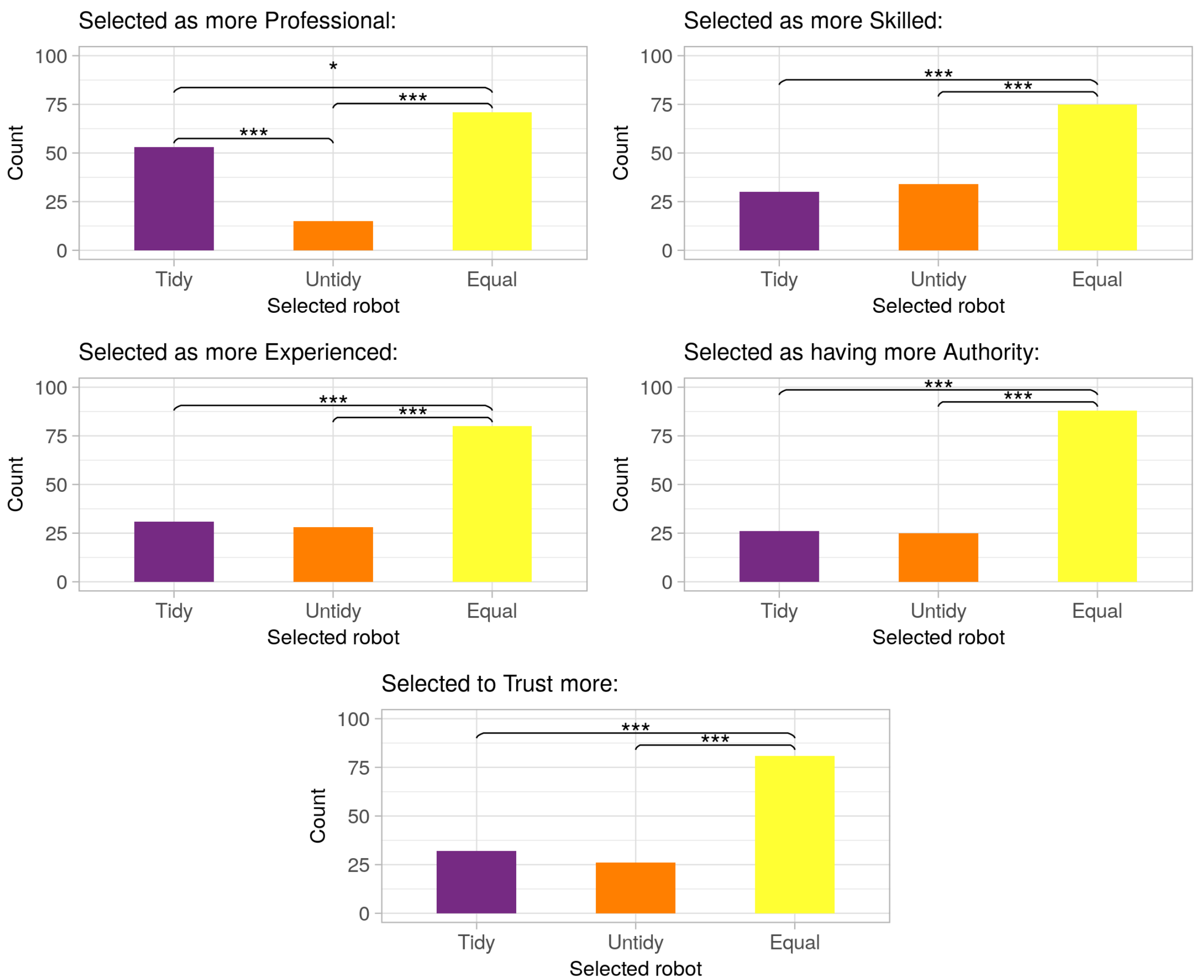


Figure 4.12: Comparisons between the two robots with only difference being their clothing style. $* = p < .05$; $*** = p < .001$, according to binomial tests.

robots.

Effects of *encounter* factor: Another factor that seemed important was the order in which each participant taught the robots (i.e., the *encounter* factor). Figure 4.14 demonstrates the participants' perceptions of the first and the second robot that they observed. Same as Figure 4.13, here there was also a correlation between improvements in the perceptions and enhancements in the behaviours of the robots. As will be shown with the models, the encounter factor could significantly affect all the measures, except the robots' perceived calmness. Using paired-samples t-tests, we detected some significant

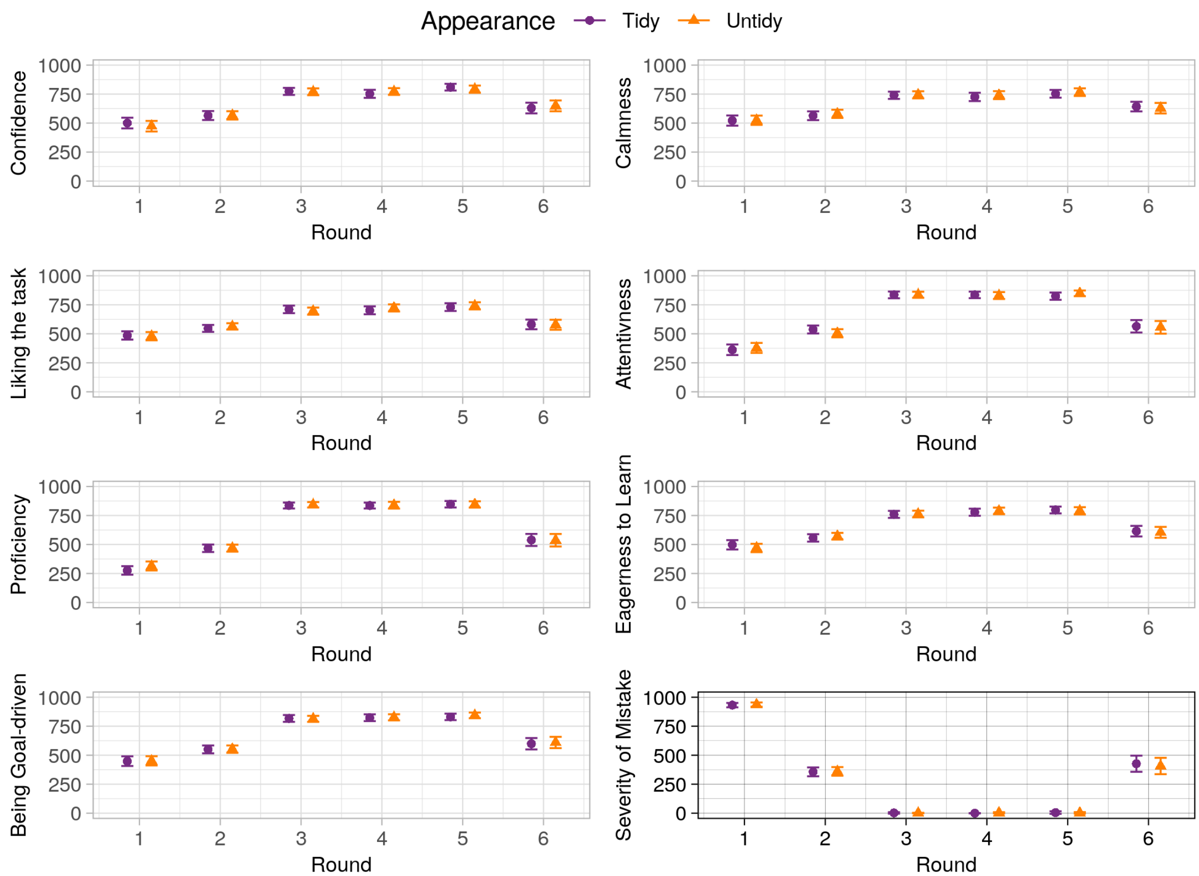


Figure 4.13: Averages of participants' perceptions of the robots' attributes, grouped by robot *appearance*. Error bars represent 95% confidence intervals and the results are pooled regardless of the condition.

differences between the first and the second robots in certain rounds and for some of the tested attributes. These are marked on Figure 4.14.

Even though the appearance of the robots was ineffective in the perceived severity of mistakes, the small errors were perceived as less severe with the robot that was taught first ($t(187) = -2.098, p < .05$). There was a same difference for the big errors, but was only close to being statistically significant ($t(182) = -1.643, p = .102$).

Effects of *condition*: As participants experienced different behaviours of the robots based on the conditions that they were assigned to, that factor was essential to investigate. Figure 4.15 presents the averages of perceptions of the robots (note that each participant

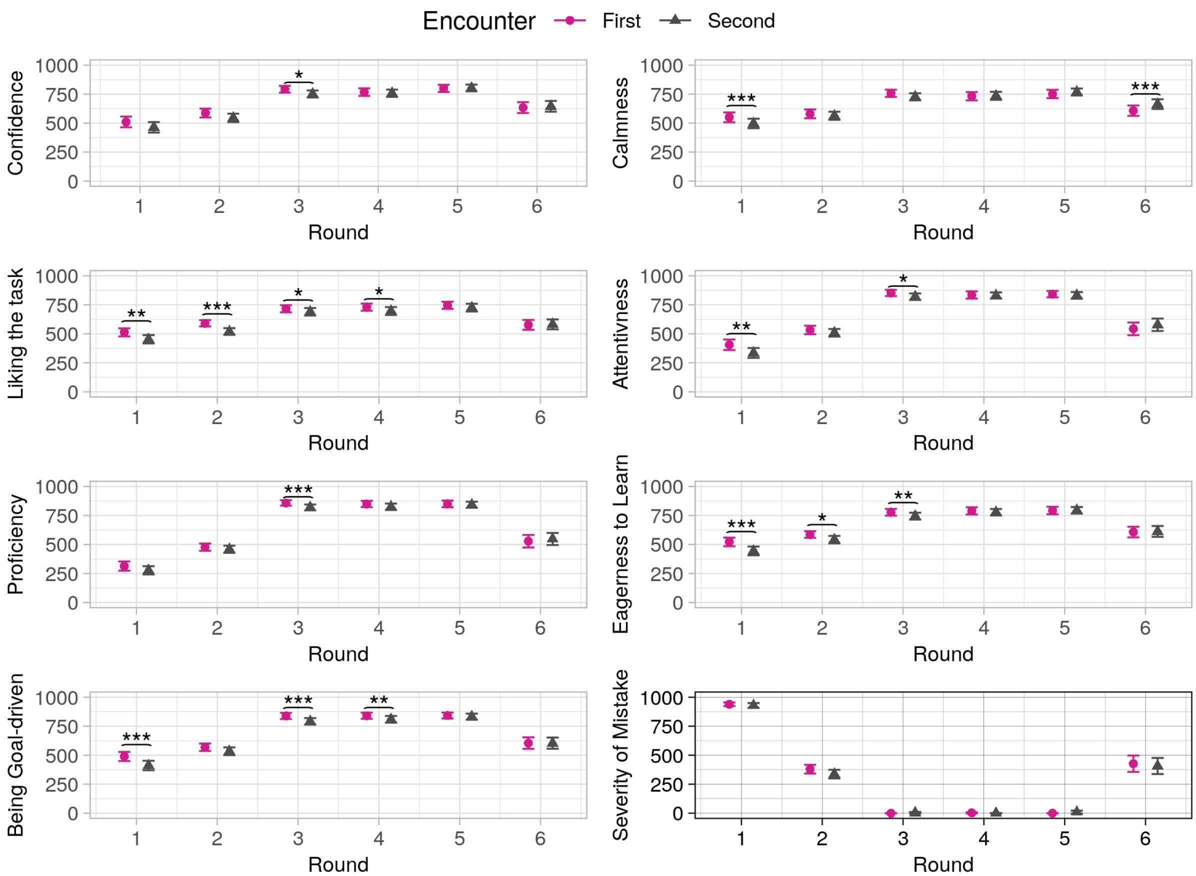


Figure 4.14: Averages of participants' perceptions of the robots' attributes, grouped by the *encounter* factor. Error bars represent 95% confidence intervals and the results are pooled regardless of the condition. * = $p < .05$; ** = $p < .01$; *** = $p < .001$, according to paired-samples t-tests.

rated two robots) according to the condition. There were the same general trends as in the previous two figures, with the exception that in the sixth rounds scores reduced in proportion to the severity of the mistake (according to *condition*). Remark that rounds 1 to 5 were identical for all the groups and therefore, there was no experimental difference among the three conditions on the left side of the dashed lines in Figure 4.15.

According to paired-samples t-tests that were adjusted using the Holm-Bonferroni method for multiple testings, all the measured attributes significantly increased from round 1 to round 2, and from round 2 to round 3. This was correlated with the improvement in

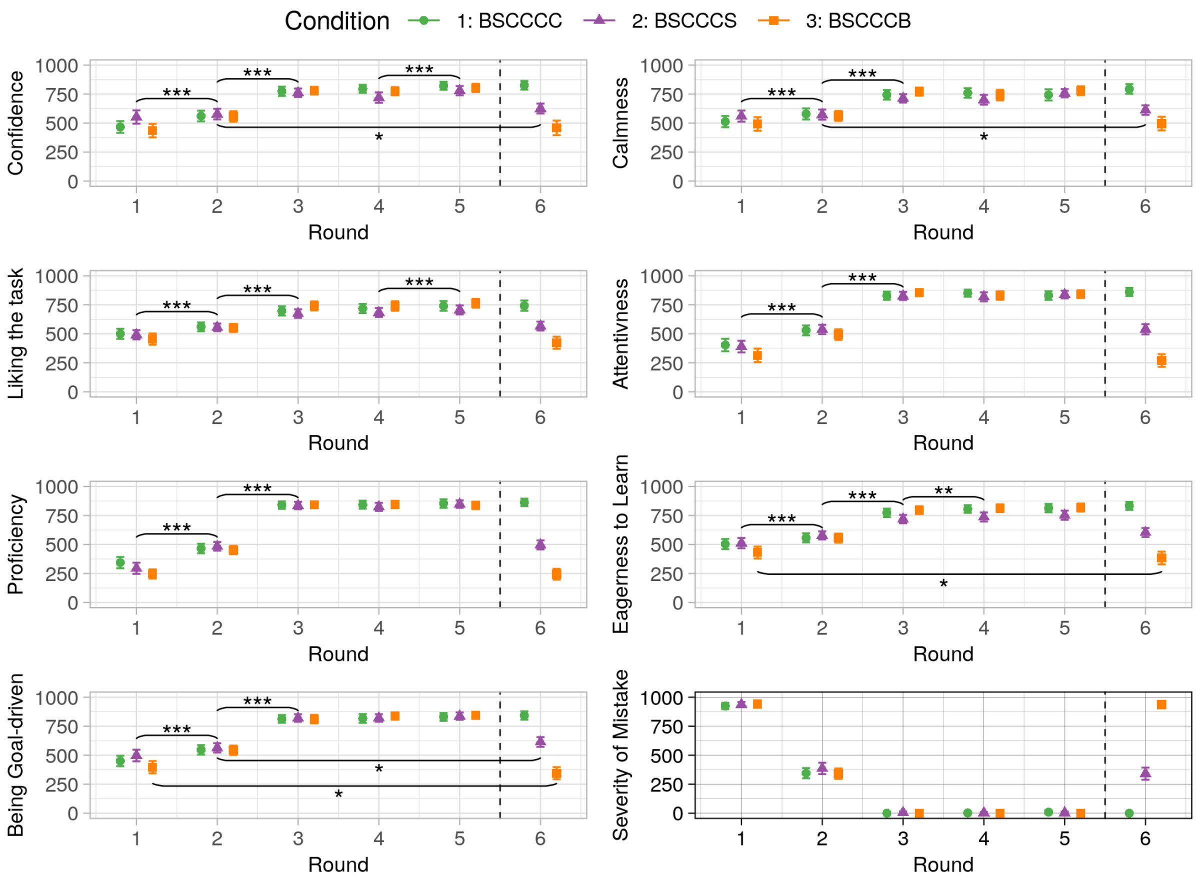


Figure 4.15: Averages of participants' perceptions of the robots' attributes, grouped by *condition*. Error bars represent 95% confidence intervals and the results are pooled regardless of the robot appearance. * = $p < .05$; ** = $p < .01$; *** = $p < .001$, according to t-tests. The comparisons marked on the left side of the dashed lines are about the entire sample, while the other comparisons are concerning participants in a single condition.

the behaviour of the robots. By also comparing round 3 with round 4, as well as round 4 with round 5, three instances were noticed in which the perceptions of the robots improved significantly, even though their behaviour was the same. The perception of confidence ($t(138) = 3.739, p < .001$) and liking the task ($t(138) = 3.997, p < .001$) were higher in round 5 compared with round 4. In addition, participants perceived higher eagerness to learn in round 4 of teaching robots compared with round 3 ($t(138) = 3.445, p < .05$).

Further paired-samples t-tests revealed that within conditions, there were some statistically significant differences in the perceptions when the same mistakes happened at the beginning (either round 1 or 2) or at the end (round 6). For participants in condition 2 and regarding the small mistakes of the robots, perceptions of confidence ($t(48) = 2.140, p < .05$), calmness ($t(48) = 2.0545, p < .05$) and being goal-driven ($t(48) = 2.304, p < .05$) were significantly higher in round 6 compared with round 2. Concerning the big errors for people in condition 3, the same effect in the opposite direction was detected; the robots were perceived as less eager to learn ($t(43) = -2.243, p < .05$) and less goal-driven ($t(43) = -2.283, p < .05$) after big mistakes in round 6 compared with round 1.

Predicting models: To further investigate changes in the perceptions of the participants regarding our trainee robots, the LMMs predicting the attributes are summarized in Table 4.5 and 4.6. As can be seen, *severity* and *encounter* were two factors that affected every measure (expect *encounter* did not affect robots' calmness). However, the appearance of the robot and participants' age and gender did not affect any of the attributes.

All the aspects we measured (expect calmness) were generally perceived as higher regarding the first robot each participant taught. Moreover, all the ratings were found to be negatively correlated with the severity of mistakes in each round. About the TIPI scales, we found that more extroverted people perceived the robots as less confident ($se = 5.99, t = -2.385, p < .05$), calm ($se = 6.76, t = -3.298, p < .01$), attentive ($se = 5.24, t = -2.121, p < .05$) and goal-driven ($se = 5.60, t = -2.726, p < .01$). Participants who had higher conscientiousness rated the robot to be more attentive ($se = 7.96, t = 3.258, p < .01$) and goal-driven ($se = 8.52, t = 2.056, p < .05$). The robots appeared to be calmer to those with higher emotional stability ($se = 7.67, t = 1.940, p = .054$). Finally, participants who were more open to experiences rated the robots as more liking the task ($se = 6.92, t = 2.439, p < .05$), proficient ($se = 5.56, t = 2.159, p < .05$) and eager to learn ($se = 6.63, t = 2.158, p < .05$). About the DT questionnaire items, people with a higher disposition of trusting peoples' competencies viewed the robot as more liking the task ($se = 8.54, t = 3.127, p < .05$) and more eager to learn ($se = 8.18, t = 3.002, p < .01$). The robot's appeared to be more confident ($se = 7.27, t = 3.121, p < .01$), calm ($se = 7.79, t = 3.742, p < .001$), attentive ($se = 6.12, t = 2.610, p < .05$), proficient ($se = 5.38, t = 2.758, p < .01$) and goal-driven ($se = 6.55, t = 2.045, p < .05$) to those who believed more in peoples' trusting stance.

Impact on trust

In this part, findings from the trust and learning evaluation questionnaires (responded by every participant after teaching each of the two robots, Step 6) are presented.

Table 4.5: Linear Mixed-effects Models predicting the robots' attributes: confidence, calmness, liking the task, and attention to the task, while learning.

Covariate	Confidence			Calmness			Liking the task			Attention to the task		
	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t
Encounter												
First (b)												
Second	-24.19	9.04	-2.677 **				-37.19	7.65	-4.858 ***	-23.53	8.92	-2.637 **
Severity	-0.32	0.01	-27.212 ***	-0.25	0.01	-22.357 ***	-0.27	0.00	-27.383 ***	-0.52	0.01	-44.870 ***
TIPI												
Extraversion	-14.29	5.99	-2.385 *	-22.30	6.76	-3.298 **				-11.10	5.24	-2.121 *
Agreeableness												
Conscientiousness										25.95	7.96	3.258 **
Emotional Stability				14.87	7.67	1.940 .						
Openness to Experiences							16.88	6.92	2.439 *			
DT												
Benevolence												
Integrity												
Competence							26.69	8.54	3.127 *			
Trusting Stance	22.69	7.27	3.121 **	29.14	7.79	3.742 ***				15.97	6.12	2.610 *

. = $p < .1$; * = $p < .05$; ** = $p < .05$; *** = $p < .001$.
(b) = baseline level

Table 4.6: Linear Mixed-effects Models predicting the robots' attributes: proficiency, eagerness to learn, and being goal-driven, while learning.

