

Learning from Demonstrations: Human Perspective and Perception

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Media and Human-Centered Computing

eingereicht von

Patrick Gietl, BSc

Matrikelnummer 01527148

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Univ.Prof. Dipl.-Inf. Dr.sc.techn. Florian Michahelles

Mitwirkung: Univ.Ass. Khaled Kassem, MSc

Wien, 26. Jänner 2025

Patrick Gietl

Florian Michahelles



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Learning from Demonstrations: Human Perspective and Perception

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Media and Human-Centered Computing

by

Patrick Gietl, BSc

Registration Number 01527148

to the Faculty of Informatics

at the TU Wien

Advisor: Univ.Prof. Dipl.-Inf. Dr.sc.techn. Florian Michahelles

Assistance: Univ.Ass. Khaled Kassem, MSc

Vienna, January 26, 2025

Patrick Gietl

Florian Michahelles



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Erklärung zur Verfassung der Arbeit

Patrick Gietl, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Ich erkläre weiters, dass ich mich generativer KI-Tools lediglich als Hilfsmittel bedient habe und in der vorliegenden Arbeit mein gestalterischer Einfluss überwiegt. Im Anhang „Übersicht verwendeter Hilfsmittel“ habe ich alle generativen KI-Tools gelistet, die verwendet wurden, und angegeben, wo und wie sie verwendet wurden. Für Textpassagen, die ohne substantielle Änderungen übernommen wurden, haben ich jeweils die von mir formulierten Eingaben (Prompts) und die verwendete IT- Anwendung mit ihrem Produktnamen und Versionsnummer/Datum angegeben.

Wien, 26. Jänner 2025

Patrick Gietl



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acknowledgements

First, I would like to express my wholehearted thanks to my fiancée, Tamara, whose constant love, support, and belief helped me stay motivated and engaged, especially during the difficult times. Whether it was her unwavering faith in my abilities, her constant patience with me talking about the thesis non-stop, or her reminders to take care of myself after long working hours, she has been my companion through every twist and turn of this journey. I am endlessly grateful to have you by my side, making every goal feel more achievable and every victory more meaningful.

My sincere thanks go to Khaled Kassem, whose expertise and willingness to share his knowledge in countless messages, e-mails, and meetings, as well as his thoughtful answers to my questions, and his guidance toward important and interesting literature, helped me greatly elevate the quality of my work.

My heartfelt thanks go to my friends and family, who shared their interest and kindly listened to my (sometimes overly detailed) explanations.

Finally, I am profoundly grateful to everyone who participated in my experiment, for generously sharing your time and thoughts.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Kurzfassung

Dank des technologischen Fortschritts wird die Integration sozialer Roboter (SRs) in unseren Alltag zunehmend realistischer. In den letzten Jahren haben sich diese bereits im Gesundheitsbereich als praktikabel erwiesen. Es gibt dennoch viele Herausforderungen die weitere Forschungen erfordern, damit soziale Roboter erfolgreich von kleineren Unternehmen oder Einzelpersonen akzeptiert und eingesetzt werden können. Bisherige Studien haben bereits gezeigt, dass sich die Aufgaben-Personalisierung positiv auf die Wahrnehmung der Benutzer*innen hinsichtlich der Benutzbarkeit der Roboter auswirken kann. Jedoch sind die klassischen Programmiermethoden für Laien eher ungeeignet. Neben anderen *Machine Learning* (ML) Programmierparadigmen stellt *Learning from Demonstration* (LfD) eine vielversprechende Methode dar, um maschinelles Lernen mithilfe menschlicher Demonstrationen zu ermöglichen. Wie auch andere Arbeiten zuvor, wurde in dieser die bestehende Lücke in der *Human-Robot Interaction* (HRI)-Literatur identifiziert, welche das fehlende Wissen über die Wahrnehmung menschlicher Instruktor*innen hinsichtlich lernender Roboter in verschiedenen Kontexten und Konfigurationen verantwortet. Diese Arbeit unternimmt einen Versuch, einen Beitrag zur Schließung dieser Lücke zu erwirken, indem sie zwei Eigenschaften eines Roboters untersucht: *Initial Proficiency* (initiale Kompetenz), die beschreibt, wie kompetent ein Roboter erscheint, bevor ein Lernvorgang gestartet wurde, und *Learning Rate* (Lernrate), die beschreibt, wie viele Demonstrationen der Roboter benötigt, um eine neue Aufgabe vollständig zu erlernen. Es wird analysiert, inwiefern diese Eigenschaften die Wahrnehmungen der Instruktor*innen hinsichtlich des Roboters und des eigenen Selbst, sowie deren Bereitschaft, den Lehrprozess fortzuführen, beeinflussen können. Dafür wurde ein kontrollierter Laborversuch entworfen und über Benutzertests ($N = 24$) evaluiert, in denen Teilnehmer*innen, in einer virtuellen- und LfD-Umgebung versucht haben, einem Roboter beizubringen, eine *Pick-and-Place*-Aufgabe zu lösen. Während die initiale Kompetenz eher eine indirekte Beeinflussung auf diverse Messungen zeigte, war der Roboter mit hoher Lernrate, im Vergleich zur einer niedrigen Lernrate, in der Lage, die Wahrnehmungen der Teilnehmer*innen im Allgemeinen positiv zu beeinflussen. Es wird daher empfohlen, effiziente Lernalgorithmen für LfD-Systeme zu priorisieren. Weitere Ergebnisse zeigen zusätzlich, dass die von den Teilnehmer*innen wahrgenommene Bereitschaft zu lehren und deren Selbstwirksamkeit vom tatsächlichen Lernerfolg des Roboters abhängen. Das führte zu der Interpretation, dass jeder sichtbare Fortschritt die Teilnehmer*innen motivierte, weitere Demonstrationen abzugeben und stärkte zudem ihr Vertrauen in deren Wirksamkeit.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Abstract

Technological advances made Social Robots (SRs) to become common in our everyday lives more feasible than ever before. In the last years, they have already shown to be practical in the healthcare domain, like for example, their productive use in the recent pandemic, which brought viable findings to light. Still, many challenges exist that need a better exploration by the scientific community to enable SRs to successfully be accepted and adopted by smaller companies or individuals. Previous work showed that task customization can be beneficial for users' perceptions of usability towards the robot. However, classic robot programming methods are not accessible to non-experts. Alongside other Machine Learning (ML) programming paradigms, Learning from Demonstration (LfD) poses a promising method for enabling machines to learn from human demonstrations in more natural ways. This study, along with other recent works, identified a currently existing research gap in the field of Human-Robot Interaction (HRI) which is responsible for the lack of knowledge when it comes to our understanding of how human instructors perceive their robotic students in various teaching settings and robotic configurations. This thesis attempts to contribute one part to fill this gap by analyzing how the two robotic traits *Initial Proficiency*, which defines how proficient the robot shows to be before being taught by a teacher, and *Learning Rate*, which defines how many demonstrations the robot needs to learn a new task by a teacher, can influence the users' perceptions of the robot and themselves, as well as their willingness to continue the teaching process. Throughout this work, a controlled lab experiment was designed and was evaluated over user tests ($N = 24$), in which participants tried to teach a pick-and-place task to a humanoid robot within a Virtual Reality (VR) and LfD environment. While initial proficiency only showed indirect effects over various measures, a fast learning robot was able to positively influence participants' perceptions' in general, compared to slow learning ones. Prioritizing efficient learning algorithms is therefore recommended for LfD teaching systems. Additionally, findings show that the participants perceived teaching motivation and self-efficacy are dependent on the actual learning success of the robot, regardless of the learning rate, which supports the suggestion that any kind of learning progress helps motivate participants to continue teaching and make them feel more confident while doing so.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Contents

Kurzfassung	ix
Abstract	xi
Contents	xiii
1 Introduction	1
1.1 Motivation	2
1.2 Goals and Research Questions	3
1.3 Expected Outcome	3
1.4 Thesis Structure	4
1.5 Contribution Statement	4
2 Related Work	5
2.1 Social Robots and their Roles in Society	6
2.1.1 Social Robots within Education	6
2.1.2 Social Robots within Healthcare	6
2.1.3 Social Robots within Tourism and Hospitality	7
2.1.4 Social Robots within Domestic Environments	8
2.2 Social Robots' Abilities and How to Teach Them	9
2.2.1 Classic Robot Programming	10
2.2.2 Machine Learning	11
2.2.3 End-User Robot Development	14
2.3 Social Robot Learning Traits and their Influence on Human Instructors	15
3 Approach	19
3.1 Study Design	19
3.1.1 Independent Variables	20
3.1.2 User Test Conditions	21
3.1.3 Tasks	23
3.1.4 Demonstration Task	23
3.1.5 Initial Proficiency Task	25
3.1.6 Resources	27
3.1.7 Participants	28

3.1.8	Measures	28
3.1.9	Expected Output of the Study	31
3.2	Implementation	31
3.2.1	Teaching Environment	32
3.2.2	The Robot Apprentice	33
3.2.3	The Human Instructor	37
3.2.4	Feedback Dialog	40
3.2.5	Logging of Events	40
3.3	Evaluation	41
3.3.1	Participants and Time Slots	41
3.3.2	Lab Experiment	42
3.3.3	Data Analysis	46
4	Results	51
4.1	Participants Personality	52
4.2	Teaching Time	52
4.3	Number of Attempts	53
4.4	Achieved Proficiency	54
4.5	Initial Observing Time	56
4.6	Robot Perception	57
4.6.1	Anthropomorphism	57
4.6.2	Animacy	59
4.6.3	Likeability	60
4.6.4	Perceived Intelligence	61
4.6.5	Perceived Safety	62
4.7	Teaching Self-Efficacy	63
4.8	Teaching Experience	64
4.8.1	Pragmatic Experience	64
4.8.2	Hedonic Experience	65
4.9	Teaching Motivation	66
4.10	Post-Hoc Explorative Analysis on Robot Success	67
4.10.1	Teaching Motivation	68
4.10.2	Teaching Self-Efficacy	69
4.11	Qualitative Results	69
4.11.1	Influence of Initial Proficiency	70
4.11.2	Influence of Learning Rate	74
4.11.3	Other Findings	77
5	Discussion & Future Work	81
5.1	Perception of Robotic Traits	81
5.1.1	Anthropomorphism	81
5.1.2	Other Traits	82
5.2	Teaching Motivation	83
5.3	Teaching Experience and Self-Efficacy	84

5.4	Expectations and the Surprise Effect	85
5.5	Implications for Design	87
5.6	Limitations	87
5.7	Future Work	89
6	Conclusion	91
6.1	How do Findings Align with Research Questions?	91
6.2	Toward Acceptable Social Robots	93
	Overview of Generative AI Tools Used	95
	List of Figures	97
	List of Tables	99
	List of Algorithms	101
	Acronyms	103
	Bibliography	105



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Introduction

After a rough day at work, you step into your favorite café, hoping for a moment of relaxation – and a good coffee, of course. Distracted by the latest messages on your phone, you barely take a glimpse at the old owner of the shop, sitting on a chair behind the counter. He’s always been very kind to you. Unaware of your surroundings you order the usual. Taking a seat, you close your eyes, trying to sort through all of your thoughts. Moments later, you recognize someone moving to your table in your peripheral. It could be your coffee. It certainly is, you hear the sound of the cup being placed on your table. You open your eyes and freeze. The hand placing your coffee has a metallic shimmer. You are carefully looking up. To your surprise, the barista is not the old kind man as usual, but a new one, a robot. “Thank you for your order. Do you need anything else?”, it says. You are wide awake now, no need for coffee anymore. It even wears an apron. “Excuse me?”. “Uh, no... no thank you”, you stammer. The robot turns around and gracefully moves towards another customer. Before you can make sense of it, the old man approaches your table, chuckling. He sits down, you look at him, still surprised by what just happened. “Ha, ha. Surprised, huh?”, you approve by nodding, “You know. It’s hard to find workers that want to do this job... willing to do the hours, so, I figured this would be a good solution”. You think about it for a few seconds, trying to get your thoughts straight. “How does it know...”, you started, but the old man interrupts excitedly “Well, I just showed it”. “You showed it, how?”, “It’s actually pretty easy, I just showed how I make coffee, served customers, you know, and it just stood there, watched me. And then, after some time... well it learned. Simple, right?”

Scenes like this, while seemingly futuristic, are becoming increasingly feasible as technology continues to evolve. From industrial to healthcare, service, and domestic domains, robots are rapidly integrating into our daily lives. Yet, teaching strategies, like the one the café owner used to train his robotic barista are not as easily achievable as of today.

Depending on their application, robots may significantly differ from each other in terms of how they perceive their surroundings, how they move, how they look, how they manipulate,

and how humans approach and interact with them [1, 2, 3]. Purely industrial robots, for example, often prioritize precision and efficiency, while often being incorporated into isolated settings. Collaborative robots, or *cobots* for short, can safely work with human colleagues in their proximity. In contrast, SRs, like the robotic barista in the café, demand nuanced movements, flexibility and acceptable social behaviors to naturally interact with people around them [4]. These requirements introduce particularly challenging issues into the field of robot development.

One promising approach to teaching robots human-like behaviors and movements is LfD, which enables human, non-expert teachers to instruct the robots by providing demonstrations to them [5]. This concept of letting machines (either virtually or physically) learn how to do things in more natural ways (ML), is not particularly new but is becoming more and more popular due to technological advances in computer hardware and open-accessible software frameworks. This type of robot feature development is especially useful, if there is the need for flexibility with tasks or where there is no clear go-to set of instructions which would always lead to a successful outcome or execution.

1.1 Motivation

Commercial robots usually come with a set of pre-programmed capabilities, which were seemingly developed by the respective manufacturer [6]. Having the option to teach a self-defined task to such a robot would be greatly beneficial for users and manufacturers in order to make them more flexible and personalized [7, 8]. As described above, LfD can be implemented as a learning mechanism for robots to achieve exactly that. In order to make the process of teaching such robots as efficient and comfortable as possible, a good understanding of how human teachers feel about their robotic students would be beneficial. Research is already done in this direction when it comes to how human teachers perceive the robot if it fails or succeeds with a given task [9, 10, 11]. However, little is found in research about how those human teachers would perceive themselves in terms of self-efficacy or teaching motivation and whether they would perceive it as more likeable, intelligent, or efficient, depending on how proficient the robot initially is or how fast the robot learns to do the given task successfully.

1.2 Goals and Research Questions

The goal of this study is to investigate whether the **initial robot proficiency**, i.e. how well a robot is able to perform a task prior to it being taught and the **robot learning rate**, i.e. how fast a robot is able to learn a new task, affects human teachers in the following ways:

- 1) Level of **anthropomorphism** of the robot.
- 2) Perception of robot **likeability**, **intelligence** and **safety**.
- 3) **Self-efficacy**: How much do teachers think they are able to teach the robot.
- 4) **User experience**: A bad user experience with the robot might lead teachers to exit the process early and could also potentially hinder them from teaching like this in the future.
- 5) **Motivation** to teach the robot: Since some ML programming paradigms require a considerable number of iterations to evoke a meaningful outcome, the motivation of teachers to continue the teaching process could be of high interest.

Out of these goals, the following research questions were formulated for and are being addressed within this study:

Research Questions

- RQ₁**: How does the initial proficiency level of a robot in a teaching environment affect human perception of the robot's capabilities and intelligence?
- RQ₂**: How does the rate at which a robot appears to learn a new skill influence the human instructor's perception of the robot?
- RQ₃**: What is the relationship between the perceived robot's proficiency level and learning rate, and the self-efficacy of the human instructor?
- RQ₄**: How do variations in the robot's initial proficiency and demonstrated learning rate impact the willingness of human instructors to continue teaching the robot?

1.3 Expected Outcome

This study aims to provide information about how human teachers may or may not be influenced by the aforementioned parameters of a robotic student and therefore also open

up a discussion about whether this influence, if any, is of significance, and if so, how much of a significance. Out of this discussion, recommendations will be formulated, if possible, on how to design certain aspects of robots within a LfD setting.

1.4 Thesis Structure

The main part of this thesis consists of the following chapters:

- **Related Work:** This chapter summarizes other relevant work regarding the themes *SRs within society, their abilities and how to teach them* and lastly *how robot learning outcome may influence their human teachers*. In the last section of the chapter, the previously indicated gap in literature will be presented more closely.
- **Approach:** This chapter starts with how the research questions were approached through a controlled lab experiment. The overall study design will be presented before going into actual implementation details. Additionally, there will be information about how relevant data was evaluated by using two sets of questionnaires and one final interview for each of the participants of the study. Finally, a concise description of how the data was further analyzed and processed will be given.
- **Results:** This chapter will give insights into the results of the data after they have been merged and analyzed. All of the relevant quantitative results will be shown, right before moving on to the qualitative results gathered by the interviews.
- **Discussion & Future Work:** This chapter will address the results shown in the chapter before in more detail, to provide meaningful insights within the context of the thesis. Quantitative and qualitative results will be mixed and an attempt at interpretation together with the findings of previous work will be made. The chapter closes with a list of the identified limitations of this work and recommendations about where to take further research.
- **Conclusion:** The final chapter of this work will summarize the whole thesis and provide answers to the research questions described above.

1.5 Contribution Statement

This thesis contributes to the research field of HRI by addressing the currently existing research gap. It proposes findings about how much the above-specified robotic properties **initial proficiency** and **learning rate** directly and in combination can affect humans in terms of **perception of the robot**, **self-efficacy** and **teaching motivation** within a robot-student, human-teacher, and LfD context.

Related Work

Due to advances within ML and pattern recognition, sensor technology, and computer science, the amount of research in robotics has heavily increased within recent years [12]. Robotics itself is a multidisciplinary field of research which includes e.g. materials science, mechatronics, computer science, biomechanics, aeronautics, (micro) electronics, and HRI [1, 13]. Advances within all of these research areas promise to automate even more complex tasks across many different fields of applications [4].

Compared to earlier versions of robots, which usually only were deployed within the industry and usually were only used for very specific tasks while also having their separate workspaces [14], nowadays robots can work semi- or even fully autonomously on more complex tasks and within the proximity of humans [2]. This trend makes the research area of HRI especially interesting with current robotics development [15]. It explores how communication and interaction between robots and humans can be understood, improved, and optimized [16]. That includes e.g. spacial, gesture, and human facial feature detection or detection and use of social cues, like emotions [17]. Robots that are able to communicate with humans to a certain degree can be considered as SRs.

Along with the many different ways to categorize robots and multiple ways to build taxonomies around different key features of robots, there is also no real agreement on how to clearly identify a SR [4, 18, 12, 19, 13]. However, Sarrica et al. stated that there is a common understanding about SRs at least taking part in social interactions [20]. Youssef et al. expanded on that by mentioning that they are able of doing services, while also being able to interact with humans by using different means of communication, including speech, gesture, or facial expressions [4]. Hegel et al. reviewed some of the most popular definitions of SRs in literature and constructed their own [21]:

“A social robot is a robot plus a social interface. A social interface is a metaphor which includes all social attributes by which an observer judges the robot as a social interaction partner.”

The specific research concerned about SRs is of higher interest for this work, as there seems to be a bigger overlap with HRI concerns than with more traditional robots. The following sections therefore are focused on SRs more specifically.

2.1 Social Robots and their Roles in Society

There is already a vast variety of different SRs being used in a multitude of different fields of applications, including education, customer service, medicine, healthcare, and also more recently, domestic environments [15]. In each of these respective environments (described in more detail below), the implementation of SRs promises to help solve several domain-specific problems, including reducing teacher workloads over tutoring robots in educational environments, reducing care staff workloads over assistive robots, help with aging in place for elderly adults, providing comfort, companionship and emotional support with hospitalized children in the healthcare sector and provide more efficient, emotionless and unbiased services to customers within the tourism and hospitality sectors.

Despite many positive advances in the field of SRs, there are still many challenges, including current technical limitations in terms of Artificial Intelligence (AI), environmental perception, low creativity, social engagement, and user acceptability in terms of self-efficacy and perceived usefulness, privacy, legal and ethical concerns, and maneuvering the *Uncanny Valley* in terms of anthropomorphism, to name a few. Besides the many challenges, SRs still carries a high potential to enhance people's lives in the future.