Covariate	**Proficiency**			**Eagerness to Learn**			**Being Goal-driven**		
	Estimate	SE	t	Estimate	SE	t	Estimate	SE	t
Encounter									
First (b)									
Second	-23.87	8.37	-2.852 **	-30.36	7.88	-3.854 ***	-37.92	8.17	-4.642 ***
Severity	-0.59	0.01	-54.544 ***	-0.34	0.01	-33.569 ***	-0.42	0.01	-39.911 ***
TIPI									
Extraversion							-15.27	5.60	-2.726 **
Agreeableness									
Conscientiousness							17.52	8.52	2.056 *
Emotional Stability									
Openness to Experiences	12.00	5.56	2.159 *	14.32	6.63	2.158 *			
DT									
Benevolence									
Integrity									
Competence				24.57	8.18	3.002 **			
Trusting Stance	14.83	5.38	2.758 **				13.39	6.55	2.045 *

* = $p < .05$; ** = $p < .05$; *** = $p < .001$.
(b) = baseline level

Preferences in cooking and laundry tasks: To investigate participants' trust, we asked them to specify whether they allow the robot to cook dinner for them alone or collaboratively, or they would prefer to do that by themselves or buy food from a restaurant. The results grouped by three different factors (i.e., *appearance*, *encounter* and *condition*) are displayed in Figure 4.16. This figure also includes the same question with regards to doing laundry, for studying the transfer of trust to other tasks. The first two choices that allowed the robot to take part in any form, are grouped together as a sign of trusting. The two other choices that totally excluded the robot are considered together as indicators of not trusting it.

It can be noticed from Figure 4.16 that the percentage of people who did not trust the robot was increasing when the final mistakes were becoming more severe from condition 1 to condition 3. According to the figure, the number of people who trusted the robots based on their appearance or according to the order in which they taught them seems constant. Utilizing GLMs, we further examined the effects of *appearance*, *encounter* and *condition* on these two measures of trust, while considering the confounding factors. The models presented in Table 4.7 confirmed the mentioned observations. Regarding both the cooking and laundry scenarios, only a significant effect of the last impression of learning (i.e., *condition*) was detected among those factors, in addition to some effects of items in the TIPI and DT questionnaires.

Pairwise comparisons adjusted using the Holm-Bonferroni method showed that small mistakes of the robots in the sixth practising rounds (condition 2) could significantly decrease the trust ($se = 0.35, z = -3.184, p < .01$). Big errors at the end (condition 3) could break trust even more compared with small errors ($se = 0.34, z = -3.826, p < .001$). In terms of confounding factors, we found that participants who had a higher disposition for trusting peoples' benevolence were significantly more tended to let the robots participate in cooking dinner for them ($se = 0.14, z = 2.334, p < .05$). The same effect was also observed with those who were more open to new experiences ($se = 0.11, z = 2.136, p < .05$). We noticed a trend, approaching significance, suggesting those with higher dispositions for trusting people's competencies may rely more on the robots for cooking ($se = 0.14, z = 1.811, p = .070$). In contrast, people with higher conscientiousness had significantly less tendency for allowing the robots to cook for them ($se = 0.14, z = -2.130, p < .05$).

While studying the transfer of trust to a laundry scenario, the adjusted pairwise tests revealed that people in condition 2 had significantly lower faith in the robots helping with that task compared with those in condition 1 ($se = 0.374, z = -2.664, p < .05$). People lost more trust in robots doing their laundry when a huge mistake happened at the end of learning the cooking task (condition 3), compared with when a small mistake was made ($se = 0.32, z = -2.558, p < .05$). Participants with a higher level of deposition to trust others' competencies were more likely to trust robots to do their laundry ($se = 0.12, z = 3.366, p < .001$). It was also detected that those with greater openness to experience trusted the robots more in this scenario ($se = 0.10, z = 2.315, p < .05$). However, participants who were more emotionally stable trusted the robots less in the laundry task ($se = 0.10, z = -2.053, p < .05$).

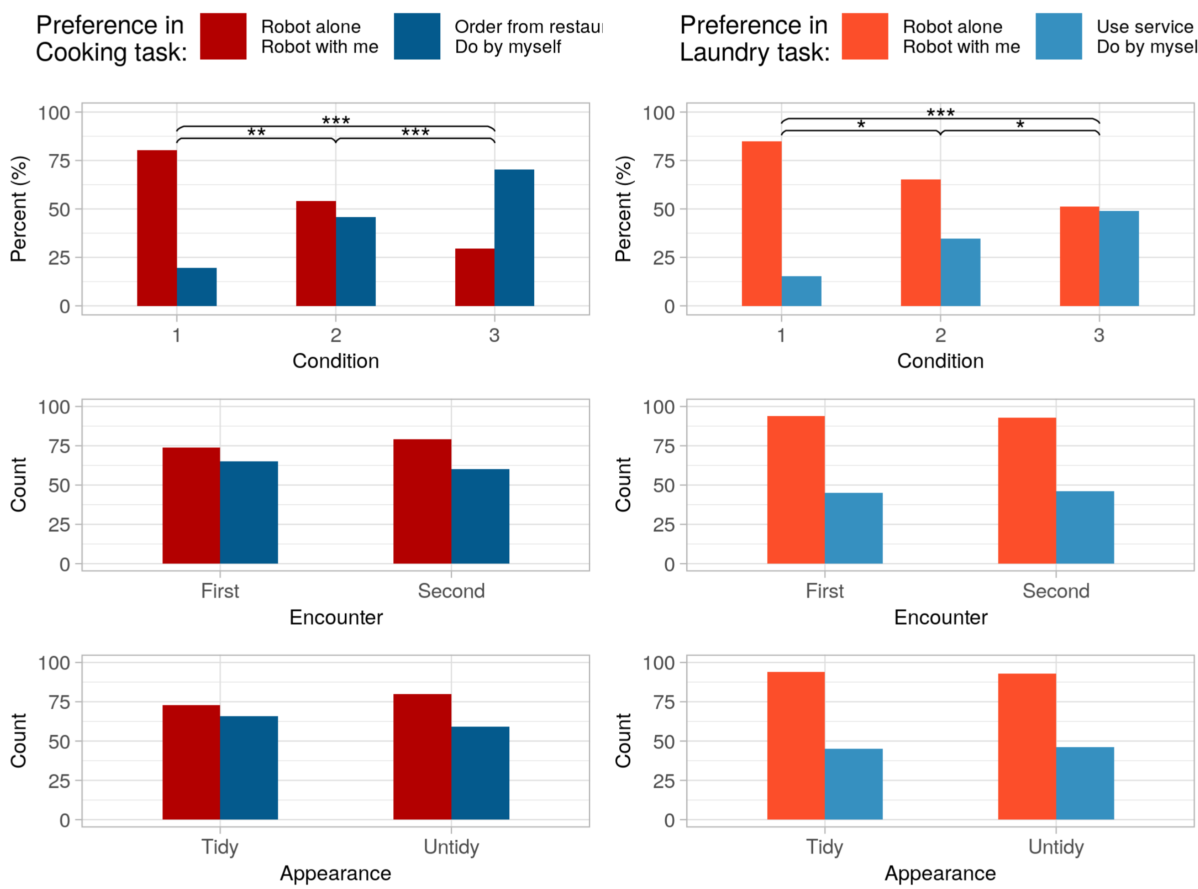


Figure 4.16: The preferences of participants for cooking and laundry tasks, grouped by *condition*, *encounter* and *appearance* factors. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Participants' attitudes: The participants also responded to four questions on continuous scales regarding the teaching experience and their opinions towards the robots, after finishing teaching each of them (Step 6). These items were the perceived realism of the teaching scenarios, perceived improvement of robots over time, expected success of robots in teaching cooking tasks to another robot, likelihood of using robots to assist with

Table 4.7: Generalized Leaner Models predicting participants' preferences in cooking and laundry tasks

Covariate	**Trust for cooking**				**Trust for doing laundry**			
	Estimate	SE	z		Estimate	SE	z	
Condition								
1 (BSCCCC) (b)								
2 (BSCCCS)	-1.12	0.35	-3.184	**	-1.00	0.37	-2.664	**
3 (BSCCCB)	-2.42	0.38	-6.336	***	-1.83	0.38	-4.858	***
TIPI								
Extraversion	-14.29	5.99	-2.385	*				
Agreeableness								
Conscientiousness	-0.29	0.14	-2.130	*				
Emotional Stability					-0.20	0.10	-2.053	*
Openness to Experiences	0.24	0.11	2.136	*	0.23	0.10	2.315	*
DT								
Benevolence	0.31	0.14	2.334	*				
Integrity								
Competence	0.26	0.14	1.811	.	0.41	0.12	3.366	***
Trusting Stance	-0.19	0.12	-1.509					

. = $p < .1$; * = $p < .05$; ** = $p < .05$; *** = $p < .001$.
B = Big error; S = Small error; C = Correct behaviour.
(b) = baseline level

chores in the future. The average ratings for each robot (i.e., *appearance* factor) and in various conditions are plotted in Figure 4.17. LMMs presented in Table 4.8 were employed to further investigate these measures.

About how realistic the teaching scenarios looked, although the ratings in condition 3 seem slightly lower than other conditions according to Figure 4.17, there was no significant effect of condition detected by the model. The only effective factors were found to be some items of the DT questionnaire. We noticed that participants with a higher disposition of trusting people's integrity ($se = 25.06, t = 2.365, p < .01$) and competence ($se = 41.13, t = 2.601, p < .05$) rated the scenarios as more realistic, but people with a higher disposition of trusting people's benevolence ($se = 24.76, t = 6.430, p < .001$) rated them as less realistic.

While the appearance of the robots could not affect any of those scales, the mistakes (i.e., *condition*) affected the rest of the measures described here. Adjusted pairwise comparisons showed that people felt significantly less improvement in the performance of the robots after the big mistake happened at the end, compared with after a small mistake happened ($se = 34.86, t = -3.265, p < .01$), and no mistake ($se = 35.49, t = -10.740, p < .001$). The small errors could also decrease this perception compared with no error ($se = 34.71, t = -7.703, p < .001$). In the measure, opposite to the realism of teaching scenario, those with a higher perception of people's benevolence perceived that robots improved more ($se = 10.49, t = 2.285, p =< .05$).

The only factor that could affect the expected success of robots in teaching the cooking tasks to another robot was *condition*. Small ($se = 41.68, t = 6.004, p < .001$) and big ($se = 42.81, t = 7.207, p < .001$) mistakes in the sixth rounds of practising could significantly decrease this measure, with no difference detected between them ($se = 42.17, t = 1.382, p = .169$). Finally, the likelihood that participants use the robots in future to assist them was found to be affected by the last mistakes and their disposition of trusting people's competencies ($se = 17.64, t = 3.102, p < .01$). When big mistakes happened in the sixth rounds, people appeared to less inclined to use the robots compared to when small mistakes happened ($se = 50.74, t = 2.303, p = .023$) and the behaviour was fully correct there ($se = 51.39, t = 6.612, p < .001$).

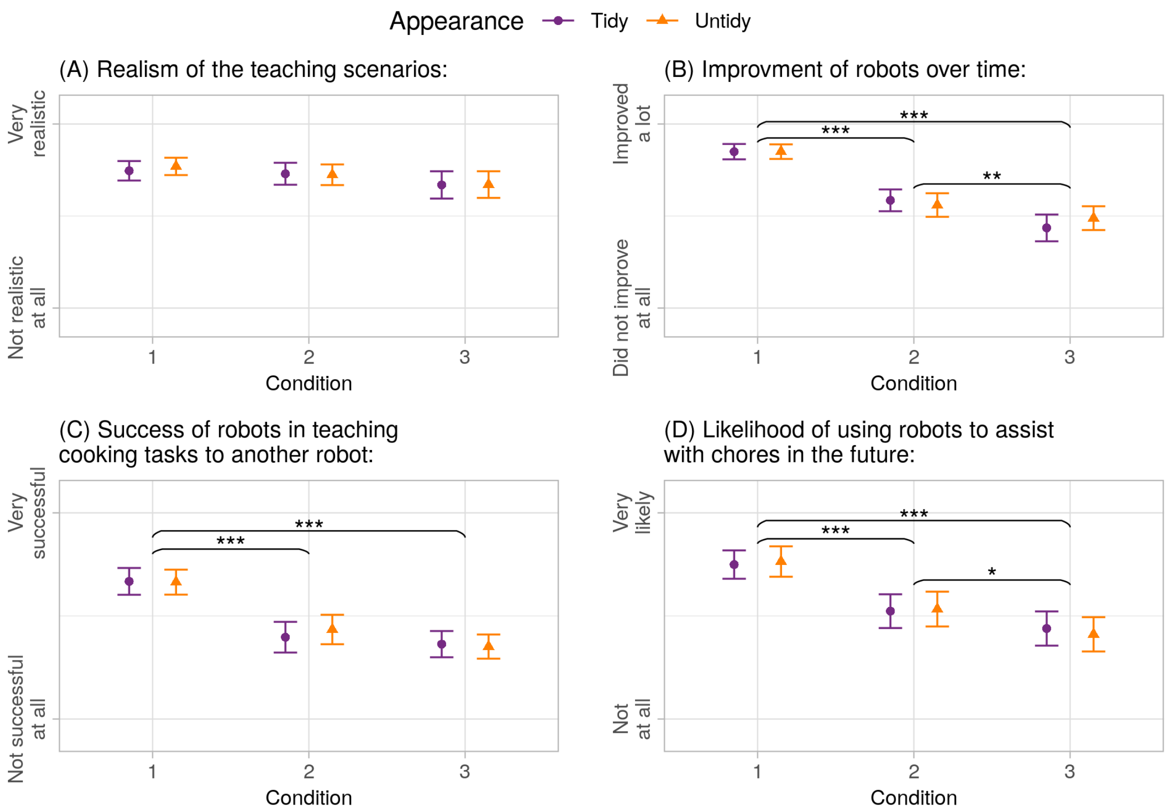


Figure 4.17: Participants' evaluation of (A) realism of the teaching scenarios, (B) improvement of the robots over time, (C) success of the robots in teaching cooking tasks to another robot, (D) likelihood of using the robots to assist with chores in the future. 95% confidence intervals are shown. * = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Table 4.8: Linear Mixed-effects Models predicting participants' teaching experience and their opinions about the robots

Covariate	**Realism of teaching scenarios**				**Improvement of robots over time**			
	Estimate	SE	z		Estimate	SE	z	
Condition								
1 (BSCCCC)								
2 (BSCCCS)					-267.33	34.71	-7.703	***
3 (BSCCCB)					-381.16	35.49	-10.740	***
DT								
Benevolence	-67.40	24.76	-2.722	**	23.97	10.49	2.285	*
Integrity	71.78	25.06	2.865	**				
Competence	41.13	15.81	2.601	*				
Trusting Stance								

Covariate	**Success of robots in teaching cooking tasks to another robot**				**Likelihood of using robots in the future for assistance**			
	Estimate	SE	z		Estimate	SE	z	
Condition								
1 (BSCCCC)								
2 (BSCCCS)	-250.26	41.68	-6.004	***	-222.95	50.00	-7.703	***
3 (BSCCCB)	-308.52	42.81	-7.207	***	-339.80	51.39	-10.740	***
DT								
Benevolence								
Integrity								
Competence					54.72	17.64	3.102	**
Trusting Stance								

. = $p < .1$; * = $p < .05$; ** = $p < .01$; *** = $p < .001$.
B = Big error; S = Small error; C = Correct behaviour.
(b) = baseline level

4.4.6 Discussion

We examined how human teachers interpreted their trainee robots, as the training progressed and they observed improvements in the robot task performance. Furthermore, trust in trainee robots under different conditions when no errors occurred in the last rounds versus when a small or a big error occurred was investigated. Our hypotheses are summarized in Table 4.9.

Giving participants experience of teaching

After teaching each robot, we asked the participants whether "the teaching scenario" looked very realistic or not at all, on a continuous scale. As observed in Figure 4.17 (A), the average of participants' evaluation of the teaching scenario was closer to being realistic, regardless of the ending faulty behaviours of the robots (i.e., no difference between *conditions* was detected). This implies that the designed virtual framework could successfully give the participants the experience of teaching a robot in all the tested situations. The usefulness of our framework for future virtual studies that need humans teach robots may also be understood from this finding.

Although a physical teaching situation could be implemented completely different (as will be noted in the Limitations section), the overall participants' experience would be the same if robots in this experiment were running a real learning system and interacting with participants live. For example, with appropriate pre-training of robot motor actions for object manipulations, instead of using pre-recorded videos, we could only leave it to participants to teach the part in which the robot needs to select the items to add, rather than asking them to demonstrate motion trajectories.

Clothing type of robots as a sign of professionality

With two different clothing styles for the robots (i.e., *tidy* or *untidy*), we aimed to investigate any possible effect of a priming on the perceived authority, as well as trust (**RQ1** [6]). The robot with a tidy appearance appeared to be significantly more proficient compared with the untidy one. People who are specialists usually dress properly and according to their occupation, so participants could correlate the neat look of the robot with its professionality, and differentiate between the two robots at least to some extent. However, there

[6] Can the professional look of a trainee robot (indicated by its clothing style) affect view of the teachers about the robot and their trust?

Table 4.9: Summary of hypotheses and their degree of support

	Impact of robot appearence:	
H1	People perceive higher levels of authority from a student robot and trust in it more when it has a professional dressing type.	Not supported
H2	Perception of liking the task, proficiency and eagerness to learn would be higher regarding the robot that has a professional clothing style.	Not supported
	Impact of robot errors:	
H3	As the robot makes smaller mistakes while learning, perceptions of its behaviours improve.	Supported
H4	People ignore a small error after they experience the robot improving in a task.	Rejected
H5	A big error after the learning process may cause a great loss of trust.	Supported
H6	There will be correlations between people's personality and their trust in the robot.	Supported
H7	There will be some relationships between people's disposition of trust to other humans and their trust in the robot.	Supported

were no other differences between the two appearance types, all over the experiment. In our study, the tidy clothing style itself could not convey anything beyond professionality, when the behaviours of two robots were the same. This suggests that the behaviours of a student robot are much more important than its clothing style for shaping human teachers' interpretations. Since we found neither the trust nor any items in the perceived attributes were affected by the robot's appearance, this study could not confirm **H1** and **H2**. In fact, both robots were perceived as equally professional, skilled, and experienced by the majority of participants, and they also had equal authority and were equally trustworthy, based on the direct comparisons.

Human teachers' perceptions changing over time

With respect to **RQ2** [7], we observed that as the robots made smaller mistakes over the first three training rounds, participants' perceptions of their behaviours improved. This was expected in **H3** concerning the robots' confidence, proficiency and being goal-driven. However, the improvements were significant in perceptions of all the attributes measured (i.e., calmness, liking the task, attentiveness and eagerness to learn as well as those in H3). One noteworthy finding was that the perceptions of confidence and liking the task were

[7] How do the perceptions of teachers about behaviours of a trainee robot change over time, while it is practising a task and appears to gradually improve?

found to be higher in the fifth rounds compared with the fourth rounds. When participants observed that the robots were capable of performing the task correctly for two consecutive rounds (i.e., the third and the fourth rounds), they felt the robots have become slightly more confident and they liked the task a bit more afterwards (the fifth rounds). More interestingly, we detected that the perception of eagerness to learn continued to increase in the fourth rounds, which was even one step after the behaviours of the robots stopped improving and remained the same compared with the previous step. This may imply that while a robot is progressing in a task, people strongly assume that the robot is becoming more eager to learn, the effect of which remains in place even after when it seems the learning has been completed.

There were also some exploratory findings regarding the changes in the perceptions that deserve a discussion. Related to the encounter factor, some measures in some particular rounds were found to be rated significantly higher concerning the first robot that each participant taught, compared with the second robot. Teachers may have had lower expectations when teaching a robot for the first time, regardless of its appearance, and as a result, rated its behaviour to be better. In line with that, we also detected that small mistakes made by the first robot were perceived as less severe than those made by the second robot.

While the behaviours of the robots were the same in the second rounds and in the sixth rounds for participants in condition 2 (i.e., robots made small errors in both cases), perceptions of confidence, calmness and being goal-driven were rated to be slightly higher in the sixth rounds compared with the second rounds. The fact that people saw the second faulty behaviour after three correct actions might be responsible for these findings; after observing the robots performing correctly, they may have become more permissive to final small faults of the robots in the mentioned form. However, when the same situation concerning the big errors was explored (people in condition 3 observed big errors in both the first and the sixth rounds), we detected that ratings of eagerness to learn and being goal-driven were lower in the sixth rounds than in the first rounds. We may argue that people were not expecting the robots to make those severe mistakes again (after they observed that their behaviours improved a lot from the first rounds) and thus, the negative impact of the big errors were quite more intense.

Last impression of a student robot affecting trust

To answer **RQ3** [8], we had hypothesized that teachers may ignore small mistakes when they detect that a robot is improving (**H4**). However, all the measures in the trust and learning evaluation form were significantly affected by even a small mistake occurring in the sixth training rounds. People felt a lesser overall improvement from robots and lost trust in them for doing the cooking task or performing another task (i.e., their laundry), when the robots made any errors after they seemed to have learned the tasks properly by making no mistakes in three consecutive rounds. The robots that made mistakes after learning were also deemed less successful in teaching cooking skills to another robot and less suited for assistance in the future. All of these, again, indicate the importance of robots' behaviours during the learning phase for setting attitudes of their users and impacting their trust.

Furthermore, a big error led to even a more considerable loss of trust for helping with the same or a new task, confirming **H5**. With such error, a larger drop in the participants' overall perceived improvement and the likelihood of using the robots in the future was detected. All those factors were more severely affected by big errors as compared with small errors. Similar to our findings, the trust had been found to be correlated with the magnitude of robot errors in [94]. We should again note that there was no difference in the expected success of the robots to teach another robot when they made a small final error of a big one. This may imply that people assume high qualification criteria for someone (or another robot) to teach a robot.

Personalities of human teachers impacted trust and perceptions

GLMs and LMMs were employed to explore the impact of different characteristics of participants on their perceptions about the robots, as well as on their trust (to study **RQ4** [9]). We had hypothesized that the personality traits of the participants (**H6**) and their disposition of trust in other people (**H7**) may affect some aspects of their trust in the robot. In our experiment, trust was found to be positively correlated with participants' belief in the benevolence of other people (same as [96]) and their openness to experience. Participants with higher conscientiousness were found to rely more on themselves/a restaurant instead of robots to cook for them. The likelihood of using the robot for a task other than cooking, when the participants were not familiar with robots' capabilities in that, was positively correlated with their level of trust in peoples' competencies, and again, their openness to

[8] How does the last impression of a student robot affect trust?

[9] Do different personalities of human teachers and their disposition to trust other people affect their perception of and their trust in a trainee robot?

experience. All these confirmed **H6** and **H7** which predicted that trust in a robot may vary based on individual differences. There were also some correlations between the above-mentioned factors and changes in the perceptions of participants regarding trainee robots while learning. Most notably, we found that more extroverted people perceived the robots as being less confident, calm, attentive and goal-driven.

4.5 Limitations and Future Work

While the participants evaluated the scenarios to be realistic, the situation could be different from a real-world study. In that case, teachers may have an opportunity to explore the robot's capabilities to better understand its limitations, using methods such as kinesthetic teaching. As a result, they might form a more realistic mental model about the abilities of the robot that can affect their trust. In a physical situation, the teaching could also involve more subtle actions and gestures such as pointing and looking at, touching the ingredients, and/or speech, whereas in our online platform this was replaced by mouse actions to do the teaching.

Furthermore, the way we implied progress in learning could be different when using a real mechanism that learns sequences of actions and trajectories. For instance, in a manipulation task, learning may reflect on trajectories becoming more smooth or the robot grasps more accurately. For the teachers, specially novice robot users, these attributes may not be as visible as the errors we introduced in this work. Thus, the changes we detected in the interpretation of the teachers from the beginning of the interaction are not easy to generalize.

Finally, depending on the task, robots may exhibit various kinds of errors in different stages of the task performance (e.g., in grasping, creating generalized trajectories and etc.). In this study, we only concentrated on object recognition failures. Future studies may consider the type of errors while learning as a factor to see how they could impact trust.

Chapter 5

Conclusion

5.1 Summary of Findings

In the following, we provide a list of the research questions studied in this thesis and a brief answers to each of them based on what we found:

How do nonverbal aspects of a humanoid robot's behaviour (different types of gaze and arm movements) influence the way human teachers perceive a trainee robot in terms of confidence, calmness, proficiency in the task, attention, eagerness to learn, being goal-driven, and liking the task? The short answer is: in several forms. We discovered that adjusting the way that a robot's gaze is split between the teacher and the task was effective in controlling perceived attention, calmness, liking the task, eagerness to learn, and being goal-driven. Additionally, the speed of performing the actions, along with the duration and frequency of the pauses, were effective in shaping perceptions of all the aspects mentioned in the research question. In Table 3.6, we provided guidelines on how each behavioural adjustment could be beneficial for a given purpose.

Can the professional look of a trainee robot (indicated by its clothing style) affect view of the teachers about the robot and their trust? We did not find any evidence suggesting that the way we changed the clothing style of a robot affected teachers' attitudes and trust. However, a robot that was neatly dressed conveyed more impression of being "professional" compared with the same robot that was dressed untidily.

How do the perceptions of teachers about behaviours of a trainee robot change over time, while it is practising a task and appears to gradually improve? Every aspect of teachers' interpretations that we measured (e.g., perceived attention, proficiency, etc.) was found to be improving when the behaviours of the robots improved (i.e., they were negatively correlated with the severity of robots' errors throughout the teaching rounds). We even detected some instances when robots appeared more confident, calm, and eager to learn by just repeating their correct practice of the task, without showing any improvement in their behaviour.

How does the last impression of a student robot affect trust? (i.e., if a robot is generally improving but an error happens in the end, how does that single mistake affect teachers' trust in the robot? Do people expect the robot to work properly afterwards and consider the faulty behaviour as an accident?) Human teachers were found to be strict about the actions of their student robots. Even a small error when practising a recently learned task had a significant negative effect on trust, and negatively impacted participants' general willingness to use the robots in the future. Larger errors had more adverse effects.

Do different personalities of human teachers and their disposition to trust other people affect their perception of and trust in a trainee robot? Yes. Similar to previous studies, we found trust in the robots to be correlated with several personality traits of participants as well as with their disposition to trust other people in a teaching interaction. More details were presented in Section 4.4.5.

5.2 Contribution to Knowledge

Our key contributions to the HRI knowledge are outlined below:

- We demonstrated that a student robot can adjust its nonverbal behavioural parameters in certain manners to convey specific information to its human teacher, who is observing its actions. Using the predictive model presented at the end of the discussion in Chapter 3, the robot designers would be informed to embed an ability in robots to adjust their behavioural factors and motion parameters whenever they need to imply something to their human teachers in a specific human-robot teaching situation. For example, a robot may want to appear to be less eager to learn for cases when it has more important tasks to do other than staying and being trained. In this case, we found it is helpful if the robot looks only at the teacher and makes long low-frequency pauses during its task performance. On the other hand, if a robot wants to motivate the teacher to move forward with teaching more material, i.e., by appearing to be as more eager to learn, one effective behavioural strategy for the robot was found to be gazing mainly at the task objects and checking the teacher occasionally while acting fast and with smooth motions.

- A virtual platform has been created based on findings from an experimental study. This allows a large group of participants to teach something (i.e., their own preference) to a robot, and helps people feel that they are acting as a teacher, even though the teaching process is virtual. Besides the experiment that was presented here, this platform can be used by other researchers in order to remotely study human teachers' attitude/behaviour/perception,etc. in situations when robots improve by learning or make errors.

- We discovered that the behaviours of a robot, even in the training phase, are very important for for shaping trust in teachers who have taught them the task. A small error of a trainee robot was found to significantly affect trust and some other aspects of the interaction. Results also suggested that the untidy clothing style of a robot may not affect trust or any other attribute of the robot, except its perceived professionality.

These findings contribute to the HRI knowledge by informing the designers and researchers about the different approaches that can be used to increase the efficiency and success of trainee robots.

5.3 Limitations and Future Work

Conducting our three experiments virtually enabled us to have a larger scale of data using the MTurk crowdsourcing framework, which lowers the risk of experimenter bias that can be caused due to direct face-to-face engagement of the participants with the experimenters [86]. Online experiments are also gaining popularity in HRI during the COVID-19 pandemic as a safe data collection method [38] and results obtained with this method have been shown to be comparable with in-person studies [61, 9]. We designed our virtual interfaces carefully to minimize the effects of confounding factors. Moreover, as discussed previously, strict inclusion criteria were used when recruiting participants and attention checks were put to identify those who were not paying enough attention to the task, which helped in improving the quality of the collected responses.

Nevertheless, our results may be different from when people teach a real robot in a live HRI scenario. As participants get involved in interacting with a physical robot, they might pay attention to different aspects of it to assess its capabilities and feel higher levels of empathy [108], which may affect perceived attributes and trust. In a recent study, researchers failed to replicate the results of their virtual study when conducting the same experiment in a real-world environment, in the context of playing a game with a robot [118]. Unlike that case, since the evaluation steps in our studies took place after prior (envisaged) teaching interactions, our participants were not distracted during the evaluation of the robot by another task (e.g., thinking about a game). All three experiments presented in this thesis, however, may possibly be conducted in the real world. Future studies should aim to validate our findings, and can also be extended in the ways noted for each study in Sections 3.4 and 4.5.

References

[1] Henny Admoni, Bradley Hayes, David Feil-Seifer, Daniel Ullman, and Brian Scassellati. Are you looking at me? Perception of robot attention is mediated by gaze type and group size. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 389–395. IEEE, 2013.

[2] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: A review. *Journal of Human-Robot Interaction*, 6(1):25, May 2017.

[3] Seyed Reza Ahmadzadeh, Ali Paikan, Fulvio Mastrogiovanni, Lorenzo Natale, Petar Kormushev, and Darwin G. Caldwell. Learning symbolic representations of actions from human demonstrations. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3801–3808. IEEE, 2015.

[4] Waqar Ali and Andrew B. Williams. Evaluating the Effectiveness of Nonverbal Communication in Human-Robot Interaction. In *HRI '20: Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 99–100. ACM, 2020.

[5] Pourya Aliasghari, Moojan Ghafurian, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. Effects of Gaze and Arm Motion Kinesics on a Humanoid's Perceived Confidence, Eagerness to Learn, and Attention to the Task in a Teaching Scenario. In *HRI '21: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, page 197–206. ACM, 2021.

[6] Pourya Aliasghari, Moojan Ghafurian, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. Impact of Nonverbal Robot Behaviour on Human Teachers' Perceptions of a Learner Robot. *Interaction Studies*, 2021 (In press).

[7] Sean Andrist, Bilge Mutlu, and Adriana Tapus. Look like me: Matching robot personality via gaze to increase motivation. In *CHI '15: Proceedings of the 33rd*

Annual ACM Conference on Human Factors in Computing Systems, pages 3603–3612. ACM, 2015.

[8] Murray R Barrick and Michael K Mount. The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, 44(1):1–26, 1991.

[9] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. Comparing the similarity of responses received from studies in Amazon Mechanical Turk to studies conducted online and with direct recruitment. *PLOS ONE*, 10(4):1–23, April 2015.

[10] Douglas Bates, Martin Mächler, Benjamin M. Bolker, and Steven C. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), Oct. 2015.

[11] Dorothy U. Behling and Elizabeth A. Williams. Influence of Dress on Perception of Intelligence and Expectations of Scholastic Achievement. *Clothing and Textiles Research Journal*, 9(4):1–7, 1991.

[12] J. Benitez, A. B. Wyman, C. M. Carpinella, and S. J. Stroessner. The authority of appearance: How robot features influence trait inferences and evaluative responses. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 397–404, 2017.

[13] Aude Billard, Sylvain Calinon, Rüdiger Dillmann, and Stefan Schaal. Robot programming by demonstration. In *Springer Handbook of Robotics*, pages 1371–1394. Springer, 2008.

[14] Aude Billard, Yann Epars, Sylvain Calinon, Stefan Schaal, and Gordon Cheng. Discovering optimal imitation strategies. *Robotics and Autonomous Systems*, 47(2-3):69–77, 2004.

[15] Susan A.J. Birch, Nazanin Akmal, and Kristen L. Frampton. Two-year-olds are vigilant of others' non-verbal cues to credibility. *Developmental Science*, 13(2):363–369, 2010.

[16] Ray L. Birdwhistell. Background to kinesics. *ETC: A Review of General Semantics*, 40(3):352–361, 1983.

[17] Mriganka Biswas, Marta Romeo, Angelo Cangelosi, and Ray B. Jones. Are older people any different from younger people in the way they want to interact with

robots? scenario based survey. *Journal on Multimodal User Interfaces*, 14(1):61–72, July 2020.

[18] Hamparsum Bozdogan. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, Sept. 1987.

[19] Cynthia Breazeal. Role of expressive behaviour for robots that learn from people. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3527–3538, Dec. 2009.

[20] Cynthia Breazeal and Brian Scassellati. Robots that imitate humans. *Trends in Cognitive Sciences*, 6(11):481–487, 2002.

[21] Andrew G. Brooks and Ronald C. Arkin. Behavioral overlays for non-verbal communication expression on a humanoid robot. *Autonomous Robots*, 22(1):55–74, Sept. 2007.

[22] Tom Bruneau. Chronemics: Time-binding and the construction of personal time. *ETC: A Review of General Semantics*, 69(1):72–92, 2012.

[23] Sylvain Calinon and Aude Billard. Learning of gestures by imitation in a humanoid robot. *Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions*, pages 153–178, 2007.

[24] Sylvain Calinon and Aude G Billard. What is the teacher's role in robot programming by demonstration?: Toward benchmarks for improved learning. *Interaction StudiesInteraction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, 8(3):441–464, 2007.

[25] Crystal Chao, Maya Cakmak, and Andrea L. Thomaz. Transparent active learning for robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–324. IEEE, 2010.

[26] Sumin Cho and Sungho Jo. Incremental online learning of robot behaviors from selected multiple kinesthetic teaching trials. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, 43(3):730–740, 2013.

[27] Josep Arnau Claret, Gentiane Venture, and Luis Basañez. Exploiting the robot kinematic redundancy for emotion conveyance to humans as a lower priority task. *International Journal of Social Robotics*, 9(2):277–292, Jan. 2017.

[28] Kerstin Dautenhahn. The art of designing socially intelligent agents: science, fiction, and the human in the loop. *Applied Artificial Intelligence*, 12(7-8):573–617, 1998.

[29] Kerstin Dautenhahn. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):679–704, April 2007.

[30] Kerstin Dautenhahn and Chrystopher L. Nehaniv. *Imitation in Animals and Arficafts*, volume 53. 2018.

[31] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*, page 251–258. IEEE Press, 2013.

[32] Munjal Desai, Kristen Stubbs, Aaron Steinfeld, and Holly Yanco. Creating trustworthy robots: Lessons and inspirations from automated systems. In *Proceedings of the AISB Convention: New Frontiers in Human-Robot Interaction*, April 2009.

[33] Giuseppe Di Cesare. The importance of the affective component of movement in action understanding. In *Modelling Human Motion: From Human Perception to Robot Design*, pages 103–116. Springer, Cham, 2020.

[34] Anca Dragan and Siddhartha Srinivasa. Generating legible motion. In *Proceedings of Robotics: Science and Systems IX*, 2013.

[35] N. J. Emery. The eyes have it: The neuroethology, function and evolution of social gaze. neuroscience and biobehavioral reviews. *Neuroscience and Biobehavioral Reviews*, 24(6):581–604, 2000.

[36] Brittney A. English, Alexis Coates, and Ayanna Howard. Recognition of gestural behaviors expressed by humanoid robotic platforms for teaching affect recognition to children with autism - A healthy subjects pilot study. In *International Conference on Social Robotics*, pages 567–576. Springer, 2017.

[37] Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H. Johnson. Eye contact detection in humans from birth. *Proceedings of the National Academy of Sciences*, 99(14):9602–9605, 2002.

[38] David Feil-Seifer, Kerstin S Haring, Silvia Rossi, Alan R Wagner, and Tom Williams. Where to Next? The Impact of COVID-19 on Human-Robot Interaction Research. *ACM Transactions on Human-Robot Interaction*, 10(1), 6 2020.

[39] Kerstin Fischer and Joe Saunders. Getting acquainted with a developing robot. In *HBU'12: Proceedings of the Third international conference on Human Behavior Understanding*, pages 125–133. Springer-Verlag, 2012.

[40] Lawrence K Frank. Tactile communication. *ETC: A Review of General Semantics*, 16(1):31–79, 1958.

[41] Frederik Funke and Ulf-Dietrich Reips. Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24(3):310–327, 2012.

[42] Moojan Ghafurian, Neil Budnarain, and Jesse Hoey. Improving humanness of virtual agents and users' cooperation through emotions. *arXiv preprint arXiv:1903.03980*, 2019.

[43] Donald Glowinski, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer. Toward a minimal representation of affective gestures. *IEEE Transactions on Affective Computing*, 2(2):106–118, April 2011.

[44] Jennifer Goetz, Sara Kiesler, and Aaron Powers. Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. RO-MAN 2003.*, pages 55–60, 2003.

[45] Samuel D. Gosling, Peter J. Rentfrow, and William B. Swann. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality*, 37(6):504–528, 2003.

[46] Edward T. Hall, Ray L. Birdwhistell, Bernhard Bock, Paul Bohannan, A. Richard Diebold, Marshall Durbin, Munro S. Edmonson, J. L. Fischer, Dell Hymes, Solon T. Kimball, Weston La Barre, S. J. Frank Lynch, J. E. McClellan, Donald S. Marshall, G. B. Milner, Harvey B. Sarles, George L Trager, and Andrew P. Vayda. Proxemics. *Current Anthropology*, 9(2/3):83–108, 1968.

[47] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y.C. Chen, Ewart J. De Visser, and Raja Parasuraman. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527, 2011.

[48] Kerstin S. Haring, Ariana Mosley, Sarah Pruznick, Julie Fleming, Kelly Satterfield, Ewart J. de Visser, Chad C. Tossell, and Gregory Funke. Robot Authority in Human-Machine Teams: Effects of Human-Like Appearance on Compliance. *Lecture Notes*

in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11575(16):63–78, 2019.

[49] Paul L. Harris and Kathleen H. Corriveau. Young children's selective trust in informants. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1179–1187, 2011.

[50] Cory J. Hayes, Maryam Moosaei, and Laurel D. Riek. Exploring implicit human responses to robot mistakes in a learning from demonstration task. *25th IEEE International Symposium on Robot and Human Interactive Communication, RO-MAN 2016*, pages 246–252, 2016.

[51] Erin Hedlund, Michael Johnson, and Matthew Gombolay. The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In *HRI '21: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, page 207–215. ACM, 2021.

[52] Joseph Henrich and James Broesch. On the nature of cultural transmission networks: Evidence from Fijian villages for adaptive learning biases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567):1139–1148, 2011.

[53] C. M. Heyes. Social learning in animals: Categories and mechanisms. *Biological Reviews of the Cambridge Philosophical Society*, 69(2):207–231, 1994.

[54] Jari K. Hietanen, Jukka M. Leppänen, Mikko J. Peltola, Kati Linna-aho, and Heidi J. Ruuhiala. Seeing direct and averted gaze activates the approach-avoidance motivational brain systems. *Neuropsychologia*, 46(9):2423–2430, 2008.

[55] Shanee Honig and Tal Oron-Gilad. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in Psychology*, 9:861, 2018.

[56] Ruud Hortensius, Felix Hekele, and Emily S. Cross. The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864, Dec. 2018.

[57] Adineh Hosseinpanah, Nicole C. Krämer, and Carolin Straßmann. Empathy for everyone?: The effect of age when evaluating a virtual agent. In *HAI '18: Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 184–190. ACM, 2018.

[58] Sandy H. Huang, Isabella Huang, Ravi Pandya, and Anca D. Dragan. Nonverbal robot feedback for human teachers. In *Proceedings of 3rd Conference on Robot Learning (CoRL)*, pages 1038–1051, 2019.

[59] Akira Ito, Shunsuke Hayakawa, and Tazunori Terada. Why robots need body for mind communication - An attempt of eye-contact between human and robot. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 473–478. IEEE, 2004.

[60] Masato Ito, Kuniaki Noda, Yukiko Hoshino, and Jun Tani. Dynamic and interactive generation of object handling behaviors by a small humanoid robot using a dynamic neural network model. *Neural Networks*, 19(3):323–337, 2006.

[61] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. Can We Trust Online Crowdworkers? Comparing Online and Offline Participants in a Preference Test of Virtual Agents. In *IVA '20: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, IVA '20. ACM, 2020.

[62] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10865–10875, 2019.

[63] Florian G. Kaiser, Karolin Glatte, and Mathis Lauckner. How to make nonhumanoid mobile robots more likable: Employing kinesic courtesy cues to promote appreciation. *Applied Ergonomics*, 78:70–75, July 2019.

[64] Tanaka Kazuaki, Ozeki Motoyuki, and Oka Natsuki. The hesitation of a robot: A delay in its motion increases learning efficiency and impresses humans as teachable. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 189–190. IEEE, 2010.

[65] Rachel Kendal, Lydia M. Hopper, Andrew Whiten, Sarah F. Brosnan, Susan P. Lambeth, Steven J. Schapiro, and Will Hoppitt. Chimpanzees copy dominant and knowledgeable individuals: Implications for cultural diversity. *Evolution and Human Behavior*, 36(1):65–72, 2015.

[66] Jaeseok Kim, Nino Cauli, Pedro Vicente, Bruno Damas, Filippo Cavallo, and Jose Santos-Victor. "iCub, clean the table!" A robot learning from demonstration approach using deep neural networks. In *18th IEEE International Conference on Autonomous Robot Systems and Competitions, ICARSC 2018*, pages 3–9. IEEE, 2018.

[67] Jaewook Kim, Sonya S. Kwak, and Myungsuk Kim. Entertainment robot personality design based on basic factors of motions: A case study with ROLLY. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 803–808. IEEE, 2009.

[68] Nathan Koenig, Leila Takayama, and Maja Matarić. Communication and knowledge sharing in human-robot interaction and learning from demonstration. *Neural Networks*, 23(8-9):1104–1112, Oct.–Nov. 2010.