2.1.1 Social Robots within Education

For example within the education sector, telepresence (social) robots can be utilized as alternatives for teachers, experts, or students if they cannot be physically in class, to still be able to interact with those who are via teleoperation. Also, the same technology, when adjusted, would be able to help teachers or experts sort of broadcast the information they want to share to multiple classrooms at once. Furthermore in education, one other idea to relieve the workload of teachers would be to implement autonomous tutor robots. Those would be able to assist with teaching during or besides regular classes, in a more autonomous way. Research shows that the implementation of SRs may be able to help counter budgetary and demographic challenges that currently exist within the educational sector [22, 23]. Still, there are many issues to be solved, e.g. intelligence of autonomous robots, correct interpretation of social context for autonomous robots, acceptability by affected people or ethical considerations, to name a few [24, 22, 25].

2.1.2 Social Robots within Healthcare

In recent years SRs gained importance within many healthcare sectors [26, 4]. This is particularly evident in the context of elderly care which demands innovative solutions to address the estimated increase in the elderly population expected in the coming decades [27]. In literature, it is advised to enable people to stay at home with the possibility of

home assistance, when needed. Ambient Assisted Living (AAL) for example proposes a variety of different technical solutions to increase health, access to caregivers, housekeeping, communication, and social environment [28]. AAL mainly includes the incorporation of sensors or actuators into homes, which then provide certain mechanisms to help with daily activities or emergencies. One challenge of AAL is to provide acceptable and easy-to-use interfaces. Here SRs may be more important in the future to improve on accessibility, as they could be connected to the system and provide a more natural interface for communication to it, like e.g. over language or social cues [29]. Although AAL proposes many beneficial solutions for aging in place, several challenges need to be addressed before these systems can be widely deployed commercially [29, 30]. While SRs may help in more automated home environments, they can still improve the well-being of elderly adults, by improving their social lives, potentially reducing loneliness and depression while also having the potential to enhance human-to-human social behavior [31, 32].

In healthcare, SRs seem to have far-reaching benefits for hospitalized children. They can help with coping with medical treatment and can provide companionship to them. They can enhance their well-being to some extent and give them a feeling of comfort, which is helpful when children must leave their usual social environment due to medical reasons [33, 34, 35].

During the recent COVID-19 pandemic physical distancing measures were introduced to contain the spread of the virus. During the early times of the outbreak, healthcare workers were especially vulnerable due to their frequent exposure to contaminated individuals, however, they were not always able to keep their distance from them. Since, compared to humans, SRs are naturally not being affected by the virus, they were being utilized in a number of applications, including logistical tasks (e.g., delivery, telepresence or telerehabilitation), health monitoring and safety enforcement (e.g., protective measures and disinfection), and social or emotional support (e.g., companionship or entertainment) [36, 26, 37].

The recent advances of SRs in the context of healthcare seem promising, however, there are still barriers that hinder the process of their integration. First, there are still technical challenges, such as the need for a better overall HRI, environmental perception, and intelligence [26]. As SRs become increasingly embedded into the routines in healthcare environments, concerns about privacy and security grow. Some robots may be able to collect sensitive medical data, raising the risk of privacy violations or data theft. Further, highly autonomous robots additionally raise concerns about accountability, which introduces the need for a clear understanding of their ethical and legal implications [32, 38].

2.1.3 Social Robots within Tourism and Hospitality

There are various applications for SRs in travel, tourism, hotel and hospitality sectors, including front desk operations, cooking, food delivery or checkout, room cleaning, luggage transportation, providing travel, shopping, and booking assistance, to name

a few [39, 40, 41, 42, 43, 44]. Given the high possibilities of customer interactions and propelled by recent technological advances, SRs present a highly transformative opportunity within these sectors [44, 45, 41]. The adoption of SRs offers many benefits, such as efficiency, unaffectedness by emotions, unbiased approach to customers, and overall reliability with offered services. Further, they can help with engaging personalized experiences or real-time language translation. They are able to streamline repetitive or tedious processes which subsequently reduces service times [39, 43, 44].

While these industries may help to propel SRs development through financial investments, current robots are not as human-like as they would need to be to achieve a high degree of customer acceptance [46]. Additionally, advances in robotics and AI are going to bring SRs to a level of anthropomorphism where the affective reaction of humans falls flat (Uncanny Valley). Business owners may then not be able to afford to sell into such robots if customers tend to reject or feel uncomfortable with them [40].

As discussed in section 2.1.2, ethical and privacy concerns also exist within the sectors of tourism and hospitality [39, 43, 44]. Additionally, some customers may feel irritated or fear being monitored by SRs. Further, the lack of emotional intelligence and creativity are limiting customer satisfaction for more complex tasks as they might feel anger or frustration when interacting with SRs that are not able to solve their issues [39, 47, 44].

2.1.4 Social Robots within Domestic Environments

As described in section 2.1.2 above, SRs may be able to be set into homes with AAL installments to provide more natural interaction possibilities with the system for elderly adults [29]. In general, there seems to be an overlap of healthcare applications for SRs with home environments. The research in terms of SRs especially focuses on special health impaired groups and how SRs may help with solving certain issues (like, e.g. help with daily activities) [48, 49]. But, SRs are not practically, and even less theoretically limited to those kinds of applications. General purpose or home assistive SRs were long imagined in science fiction and people are already accustomed to other Personal Digital Assistants (PDAs), such as smart speakers [50]. So the question emerges as to why there is a lack of development of SRs within domestic environments.

Science fiction has heavily influenced how people envision SRs in domestic environments as helpers, companions, and general purpose tools [51]. The existing gap between the expectations that emerge through fiction and the actual limited abilities of such robots experienced in reality is called *social robot paradox*, coined by Duffy and Joue [52]. Henschel et al. [51] found that robot manufacturers implement certain human characteristics into their products (anthropomorphism) while being careful not to push it too far into the uncanny valley. However, enhancing human-like features would indicate a potential for enhanced social interaction with such robots. In their investigation of the SR literature Mejia and Kajikawa [53] found, that while social sciences seem to be acknowledged to play an important role in SRs research, they remain underrepresented [51]. Due to the fact that HRI is situated between robot engineering and social sciences,

Broadbent [54] and Eyssel [55] propose that research in this field should aim for a higher incorporation of the social science principles [51].

David et al. noticed, how many works in research were expressing rapid growth of the SRs market, which they thought to be contradicting given the overall low acceptability of those robots [49]. Throughout their literature review in terms of SR acceptability, they found, e.g. that technology may be used differently as initially designed [56], that the technology of the robots are expected to be different by users, which may end up with a product not matching their needs (perceived usefulness) [57] and that a novelty effect may lead participants in various lab experiments to initially rate robots as acceptable, an effect that might not persist over time. It was found that factors like self-efficacy and prior expectations of the robot are essential for their initial acceptance and are therefore important for their adoption [58].

De Graaf et al. reviewed long-term studies on SRs in domestic environments and identified several key factors influencing SRs acceptance [58]. They included other factors that potentially increase acceptability (additional to the aforementioned), such as the belief of having the necessary skills to use a SR (self-efficacy), the perception of increasing status by the possession of a SR, the expectation that of such a robot providing enhanced enjoyable interactions and the expectation of it causing fewer privacy concerns. It is important to recognize these effects and their corresponding challenges to implement effective solutions and reach for potentially higher acceptability for such robots in people's homes.

2.2 Social Robots' Abilities and How to Teach Them

In the context of SRs, manufacturers often deliver their products with a standard set of baseline abilities, such as navigation, recognition of the environment or certain objects, basic body movements or gestures, and other basic functionalities. They represent general-purpose features that enable them to (socially) interact with the world around them. However, as such robots increasingly are adopted into all kinds of domains, the demand of people to customize or personalize such robots emerges [59, 8].

Customizations, tailoring, and personalization are especially interesting for small corporations or individuals, who may want to add, remove, or tweak certain features to meet very specific requirements. Additionally, customized robots increase their usefulness and can subsequently raise their acceptance in unique contexts [59, 60]. Additionally, users that made certain customizations to their robots may want to share their work with other people or communities, so that they can benefit from it as well.

To understand how such customizations can be made, it is helpful to understand which types of programming methods exist. While the taxonomy of programming methods described by Heimann et al. [61] is targeted at industrial robots, its classification of methods based on the interaction with the robot, still gives a good overview of how they can be programmed. Methods are classified into online and offline programming.

Online methods refer to robot programming strategies that need access to the actual robot itself, including:

- Lead through programming: One of the earliest types of online programming, where a user for example guides one of the robot's physical parts to the desired location with a teach pendant. The positions are simply saved for later playback.
- Walk through programming (also known as kinesthetic teaching [7]): This is an iteration of lead through programming, where users simply move the robot's joints to their desired locations manually and without the pendant. Compared to the previous version, this does not need a high-level understanding of the robot.
- Programming by Demonstration (PbD): This method relies on sensors for the robot to observe demonstrations, which can be performed by the user. Also, no programming skills are needed, so this method can be used by the general public [62].

Online programming methods have the disadvantage, that the robot needs to be programmed in its proximity or by actual physical manipulation of its joints.

Offline methods refer to robot programming strategies that can be implemented on some sort of abstraction of the robot, including:

- Text based programming: This method is similar to classic software development. Usually, with industrial robots, the interface allows access to previously recorded points from online strategies, like lead through programming.
- Graphical interfaces: Such interfaces provide a simplified user interface to, for example, access and fine-tune a robot task template (e.g. for a welding task).

Also, hybrid approaches are possible. An example of such a method would be to have a virtual model of the robot in a 3D environment where it can be programmed and simulated. The resulting program can then be uploaded to the actual robot. It is worth mentioning that the process of transferring the result of a virtually trained robot to a real one often does not work without issues and demands additional adjustments [63].

While all of the above methods have their advantages and drawbacks, some are inherently more useful than others for non-experts. The following sections describe programming methods in more detail.

2.2.1 Classic Robot Programming

Lead through and text based programming can be viewed as more classical approaches for defining and programming tasks for robots. Experts are needed for these methods, as they need a high level of understanding of their functions [61]. In light of customization, this seems not to be suitable for most people. It is possible to extend these methods

by incorporating graphical interfaces, which provide a more intuitive and abstract way, making programming more accessible for non-experts. Still, classic programming may not be sufficient for every scenario, when given the complex nature of tasks and variable environments in the context of SRs [7, 14].

2.2.2 Machine Learning

In the broad scientific field of AI, one of its important branches is ML, which encompasses various methods to enable computers to learn how to do tasks in ways that are closely related to how humans or animals learn [64]. This is especially useful if a computer is required to successfully solve real-world problems, such as speech, gesture, or object recognition (or pattern recognition in general), prediction of events, complex classification of data, or decision-making based on feedback from the environment. Those kinds of problems usually appear very natural to humans but can be complicated to formalize for computers. A good understanding of how humans or animals learn in nature is therefore essential. Consequently, the field of ML needs to incorporate a multitude of other scientific fields, such as psychology, neuroscience, and philosophy. Basically, ML tries to make predictions or decisions on a certain input based on prior experience.

In the context of ML this is done over *models*, which are the resulting predictors trained via a previously chosen ML approach or algorithm, specifically targeted at the problem at hand. These models may vary strongly in computational complexity, difficulty of implementation, and learning process automation, again depending on the specific problem that needs to be solved [65, 66]. The following paragraphs briefly describe a selection of ML paradigms that are relevant for this work [67, 66, 65, 64].

Supervised Learning: This is an ML approach where a model is trained to map input data to an output label, with the help of labeled training data. Labels are the ground truth for classification. Therefore, this approach requires training data that is already classified beforehand, for example by a human. Labels can be of discrete or continuous type. In a classification task, the trained model will output a label, which is one of the set of labels provided alongside the training data, for a new input. In a regression task, a trained model outputs a predicted continuous value in the range of the provided label values of the training data. The concept of supervised learning is displayed in figure 2.1.

Unsupervised Learning: Compared to supervised learning, this approach does not rely on ground truth labels. Instead, the ML algorithm identifies patterns within the training data. For example, a common method (k-means clustering) forms a certain amount of clusters of the data by analyzing predefined features. The resulting model then is able to output a specific cluster group identifier to which it predicts a new input to belong. The concept of unsupervised learning is displayed in figure 2.2.

Semi-Supervised Learning: There is also the possibility to mix both of the above methods, when just a part of the training data has labels, while the majority of the data

2. RELATED WORK

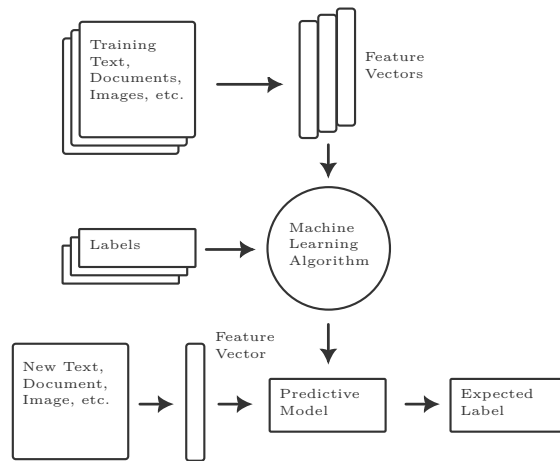


Figure 2.1: Supervised Learning: Adapted from Preeti and Dhankar [67]

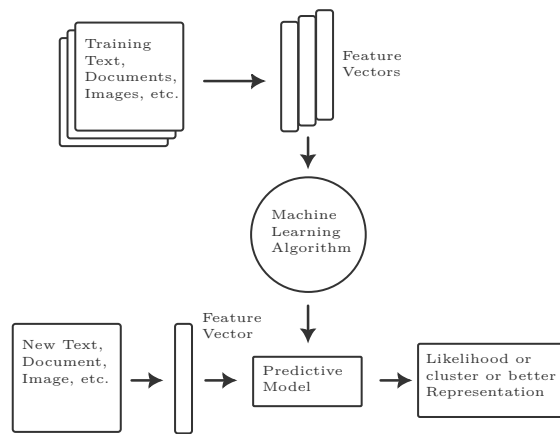


Figure 2.2: Unsupervised Learning: Adapted from Preeti and Dhankar [67]

has no labels. This may be the case if a labeling of the whole training dataset is not feasible. The resulting model has the same applications as supervised learning since at least some of the data is labeled.

Reinforcement Learning: This is an ML approach where an agent learns to perform actions in a dynamic environment to maximize a cumulative reward value. In the context of Reinforcement Learning (RL), an agent represents an entity situated in an environment that it is able to observe with one or multiple sensors and in which it can take actions over one or multiple actuators. A reward function (e.g. a set of rules) provides a reward (e.g. a numeric value) to the agent, dependent on each state of the environment (which can include the agent). The agent then collects these rewards, which it tries to maximize by exploring or exploiting various actions inside of the environment [68]. The concept of RL is displayed in figure 2.3.

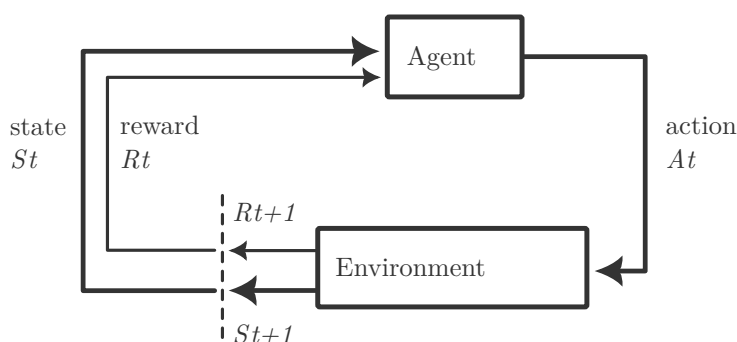


Figure 2.3: Reinforcement Learning: Adapted from Mahesh [66]

Learning from Demonstration

The training method of LfD is typically not described as a separate foundational ML paradigm, but rather as a methodology that often builds on top of them. However, it is not constrained to any specific paradigm, while still being able to integrate or enhance them. This flexibility makes it challenging to fit and align with the previous methods. Briefly described, throughout LfD a policy is taught to the agent, which is derived by analyzing demonstrations made by some entity (usually a human) that knows how to perform the desired task. Therefore it may be viewed as a subset of supervised learning, as the agent tries to approximate the function learned by the training data set, which consists of the given demonstrations. Furthermore, RL and LfD form a close relationship by aiming to teach a policy to an agent, therefore they often are compared to each other in literature [69, 68].

Within RL, the agent learns from exploration and exploitation, whereas LfD makes it learn from experience. In the context of robotics, one of the major drawbacks of RL is, that it often needs experts in the field of application, when trying to teach potentially complex real-world tasks to a robot [68]. Conversely, LfD does usually not need any expert knowledge, other than the knowledge on how to correctly perform the specific task at hand. This makes it an attractive method for humans to teach robots how to do human-like tasks. Additionally, using body movements has shown to be a more natural approach to teaching tasks to robots than formulating rule-set policies. Further, LfD is also much more suitable for teaching robots specific movements in their actual physical environment, since it only tries to learn from the provided demonstrations, whereas a traditional RL approach may lead the robot to try potentially hazardous movements.

Although LfD promises to streamline teaching processes for non-experts, it also faces some challenges. They include, for example: the issue of correspondence, the problem of limited datasets, and how policies are derived from them [69].

First, the issue correspondence describes how the movements of the teacher should be mapped to the specific body parts of the agent. Depending on the form factor of the agent and method of demonstration recording, this can either be very complex or simple.

Second, limited datasets pose the problem that a simplistic policy strictly derived from a certain set of demonstrations is not able to generalize the problem to a similar configuration of the environment, if this specific state of the environment is not represented by one of the examples in this set. Here, other ML paradigms are often utilized, as many of them are suitable to generalize policies from training data (e.g. as demonstrated in [70]). For example, including RL is a popular method to generally enhance the LfD approach or vice versa [69, 71, 68]. Demonstrations from naturally intelligent entities could in that sense serve as a baseline for achieving a high cumulative value from the RL reward function, which arguably eliminates the need for a high number of initial exploration steps which would potentially lead to low rewards without them. Off of that reference point, RL may continue to make the agent explore how different kinds of actions lead to a high reward in different environmental scenarios, which can further increase the accuracy of the resulting model. Although mixing RL with LfD has its benefits, it reintroduces the problem of requiring expert knowledge to design a usable reward function, which would eliminate one of the key advantages of LfD. However, in this combination, the problem can be circumvented by using Inverse Reinforcement Learning (IRL), which can derive such a reward function, again based on the provided demonstrations [68, 69]. It is worth mentioning, that, depending on the desired task and environment, such a reward function may not be the global optimal, since it cuts away many possible routes the agent may take in a more traditional RL approach.

Although a bigger set of demonstrations provides a better baseline for deriving policies, it should be mentioned that if humans are providing them, they may be annoyed with the need for a high repetition of the task [71]. It is therefore recommended to rely on only a few demonstrations and to make the teaching experience as pleasing as possible to maintain motivation.

2.2.3 End-User Robot Development

As mentioned above, describing and developing tasks for robots usually requires a high level of expert robot development and domain-specific knowledge [61, 72, 7, 14]. With the help LfD, the process of teaching a task to a robot can be simplified and is often considered more natural than traditional programming. However, non-experts still need some sort of tool or environment to be able to teach the robots in the first place. The field of End-User Development (EUD) encapsulates methods, techniques, and tools which allow users to customize software or hardware (e.g. with 3D printed) artifacts in a clear and simple way [73, 14, 74]. In the context of social robotics, such tools face several challenges, e.g. the need for certain context specifications which might include objects, obstacles, or intelligent entities [7].