[69] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. Behavioural responses to robot conversational failures. In *HRI '20: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, page 53–62. Association for Computing Machinery, 2020.

[70] Dana Kulić and Elizabeth Croft. Physiological and subjective responses to articulated robot motion. *Robotica*, 25(1):13–27, Aug. 2007.

[71] John E Laird, Kevin Gluck, John Anderson, Kenneth D Forbus, Odest Chadwicke Jenkins, Christian Lebiere, Dario Salvucci, Matthias Scheutz, Andrea Thomaz, Greg Trafton, et al. Interactive task learning. *IEEE Intelligent Systems*, 32(4):6–21, Aug. 2017.

[72] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80, 2004.

[73] Vera Maljkovic and Ken Nakayama. Priming of pop-out: I. Role of features. *Memory & Cognition*, 22(6):657–672, Nov. 1994.

[74] Justin Matejka, Michael Glueck, Tovi Grossman, and George Fitzmaurice. The effect of visual appearance on the performance of continuous sliders and visual analogue scales. In *CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5421–5432. ACM, 2016.

[75] D. Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3):334–359, 2002.

[76] Giorgio Metta, Paul Fitzpatrick, and Lorenzo Natale. YARP: Yet another robot platform. *International Journal of Advanced Robotic Systems*, 3(1):43–48, March 2006.

[77] Giorgio Metta, Lorenzo Natale, Francesco Nori, Giulio Sandini, David Vernon, Luciano Fadiga, Claes von Hofsten, Kerstin Rosander, Manuel Lopes, José Santos-Victor, Alexandre Bernardino, and Luis Montesano. The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks*, 23(8):1125–1134, 2010.

[78] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. The iCub humanoid robot: An open platform for research in embodied cognition. In *PerMIS '08: Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, pages 50–56. ACM, 2008.

[79] Ajung Moon, Boyd Panton, Hendrik F. Machiel Van der Loos, and Elizabeth A. Croft. Using hesitation gestures for safe and ethical human-robot interaction. In *IEEE Conference on Robotics and Automation: Workshop on Interactive Communication for Autonomous Intelligent Robots.*, pages 1–3. IEEE, 2010.

[80] AJung Moon, Chris A. C. Parker, Elizabeth A. Croft, and Hendrik F. Machiel Van der Loos. Design and Impact of Hesitation Gestures during Human-Robot Resource Conflicts. *Journal of Human-Robot Interaction*, 2(3):18–40, 2013.

[81] Yumiko Muto, Shoji Takasugi, Tomohito Yamamoto, and Yoshihiro Miyake. Timing control of utterance and gesture in interaction between human and humanoid robot. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*, pages 1022–1028. IEEE, 2009.

[82] Chrystopher L. Nehaniv, Kerstin Dautenhahn, Jens Kubacki, Martin Haegele, Christopher Parlitz, and Rachid Alami. A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication*, pages 371–377. IEEE, 2005.

[83] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[84] Aline Normoyle, Jeremy B. Badler, Teresa Fan, Norman I. Badler, Vinicius J. Cassol, and Soraia R. Musse. Evaluating perceived trust from procedurally animated gaze. In *MIG '13: Proceedings of Motion on Games*, pages 119–126. ACM, 2013.

[85] Eduardo B. Ottoni, Briseida Dogo De Resende, and Patrícia Izar. Watching the best nutcrackers: What capuchin monkeys (Cebus apella) know about others' tool-using skills. *Animal Cognition*, 8(4):215–219, 2005.

[86] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[87] Susan Perry. Conformism in the food processing techniques of white-faced capuchin monkeys (Cebus capucinus). *Animal cognition*, 12(5):705–716, 2009.

[88] Rifca Peters, Joost Broekens, and Mark A. Neerincx. Robots educate in style: The effect of context and non-verbal behaviour on children's perceptions of warmth and competence. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 449–455. IEEE, 2017.

[89] Karola Pitsch, Katrin S. Lohan, Katharina Rohlfing, Joe Saunders, Chrystopher L. Nehaniv, and Britta Wrede. Better be reactive at the beginning. Implications of the first seconds of an encounter for the tutoring style in human-robot-interaction. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 974–981. IEEE, 2012.

[90] P. L.Patrick Rau, Ye Li, and Dingjun Li. A cross-cultural study: Effect of robot appearance and task. *International Journal of Social Robotics*, 2(2):175–186, 2010.

[91] Lionel P. Robert Jr., Rasha Alahmad, Connor Esterwood, Sangmi Kim, Sangseok You, and Qiaoning Zhang. A review of personality in human–robot interactions. *Foundations and Trends in Information Systems*, 4(2):107–212, 2020.

[92] Ben Robins, Kerstin Dautenhahn, Chrystopher L. Nehaniv, N. Assif Mirza, Dorothée François, and Lars Olsson. Sustaining interaction dynamics and engagement in dyadic child-robot interaction kinesics: Lessons learnt from an exploratory study. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 716–722. IEEE, 2005.

[93] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Joe Saunders. Investigating human perceptions of trust in robots for safe hri in home environments. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, page 375–376. Association for Computing Machinery, 2017.

[94] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. How the timing and magnitude of robot errors influence peoples' trust of robots in an emergency scenario. In *9th International Conference on Social Robotics*, pages 42–52. Springer International Publishing, 2017.

[95] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. Human perceptions of the severity of domestic robot errors. In *International Conference on Social Robotics (ICSR 2017)*, pages 647–656. Springer, 2017.

[96] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. The impact of peoples' personal dispositions and personalities on their trust of robots in an emergency scenario. *Paladyn, Journal of Behavioral Robotics*, 9(1):137–154, 2018.

[97] Alessandra Rossi, Fernando Garcia, Arturo Cruz Maya, Kerstin Dautenhahn, Kheng Lee Koay, Michael L. Walters, and Amit K. Pandey. Investigating the Effects of Social Interactive Behaviours of a Robot on People's Trust During a Navigation Task. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11649 LNAI, pages 349–361. Springer, 2019.

[98] Alessandra Rossi, Patrick Holthaus, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. Getting to know pepper: Effects of people's awareness of a robot's capabilities on their trust in the robot. page 246–252, 2018.

[99] Alessandra Rossi, Sílvia Moros, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L. Walters. Getting to know kaspar: Effects of people's awareness of a robot's capabilities on their trust in the robot. page 1–6, 2019.

[100] Martin Saerbeck and Christoph Bartneck. Perception of affect elicited by robot motion. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 53–60. IEEE, 2010.

[101] Maha Salem and Kerstin Dautenhahn. Evaluating Trust and Safety in HRI: Practical Issues and Ethical Challenges. *Emerging Policy and Ethics of Human-Robot Interaction: A Workshop at 10th ACM/IEEE Int Conf on Human-Robot Interaction (HRI 2015)*, 2015.

[102] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[103] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. Would You Trust a (Faulty) Robot? Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust. In *Proceedings of the Tenth Annual*

ACM/IEEE International Conference on Human-Robot Interaction, HRI '15, page 141–148. ACM, 2015.

[104] Andrea Satrapa, Marcia Burrattino Melhado, Margarida Maria Curado Coelho, Emma Otta, Ruben Taubemblatt, and Waléria De Fayetti Siqueira. Influence of style of dress on formation of first impressions. *Perceptual and Motor Skills*, 74(1):159–162, 1992.

[105] Joe Saunders, Chrystopher L. Nehaniv, Kerstin Dautenhahn, and Aris Alissandrakis. Self-imitation and environmental scaffolding for robot teaching. *International Journal of Advanced Robotic Systems*, 4(1):14, 2007.

[106] Shane Saunderson and Goldie Nejat. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics*, 11(4):575–608, 2019.

[107] Caroline Schuppli, Ellen J.M. Meulman, Sofia I.F. Forss, Fikty Aprilinayati, Maria A. van Noordwijk, and Carel P. van Schaik. Observational social learning and socially induced practice of routine skills in immature wild orang-utans. *Animal Behaviour*, 119:87–98, 2016.

[108] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 125–132. ACM, 2015.

[109] Maia Stiber and Chien-Ming Huang. Not all errors are created equal: Exploring human responses to robot errors with varying severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, page 97–101. ACM, 2020.

[110] Dag Sverre Syrdal, Kerstin Dautenhahn, Sarah N. Woods, Michael L. Walters, and Kheng Lee Koay. Looking good? Appearance preferences and robot personality inferences at zero acquaintance. *AAAI Spring Symposium - Technical Report*, SS-07-07:86–92, 2007.

[111] Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. Taxonomy of trust-relevant failures and mitigation strategies. In *HRI '20: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, page 3–12. ACM, 2020.

[112] Horst Treiblmaier and Peter Filzmoser. Benefits from using continuous rating scales in online survey research. In *Thirty Second International Conference on Information Systems*, pages 2087–2099. AIS/ICIS, 2011.

[113] Ana Lucia Pais Ureche, Keisuke Umezawa, Yoshihiko Nakamura, and Aude Billard. Task Parameterization Using Continuous Constraints Extracted from Human Demonstrations. *IEEE Transactions on Robotics*, 31(6):1458–1471, 2015.

[114] Gill L. Vale, Emma G. Flynn, Susan P. Lambeth, Steven J. Schapiro, and Rachel L. Kendal. Public information use in chimpanzees (Pan troglodytes) and children (Homo sapiens). *Journal of Comparative Psychology*, 128(2):215–223, 2014.

[115] Sanne van Waveren, Elizabeth J. Carter, and Iolanda Leite. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, page 151–158. ACM, 2019.

[116] Fabio Vannucci, Giuseppe Di Cesare, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. A robot with style: Can robotic attitudes influence human actions? In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 952–957. IEEE, 2018.

[117] Astrid Marieke von der Pütten and Nicole C. Krämer. A survey on robot appearances. In *HRI '12: Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, page 267–268. ACM, 2012.

[118] Sebastian Wallkötter, Rebecca Stower, Arvid Kappas, and Ginevra Castellano. A Robot by Any Other Frame: Framing and Behaviour Influence Mind Perception in Virtual but not Real-World Environments. In *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction*, pages 609–618. ACM, 2020.

[119] Auriel Washburn, Akanimoh Adeleye, Thomas An, and Laurel D. Riek. Robot Errors in Proximate HRI: How Functionality Framing Affects Perceived Reliability and Trust. *ACM Transactions on Human-Robot Interaction*, 9(3), 2020.

[120] Andrew Whiten, Nicola McGuigan, Sarah Marshall-Pescini, and Lydia M. Hopper. Emulation, imitation, over-imitation and the scope of culture for child and chimpanzee. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528):2417–2428, 2009.

[121] Andrew Whiten and Erica van de Waal. The pervasive role of social learning in primate lifetime development. *Behavioral Ecology and Sociobiology*, 72(5):80, May 2018.

[122] Britta Wrede, Katharina J Rohlfing, Marc Hanheide, and Gerhard Sagerer. Towards learning by interacting. In *Creating Brain-Like Intelligence*, pages 139–150. Springer, Berlin, Heidelberg, 2009.

APPENDICES

Appendix A

Ethics Clearance Certificates

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

Principal Investigator: Kerstin Dautenhahn (Electrical and Computer Engineering)

Student investigator: Pourya Aliasghari (Systems Design Engineering)

Co-Investigator: Chrystopher Nehaniv (Systems Design Engineering)

File #: 42177

Title: Evaluating the Humans' Social Perception of Nonverbal Robot Feedback in Teaching Scenarios

The Human Research Ethics Committee is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 06/18/20 (m/d/y)

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Expiry Date: 06/17/21 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Committee

Jo cenaiko

Jo Cenaiko, Ethics Advisor, jrcenaik@uwaterloo.ca, 519-888-4567, ext. 30321

This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: Revision #1 Details.pdf

file: HITInformation_v3_20200611.docx

file: HITInformation_v3_20200611.pdf

file: StudyProcedures_v2_20200527.docx

file: StudyProcedures_v2_20200527.pdf

file: InformationandConsentForm_v3_20200611.docx

file: InformationandConsentForm_v3_20200611.pdf

file: EndMessage_v2_20200521.docx

file: EndMessage_v2_20200521.pdf

Approved Protocol Version 3 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Kerstin Dautenhahn (Electrical and Computer Engineering)

Student investigator: Pourya Aliasghari (Systems Design Engineering)

Co-Investigator: Moojan Ghafurian (Electrical and Computer Engineering)

Co-Investigator: Chrystopher Nehaniv (Systems Design Engineering)

File #: 42782

Title: Understanding Human Perceptions of the Severity of Robot Errors

---

The Human Research Ethics Committee is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 12/03/20 (m/d/y)

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Expiry Date: 12/04/21 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Committee

Heather Root, Senior Manager, Ethics, heather.root@uwaterloo.ca, 519-888-4567, ext. 30469

This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: HITInformation_v1_202011.docx

file: HITInformation_v1_202011.pdf

file: InformationandConsentForm_v2_202011.docx

file: InformationandConsentForm_v2_202011.pdf

file: EndMessage_v1_202011.docx

file: EndMessage_v1_202011.pdf

file: StudyProcedures_v1_202011.docx

file: StudyProcedures_v1_202011.pdf

Approved Protocol Version 2 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Kerstin Dautenhahn (Electrical and Computer Engineering)

Student investigator: Pourya Aliasghari (Systems Design Engineering)

Co-Investigator: Chrystopher Nehaniv (Systems Design Engineering)

Co-Investigator: Moojan Ghafurian (Electrical and Computer Engineering)

File #: 42731

Title: Evaluating a Student Robot while Learning a Task

The Human Research Ethics Committee is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 01/22/21 (m/d/y)

University of Waterloo Research Ethics Committees are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Committees are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREC) and IRB00007409 (CREC).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Expiry Date: 01/23/22 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Committee

Heather Root, Senior Manager, Ethics, heather.root@uwaterloo.ca, 519-888-4567, ext. 30469

This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: HITInformation_v2_202101.pdf

file: HITInformation_v2_202101.docx

file: StudyProcedures_v2_202101.pdf

file: StudyProcedures_v2_202101.docx

file: InformationandConsentForm_v2_202101.pdf

file: InformationandConsentForm_v2_202101.docx

file: EndMessage_v2_202101.pdf

file: EndMessage_v2_202101.docx

Approved Protocol Version 2 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.

Appendix B

Experiment 1 Online Interface

This appendix includes the screenshots form the designed online interface used for conducting Experiment 1.

Information and Consent Form

Evaluating the Humans' Social Perception of Nonverbal Robot Feedback in Teaching Scenarios

Research Team:

Principal Investigator: Professor Kerstin Dautenhahn,
Departments of Electrical and Computer Engineering/Systems Design Engineering, University of Waterloo, Canada
Phone: +1-519-888-4567 x31455
Email: kerstin.dautenhahn@uwaterloo.ca

Co-Investigator: Professor Chrystopher Nehaniv,
Departments of Systems Design Engineering/Electrical and Computer Engineering, University of Waterloo, Canada
Phone: +1-519-888-4567 x40168
Email: chrystopher.nehaniv@uwaterloo.ca

Co-Investigator: Dr. Moojan Ghafurian,
Department of Electrical and Computer Engineering, University of Waterloo, Canada
Email: moojan@uwaterloo.ca

Student Investigator: Pourya Aliasghari,
Department of Systems Design Engineering, University of Waterloo, Canada
Email: paliasghari@uwaterloo.ca

Overview:

As a part of studying human-robot interaction and due to the dominant role of nonverbal communications, it is necessary to explore how the nonverbal behaviours of robots influence humans. Additionally, researchers are working on developing intelligent robots that can learn new skills from people, even those who do not have any robotics knowledge. Motivated by the social learning process, the information exchange through feedback and interactivity is also essential in robot learning. As humans respond to a robot according to their interpretations of its behaviours, we first need to know how people naturally understand robots while observing them executing a learnt task. In this online study, we will ask you about your perceptions and feelings when you see videos of the robot performing learnt actions in front of you. You need to assume yourself in the situation that you have taught the robot at some point how to do the task, as a teacher. This experiment will be a part of a Master's thesis that will be submitted for the degree compilation of student investigator, Pourya Aliasghari.

I. Your responsibilities as a participant

Participation in this study will approximately take 25-30 minutes. Your answer to the questions will be logged. This does not require you to download anything onto your computer. The videos will be automatically paused if you switch to another window and this could increase the time needed for completing the study. The above estimation is based on your continuous engagement.

Procedure: The study will start after you give consent by checking the box and pressing the button at the end of this page. First, we will ask your demographic information (Only your gender, age, educational level, and culture). If you do not wish to answer certain questions, you may choose not to share that information. In the next step, to make you familiar with the context, we will show you a short video about teaching robots and your imaginary role as a teacher for them, to see how this situation looks like. Afterwards, you will watch 12 different videos of a humanoid robot performing one physical task in a simulation environment. After each one, which would take about 20-40 seconds, you will be asked to rate the behaviour of the robot in terms of 9 different attributes by moving some sliders. At the end, we will provide you with a code to submit the HIT and receive your remuneration.

You can complete this study only once. Your responses are valuable information that can help us in our research. Please pay attention and provide the answers as accurate as you can.

II. Your rights as a participant

Your participation in this study is voluntary. You may decline to answer any question(s) you prefer not to answer in the demographic information form. You may decide to stop any time. If this is the case, please close the page and leave this HIT. Then, contact us by email (sirrl.waterloo@gmail.com) with your Mechanical Turk ID, so we can remunerate you for the proportion of what you completed. In this case your data will be discarded and not be used in the study.

Remuneration: You will be paid $2 if you complete all the steps of the experiment. If you do not wish to continue at any point, your participation in the study will end and you will be remunerated for the proportion of what you completed (number of questions that you have answered). If you wish to stop at any time, please contact us. You will receive the amount specified on the HIT ($1), plus an additional payment based on the number of questions that you have answered. For example, if you wish to stop in the middle of the study, you will receive ($1)+($1/2)= $1.5.

To accept your HIT, you must enter a code that will be provided to you in the last page of the study or will be sent to you by us if you end the study earlier. We will not be able to accept your HIT if you do not provide a valid code.

Benefits of the study: Participation in this study may not provide any personal benefit to you. However, the study will benefit the academic community and the research team. We plan to publish the results of this study, and researchers in the field of social robotics can potentially benefit from that. We hope the outcomes of this experiment pave the way for designing more intelligent robotic systems.

Confidentiality of identities and data: You will not be identified individually in any written reports of this research. Note that when information is transmitted over the internet, privacy cannot be entirely guaranteed. There is always a risk that your responses may be intercepted by a third party (e.g., government agencies, hackers). Any confidential information (namely your Mechanical Turk ID) will be removed from our records after 1 month from your participation. You may withdraw your consent and ask us for deleting your data within this period by contacting us at sirrl.waterloo@gmail.com. The data captured will be securely stored on a password-protected lab server with access only for the researchers. Anonymized data may be published publicly in the future, and will be stored for a minimum of 7 years.

III. Questions, comments, or concerns

Ethics clearance: This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#42177). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Contact Information: If you have any questions regarding this study, or would like additional information about your participation, please contact us by email at sirrl.waterloo@gmail.com.

IV. Consent

Thank you for your interest in our research and your assistance with this project. You must be 18 years of age or older to consent to take part in this study. By accepting this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

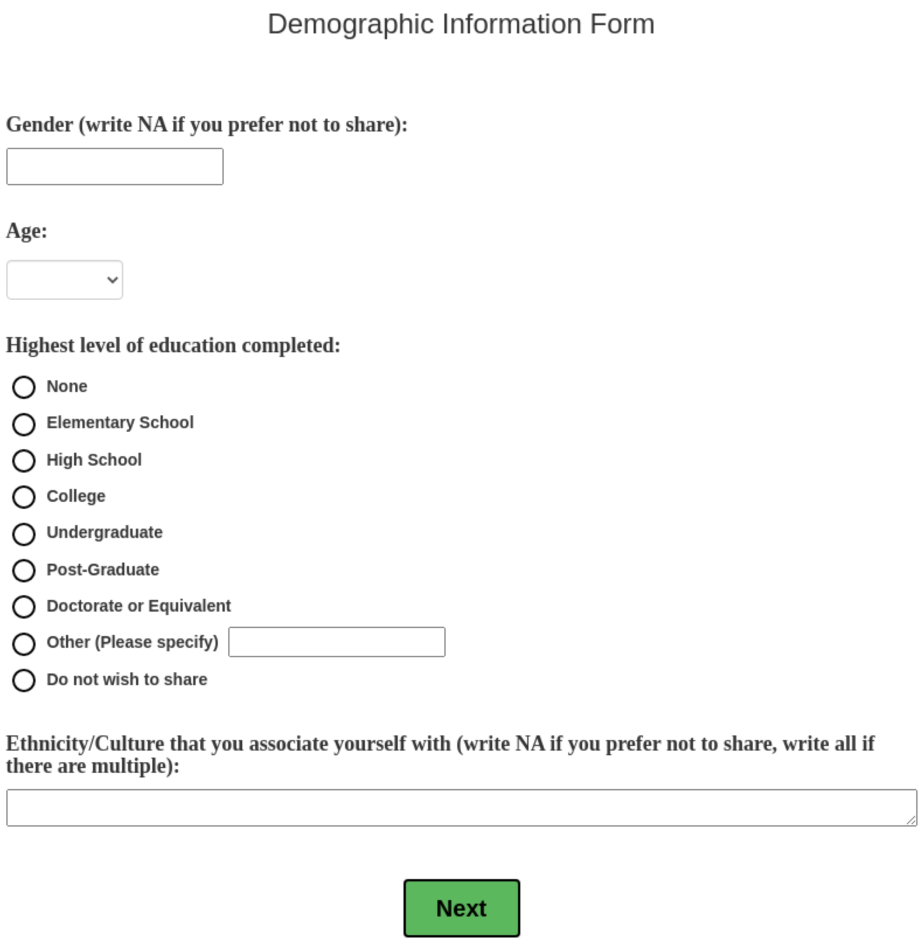
○ With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

Mechanical Turk ID:

HIT ID:

Next

Figure B.1: Information and Consent Form. You may need to zoom-in to read the text.


Demographic Information Form

Gender (write NA if you prefer not to share):

Age:

Highest level of education completed:

- None
- Elementary School
- High School
- College
- Undergraduate
- Post-Graduate
- Doctorate or Equivalent
- Other (Please specify)
- Do not wish to share

Ethnicity/Culture that you associate yourself with (write NA if you prefer not to share, write all if there are multiple):

Next

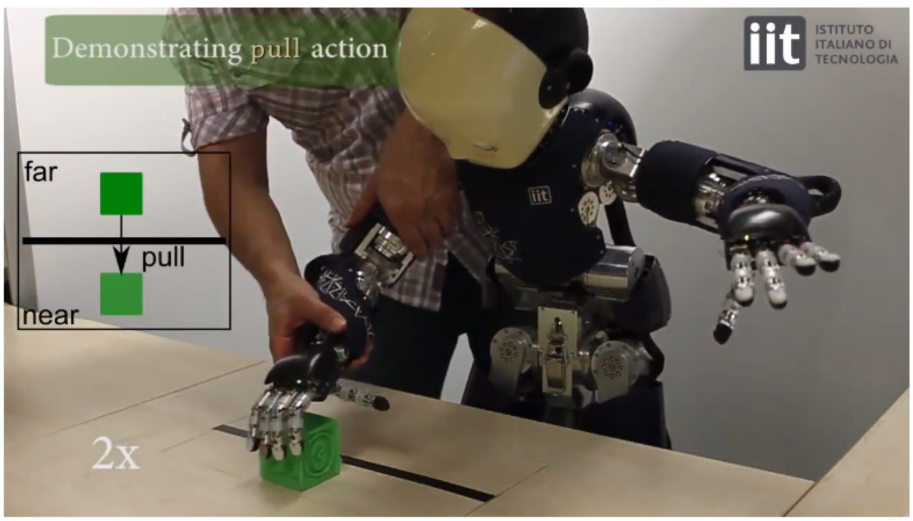
Figure B.2: Demographic Information Form asking the participants' about their gender, age, education and culture.

Robot Learning Familiarization

1/2

Modern intelligent robots are able to acquire new skills from people through observation or experience. For example, someone can grab the arms of a robot and show it a way of performing a task.

To see how a real robot learns, please watch this video showing the outcomes of two different scientific works. As you will see, a person first shows the robot (named iCub) how to do some actions, and then, the robot tries to execute the task on its own.



Source:
S. R. Ahmadzadeh, A. Paikan, F. Mastrogiovanni, L. Natale, P. Kormushev and D. G. Caldwell, "Learning symbolic representations of actions from human demonstrations," 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, 2015, pp. 3801-3808, doi: 10.1109/ICRA.2015.7139728.
J. Kim, N. Cauli, P. Vicente, B. Damas, F. Cavallo and J. Santos-Victor, ""iCub, clean the table!" A robot learning from demonstration approach using deep neural networks," 2018 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC), Torres Vedras, 2018, pp. 3-9, doi: 10.1109/ICARSC.2018.8374152.

Figure B.3: Robot Learning Familiarization step, part 1, with a video showing how humans can teach robots in real-world. The “next” button has not appeared yet.

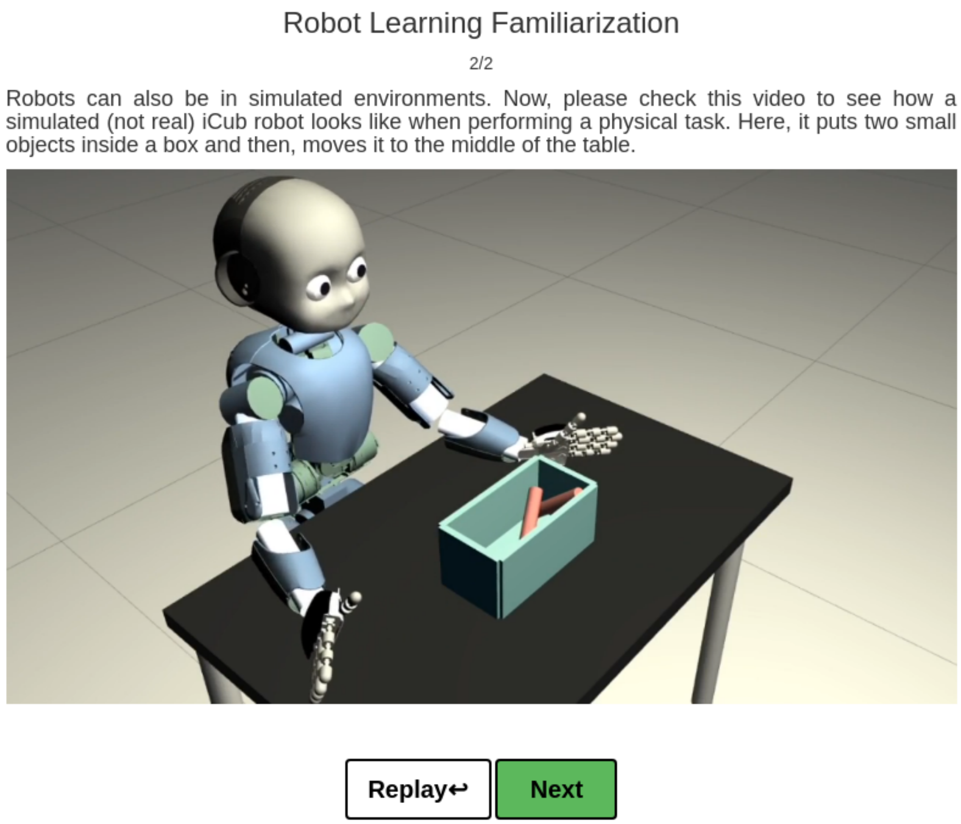


Figure B.4: Robot Learning Familiarization step, part 2, with a video showing our simulated iCub robot doing the designed task from various points of view.

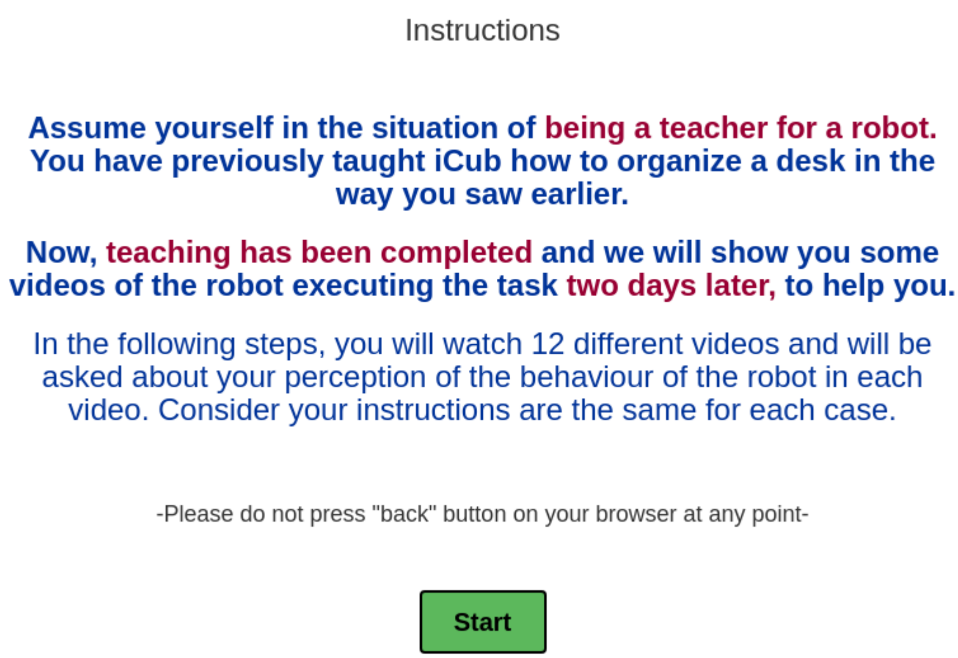


Figure B.5: Instructions, where the situation was explained to the participants using text. This example shows the description shown to people in group 2

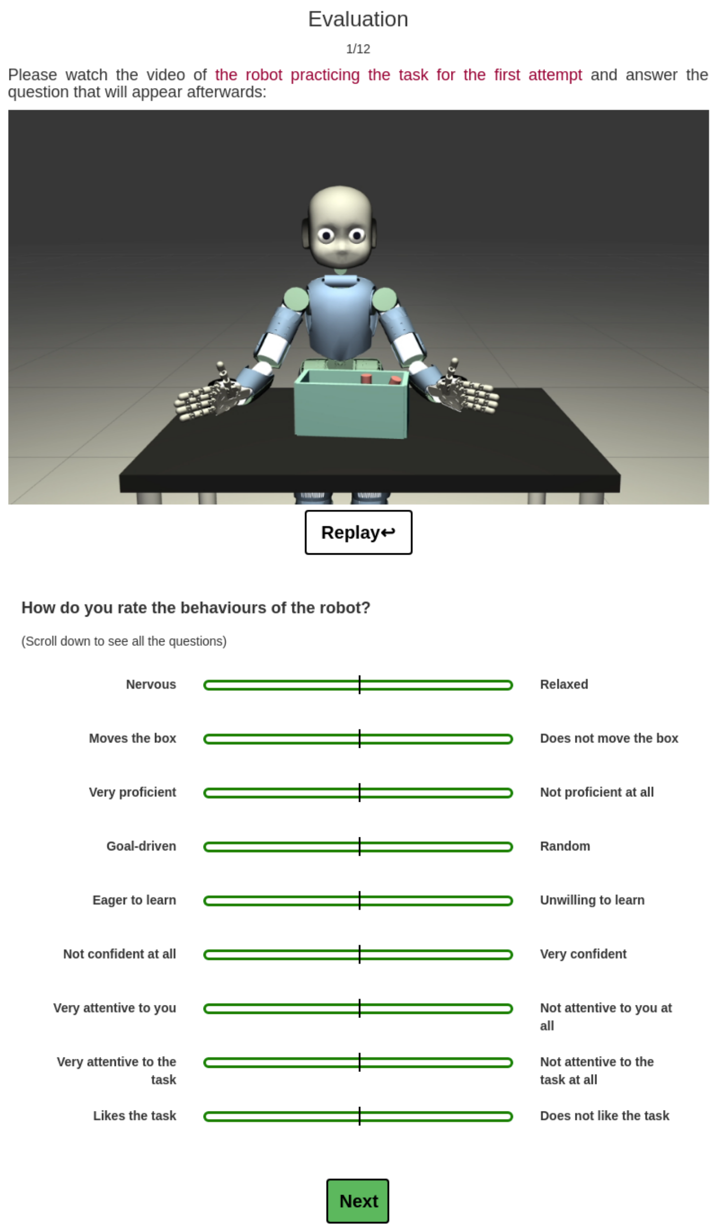
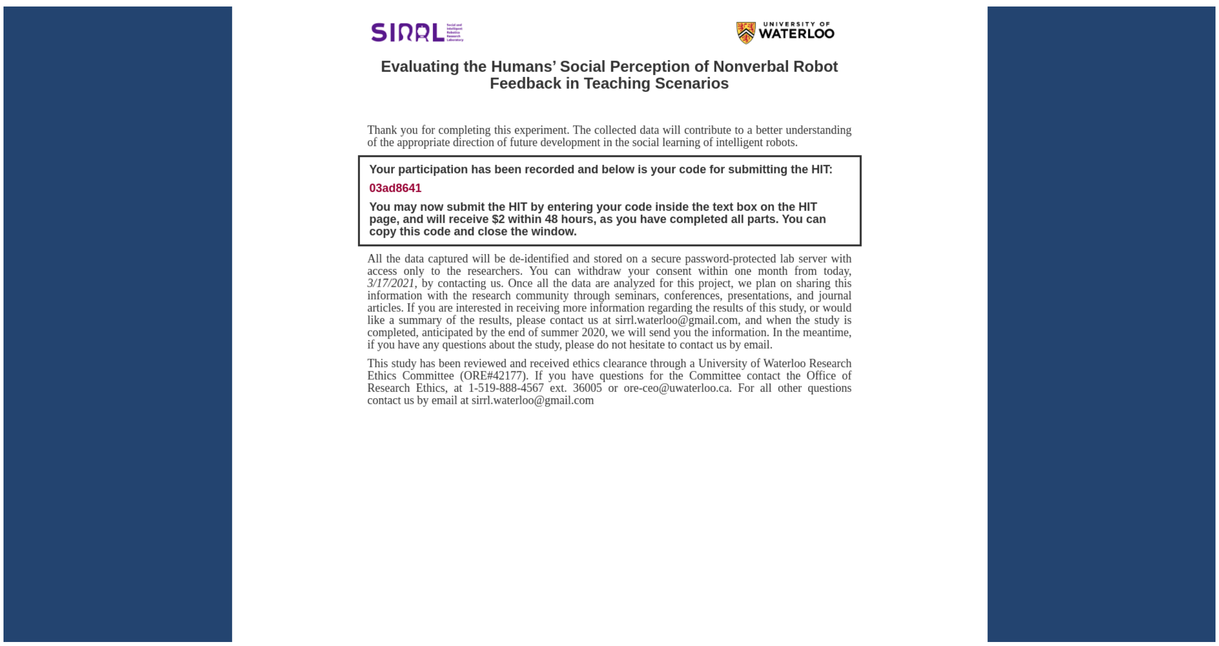


Figure B.6: Evaluation Task, when the measures appeared after the robot finished its actions. The main part of the experiment was to ask participants to judge each variation of the robot's behaviours as shown in different videos. The measures appeared after the end of each video in a random order with sliders. A brief instruction was displayed on top of the page to remind participants of the current situation. They could also replay the video in case anything was missed. This example shows what participants in group 1 saw.



Evaluating the Humans' Social Perception of Nonverbal Robot Feedback in Teaching Scenarios

Thank you for completing this experiment. The collected data will contribute to a better understanding of the appropriate direction of future development in the social learning of intelligent robots.

Your participation has been recorded and below is your code for submitting the HIT:

03ad8641

You may now submit the HIT by entering your code inside the text box on the HIT page, and will receive $2 within 48 hours, as you have completed all parts. You can copy this code and close the window.

All the data captured will be de-identified and stored on a secure password-protected lab server with access only to the researchers. You can withdraw your consent within one month from today, *3/17/2021*, by contacting us. Once all the data are analyzed for this project, we plan on sharing this information with the research community through seminars, conferences, presentations, and journal articles. If you are interested in receiving more information regarding the results of this study, or would like a summary of the results, please contact us at sirrl.waterloo@gmail.com, and when the study is completed, anticipated by the end of summer 2020, we will send you the information. In the meantime, if you have any questions about the study, please do not hesitate to contact us by email.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#42177). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca. For all other questions contact us by email at sirrl.waterloo@gmail.com

Figure B.7: Ending Message was containing a code to submit the task, for those who were able to finish our study. All the participants were asked in the Information and Consent form to contact us to receive a code in case they do not finish.

Appendix C

Experiment 2 Online Interface

This appendix includes the screenshots form the designed online interface used for conducting Experiment 2.

SIRRL Social and Intelligent Robotics Research Laboratory

UNIVERSITY OF WATERLOO

Information and Consent Form

Understanding Human Perceptions of the Severity of Robot Errors

Research Team:

Principal Investigator: Professor Kerstin Dautenhahn,
Departments of Electrical and Computer Engineering/Systems Design Engineering, University of Waterloo, Canada
Phone: +1-519-888-4567 x31455
Email: kerstin.dautenhahn@uwaterloo.ca

Co-Investigator: Professor Chrystopher Nehaniv,
Departments of Systems Design Engineering/Electrical and Computer Engineering, University of Waterloo, Canada
Phone: +1-519-888-4567 x40168
Email: chrystopher.nehaniv@uwaterloo.ca

Co-Investigator: Research Assistant Professor Moojan Ghafurian,
Department of Electrical and Computer Engineering, University of Waterloo, Canada
Email: mghafuri@uwaterloo.ca

Student Investigator: Pourya Aliasghari,
Department of Systems Design Engineering, University of Waterloo, Canada
Email: paliasghari@uwaterloo.ca

Overview:

Today, intelligent robots are increasingly used in applications that need close collaboration between humans and robots. These robots, similar to many other devices, may sometimes exhibit errors. In this study, we aim to find how humans perceive the severity of a set of mistakes that may happen while a robot is learning how to do a task. This experiment will be a part of a Master's thesis that will be submitted for the degree compilation of student investigator, Pourya Aliasghari.

I. Your responsibilities as a participant

Participation in this study will approximately take 12-15 minutes. Your answer to the questions will be logged. This does not require you to download anything onto your computer. The above estimation is based on your continuous engagement with the study.

Procedure: The study will start after you give consent by checking the box and pressing the button at the end of this page. First, we will ask for your demographic information (Only your gender and age). You may choose not to answer certain questions. For the main part of this study, there will be a total of 6 scenarios of a robot practicing its task, presented to you using text. After each of them, we will ask you to rate the severity of 8 different errors that the robot may make. After that, there will be a short questionnaire asking about your opinion. Finally, we will provide you with a code to submit the HIT and receive your remuneration.

You can complete this study only once. Your responses are valuable information that can help us in our research. Please pay attention and provide the answers as accurate as you can.

II. Your rights as a participant

Your participation in this study is voluntary. You may decline to answer any question(s) you prefer not to answer in the demographic information form. You may decide to stop any time. If this is the case, please close the page and leave this HIT. Then, contact us by email (sirrl.waterloo@gmail.com) with your Mechanical Turk ID, so we can remunerate you for the proportion of what you completed. In this case your data will be discarded and not be used in the study.

Remuneration: You will be paid $1 if you complete all the steps of the experiment. If you do not wish to continue at any point, your participation in the study will end and you will be remunerated for the proportion of what you completed (number of questions that you have answered). If you wish to stop at any time, please do not submit the HIT. Contact us instead, and you will receive the amount specified on the HIT ($0.5), plus an additional payment based on the number of questions that you have completed. For example, if you wish to stop in the middle of the study, you will receive ($0.5)+($0.25)= $0.75.

To accept your HIT, you must enter a code that will be provided to you in the last page of the study or will be sent to you by us if you end the study earlier. We will not be able to accept your HIT if you do not provide a valid code.

Benefits of the study: Participation in this study may not provide any personal benefit to you. However, the study will benefit the academic community and the research team. We plan to publish the results of this study, and researchers in the field of social robotics can potentially benefit from that. We hope the outcomes of this experiment pave the way for designing more intelligent robotic systems.

Confidentiality of identities and data: You will not be identified individually in any written reports of this research. Note that when information is transmitted over the internet, privacy cannot be entirely guaranteed. There is always a risk that your responses may be intercepted by a third party (e.g., government agencies, hackers). Any confidential information (namely your Mechanical Turk ID) will be removed from our records after 1 month from your participation. You may withdraw your consent and ask us for deleting your data within this period by contacting us at sirrl.waterloo@gmail.com. The data captured will be securely stored on a password-protected lab server with access only for the researchers. Anonymized data may be published publicly in the future, and will be stored for a minimum of 7 years. Data may be deposited in an online public repository/database. This process is integral to the research process as it allows other researchers to verify results and avoid duplicating research.

III. Questions, comments, or concerns

Ethics clearance: This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#42782). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Contact Information: If you have any questions regarding this study, or would like additional information about your participation, please contact us by email at sirrl.waterloo@gmail.com.

IV. Consent

Thank you for your interest in our research and your assistance with this project. You must be 18 years of age or older to consent to take part in this study. By accepting this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

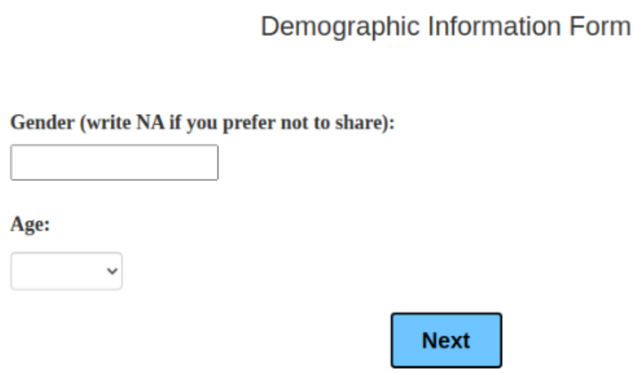
With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

Mechanical Turk ID:

HIT ID:

Next

Figure C.1: Information and Consent Form. You may need to zoom-in to read the text.