Ajaykumar et al. [7] identified a number of different types of approaches for robotic EUD tools, which include programming via visual, i.e. by using Graphical User Interfaces (GUIs), Augmented, Virtual or Mixed Reality (XR), i.e. by using a simulated virtual representation of a robot, demonstration, i.e. by using LfD, natural language or tangible methodologies. Depending on the use case and target users, several robot capabilities were

able to be programmed by these tools and methodologies, including (social) interactive behavior, object manipulation, execution of motions (which includes gestures), navigation or sensing of the environment, and providing audiovisual feedback. Additionally, the authors of this survey identified a set of measures that are usually incorporated into EUD evaluation. They included, for example, task success (i.e. by checking whether user-authored programs were executed by the robot successfully), programming time and progress (i.e. how efficiently a user is able to use the programming tool to teach the robot), user perceptions of the system, for example by measuring perceived usability over the System Usability Scale (SUS), or user experience over the User Experience Questionnaire (UEQ). Additional measures also include evaluation for willingness to use the system and the users' perceptions and other subjective evaluations of the robot itself and how the proposed EUD system may affect them. Furthermore, objective measures, such as the number of interactions of the users with the robots might provide additional insights into how engaging the trained robot and the system as a whole are. Finally, some sort of user instruction or training (e.g. over an in-application tutorial) seems to be a common practice, before the actual system is evaluated, since some might need users to get comfortable with the procedure or special hardware.

Within the identified approaches, the use of XR in combination with LfD seems to be especially interesting for SR experimental settings, since there is the possibility to test robotic features that do not yet exist or if testing a physical robot is not feasible or even dangerous [7, 72].

2.3 Social Robot Learning Traits and their Influence on Human Instructors

As described in section 2.2.2, it is generally recommended to keep the number of needed demonstrations by humans low in an LfD approach. However, this may not always be possible with every teaching scenario, as they vary in complexity and their required accuracy. Also, a human instructor might expect a robot to behave differently, based on certain observations [9, 75]. It is therefore important to understand how teachers perceive the robotic student and themselves in different robot teaching settings to avoid frustration or mistrust and instead enable a productive, engaging, and effective teaching experience, potentially increasing their acceptability [11] and increasing the success of teaching robots custom tasks.

Hedlund et al. [9] conducted a lab experiment, where human non-experts tried to teach different tasks to a physical robotic arm over different instruction methods in a LfD setting. Within the study, participants were asked to try to teach three different tasks to this robot sequentially: first, pick-and-place, second, rod insertion and third, object retrieval. Each participant was trying to teach the tasks to the robot over one of three instruction methods: kinesthetic teaching (see section 2.2), teleoperation (by using a joystick controller), or motion capture (by using a camera and a glove with an attached April Tag for the participant). Participants were able to train each of the tasks before the

actual actual LfD process was started. Then one official demonstration (also referred to as *One-Shot Learning* [76]) was provided by them and the robot presented if it successfully learned the task or not. Participants were made to believe that the demonstration was incorporated into a ML procedure, so the successful or failing output of the robot was shown over a set of pre-recorded trajectories. Thus, the success or failure of the robot for each of the tasks was pre-defined for each of the participants in a counterbalanced sequence.

The goal of this study was to evaluate how the success and failure of the robot in the different settings of instruction methods and over the three tasks influences the human instructor's perception of workload and perception of trust and impression of the robot and themselves. Results showed that, if the robot was failing at the task, participants trusted the robot and themselves less. Further, participants reported a lower impression of themselves when they were asked to teach over kinesthetic teaching compared to the other two methods. No differences were measured over the different tasks. Additionally, participants expressed a higher perceived workload, when the robot was failing compared to when it was successful.

A similar lab experiment was conducted by Moorman et al. [10], where the goal was to evaluate how certain robotic task learning settings were able to influence different target users and their perception of the robot. Within this study, participants faced a physical robot trying to learn a cutting task over one of three different simulated learning methods, which consisted of RL (i.e. participants were only observing a trial and error learning process), Interactive RL (by enabling users to give binary feedback to the robot during the RL learning process) and LfD (over kinesthetic teaching) approaches. Additionally, one learning condition was shown to all of the users, where they only initially observed the robot downloading a model on how to do the task from "the cloud" and then subsequently observed the robot executing the task based on this model. Further, participants were divided into two groups, where one of them participated in person and the other one observed and interacted with the robot remotely.

Results showed that, with increasing involvement of human guidance throughout the different learning methods, the robot is perceived as less human-like. The authors suggested that this may be due to participants assigning personality to the robot when it tries to learn on its own. Further, reliability and ease of use were impacted by the perceived success of the robot while the learning method had no effect. This was interpreted as an indication that users may prefer robot learning methods that promise learning success. Finally, the physical presence of participants positively affected the perception of safety, reliability, attitude, and trust towards the robot compared to remote participants. According to the authors, this may be due to the presence of the researcher during the study and due to participants increased perceived usability when they are co-located with the robot.

Another interesting study is proposed by Wang et al. [77] which is concerned about how natural human feedback for robots learning over RL approaches might unfold over different robotic competency levels in terms of perceived usability, workload, and trust.

The competency of the robot will be configured as either consistently low (i.e. that the robot shows an overall low competency and not learning from feedback), consistently high (i.e. that the robot shows an overall high competency and does not need much feedback), decreasing (i.e. the robot shows high competency during the initial phase and then low competency in the following phase) and increasing (i.e. the robot shows low competency during the initial phase and then high competency in the following phase). The authors plan to conduct a study with participants who observe a physical robotic arm learning a pick-and-place task. Participants will be offered to provide feedback to the robot (whenever they see fit) by first interrupting the robot and then subsequently by providing corrections over kinesthetic teaching. The goal of the study is to evaluate how these different competency configurations affect the participants' feedback and how they affect the participants' perception of trust.

As it stands, recent works seem to be more interested in understanding how users perceive their robotic students under various conditions, such as different competency levels [77], different modes of operation, or different methods of learning [9, 10]. All of these studies have identified the existing research gap in the field of HRI and acknowledge that there is still more to explore to strive for better customization of robots and to potentially achieve higher acceptability of SRs in the long run. The work in this thesis hits a similar chord as the aforementioned studies and contributes additional findings that align with their direction.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Approach

A controlled lab experiment was designed to evaluate the research questions defined in section 1.2. The experiment primarily consisted of user tests for measuring quantitative data over questionnaires, mainly focusing RQ₁₋₃, and then subsequently also semi-structured interviews as a follow-up for qualitative inputs, mainly considering RQ₄.

In this chapter the overall approach will be reviewed in detail within three sections: Beginning with the description of the study design itself, this section introduces how the relevant data can be collected. The second part is about how all of the requirements for the lab experiment were gathered and implemented, delivering details about specific decisions made for the user tests and interviews. Finally, the evaluation process will be presented, showing how all of the outputs of the experiment were processed to be able to aggregate results.

3.1 Study Design

As mentioned above, the study design was compiled into a mix of questionnaires, user tests, and interviews, effectively resulting in a mixed-methods design (as seen in figure 3.1). The main goal of the study was to evaluate how human instructors in VR and LfD environments are being affected by the two independent variables, *Initial Proficiency* and *Learning Rate* (refer to section 3.1.1). The design was constructed such that the findings of the quantitative data (derived from user tests together with questionnaires) were able to be contextualized with participants' reflections (derived from semi-structured interviews). The quantitative part of the study design consisted of one onboarding questionnaire, which included items to capture demographic information and personality traits, then the user tests themselves, during which several measures were recorded, including the time participants spent doing various things, the number of demonstrations participants are providing, or the current state of the robot, and finally, another questionnaire, which included items to measure participants' perceptions of the robot, and their teaching

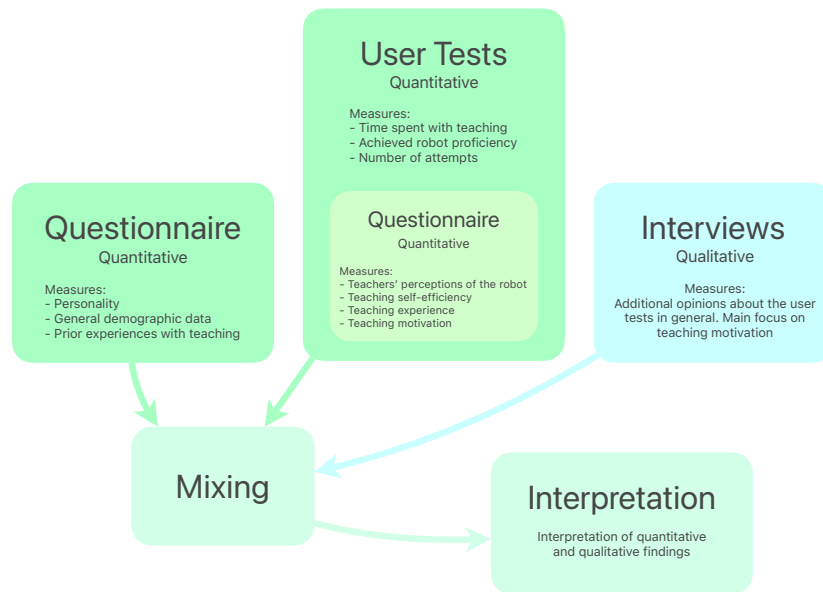


Figure 3.1: Study Design Chart

experience, self-efficacy, and motivation scales (refer to section 3.1.8). The qualitative part of the study design mainly consisted of the semi-structured interviews which concluded each of the participants’ sessions. Interview questions were mainly concerned with how different robot configurations affected them in terms of their teaching motivation. The experimental design followed a within-subject approach, where each of the participants interacted with the robot across four different configurations. The four test conditions were assigned to participants in a Latin-square counterbalanced sequence. The LfD robot teaching process was conducted in an VR environment, where participants tried to teach a pre-defined task to the robot (refer to section 3.1.3). The learning progress was simulated, so it was known beforehand which of the conditions made the robot act in certain ways. Subsequently, this ensured comparability between the individual results of the participants.

3.1.1 Independent Variables

Since the study aims to measure potential differences across various variables for the two robotic properties: initial proficiency and learning rate, exactly these two had been selected to be the independent variables within the user tests.

Initial proficiency: Within this work, this relates to the robot’s ability or accuracy in being able to successfully execute a fixed predefined task before an attempt to teach it had been made. The initial proficiency of a robot was presented to each of the participants

for each of the four test conditions respectively. It is important to mention that the two tasks involved within each of the tests, i.e. the task over which the robot is presenting its initial proficiency to the participant, and the actual task that the participants attempted to teach to the robot, differ from each other. Please refer to section 3.1.3 for more specific information on tasks.

Learning rate: Within this work, this relates to the number of demonstrations the robot needs to observe (i.e. teaching iterations a participant needed to conduct to the robot within a single, particular test) to be able to successfully execute the task.

3.1.2 User Test Conditions

To be able to measure potential differences within dependable variables, both initial proficiency and learning rate need to be varied across at least two states respectively. This led to the following definitions of states: Low initial proficiency (L), high initial proficiency (H), slow learning rate (S), and fast learning rate (F). Combining these leads to four distinct test conditions, which are: LS, LF, HS, and HF. To make it easier to refer to these conditions they were encoded into a single letter format: A, B, C, and D. Please also refer to table 3.1.

Test Condition		Initial proficiency	
		<u>L</u> ow	<u>H</u> igh
Learning rate	<u>S</u> low	A	C
	<u>F</u> ast	B	D

Table 3.1: User Test Conditions

Counterbalancing: Since there are four conditions to test for each of the participants, it is good practice to counterbalance them to avoid the effects of accumulated practice or carryover effects more generally [78]. That means, that every participant will not get the same sequence of test conditions, but different ones. The most straightforward approach would be to test every possible permutation of condition sequences, which would be optimal to avoid carryover effects within the measured data, but also would have the negative side-effect that the number of possible permutations grows factorially. With only four conditions this results in $4! = 24$ possible sequences which would also mean that the number of participants needs to be a multiple of 24.

To reduce this number, a balanced Latin square approach was chosen. A Latin square design implies the use of a n^2 matrix, where n is the number of conditions. The cells of this matrix can be filled in with test conditions (A, B, C, and D in this case) in such a way that each row of this matrix contains every single condition. Each of the resulting

rows represents one sequence of conditions. Thus, the number of possible sequences effectively is reduced to n , the number of conditions, i.e. four sequences in this case.

There are several possibilities on how to fill the cells of the Latin square with the actual conditions. In the balanced Latin square approach, one simple solution is to use the following instructions. Assuming we have a finite set of conditions $C = \{1, 2, 3, \dots, n\}$ where $n = |C|$ is even, the first row c_1 for the matrix can be evaluated as such (modified version of Edwards [79]):

$$c_1 = 1, 2, n, 3, n - 1, 4, n - 2, \dots, \frac{n}{2} + 1 \quad (3.1)$$

which can be defined as a sequence:

$$(c_{1,k})_{k=1,\dots,n} = \begin{cases} k, & \text{if } k \leq 2 \\ \lceil \frac{k+1}{2} \rceil, & k > 2 \text{ and } k \text{ is even} \\ n - \lfloor \frac{k-3}{2} \rfloor, & k > 2 \text{ and } k \text{ is odd} \end{cases} \quad (3.2)$$

Then the subsequent rows $c_{r,k}$ with $r > 1$ are built upon their predecessor, like such:

$$(c_{r,k})_{r=2,\dots,n, k=1,\dots,n} = (c_{r-1,k} \bmod n) + 1 \quad (3.3)$$

Applying this to a set of four conditions: $C = \{1, 2, 3, 4\} \implies n = 4$, this results to the following matrix:

$$\begin{pmatrix} 1 & 2 & 4 & 3 \\ 2 & 3 & 1 & 4 \\ 3 & 4 & 2 & 1 \\ 4 & 1 & 3 & 2 \end{pmatrix} \quad (3.4)$$

Which can be substituted with the actual conditions of the user tests $X = \{A, B, C, D\}$ while preserving the same order of elements, i.e. $1 \rightarrow A$, $2 \rightarrow B$, $3 \rightarrow C$ and $4 \rightarrow D$, this results in the final Latin square as defined in table 3.2. This table was used during

Condition Sequences (S_i)				
Sequence 1 (S_1)	A	B	D	C
Sequence 2 (S_2)	B	C	A	D
Sequence 3 (S_3)	C	D	B	A
Sequence 4 (S_4)	D	A	C	B

Table 3.2: Latin square balanced condition sequences

the experiment to read the sequence of conditions for each of the participants. Each participant was assigned a unique ID $p = [1, m], p \in \mathbb{N}$, which was just a simple integer

sequence number and m being the last participant ID. This sequence was sufficient, with me being the only researcher conducting all of the user tests. The Sequence number i for each participant was derived as such: $i = ((p - 1) \bmod 4) + 1$, i.e. for the first and after each set of four participants the sequences from table 3.2 were cycling from top to bottom.

3.1.3 Tasks

There were two tasks that participants were confronted with within the user tests. One task was the one that the participants were trying to demonstrate to the robot. The other task was done solely by the robot, which it performed in front of the participants to demonstrate its initial skill (see section 3.1.2). Both will be described in more detail below.

3.1.4 Demonstration Task

In robot LfD research there is a variety of different tasks which for example include: pick and place, peg insertion, polishing, grasping, or assembly operations. These are found often in manufacturing-themed works. More specific ones can be found for example within the healthcare sector, which include: feeding, specific tasks for physical rehabilitation, surgery, and assisting or handling objects. Besides these examples, there can also be found tasks which are about basic movements for specific mobile robots, which include: flips, rolls, or even more complicated flight maneuvers for aerial robots, valve turning or marine data collection for underwater robots, walking or gait optimization for bipedal or quadrupedal robots (locomotion) [80].

As this work already focused towards SRs, the type of robot was chosen to be a humanoid SR for the experiment. Furthermore, as described above, the user tests will be conducted within a VR environment, so the robot will be presented digitally. This way, certain basic robotic movements can simply be done over pre-recorded animations. Many of the above examples of tasks are too specific and could lead to different and strong opinions by participants when asking certain questions about the robot's abilities and safety etc. However, some tasks are more abstract, like for example: pick and place, grasping, or peg insertion. The first one among these three seemed to be the most straightforward one to choose since this can be done with all kinds of basic shaped objects and is often incorporated into related user studies [77, 81, 72, 7, 14].

The final pick-and-place task was defined as such: Initially, three cubes are being generated within a pickup area for both, the robot and the teacher respectively. The teacher stands opposite the robot while demonstrating to it, how to pick and place the three cubes into a fixed target area while maintaining a specific final order to them. The correct target order of the three cubes is instructed to the teacher (please refer to figure 3.2).

Within this task, the learning rate of the robot was of importance (see section 3.1.2). This defined how many demonstrations were needed as input from the human teacher to successfully teach the robot. This implied that the robot must not have accomplished

3. APPROACH

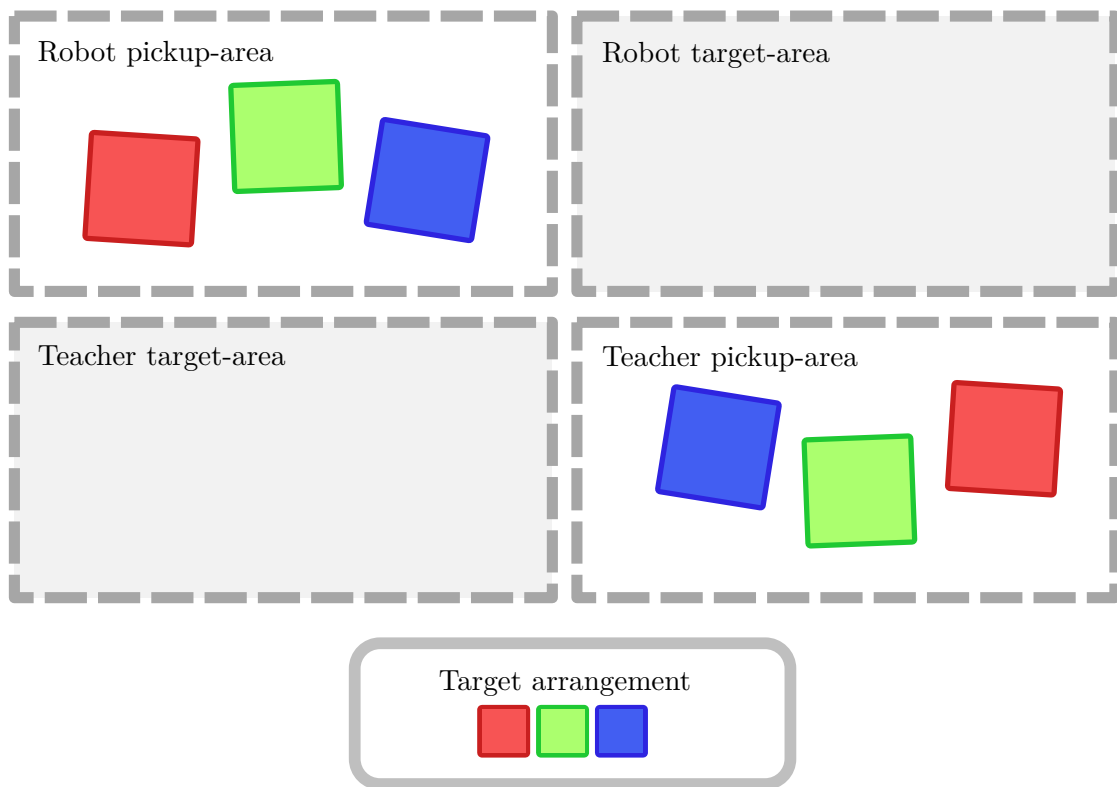


Figure 3.2: Initial setup for pick-and-place task in top-down view

the task before it observed a certain amount of demonstrations depending on whether its learning rate was set to slow or fast. To avoid having the robot achieve the goal of the task by accident, the actual learning of it was simulated instead of using actual machine learning. This ensured that the robot's movements and learning outcomes were the same across each of the user tests for all of the participants. That way the results from each of the participants were comparable more accurately, excluding any kind of randomness that an actual machine learning strategy would have introduced.

For the simulated learning to happen, the following questions had to be answered:

- How does the robot learn?
- How does the robot show to the teacher its increasing proficiency at the task?

How does the robot learn? Within each single user test and for each single condition the robot started without having any proficiency with the task. When a participant provided a valid demonstration (i.e. the cubes on the participant's side were correctly picked and placed into the target area while also having them in the correct final order), the task proficiency of the robot increased. The task proficiency itself can be represented as a percentage value between 0% (i.e. no proficiency at the task) and 100% (i.e. being

able to correctly do the task). The extent to which this level increased for each valid participant demonstration, depended on the learning rate setting of the robot. For the test-conditions A and C (see table 3.1) the learning rate was set to slow, which means the robot increased its task proficiency by only 11% for each valid demonstration provided. For the other two conditions B and D, the learning rate was set to be fast, increasing the robot task proficiency by 34% for each valid demonstration provided. This resulted in a slow learning configured robot being able to accomplish the task after ten iterations and a fast learning configured robot already after only three iterations of valid demonstrations.