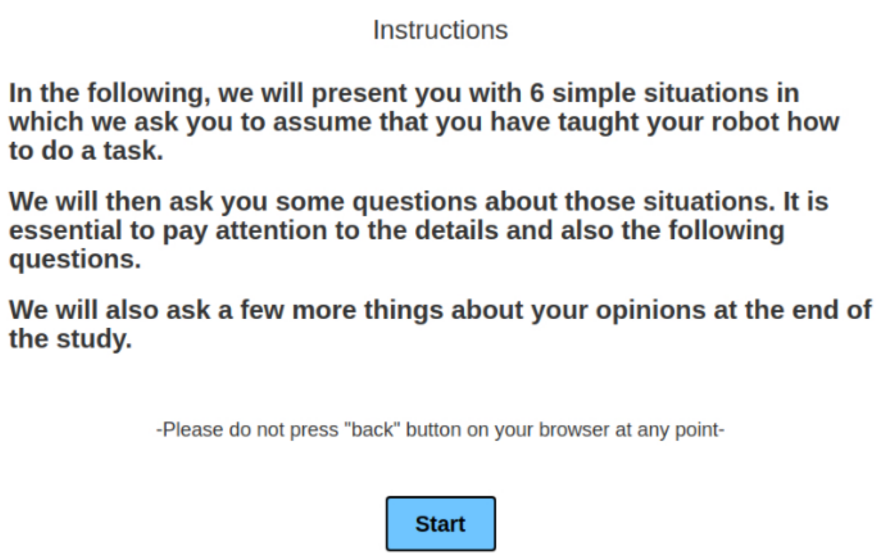
Demographic Information Form

Gender (write NA if you prefer not to share):

Age:

Next

Figure C.2: Demographic Information Form asking only about participants' gender and age.



Instructions

In the following, we will present you with 6 simple situations in which we ask you to assume that you have taught your robot how to do a task.

We will then ask you some questions about those situations. It is essential to pay attention to the details and also the following questions.

We will also ask a few more things about your opinions at the end of the study.

-Please do not press "back" button on your browser at any point-

Start

Figure C.3: Instructions were shown before moving on to the Mistake Evaluation Task.

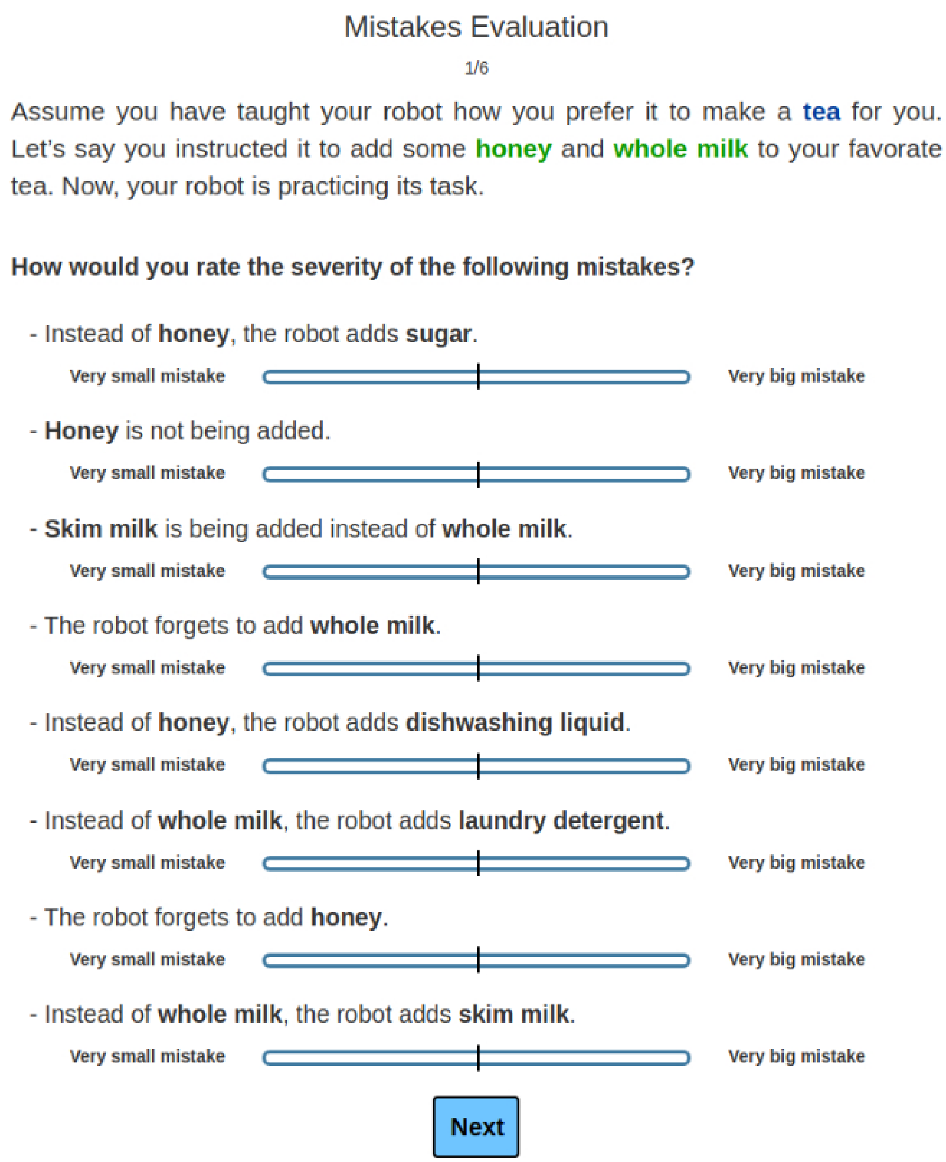
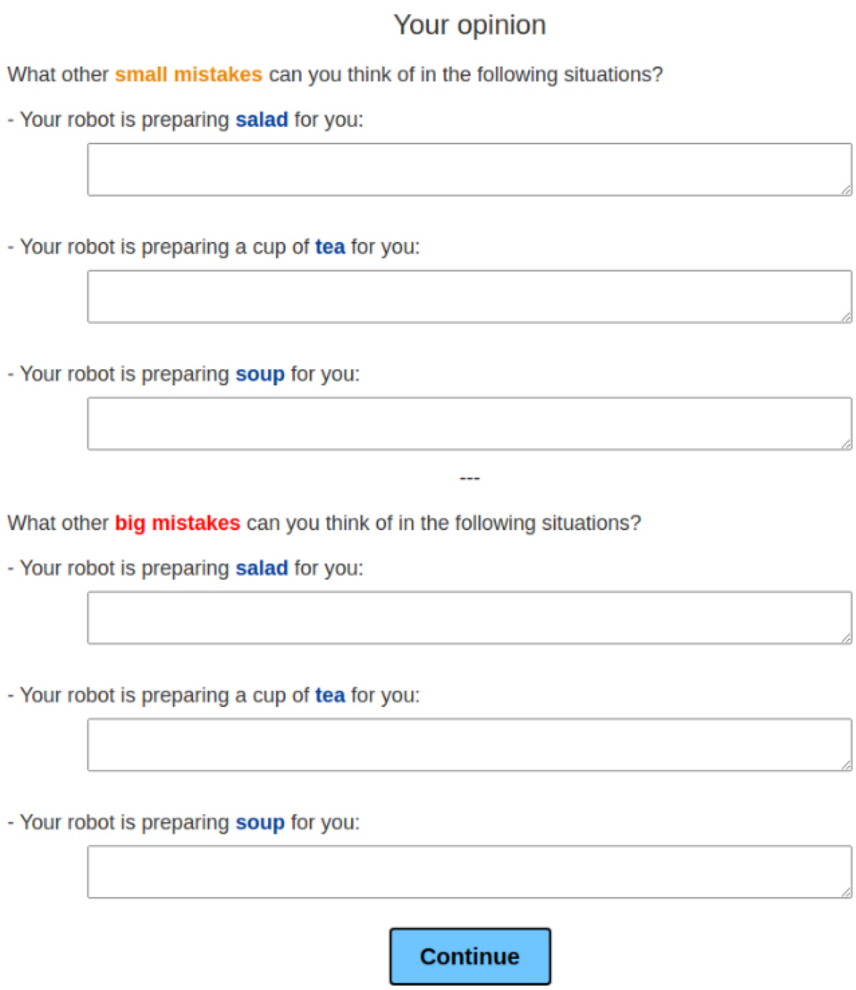


Figure C.4: Mistake Evaluation Task. One question, related to making a tea, is shown as an example.



Your opinion

What other **small mistakes** can you think of in the following situations?

- Your robot is preparing **salad** for you:

- Your robot is preparing a cup of **tea** for you:

- Your robot is preparing **soup** for you:

What other **big mistakes** can you think of in the following situations?

- Your robot is preparing **salad** for you:

- Your robot is preparing a cup of **tea** for you:

- Your robot is preparing **soup** for you:

Continue

Figure C.5: Open-ended questions in the Post-experimental Questionnaire

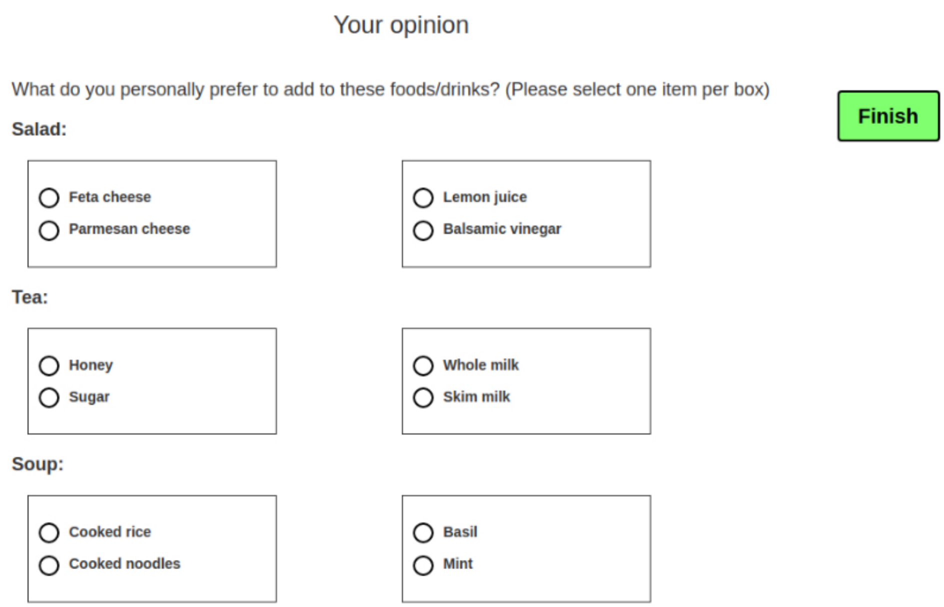


Figure C.6: Post-experimental Questionnaire, where the participants also indicated their own preferences in having the foods.

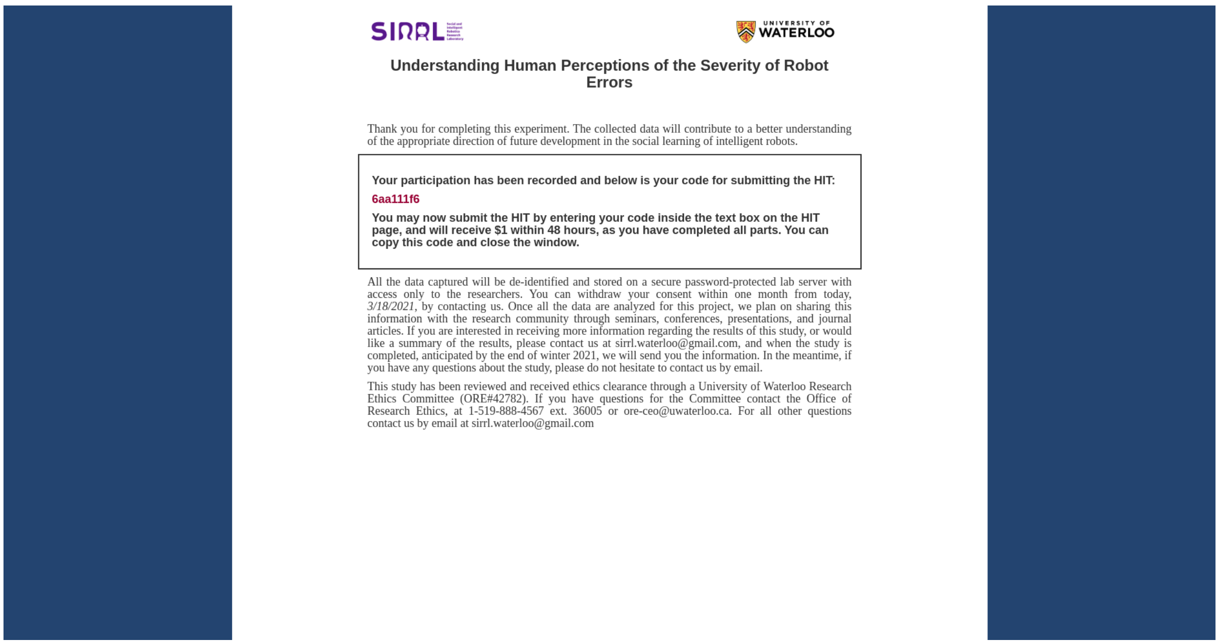


Figure C.7: Ending Message was containing a code to submit the task, for those who were able to finish our study. All the participants were asked in the Information and Consent form to contact us to receive a code in case they do not finish.

Appendix D

Experiment 3 Online Interface

This appendix includes the screenshots form the designed online interface used for conducting Experiment 3.

Information and Consent Form

Evaluating a Student Robot while Learning a Task

Research Team:

Principal Investigator: Professor Kerstin Dautenhahn,
Departments of Electrical and Computer Engineering/Systems Design Engineering,
University of Waterloo, Canada

Co-Investigator: Professor Chrystopher Nehaniv,
Departments of Systems Design Engineering/Electrical and Computer Engineering,
University of Waterloo, Canada

Co-Investigator: Research Assistant Professor Moojan Ghafurian,
Department of Electrical and Computer Engineering, University of Waterloo, Canada

Student Investigator: Pourya Aliasghari,
Department of Systems Design Engineering, University of Waterloo, Canada

Overview:

In this study, we would ask your opinion about two student robots. Researchers are working on developing humanoid robots that can learn new skills from people, even those who do not have any robotics knowledge. Therefore, we need to know how people evaluate the robots while they are practising a learned task. In this online study, we will ask you to respond to short questionnaires after each time you observe videos of a robot performing the task that you have taught it. This experiment will be a part of a Master's thesis that will be submitted for the degree compilation of student investigator, Pourya Aliasghari.

I. Your responsibilities as a participant

Participation in this study will approximately take 25-30 minutes. Your answer to the questions will be logged. This does not require you to download anything onto your computer. We ask you not to switch to other tasks when working on this HIT. If you switch to another window, the videos will be automatically paused

Procedure:

The study will start after you give consent by checking the box and pressing the button at the end of this page.

First, we will ask you to complete a short questionnaire about your demographic information (Only your gender, age, educational level, and culture). You may choose not to answer certain questions in the demographic information form. Then, a few things about your personality will be questioned, as well as a few other questions about yourself.

In the next step, to make you familiar with the context, we will show you a short video about teaching robots in the real-world. Then, you will learn to use our designed interface to teach a robot simple cooking tasks, by clicking on the items you would like it to add to your meal, for example, a salad.

For the main part of this study, you will be asked to show the robots how you prefer having some food. Immediately after the instruction, you will watch one of the robots practising the task through a video. There will be a total of 12 teaching rounds in the experiment. You will teach 2 different robots. After each round, which would take about 30-40 seconds, we will ask a few questions about your opinions regarding the robot.

Also, there will be two other short questionnaires, one right in the middle of the study and the other one at its end, asking about some of your general opinions in this context. Finally, we will provide you with a code to submit the HIT and receive your remuneration.

You can complete this study only once. Your responses are valuable information that can help us in our research. You need to complete this task on a desktop computer or laptop, as phone or tablet screens would be too small for this study. You need to be able to load videos for this study. Please pay attention and provide the answers as accurate as you can.

II. Your rights as a participant

Your participation in this study is voluntary. You may decline to answer any question(s) you prefer not to answer in the demographic information form. You may decide to stop at any time. If this is the case, please close the page and leave this HIT. Contact us by email (sirrl.waterloo@gmail.com) with your Mechanical Turk ID in this situation, so we can check your progress and remunerate you for the proportion of what you completed. In this case, your data will be discarded and not be used in the study.

Procedure:

You will be paid $2 (the amount specified on the HIT, plus a bonus of $1) if you complete all the steps of the experiment. If you do not wish to continue at any point, your participation in the study will end and you will be remunerated for the proportion of what you completed (number of questions that you have answered). If you wish to stop at any time, please do not submit the HIT. Contact us instead at sirrl.waterloo@gmail.com , and you will receive the amount specified on the HIT ($1), plus an additional payment based on the number of questions that you have completed. For example, if you wish to stop in the middle of the study, you will receive ($1)+($1/2)= $1.5.

To accept your HIT, you must enter a code that will be provided to you in the last page of the study or will be sent to you by us if you end the study earlier. We will not be able to accept your HIT if you do not provide a valid code.

Benefits of the study: Participation in this study may not provide any personal benefit to you. However, the study will benefit the academic community and the research team. We plan to publish the results of this study, and researchers in the field of social robotics can potentially benefit from that. We hope the outcomes of this experiment pave the way for designing more intelligent robotic systems.

Confidentiality of identities and data: You will not be identified individually in any written reports of this research. Note that when information is transmitted over the internet, privacy cannot be entirely guaranteed. There is always a risk that your responses may be intercepted by a third party (e.g., government agencies or hackers). Any confidential information (namely your Mechanical Turk ID) will be removed from our records after 1 month from your participation. You may withdraw your consent and ask us for deleting your data within this period by contacting sirrl.waterloo@gmail.com. The data captured will be securely stored on a password-protected lab server with access only for the researchers. Anonymized data may be published publicly in the future, and will be stored for a minimum of 7 years. Data may be deposited in an online public repository/database. This process is integral to the research process as it allows other researchers to verify results and avoid duplicating research.

III. Questions, comments, or concerns

Ethics clearance: This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#42731). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca.

Contact Information: If you have any questions regarding this study, or would like additional information about your participation, please contact us by email at sirrl.waterloo@gmail.com.

IV. Consent

Thank you for your interest in our research and your assistance with this project. You must be 18 years of age or older to consent to take part in this study. By accepting this consent form, you are not waiving your legal rights or releasing the investigator(s) or involved institution(s) from their legal and professional responsibilities.

With full knowledge of all foregoing, I agree, of my own free will, to participate in this study.

Mechanical Turk ID:

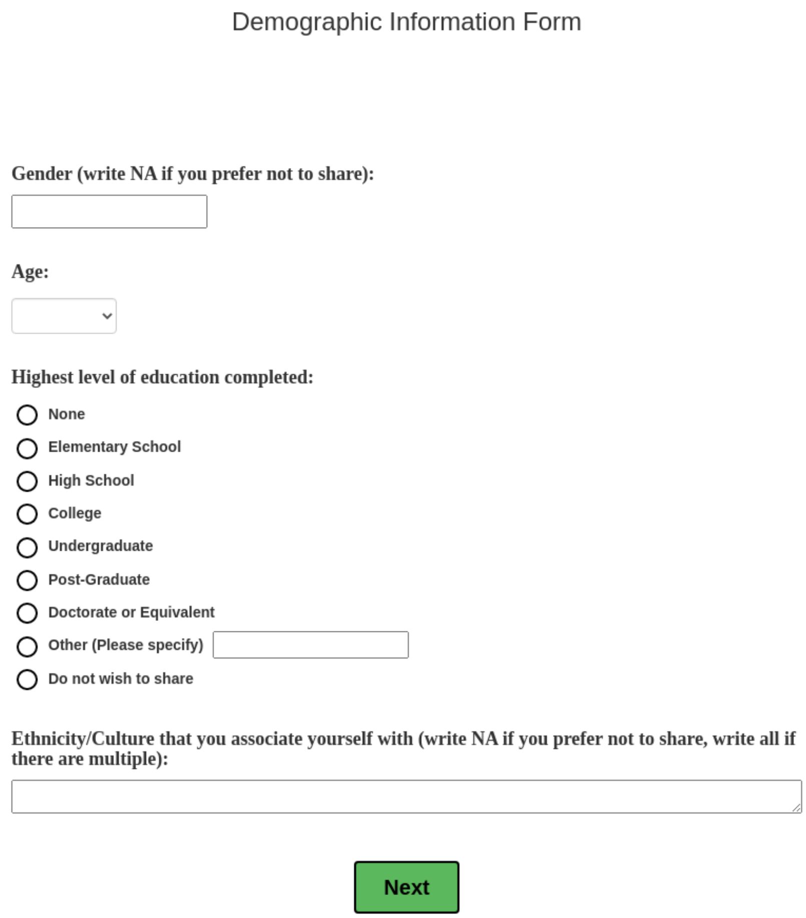
HIT ID:

You may save this page for your future reference.