How does the robot show to the teacher its increasing proficiency at the task?

Since the task proficiency of the robot was being tracked for each user test and increased for each valid demonstration, this property was also then used by the robot to show the teacher how proficient it was at the task. For this, the robot tried to do the task by itself after each demonstration it observed from the teacher. Depending on the current task proficiency, the robot then either executed the task with multiple mistakes at once (i.e. a combination of single mistakes), a single mistake or no mistake (please refer to figure 3.3 and 3.4). So, if the robot made a mistake it was one of the following: Cubes in the wrong order (M_1), cubes misplaced (M_2), or the combination of these two (M_{1+2} , see table 3.3). Additionally, if the robot made a M_2 mistake, the distance of misplacement also depends on whether the robot has low or high task proficiency. If it is low, the distance is greater, and vice versa.

Demonstration task mistake table				
Proficiency	< 33.34	≥ 33.34	≥ 66.68	≥ 100
Range (%)		< 66.68	< 100	
Mistake	M_{1+2}	M_1	M_2	-

Table 3.3: Mistakes made by the robot by task proficiency level

The proficiency increasing step sizes of 11% and 34% were roughly picked because, first the participants should be well able to distinguish between slow and fast learning robots, and second the individual experiment runs (i.e. briefing, questionnaires, four user test, and interview) should fit into roughly one-hour time slots.

3.1.5 Initial Proficiency Task

This is a task that the robot executes on its own at the very beginning of each of the user tests (i.e. also for each of the individual test conditions). This was done to suggest to the participants how much proficiency the robot had with a different task, effectively introducing them to the first mentioned independent variable, described in section 3.1.2. Since the task must be different from the demonstration task, another more abstract task was chosen, which is the task of drawing a rectangular shape. This task is different

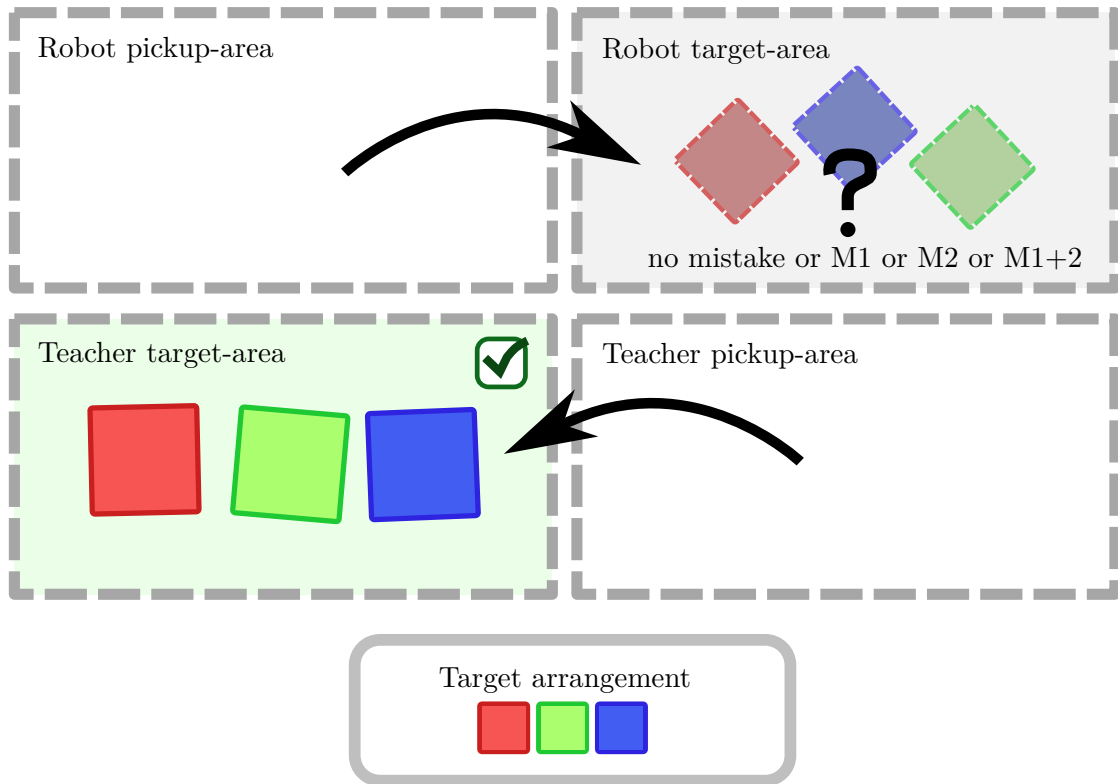


Figure 3.3: Valid demonstration, robot task execution outcome depends on proficiency

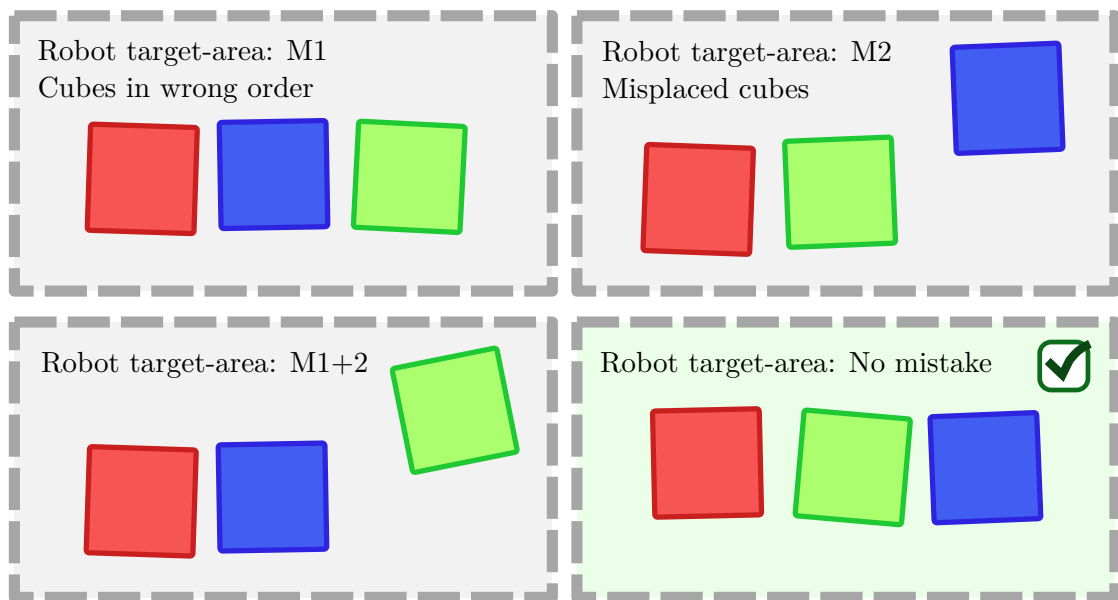


Figure 3.4: Types of mistakes

from the demonstration task enough so to avoid having participants interpreting it as being interconnected, while also still sharing similarities between them. For example, the 2D shape of a cube on a flat surface is a rectangle, so if the robot understands how this shape is drawn, this might be able to suggest to the teachers that there is some understanding of how to place cubes onto flat surfaces. While this suggestion is not required for participants to get to answer any of the research questions, still it seemed that this may have led to higher chances of participants making guesses on how the robot would perform on the pick-and-place task afterward which also would not be of any discount to the results. Also, like the demonstration task, this task was not done with actual machine learning to prevent difficult-to-compare results.

For the test-conditions A and B (see table 3.1), the initial proficiency was set to low, meaning the robot was not able to correctly draw a rectangle, whereas in test-conditions C and D it was high, meaning that the robot was able to perfectly draw a rectangle. The path that the robot tracked with its hand to draw the rectangle was fixed for each of the test conditions (please refer to figure 3.5). Furthermore, there is no learning process running for the robot as it presents its skills on this task, leaving it always drawing the same shape over and over again within a given test condition until the teacher decides to move on to teaching the actual demonstration task.

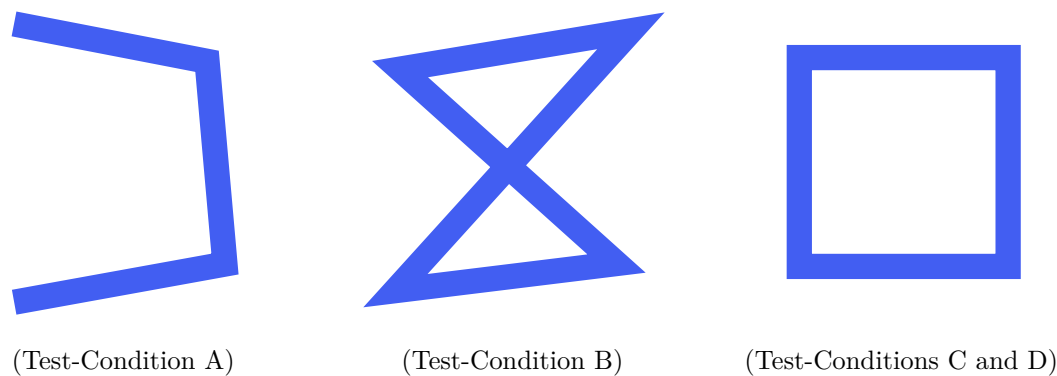


Figure 3.5: Different paths for the initial proficiency drawing task

For condition A the path was set deliberately to be incomplete and with offsets to the edges, resulting the robot in drawing an open imperfect shape. For condition B the order of the edges of the drawing path was swapped with also having offsets to the edges as well, resulting the robot in drawing more of an hourglass shape. With conditions C and D, the robot was given a path for drawing a perfect enclosed rectangular shape.

3.1.6 Resources

For the study to be conducted, the following was necessary: A computer capable of running Unity 2022.3.21f1, a Meta Quest 2 Head-mounted Display (HMD) with one default hand-held controller, an audio recorder for recording interviews, a notebook for letting participants fill in the questionnaires digitally and a lab room with enough space

for participants to do the necessary interactions within the VR environment in a standing position, requiring about $4m^2$ of room without any obstacles.

3.1.7 Participants

Due to the Latin square balanced condition sequences for the user-tests (see section 3.1.2), the study needed to have a multiple of four participants, and aimed for at least 20. People of all ages, genders, and ethnic groups were welcome to join the study. Each of them would take about one hour in the lab to complete the experiment, including briefing, the four individual user tests, questionnaires, and an interview. People who had no experience with VR before were expected to need a bit more time to get used to it, which needed to be accounted for while planning the time slots. Furthermore, short-sighted people with bigger glasses were not able to join the study, as they would not fit into the relatively small interface gap inside of the HMD.

3.1.8 Measures

For this study, the required measures have been made over multiple sources, which consisted of: an onboarding questionnaire, a robot assessment questionnaire, and event-logging with timestamps from each of the user tests and interviews. Each of these sources will be described in more detail below.

Onboarding questionnaire: Participants filled in this questionnaire right after the briefing and right before doing the user tests. The questionnaire measured:

- Age: text field
- Gender: multiple choice with an optional text field.
- Highest level of education: single choice with an optional text field.
- Employment status: single choice with an optional text field.
- Experience with teaching: text field.
- Technical skills in terms of machine learning: 1 to 5 Likert-scale.
- Education related to robotics, computer science, or similar: text field.
- Experience with VR: single choice.
- Comfort with new technologies: 1 to 5 Likert-scale
- Interest about humanoid robots: 1 to 5 Likert-scale
- Personality traits according to the Ten-Item Personality Inventory (TIPI) [82].

These measurements have been made to be able to provide overall demographic information about the people who participated in the experiment as well as for potentially being able to reason about certain study results, if applicable.

Robot assessment questionnaire: This questionnaire was filled in by participants right after completing each of the four individual user tests and measured:

- Participants' impressions of the robot, including anthropomorphism, animacy, likability, perceived intelligence, and perceived safety, according to the Godspeed series [83].
- Self-efficacy of teaching by using a subset of the Self-efficacy in Human-Robot Interaction Questionnaire (SE-HRI) [84].
- User-Experience about teaching the robot by using a User Experience Questionnaire (short version) (UEQ-S) set of questions [85].
- Motivation to continue teaching the robot: 1 to 7 Likert-scale.
- Reason that made the participant stop teaching (qualitative): Free text input.

The measures of the Godspeed series, SE-HRI and UEQ-S were included to reliably being able to gather data mainly concerned about RQ₁₋₃ while the last two items were created manually and provide information regarding RQ₄, which is about teaching motivation.

The task for the participants was to teach the robots how to do the pick-and-place task, as much as they wanted to, so they were able to break up each of the four teaching sessions without having to teach until the robot accomplished the task. This was one of the key mechanics to be able to check whether different configurations of the independent variables made a difference for example in the participants' motivation to continue teaching.

Event logging & screen-recording During each of the user tests, certain events have been logged aiming to gather vital information about the participants' behaviors. Each of the events had a timestamp attached to it, to be able to reconstruct the timeline, if needed, and, which is of more importance, to automatically extract or summarize temporal data. The following measures were extracted from event log files:

- Time needed from beginning to end of the test.
- Time the participant took to observe the robot doing its initial task.
- Sum of time the participant spent interacting with the cubes.
- Mean time of each demonstration within the test the participant spent interacting with the cubes.

3. APPROACH

- Sum of demonstrations given to the robot.
- Final robot task proficiency at the end of the test.

These measures were made for each individual test case A, B, C, and D. Additionally all of these were summarized over all of the test cases as well, e.g. the sum of times needed from beginning to end for all of the test cases and so forth. Please refer to section 3.2.5 for more detailed information on how this was done.

As an addition to logging, screen recordings of each of the user tests were made as a sort of backup of a trustworthy source for the data. This was helpful in one or two cases, where the logging contained a few more events than it should have due to unforeseen actions of participants inside the VR environment. The data itself was consistent, but the additional entries caused problems with the automatic data extraction script. With the help of the recordings it was simple to repair the affected log files.

Interviews: Each of the participants took part in a semi-structured debriefing interview after all of the test conditions had been completed. It was mainly concerned about RQ₄ and included the following open questions:

- Q₁. Please describe your overall experience with teaching all the different robots.
- Q₂. When seeing the robot's initial skill level, which means how well it did in drawing the rectangle. Did this fact affect your willingness to teach the robot with the new task? (If yes: In which way?)
- Q₃. When seeing how slow or fast the robot's learning capabilities have been. Did this rate of learning affect your willingness to teach the robot with the new task? (If yes: In which way?)
- Q₄. Considering only those robots that had a bad initial skill level while drawing the rectangle.
 - A. Did your motivation to continue teaching change when you saw how fast the robot was learning the new task with the fast learning robot? Were you positively surprised by how fast it was?
 - B. Did your motivation to continue teaching change when you saw how slow the robot was learning the new task with the slow learning robot? Did you expect this slow rate of learning?
- Q₅. Considering only those robots that had a good initial skill level while drawing the rectangle.
 - A. Did your motivation to continue teaching change when you saw how fast the robot was learning the new task with the fast learning robot? Did you expect this fast rate of learning?

B. Did your motivation to continue teaching change when you saw how slow the robot was learning the new task with the slow learning robot? Were you negatively surprised by how slow it was?

Q₆. Which robot did you find the most demanding in terms of your engagement and patience and why? Discuss each of the robots and ask why it is in this place on the table.

Finally, participants were asked if they wanted to add anything else to the discussion that had not been mentioned before. The time consumption of single interviews was aimed at about ten minutes. The last three questions Q₄₋₆ were special in a sense because they only have been asked to the participants in conjunction with the data that was gathered from the very last portion of the robot assessment questionnaire which was about the participant's self-rated motivation to continue teaching the robot (Likert-scale from 1 to 7). All four results regarding this part were manually noted down in a table by me as the user tests were conducted. So, by the time the interview started I had a filled table with all the motivation rankings the participants provided for all the test cases A, B, C, and D. With this data the questions Q₄₋₆ were used to ask participants why they provided certain ratings they way they did. For example, if the teaching motivation rating for condition D was much lower than for condition B (both were fast learning), then these questions were used. On the other hand, if the ratings were clear from the beginning, then the questions were either completely omitted or shortened to some extent.

3.1.9 Expected Output of the Study

An overall better rating of the perceived robot likability, intelligence, and safety along with higher ratings regarding participant motivation and self-efficacy for user tests with fast learning rate and high initial proficiency have been anticipated. The results of the study and their interpretation may influence or inform certain design choices regarding HRI research as described within section 1.5.

3.2 Implementation

The framework for conducting the user studies was provided by TU Wien as a foundation. All the specific requirements that were needed in addition were built on top of that framework. It provided the following features at the time of implementation: a humanoid robot model with basic movement animations (idle state and hand waving gesture), a table model that provided a surface for object manipulation testing, two pick-and-place example scenarios including cube objects ready for manipulation and finally a VR headset integration provided by various official Unity XR plugins. The framework was built on Unity 2022.3.21f1 which also then has been used to implement extending features required for the user-study in this work. All of the features which had been implemented in addition to this work are described in detail below.

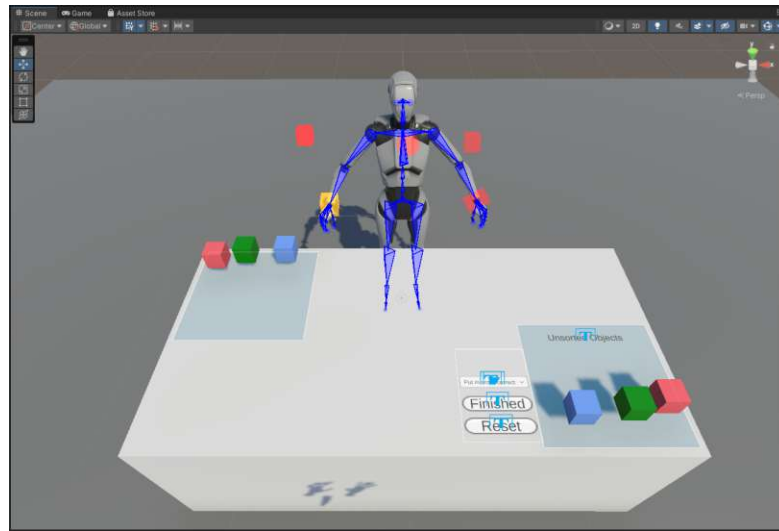


Figure 3.6: Humanoid robot sample scene provided by TU Wien HRI-testing framework

3.2.1 Teaching Environment

A basic room, about the size of $14m \times 6m \times 24m$ (W x H x D) was designed around the table. The table itself also has been re-designed and had about the size of $1.4m \times 0.85m \times 1.2m$ (W x H x D). For the pick-and-place demonstration mechanic, four individual cube areas have been put on top of the table as seen in figure 3.7. Both of the areas with the name ‘Unordered Objects’ represented the space in which the cubes would appear when a reset is required. The other two areas were for task validation which is described in more detail in section 3.2.3. Finally, a whiteboard was added to the scene which was able to show in which order the cubes needed to be put down into the target area for the teacher to have

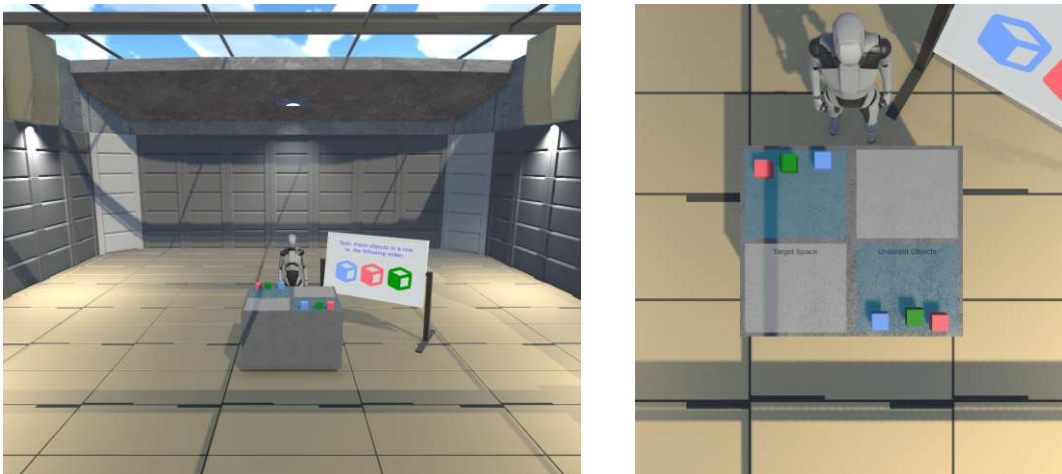


Figure 3.7: Teaching environment

a valid placement. Both, the room and the whiteboard had been modeled in Blender.

3.2.2 The Robot Apprentice

The humanoid robot model was imported from the Unity Starter Assets Package provided by Unity Technologies. It comes out of the box with basic animations, materials, textures, and sound effects, whereas the latter was not utilized within the framework or the extensions of this work (user tests were conducted without in-application sound effects or music).

Movement

The TU Wien HRI-testing framework provided a test scene with the robot model in the center standing behind a table (see figure 3.6), with the robot model having a hand waving gesture and an already default attached idle animation (slight natural body movements). Additionally, the model within this scene has inverse kinematic constraints attached to its arms and head. For these constraints, there was also already an attached empty target object (i.e. without a rendered shape). That way the hands, arms, and head of the robot could simply be moved programmatically or within the Unity editor by moving their respective target objects to the desired location.

Robot Initial Proficiency

This property was shown over a drawing task according to section 3.1.3. In Unity, a simple rectangle mesh, containing exactly four vertices was added to the scene. Within a script, these vertices were able to be retrieved over a simple function. Another script that managed the drawing task then was able to retrieve those vertices, and make changes to or reorder them depending on the configuration of the learning rate and initial proficiency configurations. Then the list of vertices is traversed. The target object controlling the kinematics of the right robot arm and the target object controlling the gaze direction of the head were then both transformed towards the first vertex in the list over a fixed period of time to produce an actual motion effect. Then the same movement mechanic is done towards the next vertex in the list. If two vertices have been traversed by the robotic hand, a line renderer was used to compute a 2D line between them. In this way, the impression occurs that the robot is actually drawing a rectangle (see figure 3.10). When the user test launched, participants were able to observe the robot doing this task. In this stage of the test, only if they were looking at the robot for at least ten seconds, the ‘Switch Task’ UI button appeared floating above the table. If participants pressed this button, the environment switched to the pick-and-place task.

Errors: As mentioned earlier the robot was also able to draw different shapes to convey to the teachers that the robot made a mistake or that it did not learn how to properly draw the rectangle. Those shapes were only being drawn by the robot if the initial proficiency was set to low, i.e. only with test conditions A and B. Participants were told

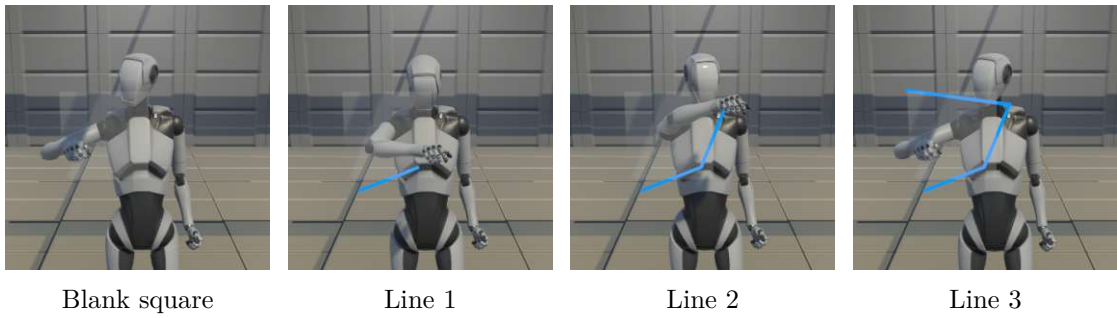


Figure 3.8: Test-Condition A: Robot missing a side of the rectangle, edges are offset

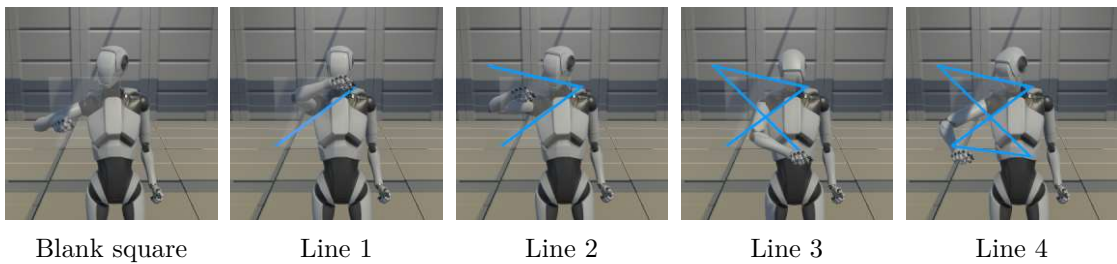


Figure 3.9: Test-Condition B: Robot drawing hourglass instead of rectangle

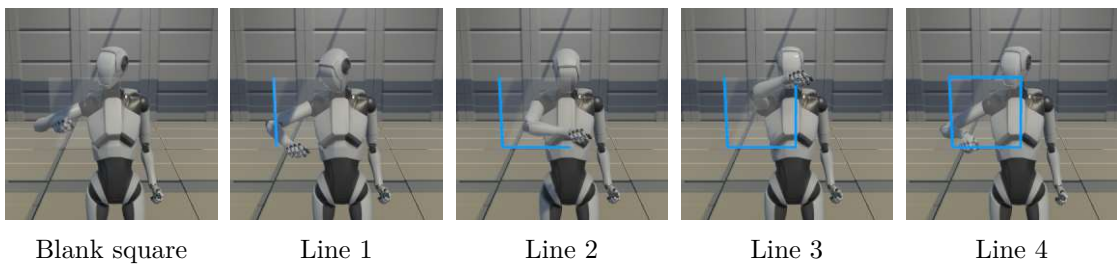


Figure 3.10: Test-Conditions C and D: Robot drawing a correct rectangle

that each of the test cases had differently trained robots, so in order to not confuse them by showing the same erroneous rectangle twice, two different shapes were used (see figures 3.8 and 3.9), like described in section 3.1.5.

Robot Learning Rate

The robot learning rate defined how big of a step the robot task proficiency p increased after each demonstration had been made by the human teacher. For each user test it was initialized with $p = 0.0$. If a teacher provided a valid demonstration, the value has been increased either by 0.11, if the learning rate was set to slow, or 0.34, if the learning rate was set to fast. The maximum value for p was set to 1.0 representing 100% proficiency. If the value reaches this maximum, the robot will not make any more mistakes.

Algorithm 3.1: Demonstration task: Main loop

```

Data: Global  $LR$ , Global  $E$ 
//  $LR$  - Learning rate setting:  $LR \in \{0, 1\}$ 
//  $E$  - List of possible errors:  $E = \{0, 1, 2\}$ 
//  $c$  - Defines whether to continue or not:  $c \in \{\perp, \top\}$ 
//  $p$  - Current robot task proficiency:  $p \in [0, 1], p \in \mathbb{R}$ 
//  $L$  - List of target locations for each cube
1  $c \leftarrow \top$ ;
2  $p \leftarrow 0$ ;
3  $i \leftarrow 0$ ;
  // Start main loop
4 do
5    $e \leftarrow$  Call Next-Error( $p$ );
6   Call Reset-Cubes-Location();
  /* Waits until the teacher places all the cubes and hits
     the finish button, returns target locations for each
     cube already prepared for the robot target area. */
7    $L \leftarrow$  Call Wait-For-Demonstration-Finished();
8   if  $LR = 0$  then
9     |  $p \leftarrow p + 0.11$ ;
10  else
11    |  $p \leftarrow p + 0.34$ ;
12  end
13  if  $p > 1.0$  then
14    |  $p \leftarrow 1.0$ ;
15  end
16  Call Robot-Execute-Task( $L, e, p$ );
17   $i \leftarrow i + 1$ ;
18  if  $i > 3$  then
19    |  $c \leftarrow$  Call Show-Feedback-Dialog();
20  end
21 while  $c = \top$ ;

```

Explained in more detail, algorithm 3.1 shows a simplified version of the demonstration-task loop. First, the next error is picked from the list of possible errors which is based on the current robot task proficiency value. More details about errors are provided in the upcoming paragraph. Next, the code moves all of the cubes back to their original locations and into their respective pickup areas. Then it waits for the teacher to complete with a demonstration. *Wait-For-Demonstration-Finished()* internally then fetches where in the teacher's target area the cubes have been placed, and then calculates and returns the coordinates where the robot should place its cubes into its target-area, for it to be a correct reproduction of the demonstration. Also after the teacher finished a demonstration the task-proficiency increased, with a step size depending on the current learning rate

Algorithm 3.2: Function: Robot-Execute-Task

```
Data: Global  $C$  – List of cube objects
Input:  $L$  – List of target locations for each cube, i.e. 3D-coordinates
Input:  $e$  – Error type:  $e = -1$  if no error should be made
Input:  $p$  – Current robot task proficiency:  $p \in [0, 1], p \in \mathbb{R}$ 
//  $L_i$  – List item of  $L$ :  $L_i = (x, y, z)$  where  $x, y, z \in \mathbb{R}$ 
//  $C_i$  – List item of  $C$ : Single cube at index  $i$ 
// Start function
1 assert  $|C| = |L|$ ;
2 if  $e = 0$  then
    /* Picks two random target coordinates and swaps them.
       This way the cubes are in the wrong order after the
       robot places them into the target area. */
3    $L \leftarrow$  Call Randomly-Swap-Two( $L$ );
    /* Randomly picks one of the target coordinates and
       applies an offset to the z-axis. This results in one
       cube being put too far away from the other ones. The
       magnitude of the offset is dependent on  $p$ . The
       higher  $p$  is, the smaller the degree of misplacement.
       */
4    $L \leftarrow$  Call Randomly-Misplace-Single( $L, p$ );
5 end
6 if  $e = 1$  then
7    $L \leftarrow$  Call Randomly-Swap-Two( $L$ );
8 end
9 if  $e = 2$  then
10   $L \leftarrow$  Call Randomly-Misplace-Single( $L, p$ );
11 end
    /* Loops over the individual cubes and makes the robot
       pick cube  $C_i$  and place it to the target-coordinate  $L_i$  */
12 for  $i \leftarrow 0$  to  $|L|$  do
13  | Call Robot-Pick-And-Place-Cube( $C_i, L_i$ );
14 end
```

setting. Algorithm 3.2 then takes the error type from before as input and handles it accordingly. It does so by swapping or manipulating the target coordinates if needed. Then, the robot model will be told to pick and place each of the individual cubes to their respective target locations. Finally, the teacher will be asked whether the teaching session should continue or not over *Show-Feedback-Dialog()*. Please refer to section 3.2.4 for more detailed information.

Errors The robot has been making mistakes according to the definitions in table 3.3 and figure 3.4, if the current robot task proficiency was $p < 1.0$. Algorithm 3.3 shows how the error type was picked from the list of possible errors $E = \{0, 1, 2\}$ which depended on p . Each of the elements of E represents an error type: $0 \rightarrow M_{1+2}$, $1 \rightarrow M_1$ and $2 \rightarrow M_2$.

Algorithm 3.3: Function: Next-Error

Data: Global E – List of possible errors (integer-encoded): $E = \{0, 1, 2\}$
Input: p – Current robot task proficiency: $p \in [0, 1], p \in \mathbb{R}$
Output: e – Integer-encoded error type: $e \in E \cup \{-1\}$
 // $e = -1 \rightarrow$ no error
 // Start function
 1 **if** $p \geq 1$ **then**
 2 | **return** -1 ;
 3 **end**
 4 $i \leftarrow \lfloor |E| \cdot p \rfloor$;
 5 $e \leftarrow E_i$;
 6 **return** e ;

3.2.3 The Human Instructor

Participants were asked to provide demonstrations to the robot within the user tests. A Meta Quest 2 HMD together with its right hand-held controller was used to enable them to do so within a VR environment. Usually, no special locomotion technique was needed as they only would generally have to stand in a specific area within the scene to be able to do the task. Also, the physical space within the lab room was enough for them to adjust to smaller spatial corrections if needed. In rare cases, when the mapped location from the VR headset was off a few meters within the VR environment, I made a live location reset of the participant within the scene through the unity editor.

User Input

The participants were able to do all the necessary actions over the hand-held controller's trigger and grip button. The first one allowed them to press User Interface (UI) buttons while the latter one enabled the pick-and-place mechanic for objects. In order to pick up an object, participants simply had to target a draggable one (which was only cubes

3. APPROACH

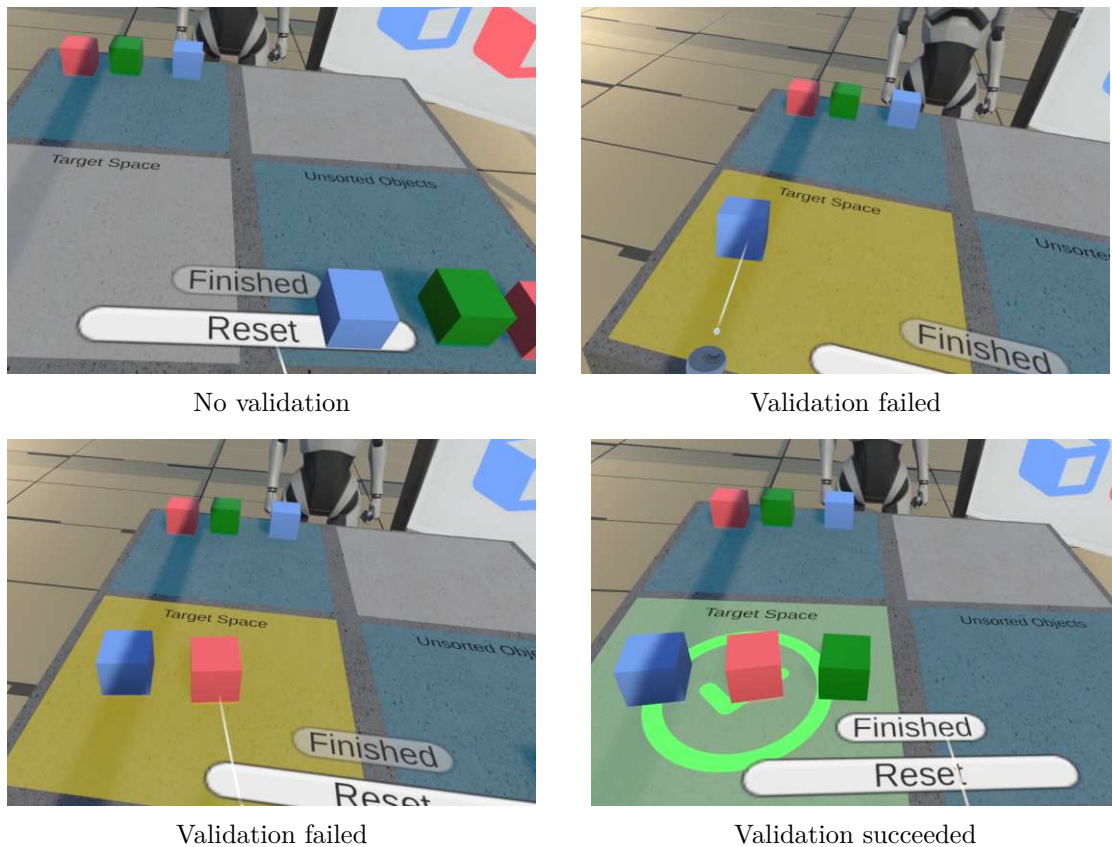


Figure 3.11: Individual cube placement and validation of demonstration

in this case) and then press and hold the grip button of the controller. While holding the object, participants were able to relocate it to their liking. Once they let go of the button, the object would simply drop from where it has been relocated to.

Task Demonstrations and Validation

Demonstrations were provided by teachers by moving each of the cubes from the pickup area into the target area. The final row order of the placed cubes had to match with the description on the whiteboard. Only if the demonstration was valid, the teacher was able to pass it to the robot by pressing the 'Finished' UI button. Algorithm 3.4 shows in a simplified version how the cube order was validated. It assumes to have an already sorted list of the locations of the teacher's cubes as input, i.e. the cube which should be the outer left according to the task description on the whiteboard is the first one in the list, the one which should be in the middle is the second and the one that should be right is the last. The list of cube coordinates is traversed and various conditions are checked, such as whether each of the cubes is inside of the target area, whether they are in a row formation, or whether they are in close enough proximity. Figure 3.11 presents

a sequence of cube placements and their validation states. If the target area was colored gray, this meant that no validation had been made so far, an orange color meant that the validation failed and if the area was rendered green with a big tick symbol in the center, this meant that the validation succeeded.

Algorithm 3.4: Validation of cube placement

Data: Global c_s – Side length of cube objects: $c_s \in \mathbb{R}$
Input: V – List of teacher’s cube locations, sorted according to the target order
Output: v – Placement of cubes valid: $v \in \{\top, \perp\}$
 // V_i – List item of V : $V_i = (x, y, z)$ where $x, y, z \in \mathbb{R}$
 // Start function

```

1  $\bar{z} \leftarrow \frac{1}{|V|} \sum_{i=1}^{|V|} V_{i,z};$ 
2  $v \leftarrow \top;$ 
3 for  $i \leftarrow 1$  to  $|V|$  do
4   if  $\neg \text{Call } \textit{Is-Cube-In-Target-Area}(V_i)$  then
5     |  $v \leftarrow \perp;$ 
6   end
7   /* Checks if the order of placement is correct */
8   if  $V_{i+1} \neq \textit{None} \wedge V_{i+1,x} < V_{i,x}$  then
9     |  $v \leftarrow \perp;$ 
10  end
11  /* Checks if cubes are in proximity (i.e. space between
12  them does not exceed 7.5 cm) */
13  if  $V_{i+1} \neq \textit{None} \wedge |V_{i,x} + \frac{c_s}{2} - (V_{i+1,x} - \frac{c_s}{2})| > 0.075$  then
14    |  $v \leftarrow \perp;$ 
15  end
16  /* Checks if cubes are placed within a row on the x-axis
17  (i.e. from left to right from the teacher’s point of
18  view) */
19  if  $|\bar{z} - V_{i,z}| > \frac{c_s}{3}$  then
20    |  $v \leftarrow \perp;$ 
21  end
22  /* Checks if cubes are not stacked on top of each other
23  in some way */
24  if  $V_{i+1} \neq \textit{None} \wedge (V_{i+1,y} < V_{i,y} - V_{i,y} \cdot 0.05 \vee V_{i+1,y} > V_{i,y} + V_{i,y} \cdot 0.05)$  then
25    |  $v \leftarrow \perp;$ 
26  end
27 end
28 return  $v$ 

```

3.2.4 Feedback Dialog

After having completed a fixed set of four demonstrations and when the robot finished placing the cubes, a dialog appeared, as seen in figure 3.12, in front of the participants, asking them whether or not they want to continue teaching the robot or not. After the ‘Continue’ UI button was clicked once, the dialog appeared after each successive demonstration. This mechanic let participants discontinue their efforts more confidently for each of the different test conditions.



Figure 3.12: Feedback dialog asking the participant to continue or stop with teaching

3.2.5 Logging of Events

In addition to questionnaires, the logging mechanic during the user tests has been one of the main sources for the quantitative data analysis, as already mentioned in section 3.1.8. The following events have been logged during each of the user-tests:

- Application events:
 - Application started.
 - Robot initial proficiency and learning rate settings on application start.
 - Application ended.
- Demonstration events:
 - Increased robot current task proficiency.
 - Robot finished with task execution.
 - Next robot mistake type chosen (i.e. M_1 , M_2 or M_{1+2}).
- UI events:

- ‘Finished’ button pressed.
- ‘Reset’ cubes button pressed.
- ‘Switch task’ button pressed.
- ‘Continue’ feedback dialog button pressed.
- ‘Exit’ feedback dialog button pressed.