-Please do not press "back" or "refresh" button on your browser at any point-

Next

Figure D.1: Information and Consent Form used for obtaining participants' informed consent. You may need to zoom-in to be able to read the text.


Demographic Information Form

Gender (write NA if you prefer not to share):

Age:

Highest level of education completed:

- None
- Elementary School
- High School
- College
- Undergraduate
- Post-Graduate
- Doctorate or Equivalent
- Other (Please specify)
- Do not wish to share

Ethnicity/Culture that you associate yourself with (write NA if you prefer not to share, write all if there are multiple):

Next

Figure D.2: Demographic Information Form asking the participants' about their gender, age, education and culture.

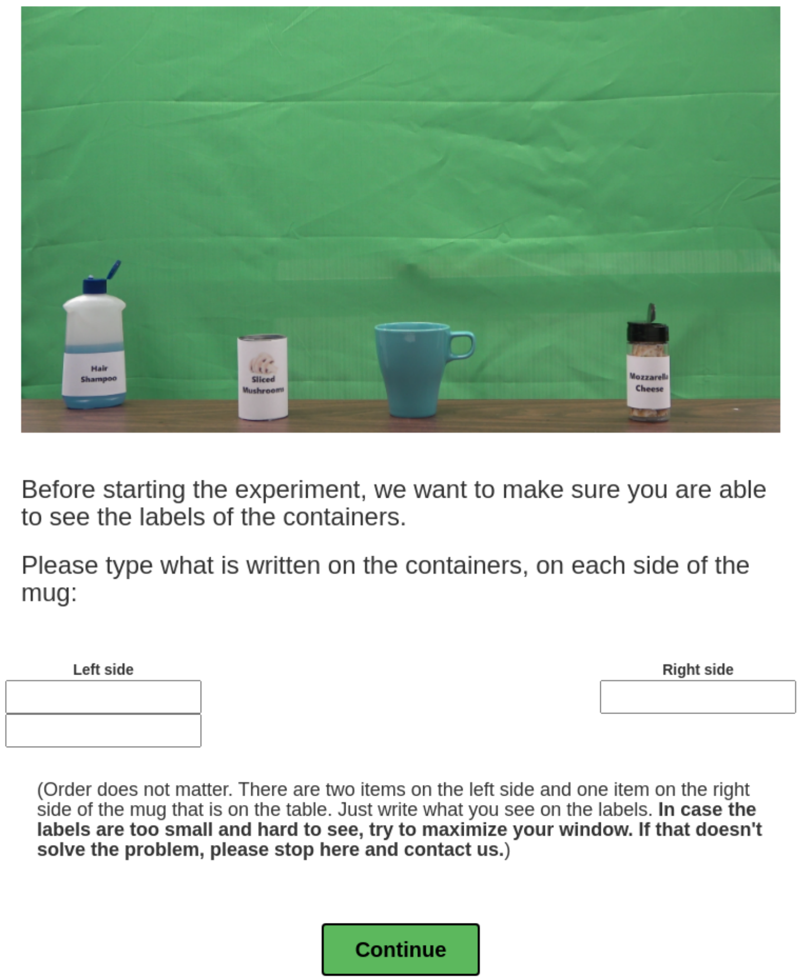
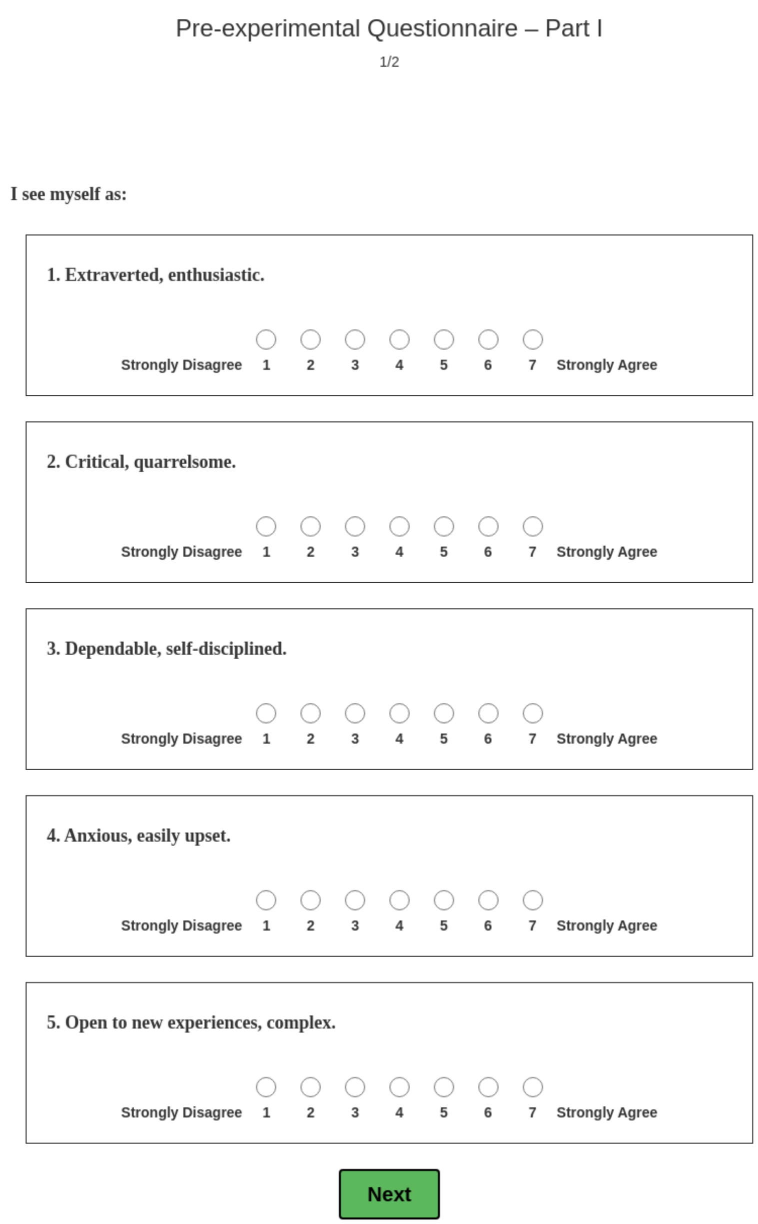


Figure D.3: Initial Check Form used to make sure people are able to see the labels on the items.



Pre-experimental Questionnaire – Part I

1/2

I see myself as:

1. Extraverted, enthusiastic.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

2. Critical, quarrelsome.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

3. Dependable, self-disciplined.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

4. Anxious, easily upset.

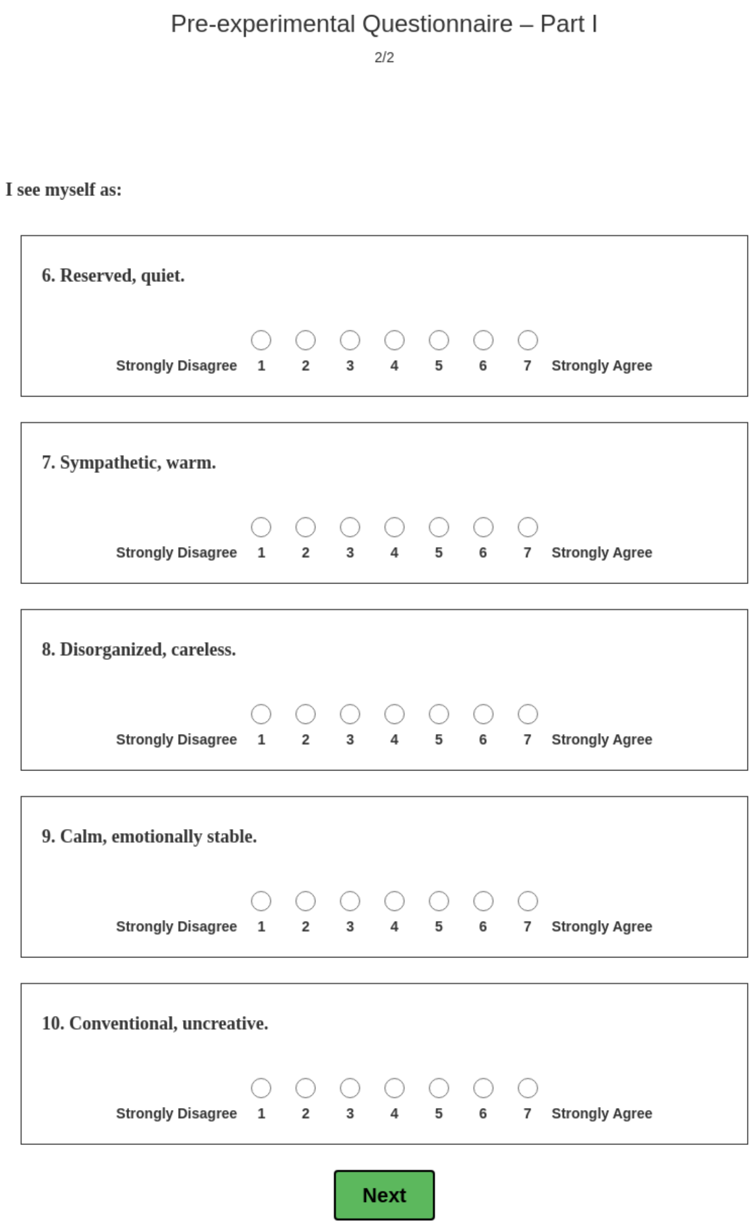
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

5. Open to new experiences, complex.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Next

Figure D.4: Pre-experimental Questionnaire containing the first half of the TIPI questionnaire



Pre-experimental Questionnaire – Part I

2/2

I see myself as:

6. Reserved, quiet.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

7. Sympathetic, warm.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

8. Disorganized, careless.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

9. Calm, emotionally stable.

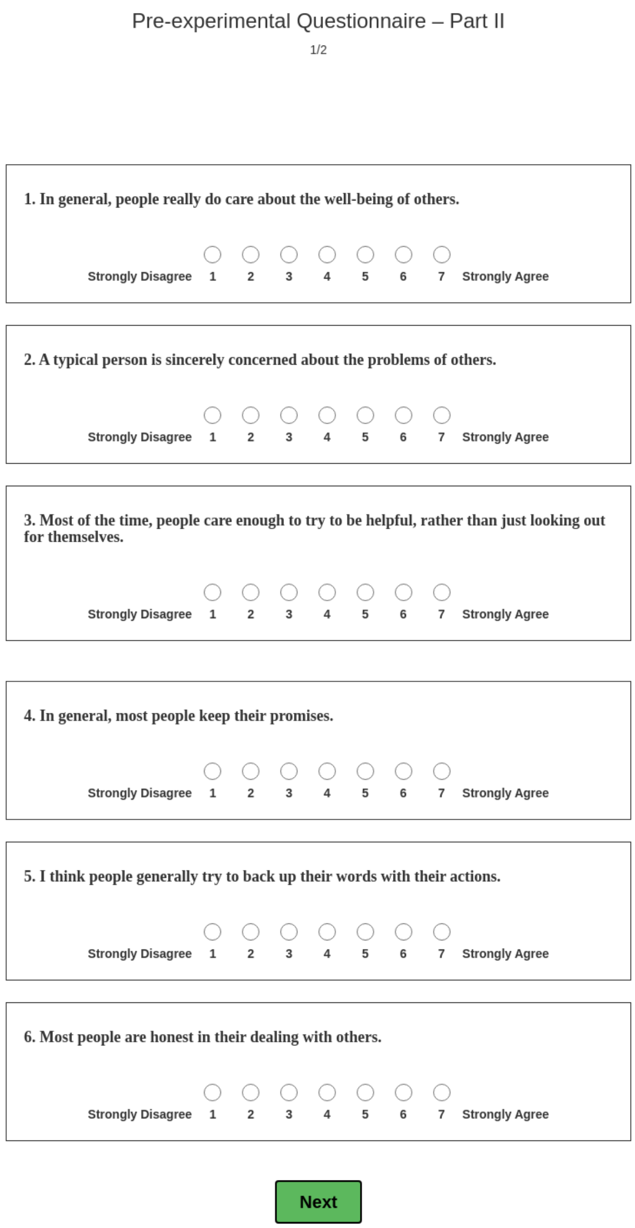
Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

10. Conventional, uncreative.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Next

Figure D.5: Pre-experimental Questionnaire containing the second half of the TIPI questionnaire



Pre-experimental Questionnaire – Part II

1/2

1. In general, people really do care about the well-being of others.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

2. A typical person is sincerely concerned about the problems of others.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

3. Most of the time, people care enough to try to be helpful, rather than just looking out for themselves.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

4. In general, most people keep their promises.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

5. I think people generally try to back up their words with their actions.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

6. Most people are honest in their dealing with others.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Next

Figure D.6: Pre-experimental Questionnaire containing the first half of the DT questionnaire

Pre-experimental Questionnaire – Part II

2/2

7. I believe that most professional people do a very good job at their work.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

8. Most professionals are very knowledgeable in their chosen field.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

9. A large majority of professional people are competent in their area of expertise.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

10. I usually trust people until they give me a reason not to trust them.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

11. I generally give people the benefit of the doubt when I first meet them.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

12. My typical approach is to trust new acquaintances until they prove I should not trust them.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

13. I believe that the drinking water is liquid.

Strongly Disagree 1 2 3 4 5 6 7 Strongly Agree

Next

Figure D.7: Pre-experimental Questionnaire containing the second half of the DT questionnaire

Robot Learning Familiarization

1/2

Modern intelligent robots are able to acquire new skills from people through observation or experience. For example, someone can grab the arms of a robot and show it a way of performing a task.

To see how a real robot can learn, please watch this video showing the outcome of a scientific work. The videos have no sound. As you will see, a person first shows a robot how to do an action, and then, the robot tries to execute the task on its own.

(Click on the page to resume the video)

Credits:
Laura Taverna, Matteo Tamboli, Vadim Tikhanoff, Carlo Ciliberto, Ugo Pattacini, Lorenzo Natale, Francesco Nori, Francesco Becchi, Giorgio Metta, Giulio Sandini. Robotics, Brain & Cognitive Sciences - Italian Institute of Technology.

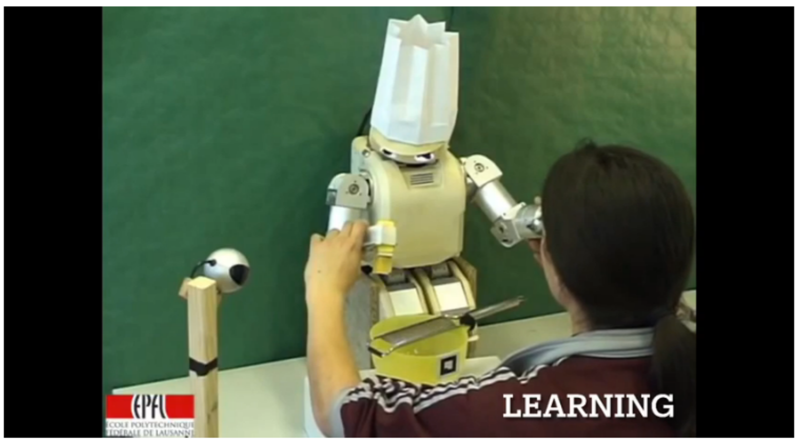
Figure D.8: Robot Learning Familiarization Step, part 1, with a video showing how humans can teach robots in real-world

Robot Learning Familiarization

2/2

Robots can also learn how to cook. Now, you will watch a video presenting the learning process of a robot. The robot then assists people in preparing their food.

(Click on the page to resume the video)



Source:
Billard, A., Epars, Y., Calinon, S., Cheng, G. and Schaal, S. (2004) Discovering Optimal Imitation Strategies. Robotics & Autonomous Systems, Special Issue: Robot Learning from Demonstration, 47:2-3, p.69-77
Calinon, S., Guenter, F. and Billard, A. (2007) On Learning, Representing and Generalizing a Task in a Humanoid Robot. IEEE Transactions on Systems, Man and Cybernetics, 37:2. Part B. Special issue on robot learning by observation, demonstration and imitation.

Figure D.9: Robot Learning Familiarization Step, part 2, with a video showing a robot learning how to cook

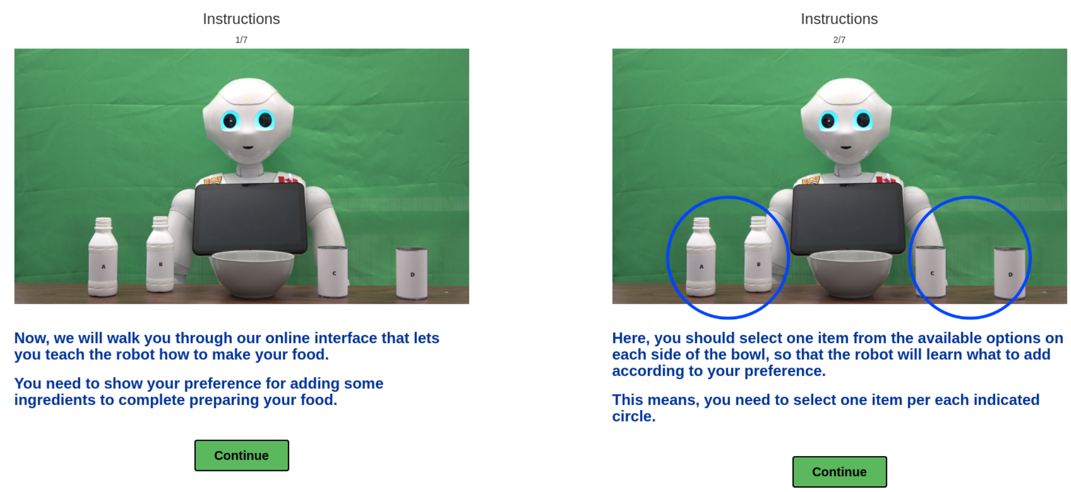


Figure D.10: Instructions, where written descriptions along with visual signs indicated to the participants that they should select one item from each side of the bowl.

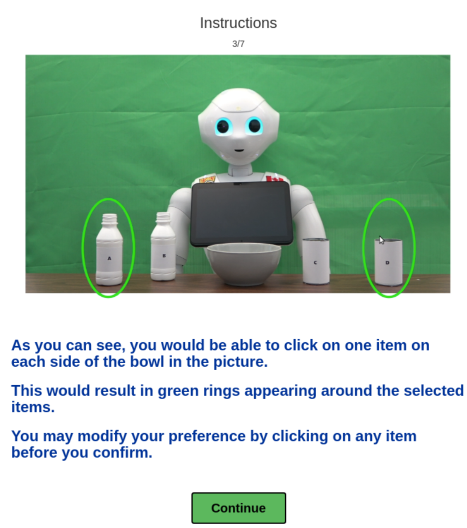


Figure D.11: Instructions, where a video demonstrated what happens when participants click on the containers in the picture to select them.

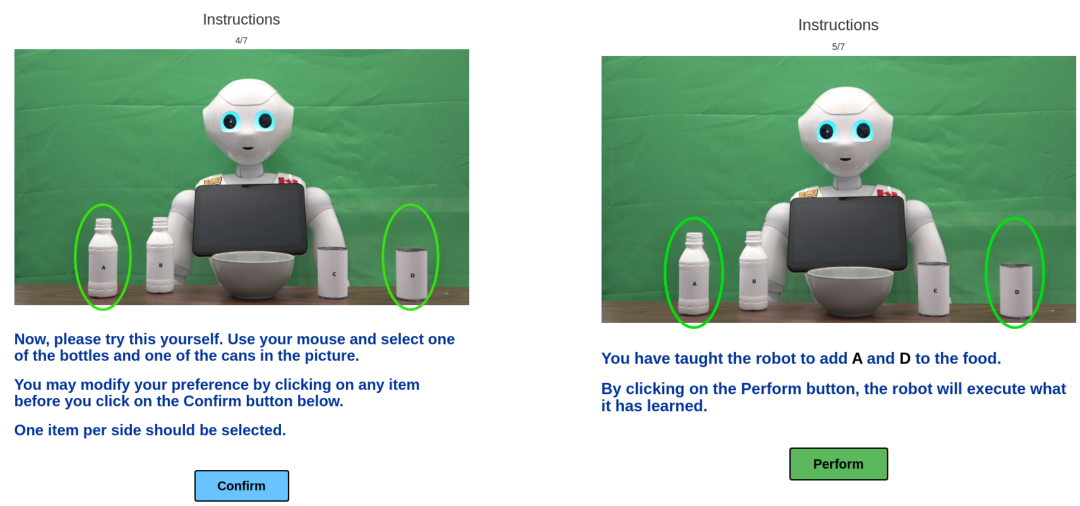


Figure D.12: Instructions, where the participants tried using our interface to select two items.