The respective log files were created at the application start if needed. The naming followed the format: `Log_p<participant-ID>_<yyyy-MM-dd>.txt`, where the current participant ID was taken from a public variable editable within the unity editor, and the current date is appended, e.g. as such: `Log_p01_2024-07-03.txt`. If a file with the same name already existed (i.e. for the same participant and date), then all of the logs during user tests were just appended to this file, otherwise, it has been created. Each of the log entries was appended to the file over an internal logging object, which contained the properties: event-type (which referred to one of the items in the listing above, encoded as a number), timestamp (date and time of the event) and message (which contained human-readable text and sometimes vital payload data). These logging objects were transformed into a JSON format before they were finally written into the file. This was done to simplify the later data processing of these files in the evaluation phase of the study, which is described in section 3.3.3.

3.3 Evaluation

3.3.1 Participants and Time Slots

Participants were recruited based on availability, accessibility, and mailing lists. Once the communication channel to potential participants has been established, further details were provided to them by sending a *Termino* link to them. *Termino* is a publicly available online platform designed for appointment coordination with a special interest in data protection and GDPR compliance. On the website, potential participants were able to read a short description of what the study was about, where the lab room was located, and how long it would take them to participate. Time slots were provided to them, giving them options usually for the current or also even for the following week. If they found a fitting time slot, they were able to book it by providing an e-mail address.

Each week had up to about ten time slots, usually targeted at afternoons or early evenings, scattered over the weekdays from Monday to Saturday. Occasionally participants reached out to negotiate special time slots when there was none provided that would fit into their schedule. Time slots were scheduled with 40 minutes buffer time between them, in cases where they have been added back-to-back. Over roughly two months (July and August 2024), a total of 24 participants, 11 male and 13 female (self-identified), had volunteered to take part.

3.3.2 Lab Experiment

All of the participants taking part in the experiment went through three stages, consisting of onboarding, user tests, and interviews. Table 3.4 shows an outline of these stages, how much estimated time they needed, and which data was collected.

Lab experiment			
Stage		Estimated time needed	Data collection
Onboarding	Welcome & Consent Form	~ 3 min	-
	Onboarding Questionnaire	~ 5 min	Quantitative: General demographic data, TIPI [82]
	Briefing	~ 2 min	-
User Tests		~ 30 min	Quantitative: Logged event data with timestamps. Data from robot assessment questionnaire including Godspeed [83], UEQ-S [85] and a subset of SE-HRI [84] Qualitative: Open question from the concluding part of the robot assessment questionnaire
Debriefing Interview		~ 10 min	Qualitative: Data on participant's motivation about teaching the robot

Table 3.4: Lab experiment summary: Estimated time and data collection

Onboarding

The onboarding stage of the experiment included the welcoming of a participant, an introduction to the project, a consent form for them to sign, a questionnaire for them to fill out, and finally a briefing that explained what they were being asked to do within the experiment.

Consent form: Once participants got to the lab room within their respective time slots, they first have been asked to sign a consent form and with it give the necessary permissions to collect, store, and process all of the relevant data for the study, including photos, video and audio recordings during the individual parts of the experiment. Although all participants signed this document, if pictures or videos were being made of them, for example during the user tests, they were asked again for their permission, since this data was not mandatory for data analysis. Additionally, the document contained a short project description and informed the participants that they were able to speak freely during the experiment and that they were able to leave it at any time without consequences.

Onboarding questionnaire: This questionnaire gathered demographic data and asked participants about certain personality traits. Participants were aged between 22 and 40 ($M = 29.54, SD = 4.23$). When asked about their highest level of education, participants responded with ‘Bachelor’s degree’ (9), ‘Master’s degree’ (8), ‘Matura (Austria) or high school’ (6) and ‘Dipl. Ing. (FH)’ (1). Their current state of employment were ‘employed part-time’ (9), ‘employed full-time’ (4), ‘employed part-time & student’ (4), ‘self-employed’ (3), ‘student’ (2) and ‘unemployed’ (2). When asked about whether they have experience with teaching, participants responded with having experience as a tutor or school teacher (11), no experience (9), experience with teaching children outside of school (3), and corporate training (1). Participants were asked on a Likert scale from 1 (novice) to 5 (expert) about their technical skills in terms of machine learning or artificial intelligence ($M = 2.54, SD = 1.25$). When asked whether they had any kind of formal education about computer science or robotics, participants responded with currently studying or already having a university degree in computer science (15), having basic or advanced education in software developing (3), and, no formal education (6). Participants had minimal (17), moderate (4), extensive (1), or no (2) experience with VR. Further, they were asked about how comfortable they are with new technologies on a Likert scale from 1 (disagree strongly) to 5 (agree strongly) ($M = 4.25, SD = 0.85$) and on the same scale from 1 (no interest) to 5 (high interest), they have been asked how interested they were about humanoid or collaborative robots ($M = 3.67, SD = 0.87$).

Additionally, as mentioned in section 3.1.8, participants have filled in the TIPI. Results are provided in the next chapter in section 4.1.

Briefing During the briefing, participants were informed that the study was about certain aspects of robots learning from demonstration together with a short explanation about what that means, and that the tests would be conducted within a VR application. They have been told that they would try to teach a certain pre-defined task to a humanoid robot within said application. Subsequently, it was communicated that the robot is learning from them with the help of some ML algorithm which would run in the background during the process of teaching. The initial proficiency task was brought to their attention, by mentioning that the robot would demonstrate to them what it learned on another task. Afterward, the actual task, that they should teach the robot,

was mentioned to them, and that it would be necessary to provide correct demonstrations to have it learn. Additionally, they were told that they would face the robot four times altogether with each time having a different configuration of the ML algorithm behind the scenes, and that they should be aware of having to provide different numbers of demonstrations for each of the individual tests to have the robot learn how to successfully execute the task. To avoid confusion, they were notified that each of the individual tests would start afresh, essentially saying that the results or learning outcomes were not dependent on each other. Finally, they were told that after some time there would be a dialog appearing in front of them, asking them whether or not they wanted to continue teaching or not and that they were free to decide on their own about which option to choose, and if they decided to discontinue the process, that any possible reason would be a valid one.

Questions from participants regarding the information above were answered during or after the briefing if needed. Any questions regarding study details, which occasionally arose due to the demographics of the participants were deferred to the very end of the experiment. Participants were asked about their experience with VR and were given a short introduction on how to attach the HMD and how to use the hand-held controller.

Tutorial

Once participants knew how to put on the HMD they were first confronted with a tutorial within the VR application. They have been put in front of a table with the humanoid robot standing behind it, just like with the actual pick-and-place teaching scenario, but without any cubes or the usual UI. A dialog guided them through the process step by step. Figure 3.13 shows example screenshots of two different stages of the tutorial.

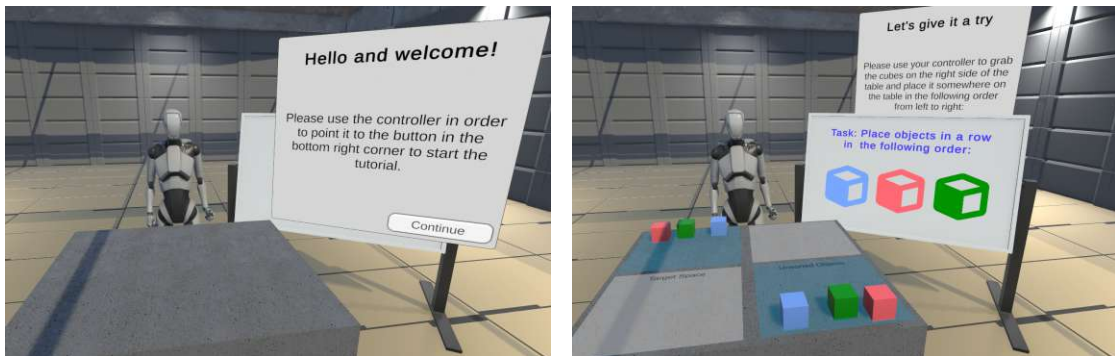


Figure 3.13: Example Screenshots from the tutorial mode

User Tests

After the participants have finished the tutorial, they have been asked whether there are any remaining questions concerning the user input or the pick-and-place task. Once everything was clear, the user tests were started with the first set of conditions according

to table 3.2. Just as with the tutorial, participants started off standing behind the empty table. On rare occasions the mapped location of the VR headset in the virtual environment had a few meters offset on the x-axis or z-axis, i.e. participants were too far off from the table and would not be able to do the test in such cases successfully. Since this bug was known beforehand (see section 3.2.3), a location reset option was built into the unity-editor, only visible to me, which relocated participants back to the default location, if needed.

In the default case, or after participants got re-located, the robot started with the rectangle drawing task, as described within sections 3.1.5, 3.2.2 and as shown in figure 3.10. In order to continue, participants had to spend some time looking into the direction of the robot, essentially observing it doing its drawing task. After some time a UI button appeared in front of the participants and they were able to switch to the pick-and-place task. When this action was triggered, the interactable cubes, UI elements, and all of the four picking and placing areas were made visible on top of the table and participants were able to start providing demonstrations as described in sections 3.1.4 and 3.2.2. Participants spent the majority of the time within the VR application within this stage of the user tests.

During this time, usually, they straightforwardly provided demonstrations, i.e. by randomly picking and placing the cubes into the target area until the task was completed. Participants tended to misplace the cubes within their first few tries, violating one or more of the validation rules described in algorithm 3.4. In this case, participants were not able to finish their demonstrations. The most common error at that point was that cubes had been placed too far from each other on either the x-axis or the z-axis in Unity's coordinate system. The second most common error was that cubes have been placed in the wrong order, which happened in later demonstrations or user tests more often than the first error. Participants sometimes asked what to do in such situations, or figured it out by themselves.

Some of the participants tried different teaching strategies, including demonstrations with slower or more precise movements, demonstrations with specific placement orders of the cubes, or introducing very specific or seemingly easy-to-get motions for the robot to observe. These were partly observed during the user tests or in their respective screen recordings. Participants sometimes mentioned how they tried different strategies in the later interviews.

After participants provided four valid demonstrations, they were asked whether they wanted to continue with teaching or not according to section 3.2.4 and as shown in figure 3.12. When this dialog appeared the first time, it is worth mentioning that the robot already learned how to do the task successfully in test cases B and D, as the robot was configured with a high learning rate (see table 3.1). Participants have not been encouraged to make a choice by this dialog to either continue or discontinue with the teaching process before the fourth demonstration has been provided to them. Due to the faster learning process of the robot, participants stopped earlier in test cases B and D than with test cases A and C, where the robot was configured to be slow learning.

If participants decided to discontinue the teaching process within a given user test, they were asked to take off the HMD to fill out the robot assessment questionnaire as described in section 3.1.8. Once this was completed as well, participants repeated the whole procedure of teaching the robot within the VR application and filling out the questionnaire for the remaining test conditions. Each time that one of the four questionnaires was completed, the results of one of the questions were noted down in a table which in the end represented the participant's self-rated motivation scores for each of the test cases on a Likert scale from 1 to 7. This served as a foundation for some of the questions for the finalizing interview as described in section 3.1.8.

Debriefing Interview

After all of the user tests and their respective questionnaires were completed, participants were asked to take part in a short finalizing and audio-recorded interview as described in section 3.1.8. In summary, Q₁ served as an entry question and participants often already had something on their mind that they wanted to share. Q₂ and Q₃ were specifically asking about whether they were affected by the robot's initial proficiency and learning rate. Especially with the robot's initial proficiency drawing task, participants felt that this did not impact them in terms of their motivation, however, sometimes these explanations did not fit with the ratings from the table which was filled with their self-rated motivation results from the robot-assessment questionnaires. In this case, or if the numbers within the table seemed to be unclear in terms of consistency, Q₄₋₆ were additionally brought into the interview. As with the learning rate of the robot, a majority of participants reported a feeling that their motivation was inhibited by a slow learning configuration compared to a fast learning configuration.

3.3.3 Data Analysis

Quantitative Data

Since both, the onboarding and robot assessment questionnaires have been filled out over Google Forms, the resulting data was made available and downloaded as comma-separated plain text data (CSV) by the online tool, which is available out of the box. For each participant, there was one file for the onboarding and four (i.e. one for each of the test cases) for the robot assessment questionnaire results. All five files were merged into a single one which contained all of the questions within the columns, whereas the robot assessment questions were suffixed with their respective test-case names `_A`, `_B`, `_C`, or `_D`. This was done to have the results for each of the participants in a single row, making data easily accessible for statistical analysis.

Further, the response values of participants from the individual questions from the TIPI, Godspeed, SE-HRI, and UEQ-S items have been summarized into values that represented their scales. These scales namely have been *Extraversion*, *Agreeableness*, *Conscientiousness*, *Emotional stability*, and *Openness to experiences*, for the TIPI part of the questionnaire. According to Gosling et al. [82], for every scale, the response values

have first been flipped for reverse-scored items, and then summarized by forming the mean of the values for every item of the scale. The same procedure was made for the Godspeed, SE-HRI and UEQ-S scales. Additionally, as for these scales, values existed for each item and each of the test cases A, B, C, and D, the summarizing value for the scales were again suffixed with `_A`, `_B`, `_C`, and `_D`. Thus, the Godspeed items were summarized into *Anthropomorphism_A/B/C/D*, *Animacy_A/B/C/D*, *Likeability_A/B/C/D*, *Perceived_intelligence_A/B/C/D* and *Perceived_safety_A/B/C/D* [83]. The subset of the SE-HRI items were summarized into a single scale *Self-efficacy_A/B/C/D* [84] and the items of the UEQ-S were summarized into the two scales *Pragmatic_quality_A/B/C/D* and *Hedonic_quality_A/B/C/D* [85].

Additionally, the event logs from each of the participants were processed with the help of a custom Python script. All of the log files (one per participant) were put into a folder out of which the script would loop over and read them one after another to process the contents. Each of the calculated parameters was again labeled with a suffix to refer to the test case it corresponds to. Again, the used suffixes were: `_A`, `_B`, `_C`, and `_D`, representing the test cases, and `_ALL` which represents the sum or average value of all of the test cases. The following data was calculated:

- ‘`time_seconds_A/B/C/D/ALL`’: How long did the test take in seconds, measured from application start and end events (see section 3.2.5).
- ‘`time_observing_initial_task_seconds_A/B/C/D/ALL`’: Represents how much time a participant spent observing the robots initial drawing task in seconds. The application start and ‘Switch task’ UI button events were required for this measurement.
- ‘`time_interacting_seconds_A/B/C/D/ALL`’: Represents how much time participants spent interacting with the cubes during the pick-and-place task, or, the time devoted to providing demonstrations to the robot. The three ‘Switch task’, ‘Finished’, and ‘Reset’ cubes UI button events were needed for the calculation.
- ‘`interacting_repetitions_A/B/C/D/ALL`’: Represents how many demonstrations the participants provided to the robot before they discontinued the teaching process. The number of repetitions is the same as the number of appearances of the ‘Finished’ UI button event.
- ‘`mean_time_per_interaction_A/B/C/D/ALL`’: Represents the average time in seconds that a participant needed to finish with a demonstration. This resulted by the division of ‘`time_interacting_seconds_A/B/C/D/ALL`’ by the corresponding ‘`interaction_repetitions_A/B/C/D/ALL`’.
- ‘`robot_proficiency_score_A/B/C/D`’: Represents the proficiency value that the robot had by the end of each user test. This was simply done by reading the payload of the *increased robot current task proficiency* event.

- ‘mean_robot_proficiency_score_ALL’: Represents the average of the previous values, i.e. the sum of the individual proficiencies from ‘robot_proficiency_score_A/B/C/D’ divided by the number of test cases.

All of the results were written into a separate CSV file which followed the same structure as the one containing the questionnaire results. Again, the already merged file was merged with the file which resulted from the event data processing, resulting in a single file with all of the quantitative results gathered throughout the experiment.

Parameter selection for statistics: Nearly all of the parameters above have been consulted for statistical data analysis, except for ‘time_seconds_A/B/C/D/ALL’, ‘mean_time_per_interaction_A/B/C/D/ALL’ and ‘mean_robot_proficiency_score_ALL’. Those parameters have been dropped because they already are indirectly represented within other parameters.

Statistical analysis: The upcoming chapter first shows statistical results from the quantitatively measured and pre-processed data (some items were summarized into representative scales like described above). Except for the TIPI scales, from all the other measurements of the selected parameters, a barplot which shows mean values and standard errors, a boxplot, and a violin plot, which shows individual and combined mean values were made for visual representation. These plots were grouped by the two independent variables learning rate (grouped on x-axes) and learning rate (grouped by color). Additionally, descriptive statistics are provided to offer a comprehensive view of the data for each measurement and to support the detailed description of the results.

For items that were summarized into their representative scales, i.e. the items from Godspeed, SE-HRI and UEQ-S, Cronbach’s α [86] was used to present how reliable these scales were representing their underlying construct.

The Aligned Rank Transform (ART) [87] was applied to all of the quantitative data before proceeding with calculating each of the ANOVA results. It is commonly used in Human-Computer Interaction (HCI) or HRI when non-parametric data is evaluated. Applying ART ensures that the data will meet expected assumptions for factorial analyses, such as with ANOVA. It is important to mention that therefore all of the ART ANOVA results in the following chapter are being calculated off of the transformed instead of the raw data.

If the ART ANOVA results showed that there was a statistically significant interaction effect of the two independent variables initial proficiency and learning rate, post-hoc contrast tests were conducted by using the ART procedure for multifactor contrast tests (ART-C) [88], which was specifically designed to perform contrast tests for ART ANOVA results. Internally, ART-C utilized Tukey’s Honestly Significant Difference (HSD) test to conduct multiple comparisons and to adjust for Type-I errors, revealing significant differences with some measures. While Bonferroni p-value adjustments were also considered, they turned out to be too conservative and resulted in non-significant

differences. These between-group contrast tests were made to help understand which of the individual test cases are significantly different from each other.

Qualitative Data

Interview recordings have first been transcribed with the help of OpenVINO™ AI plugins for Audacity. The plugin provided an audio transcription tool directly within Audacity’s UI by using features of OpenAI Whisper at its core. OpenAI Whisper is a speech recognition model trained on a large dataset, capable of various kinds of language processing, including transcription of audio files to text. Once an interview was transcribed by Whisper, the plugin inserted the text as labels beyond the audio track. Manual corrections to these labels were made subsequently to remove any kind of mistakes that the tool made. Additionally, filler words or disfluencies were also removed from the transcript to improve readability.

Besides the above, annotations to the text were made to contextualize it, including [I] and [P], indicating whether I (the interviewer) or the participant was talking, and also [A], [B], [C], [D] or the combination of them, if a part of the text referred to specific user tests. For example for a participant with the test-sequence S_1 (refer to table 3.2): ‘The first robot...’ would refer to test case [A], whereas ‘The second and the third robot...’ would refer to test cases [B, D]. After all of the interviews were transcribed and annotated, they were exported into a text file, making them easier to read for further processing.

In order to potentially extract meaningful results from the qualitative data, a template thematic analysis [89] approach was utilized. The interview questions already encoded the initial themes. The interview question Q_2 encoded a sort of expectation theme, in which participants were expected to answer if they perceived any sort of bias towards the robot doing the initial rectangle drawing task. Q_3 encoded a sort of motivational theme, where participants were expected to give details about how they were affected by the robot’s learning rate. Q_{4-6} encoded a sort of interaction theme where participants were expected to provide answers about how the interaction of the two robotic traits, initial proficiency, and learning rate influenced their motivation.

The transcripts have been transferred to a Miro board. Miro offers an infinite canvas, a collaborative design tool with various features and templates. A canvas that is given a title and which is created as blank or copied from a template is called a Miro board. For the qualitative analysis of this study, a simple table template has been used to create the board. The table was modified in such a way that the row header contained the participant ID and their assigned test sequences, the columns represented the qualitative questions asked during the experiment, starting with interview questions Q_1 through Q_6 , then with an additional column for other observations made during the interviews and finally the individual answers given by them within each of the robot assessment questionnaires when asked what made the participant stop teaching. Consequentially, each of the results for the questions of a participant was put into the respective cells of

3. APPROACH

the table, as shown in figure 3.14. Transcribed texts from the interviews have been put into conversation bubbles to further improve readability.







Participant Information	Interview Q1	Interview Q2	Interview Q3	Interview Q4-Q6	Other mentions or observations	Robot Assessment Questionnaire
<p>p16</p> 	<p>Please describe your overall experience with teaching all the different robots.</p> 	<p>When seeing the robot's initial skill level, which means how well it did with drawing a rectangle. Did this fact affect your willingness to teach the robot with the new task?</p> 	<p>When seeing how slow or fast the robot's learning capabilities. Did this rate of learning affect your willingness to teach the robot with the new task?</p> 	<p>Discuss each of the robots ask why it is in this place of the list (Compare with motivation table). Which robot did you find the most demanding in terms of your engagement and patience and why?</p> 		<p>What was it that made you stop teaching this robot?</p> 

Figure 3.14: Miro board table structure for thematic analysis. Example from the participant with ID 16. Conversation bubbles have been cut off to shorten the image.

The table on the Miro board was filled out with results from all 24 participants. The transcripts were then read once to find common themes among them. Potential themes have been noted down in a separate text file by a colleague researcher and me individually. Then, the individually found themes were merged and then selected, or eradicated based on how prominent they had been in the data.

Finally, the remaining common themes were transferred back to the Miro board into a separate table. Each theme was represented by a table column. Each cell under a given column was first made an empty placeholder, then all of the texts in the first table were re-read and if texts would fit into a common theme, they were copied and pasted into a before-mentioned placeholder cell. The participant ID was additionally added to such a cell to be able to back-track the information if needed. After this process, the common themes table contained all the data to extract information needed for the research question of this study. Results are provided in the next chapter.

CHAPTER 4

Results

This chapter focuses on the results of the experiment of both, quantitative and qualitative measures. First, quantitative results are presented from the onboarding questionnaire, which covers participants' personality traits. Subsequently, ART ANOVA statistical results from the log events of the user tests are shown, which included measurements for how much time participants spent teaching and how much time they spent observing the robot doing its initial task, the number of demonstrations participants provided to the robot and how far they managed to increase the robots' proficiency score after discontinuing the teaching process. Following with statistical results of the robot assessment questionnaire, which contained scales for how participants perceived the robot in terms of anthropomorphism, animacy, likeability, intelligence, and safety (Godspeed), scales on how participants rated their self-efficacy when teaching the robot (SE-HRI) and scales for how good of pragmatic and hedonic user experience participants had with the robot (UEQ-S). The reliability of the data is also presented for each of these scales, by using Cronbach's α . Finally, statistical results are shown of how motivated participants have been to continue teaching the robot. Besides the results of participants' personality traits, results show, for each of the scales, whether or not there have been statistically significant differences between the robot configured with a low versus high initial proficiency and the robot configured with a slow versus a fast learning rate. Additionally, results show whether or not there has been measured a statistically significant interaction effect between initial proficiency and learning rate, and if this was the case for a given measurement, then the result of subsequent contrast tests are presented that show which of the individual test cases differ from each other.

Second, qualitative results that were yielded from thematically analyzing participants' interview transcripts are presented, which were mainly concerned with how different robot configurations affected the participants' motivation to teach the robot. Along with the represented main concern of the interviews, other topics were brought up by participants, for example choosing different teaching strategies for different configurations

of the robot, perceiving the robot’s movement differently in different test cases, or their attitude towards the time aspect of the teaching process.

4.1 Participants Personality

The results from the TIPI [82] which had been collected through the personality questionnaire, as described in section 3.3.2, were summarized into *Extraversion* ($M = 4.44$, $SD = 1.28$), *Agreeableness* ($M = 4.83$, $SD = 1.01$), *Conscientiousness* ($M = 4.88$, $SD = 1.27$), *Emotional Stability* ($M = 4.96$, $SD = 1.00$), and *Openness to Experiences* ($M = 5.54$, $SD = 0.85$). Results are shown with the help of parallel coordinates in figure 4.1.

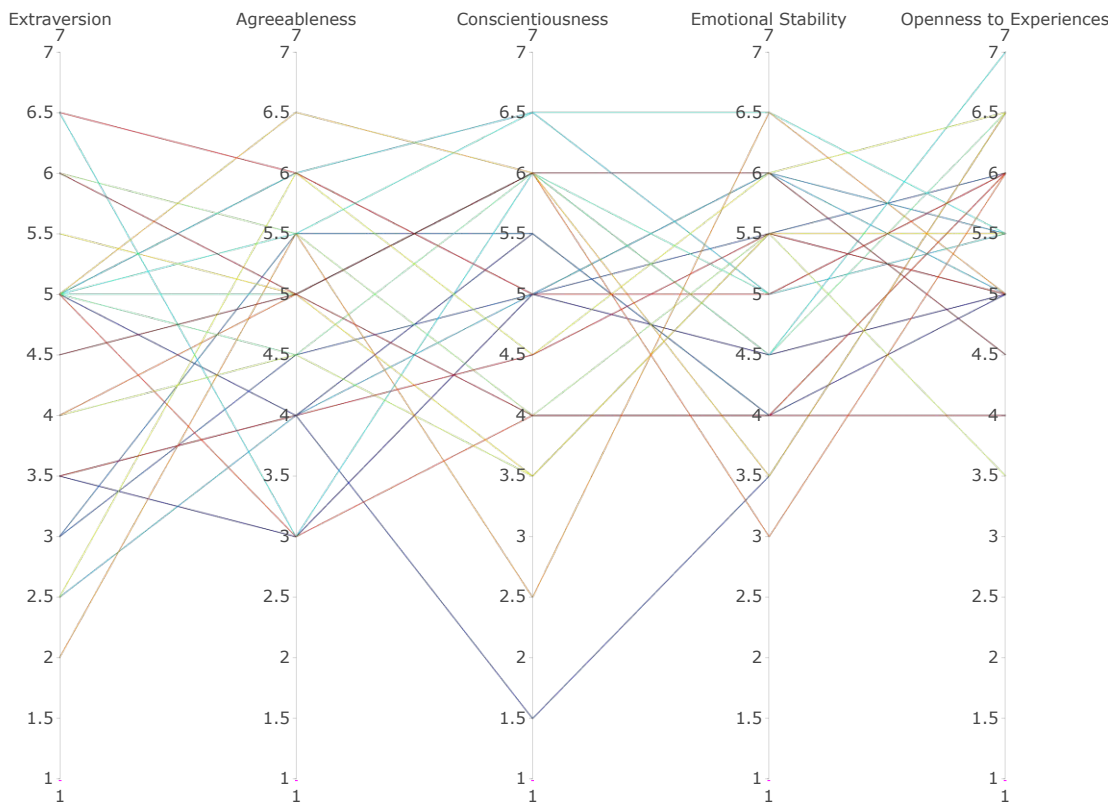


Figure 4.1: TIPI results for each participant on parallel coordinates

4.2 Teaching Time

The teaching time refers to the measurement ‘time_interacting_seconds’, described in section 3.3.3. The data in table 4.1 and its representations in figure 4.2 suggest that participants spent similar time teaching for low ($Md = 114.5$, $IQR = 61$) and high

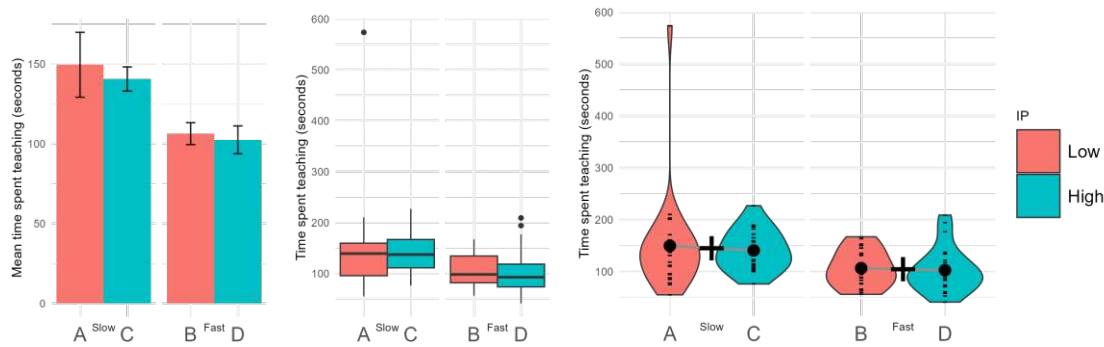


Figure 4.2: Overview of results for teaching time (seconds): A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	149.54	99.59	95.50	139.0	159.50	64.00	24
B	Low	Fast	106.42	33.82	81.50	98.0	134.25	52.75	24
C	High	Slow	140.62	37.10	111.00	137.0	166.75	55.75	24
D	High	Fast	102.50	42.74	73.50	92.5	118.25	44.75	24
A+B	Low	All	127.98	76.73	88.00	114.5	149.00	61.00	48
C+D	High	All	121.56	44.03	91.75	113.5	147.25	55.50	48
A+C	All	Slow	145.08	74.48	107.50	139.0	162.00	54.50	48
B+D	All	Fast	104.46	38.18	76.75	94.0	131.75	55.00	48

Table 4.1: Teaching time: Summary for individual and combined test cases

($Md = 113.5$, $IQR = 55$) configurations for the robot's initial proficiency (IP). The lack of statistical significance for the main effect for initial proficiency supported this finding ($F = 0.05$, $p = 0.82$, $\eta^2 = 0.0008$). When comparing participants in terms of learning rate (LR), the data shows a difference between slow ($Md = 139$, $IQR = 54.5$) and fast ($Md = 94$, $IQR = 55$) configurations of the robot. This was also supported by the statistically significant main effect for learning rate ($F = 21.95$, $p < 0.001$, $\eta^2 = 0.24$). There has been no statistically significant finding on the interaction effect of IP * LR ($F = 0.54$, $p = 0.47$, $\eta^2 = 0.0078$).

4.3 Number of Attempts

Number of attempts refers to the measurement 'interacting_repetitions' as mentioned in section 3.3.3. The violin plot in figure 4.3 visually shows a hard lower bound with four attempts, which was naturally embedded into the data, as the feedback dialog only

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	7.88	3.84	5.75	7.0	10.00	4.25	24
B	Low	Fast	5.12	1.48	4.00	5.0	5.25	1.25	24
C	High	Slow	7.79	1.93	6.75	7.5	9.00	2.25	24
D	High	Fast	5.33	1.74	4.00	5.0	5.25	1.25	24
A+B	Low	All	6.50	3.20	4.00	5.5	8.00	4.00	48
C+D	High	All	6.56	2.20	5.00	6.0	8.00	3.00	48
A+C	All	Slow	7.83	3.01	6.00	7.0	9.25	3.25	48
B+D	All	Fast	5.23	1.60	4.00	5.0	5.25	1.25	48

Table 4.2: Number of demonstrations: Summary for individual and combined test cases

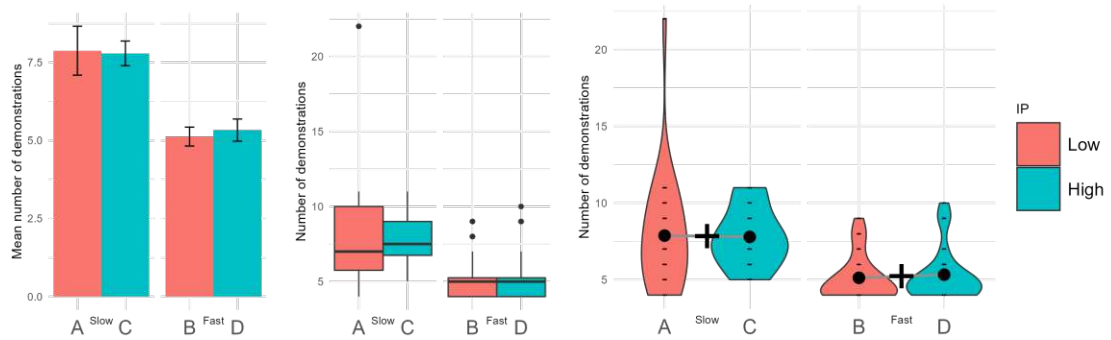


Figure 4.3: Overview of results for number of attempts: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

appeared after the fourth demonstration for participants, as described in section 3.2.4. Both the data in table 4.2 and the plots in figure 4.3 suggest that participants took a similar amount of attempts when comparing low ($Md = 5.5$, $IQR = 4$) and high ($Md = 6$, $IQR = 3$) robot initial proficiency. The main effect for initial proficiency supports this finding as there was no statistical significance ($F = 0.3$, $p = 0.59$, $\eta^2 = 0.0043$). Further, the data and plots suggest that there is a difference between slow ($Md = 7$, $IQR = 3.25$) and fast ($Md = 5$, $IQR = 1.25$) learning configurations which is consistent with the statistically significant main effect for learning rate ($F = 48.18$, $p < 0.001$, $\eta^2 = 0.41$). Finally, the interaction effect of IP * LR has not been statistically significant.

4.4 Achieved Proficiency

The achieved proficiency refers to the measurement ‘robot_proficiency_score’ described in section 3.3.3. The plots in figure 4.4 show that in fast learning configurations, participants always managed to successfully teach the robot. This is because, first the robot already

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	0.77	0.21	0.63	0.77	1.00	0.37	24
B	Low	Fast	1.00	0.00	1.00	1.00	1.00	0.00	24
C	High	Slow	0.82	0.15	0.74	0.83	0.99	0.25	24
D	High	Fast	1.00	0.00	1.00	1.00	1.00	0.00	24
A+B	Low	All	0.88	0.19	0.77	1.00	1.00	0.23	48
C+D	High	All	0.91	0.14	0.85	1.00	1.00	0.15	48
A+C	All	Slow	0.79	0.19	0.66	0.77	0.99	0.33	48
B+D	All	Fast	1.00	0.00	1.00	1.00	1.00	0.00	48

Table 4.3: Achieved proficiency: Summary for individual and combined test cases

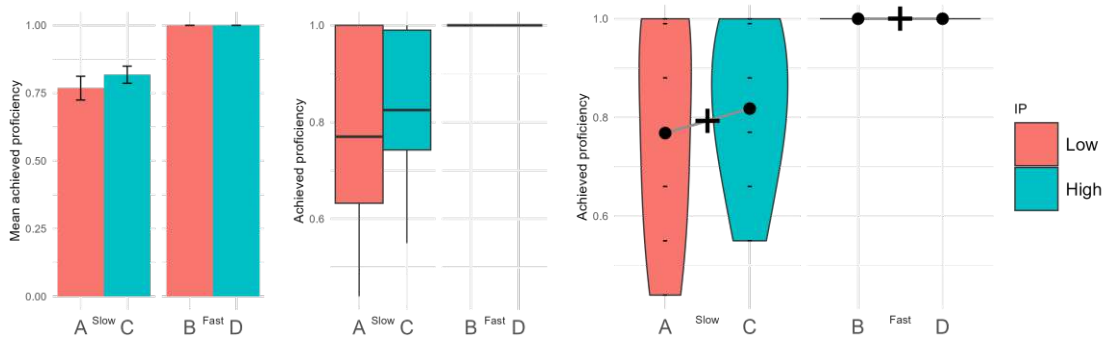


Figure 4.4: Overview of results for achieved robot proficiency: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

reaches 100% of simulated proficiency after the fourth demonstration has been given to it, as described in subsection *User Tests*, which is contained in section 3.3.2, and second, due to participants not being asked whether or not they want to continue with teaching before the fourth attempt was made, as described in section 3.2.4. Likewise, for slow learning configurations, the minimum value for achieved proficiency was bound to 44% (four attempts with a step size of 11% as described in section 3.1.4). Within these test cases, participants still were somewhat close to reaching full robot proficiency on average ($M = 79\%$, $SD = 19\%$) as seen in table 4.3, which is naturally still lower than within fast learning test cases, because of the explanations above. The statistically significant main effect for learning rate supports this observation ($F = 58.34$, $p < 0.001$, $\eta^2 = 0.46$). It was found that the achieved proficiency was lower with low ($M = 88\%$, $SD = 19\%$) than with high ($M = 91\%$, $SD = 14\%$) initial proficiency, which was also supported by the statistically significant main effect ($F = 7.71$, $p < 0.01$, $\eta^2 = 0.1$). Further, there was a statistically significant interaction effect of IP * LR ($F = 7.71$, $p < 0.01$, $\eta^2 = 0.1$). The

contrast tests showed that statistically significant between-group differences exist if there is a difference in the learning rate configuration, so differences occur in the following sets of test cases: D-C, D-A, C-B, and B-A ($p < 0.0001$). This conversely means that there are no significant differences in the sets of: D-B and C-A. Finally, this explains why there was a statistically significant main effect for IP when most certainly it had no significant effect on the results for the measured achieved proficiency.

4.5 Initial Observing Time

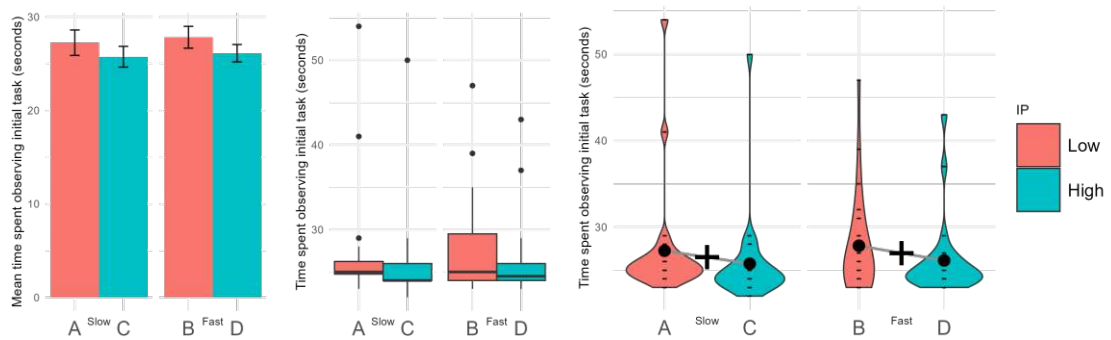


Figure 4.5: Overview of results for initial observing time (seconds): A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	27.25	6.67	24.75	25.00	26.25	1.50	24
B	Low	Fast	27.83	5.72	24.00	25.00	29.50	5.50	24
C	High	Slow	25.75	5.45	24.00	24.00	26.00	2.00	24
D	High	Fast	26.12	4.58	24.00	24.50	26.00	2.00	24
A+B	Low	All	27.54	6.15	24.00	25.00	27.25	3.25	48
C+D	High	All	25.94	4.98	24.00	24.00	26.00	2.00	48
A+C	All	Slow	26.50	6.07	24.00	25.00	26.00	2.00	48
B+D	All	Fast	26.98	5.20	24.00	25.00	27.00	3.00	48

Table 4.4: Time spent observing initial task: Summary for individual and combined test cases

The measurement of ‘time_observing_initial_task_seconds’ (see section 3.3.3) refers to initial observing time. There has been measured a higher amount of time spent observing the robot doing its initial task when initial proficiency was low ($Md = 25$, $IQR = 3.25$) than when it was high ($Md = 24$, $IQR = 2$), which was supported by the statistically

significant main effect for initial proficiency ($F = 7.53$, $p < 0.01$, $\eta^2 = 0.1$). There has not been detected any significant difference between slow ($Md = 25$, $IQR = 2$) and fast ($Md = 25$, $IQR = 3$) learning configurations. The lack of a statistically significant main effect for learning rate supports this observation ($F = 1.21$, $p = 0.28$, $\eta^2 = 0.02$). Both, table 4.4 descriptively and figure 4.5 visually represent those findings. Finally, there has not been a statistically significant interaction effect of IP * LR ($F = 0.005$, $p = 0.94$, $\eta^2 = 0.00008$). The lack of a significant difference between slow and high learning rate configurations makes sense, since the measurement of the observation time for the initial task was already complete when the learning rate took effect at a later stage of the tests (please refer to section 3.3.2 for a detailed description of the user tests and its individual steps).