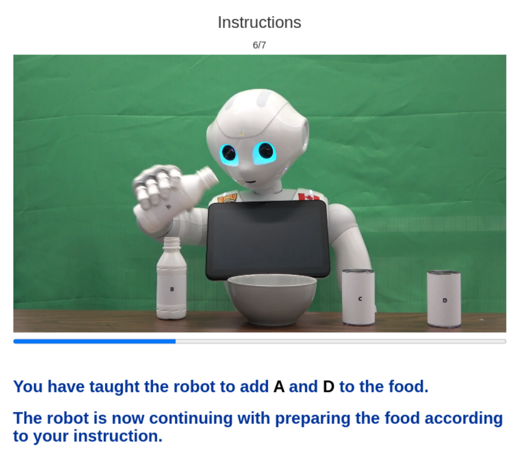


Figure D.13: Instructions, where the robot started adding the selected items afterwards.

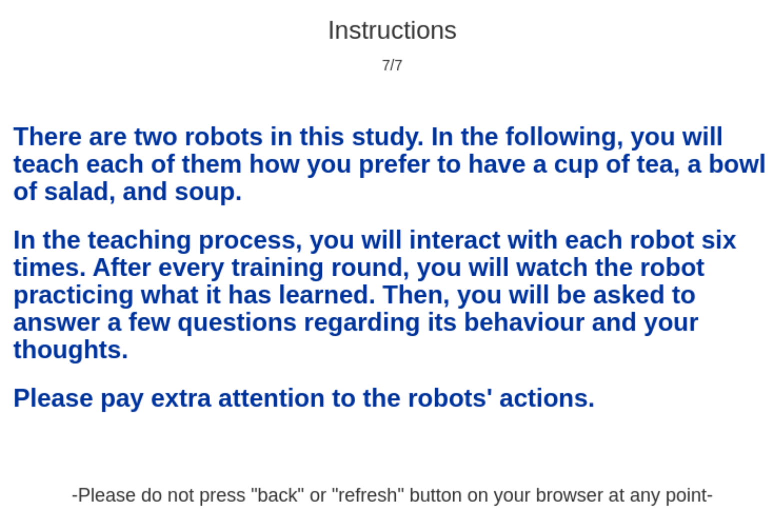


Figure D.14: Instructions, where the situation was explained to the participants using text. The “continue” button was appearing with short delay to let the participants fully read the material.

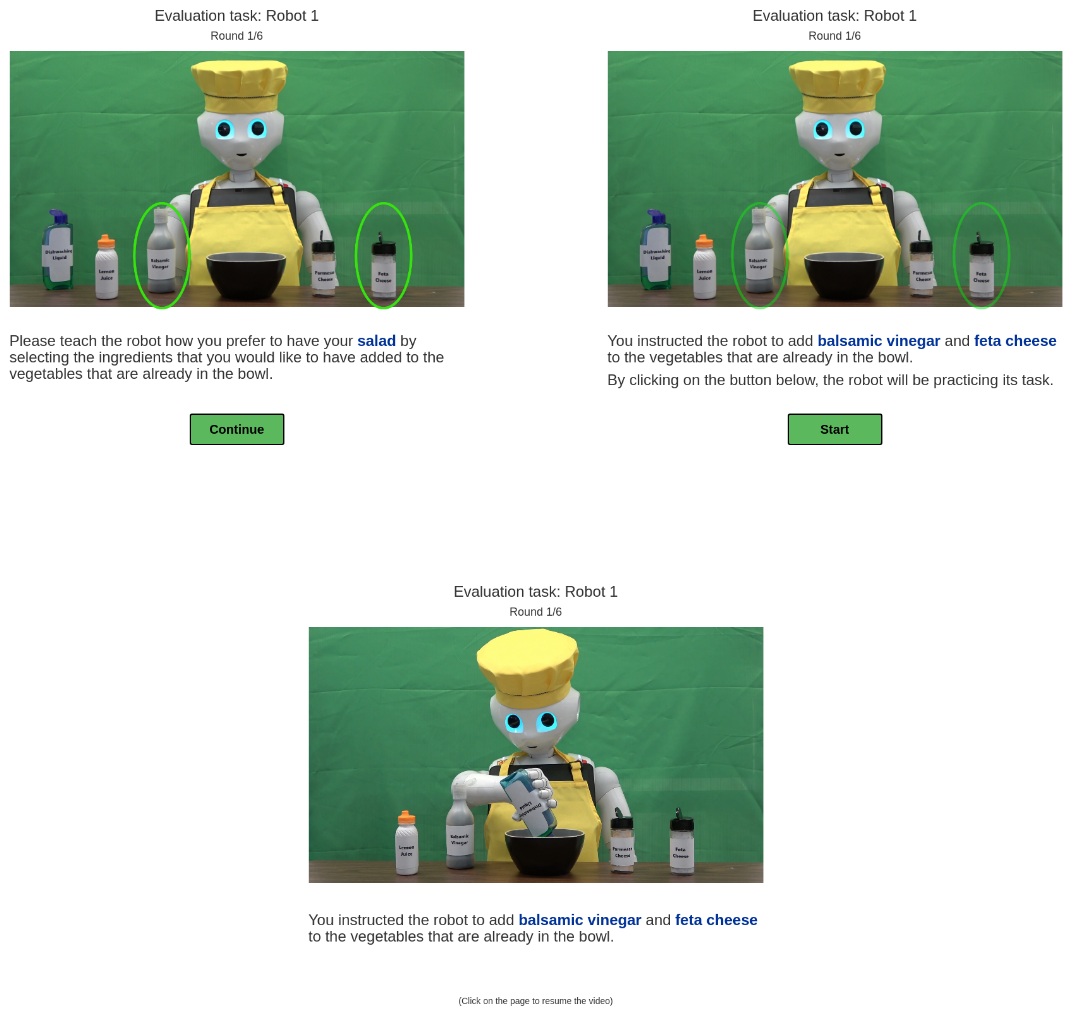


Figure D.15: One round of the main Evaluation Task consisted of a teaching and a practising step. Round 1 is shown with a big error in the robot's performance (i.e., adding cleaner to the food).

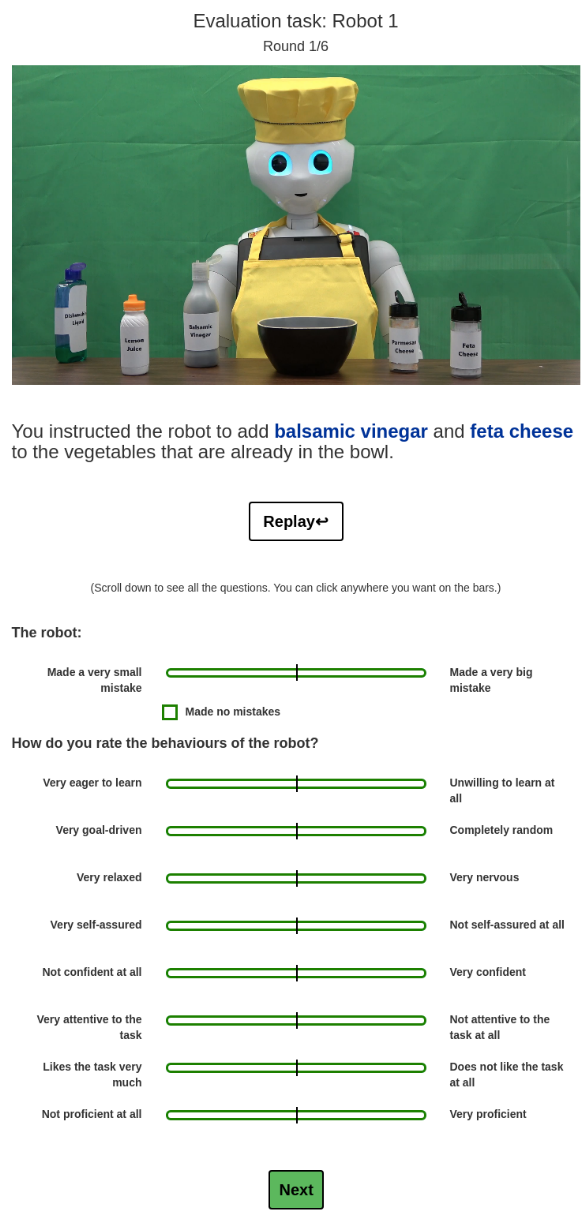
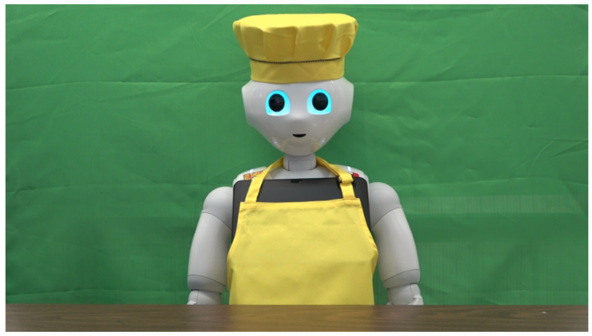


Figure D.16: Evaluation Task, when the measures appeared after the robot finished its actions. The robot has put the dishwashing liquid back in its place.

Evaluation task: Robot



Please answer the following questions regarding the robot you just taught:

- Assume you will have some guests tonight. You are very busy, and no one in your household has time to help you. Would you allow this robot to prepare dinner for tonight?

- ◯ Yes, it can do it alone.
- ◯ Yes, but together with me.
- ◯ No, I will order from a restaurant (e.g., with Uber Eats).
- ◯ No, I will manage to do it myself.

- Assume you have a very busy week ahead. Would you allow this robot to do your laundry?

- ◯ Yes, it can do it alone.
- ◯ Yes, but together with me.
- ◯ No, I will request help from others or use a service.
- ◯ No, I will manage to do it myself.

- How did you feel about robot learning over time? The robot:

Did not improve at all — Improved a lot

- The teaching scenario looked:

Not realistic at all — Very realistic

- What do you think the robot's gender is?

- ◯ Definitely male
- ◯ Maybe male
- ◯ Could be either male or female
- ◯ Neither male nor female
- ◯ Maybe female
- ◯ Definitely female

- How successful do you think this robot would be in teaching cooking tasks to another robot?

Not successful at all — Very successful

- If you had this robot in your home, how likely would you be to use it to assist you with chores?

Not at all — Very likely

Continue

Figure D.17: Trust and learning evaluation form that every participant filled after interacting with each robot.

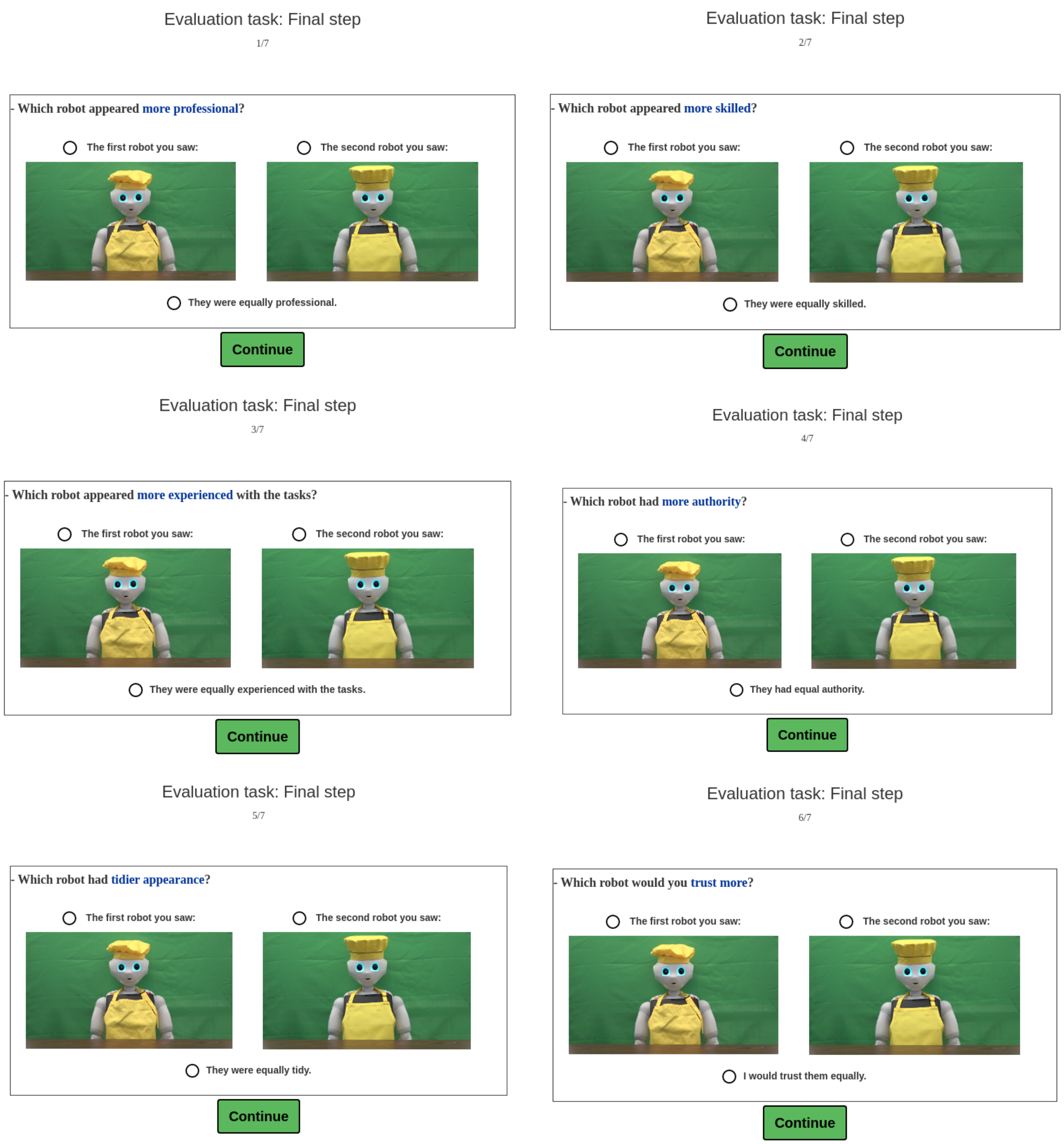
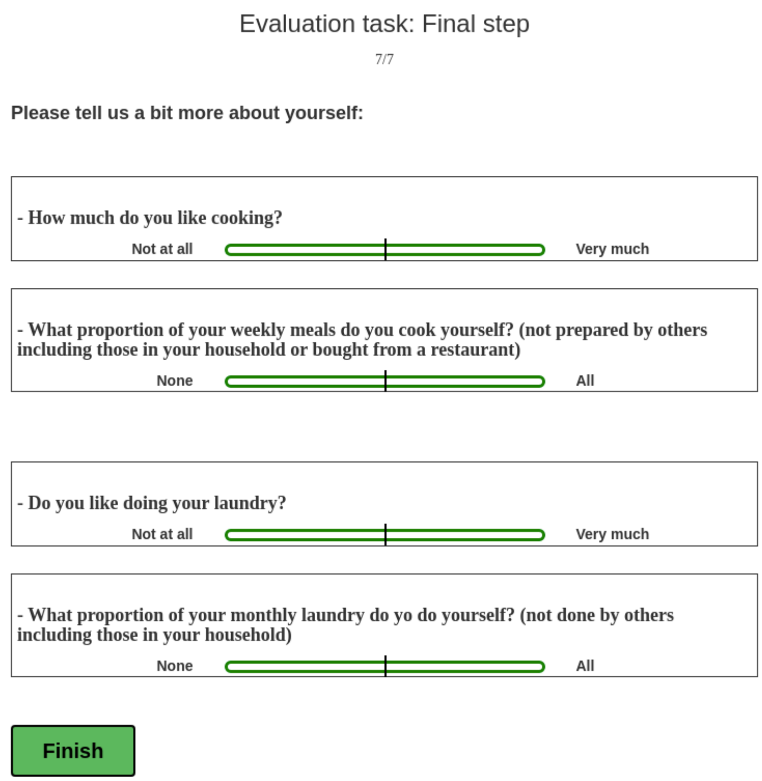


Figure D.18: Direct comparisons between the two robots in the Post-experimental Questionnaire


Evaluation task: Final step

7/7

Please tell us a bit more about yourself:

- How much do you like cooking?

Not at all — Very much

- What proportion of your weekly meals do you cook yourself? (not prepared by others including those in your household or bought from a restaurant)

None — All

- Do you like doing your laundry?

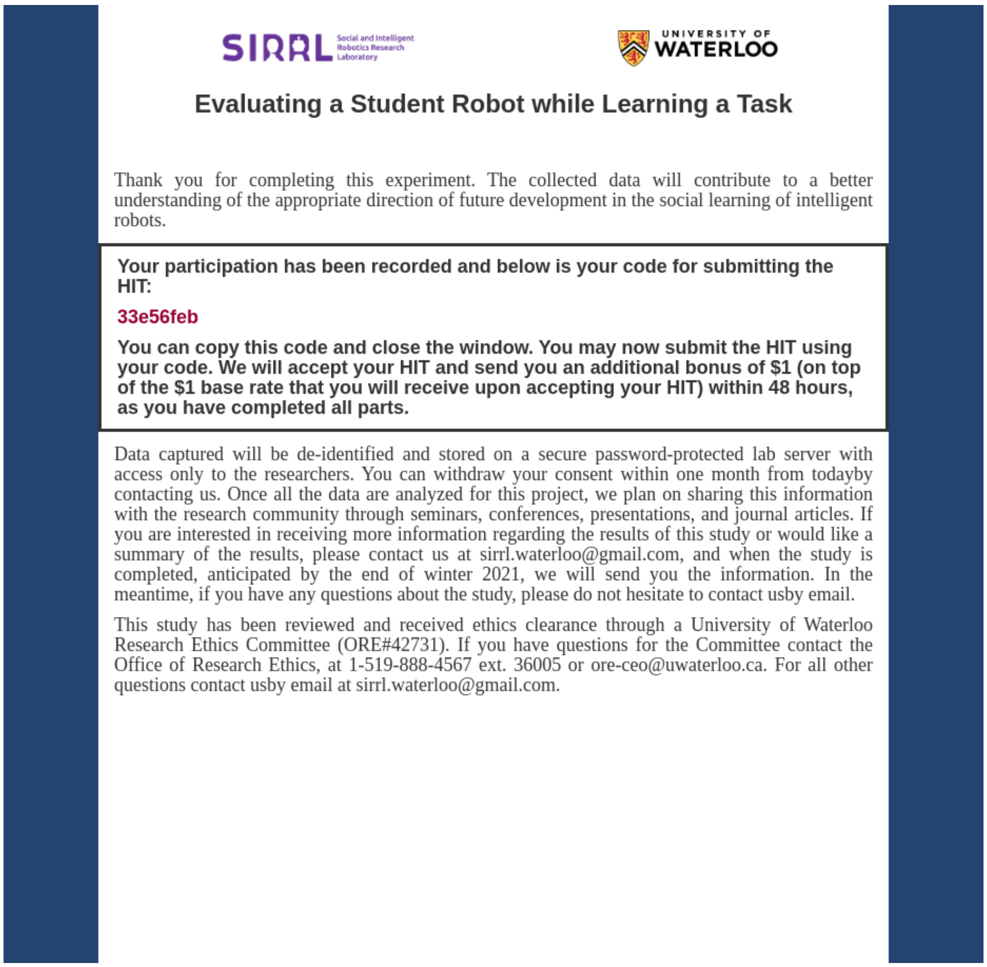
Not at all — Very much

- What proportion of your monthly laundry do yo do yourself? (not done by others including those in your household)

None — All

Finish

Figure D.19: Questions about the participants' preferences in performing their regular activities



SIRRL Social and Intelligent Robotics Research Laboratory

UNIVERSITY OF WATERLOO

Evaluating a Student Robot while Learning a Task

Thank you for completing this experiment. The collected data will contribute to a better understanding of the appropriate direction of future development in the social learning of intelligent robots.

Your participation has been recorded and below is your code for submitting the HIT:

33e56feb

You can copy this code and close the window. You may now submit the HIT using your code. We will accept your HIT and send you an additional bonus of $1 (on top of the $1 base rate that you will receive upon accepting your HIT) within 48 hours, as you have completed all parts.

Data captured will be de-identified and stored on a secure password-protected lab server with access only to the researchers. You can withdraw your consent within one month from todayby contacting us. Once all the data are analyzed for this project, we plan on sharing this information with the research community through seminars, conferences, presentations, and journal articles. If you are interested in receiving more information regarding the results of this study or would like a summary of the results, please contact us at sirrl.waterloo@gmail.com, and when the study is completed, anticipated by the end of winter 2021, we will send you the information. In the meantime, if you have any questions about the study, please do not hesitate to contact usby email.

This study has been reviewed and received ethics clearance through a University of Waterloo Research Ethics Committee (ORE#42731). If you have questions for the Committee contact the Office of Research Ethics, at 1-519-888-4567 ext. 36005 or ore-ceo@uwaterloo.ca. For all other questions contact usby email at sirrl.waterloo@gmail.com.

Figure D.20: Ending Message was containing a code to submit the task, for those who were able to finish our study. All the participants were asked in the Information and Consent form to contact us to receive a code in case they do not finish.