4.6 Robot Perception

Participants perceived a robot to be more anthropomorphic, animated, likeable, and intelligent when it was configured as fast learning compared to when configured as slow learning. When a robot was configured with low initial proficiency and as slow learning it received a lower overall anthropomorphism rating than with other configurations. Further, participants perceived a robot as more animated, when a robot was configured with low initial proficiency and with a fast learning rate than with other configurations.

For most of the scales from the Godspeed questionnaire, there has been measured high reliability by using Cronbach's α . High reliability was measured for Anthropomorphism (0.9063), Animacy (0.8957), Likeability (0.931), and Perceived Intelligence (0.9333), whereas for Perceived Safety (0.6702) there was only measured low reliability (please refer to table 4.5).

Scale	Cronbach's α	95% Confidence Interval	
		Lower Bound	Upper Bound
Anthropomorphism	0.9063	0.8765	0.9362
Animacy	0.8957	0.8631	0.9283
Likeability	0.9310	0.9092	0.9528
Perceived Intelligence	0.9333	0.9129	0.9537
Perceived Safety	0.6702	0.5498	0.7906

Table 4.5: Godspeed questionnaire: Cronbach's α and 95% confidence intervals for the individual scales.

4.6.1 Anthropomorphism

After comparing Godspeed's anthropomorphism scale for initial proficiency, there has been measured only a slight difference between low ($Md = 2$, $IQR = 1.45$) and high ($Md = 2.2$, $IQR = 1.45$), however, no statistically significant main effect was measured

4. RESULTS

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	1.75	0.73	1.20	1.60	2.05	0.85	24
B	Low	Fast	2.47	0.88	1.80	2.40	3.20	1.40	24
C	High	Slow	2.29	0.95	1.60	2.10	3.25	1.65	24
D	High	Fast	2.39	0.85	1.75	2.30	3.00	1.25	24
A+B	Low	All	2.11	0.88	1.40	2.00	2.85	1.45	48
C+D	High	All	2.34	0.89	1.60	2.20	3.05	1.45	48
A+C	All	Slow	2.02	0.88	1.40	1.80	2.45	1.05	48
B+D	All	Fast	2.43	0.86	1.80	2.30	3.20	1.40	48

Table 4.6: Anthropomorphism: Summary for individual and combined test cases

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	1.83	0.81	1.17	1.58	2.21	1.04	24
B	Low	Fast	2.55	0.87	2.12	2.58	3.17	1.04	24
C	High	Slow	2.14	0.81	1.50	2.08	2.67	1.17	24
D	High	Fast	2.43	0.75	1.83	2.33	3.00	1.17	24
A+B	Low	All	2.19	0.91	1.33	2.17	2.88	1.54	48
C+D	High	All	2.28	0.78	1.67	2.25	2.88	1.21	48
A+C	All	Slow	1.98	0.81	1.33	1.83	2.50	1.17	48
B+D	All	Fast	2.49	0.81	1.96	2.42	3.04	1.08	48

Table 4.7: Animacy: Summary for individual and combined test cases

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	2.82	0.86	2.35	2.80	3.25	0.90	24
B	Low	Fast	3.39	0.92	3.00	3.60	3.80	0.80	24
C	High	Slow	3.00	1.03	2.40	2.90	3.50	1.10	24
D	High	Fast	3.31	0.81	2.95	3.20	3.70	0.75	24
A+B	Low	All	3.10	0.93	2.60	3.10	3.65	1.05	48
C+D	High	All	3.15	0.93	2.60	3.10	3.65	1.05	48
A+C	All	Slow	2.91	0.94	2.40	2.80	3.40	1.00	48
B+D	All	Fast	3.35	0.86	3.00	3.40	3.80	0.80	48

Table 4.8: Likeability: Summary for individual and combined test cases

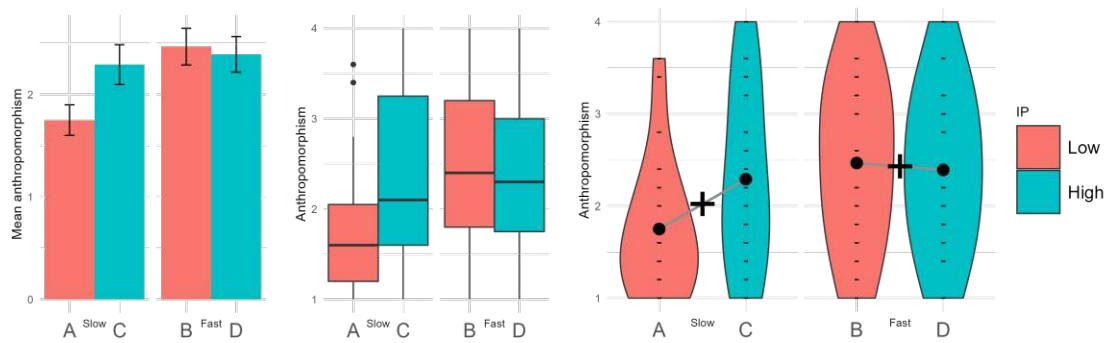


Figure 4.6: Overview of results for anthropomorphism: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

($F = 3.92$, $p = 0.052$, $\eta^2 = 0.05$). Conversely, there has been a higher measured difference between slow ($Md = 1.8$, $IQR = 1.05$) and high ($Md = 2.3$, $IQR = 1.4$) learning rate configurations and a statistically significant main effect ($F = 14.8$, $p < 0.001$, $\eta^2 = 0.18$). Further, there has been measured a statistically significant interaction effect for IP * LR ($F = 7.89$, $p < 0.01$, $\eta^2 = 0.1$). The contrast test that followed resulted in statistically significant between-group differences for the following sets of test cases: D-A ($p < 0.001$), C-A ($p < 0.01$), and B-A ($p < 0.001$), which shows that participants rated the robot lower in terms of anthropomorphism when it was configured to be with low initial proficiency and with a slow learning rate. These results are visually supported by the plots in figure 4.6.

4.6.2 Animacy

Regarding Godspeed's animacy scale, only a slight difference has been measured between low ($Md = 2.17$, $IQR = 1.54$) and high ($Md = 2.25$, $IQR = 1.21$) initial proficiency configurations of the test cases (please refer to table 4.7 and figure 4.7 for details). Also, there has not been measured a statistically significant main effect for initial proficiency ($F = 0.87$, $p = 0.35$, $\eta^2 = 0.01$). Compared to initial proficiency, there has been measured a higher difference between slow ($Md = 1.83$, $IQR = 1.17$) and fast ($Md = 2.42$, $IQR = 1.08$) learning rate configurations with an additional statistically significant main effect for learning rate ($F = 30.08$, $p < 0.001$, $\eta^2 = 0.3$). Further, there has been measured a statistically significant interaction effect for IP * LR ($F = 6.47$, $p < 0.05$, $\eta^2 = 0.09$). The between-group contrast tests measured statistically significant differences with the following sets of test cases: D-A ($p = 0.0001$), C-B ($p < 0.05$), and B-A ($p < 0.0001$), which shows that a robot configured as low initial proficiency and as fast learning, has been rated higher than with other configurations in terms of animacy. The plots in figure 4.7 visually support these findings.

4. RESULTS

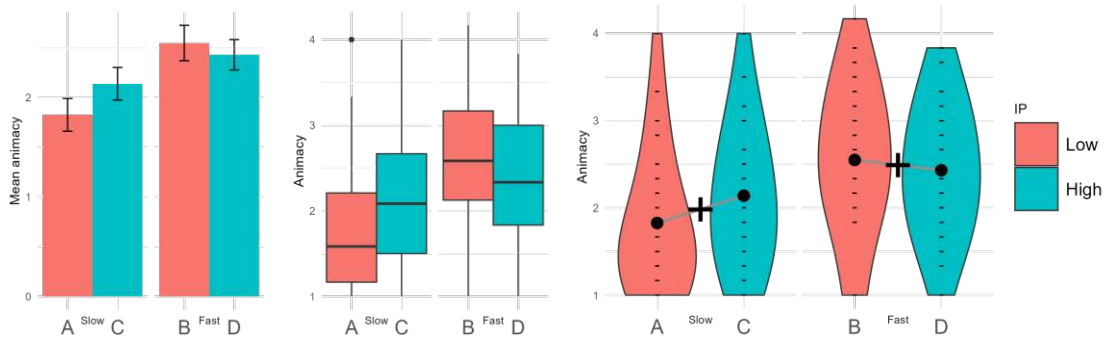


Figure 4.7: Overview of results for animacy: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

4.6.3 Likeability

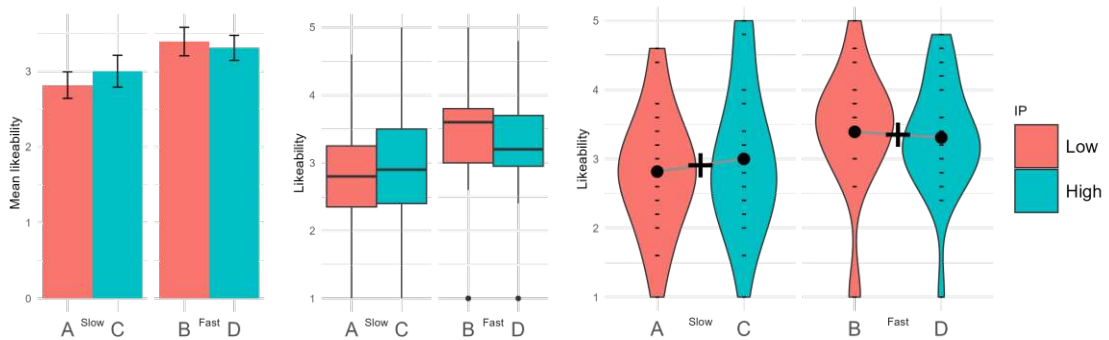


Figure 4.8: Overview of results for likeability: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

For Godspeed's likeability there was only measured very little difference between low ($M = 3.1$, $SD = 0.93$) and high ($M = 3.15$, $SD = 0.93$) initial proficiency configurations of user tests and there has also been no statistically significant main effect for initial proficiency ($F = 0.28$, $p = 0.6$, $\eta^2 = 0.004$) (please refer to table 4.8 and figure 4.8 for details). A higher overall difference has been measured for the two different learning rate configurations slow ($Md = 2.8$, $IQR = 1$) and high ($Md = 3.4$, $IQR = 0.8$) with an additional measured statistically significant main effect for learning rate ($F = 29.72$, $p < 0.001$, $\eta^2 = 0.3$). Finally, there has been measured no statistically significant interaction effect for IP * LR.

4.6.4 Perceived Intelligence

Participants did not perceive the robot in low initial proficiency ($Md = 2.4$, $IQR = 1.65$) configurations of user tests less intelligent than in the high initial proficiency configurations ($Md = 2.4$, $IQR = 1.45$) (please refer to table 4.9 and figure 4.9 for details). There has also been no statistically significant main effect for initial proficiency ($F = 0.2$, $p = 0.66$, $\eta^2 = 0.003$). Comparing learning rate, participants perceived the robot in slow learning configurations as less intelligent ($Md = 2$, $IQR = 1$) than within fast learning configurations ($Md = 3.2$, $IQR = 1.2$), which also is supported by the statistically significant main effect for learning rate ($F = 44.81$, $p < 0.001$, $\eta^2 = 0.39$). There has not been measured a statistically significant interaction effect for IP * LR ($F = 1.69$, $p = 0.2$, $\eta^2 = 0.02$).

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	1.99	0.86	1.35	1.90	2.40	1.05	24
B	Low	Fast	3.07	1.04	2.35	3.20	3.60	1.25	24
C	High	Slow	2.17	0.87	1.40	2.10	2.60	1.20	24
D	High	Fast	2.90	0.88	2.40	3.20	3.60	1.20	24
A+B	Low	All	2.53	1.09	1.75	2.40	3.40	1.65	48
C+D	High	All	2.54	0.94	1.80	2.40	3.25	1.45	48
A+C	All	Slow	2.08	0.86	1.40	2.00	2.40	1.00	48
B+D	All	Fast	2.98	0.96	2.40	3.20	3.60	1.20	48

Table 4.9: Perceived intelligence: Summary for individual and combined test cases

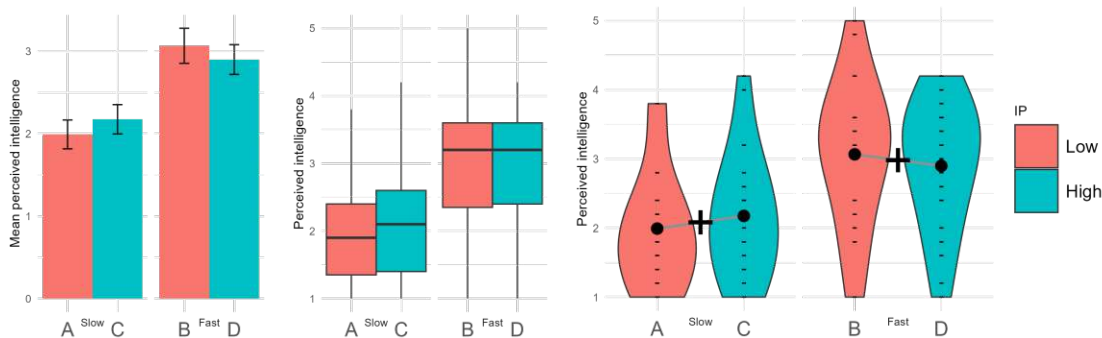


Figure 4.9: Overview of results for perceived intelligence: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

4.6.5 Perceived Safety

After comparing the responses from the Godspeed's perceived safety scale on the two independent variables, the result showed that neither participants perceived the robot less or more safe compared between low ($Md = 3.67$, $IQR = 1.08$) and high ($Md = 3.33$, $IQR = 1.08$) initial proficiency configured test cases, nor did they perceive the robot less or more safe compared between slow ($Md = 3.33$, $IQR = 1.33$) and fast ($Md = 3.5$, $IQR = 1$) learning rate configured test cases (please refer to table 4.10 for details). These findings were supported, first visually by the plots in figure 4.10, and second, by the lack of a statistically significant main effect for initial proficiency ($F = 0.19$, $p = 0.67$, $\eta^2 = 0.003$), learning rate ($F = 1.87$, $p = 0.18$, $\eta^2 = 0.03$) or the interaction effect for IP * LR ($F = 0.28$, $p = 0.6$, $\eta^2 = 0.004$).

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	3.46	0.83	3.00	3.33	4.08	1.08	24
B	Low	Fast	3.67	0.67	3.00	3.67	4.08	1.08	24
C	High	Slow	3.46	0.87	3.00	3.33	4.33	1.33	24
D	High	Fast	3.56	0.69	3.00	3.33	4.00	1.00	24
A+B	Low	All	3.56	0.75	3.00	3.67	4.08	1.08	48
C+D	High	All	3.51	0.78	3.00	3.33	4.08	1.08	48
A+C	All	Slow	3.46	0.84	3.00	3.33	4.33	1.33	48
B+D	All	Fast	3.61	0.67	3.00	3.50	4.00	1.00	48

Table 4.10: Perceived safety: Summary for individual and combined test cases

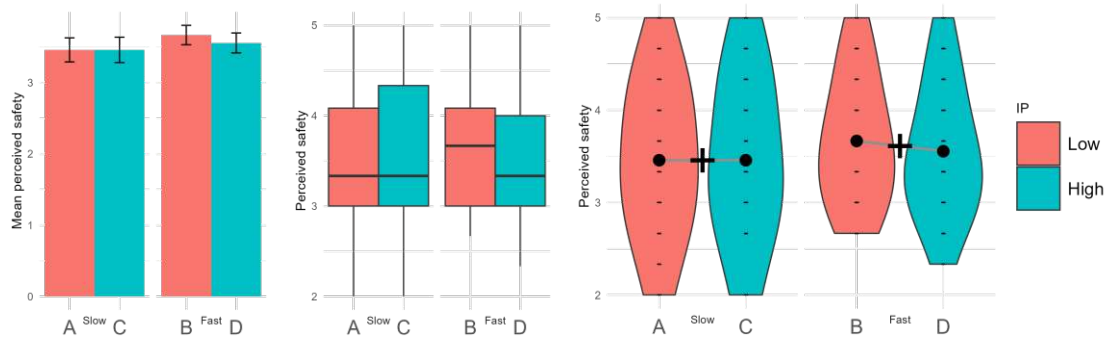


Figure 4.10: Overview of results for perceived safety: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

4.7 Teaching Self-Efficacy

Although only a subset of items from the SE-HRI made it into the robot-assessment questionnaire, the scale for self-efficacy and its response values are of high reliability (0.9404 with a 95% confidence interval [0.9207, 0.9601]), which was measured by using Cronbach's α .

Participants did not observe themselves as more or less efficient when comparing their responses between low ($Md = 4$, $IQR = 3$) and high ($Md = 4.38$, $IQR = 3$) initial proficiency configured test cases, which was also reflected by the lack of a statistically significant main effect for initial proficiency ($F = 0.27$, $p = 0.61$, $\eta^2 = 0.004$) (please refer to table 4.11 and figure 4.11 for details). On the opposite, participants observed themselves as more efficient with the robot being configured with a fast ($Md = 4.88$, $IQR = 1.38$) learning rate compared to it being configured with a slow ($Md = 2.38$, $IQR = 2.12$) learning rate. This has also been validated with a measured statistically

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	2.83	1.44	1.75	2.38	3.75	2.00	24
B	Low	Fast	4.89	1.20	4.38	5.00	6.00	1.62	24
C	High	Slow	3.00	1.66	1.69	2.50	4.31	2.62	24
D	High	Fast	4.44	1.46	4.44	4.62	5.25	0.81	24
A+B	Low	All	3.86	1.67	2.25	4.00	5.25	3.00	48
C+D	High	All	3.72	1.71	2.00	4.38	5.00	3.00	48
A+C	All	Slow	2.92	1.54	1.75	2.38	3.88	2.12	48
B+D	All	Fast	4.66	1.34	4.44	4.88	5.81	1.38	48

Table 4.11: Self-efficacy: Summary for individual and combined test cases

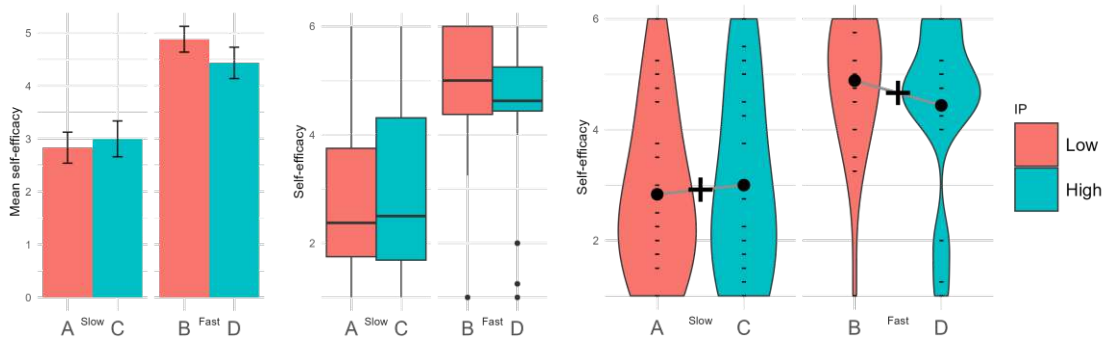


Figure 4.11: Overview of results for self-efficacy: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

significant main effect for learning rate ($F = 68.44$, $p < 0.001$, $\eta^2 = 0.5$). No statistical significance was measured for the interaction effect for IP * LR ($F = 1.08$, $p = 0.3$, $\eta^2 = 0.02$). Figure 4.11 visually supports these findings.

4.8 Teaching Experience

Participants valued a fast learning configured robot more than a slow one in terms of pragmatic and hedonic user experiences. Additionally, participants also seemed to rate a robot higher in terms of hedonic user experience, when it was configured with low initial proficiency and configured to be fast learning.

Both of the scales and their response values from the UEQ-S questionnaire have been measured with high reliability according to the Cronbach's α values. High reliability was measured for Pragmatic quality (0.9012) as well as for Hedonic quality (0.8658), as seen in table 4.12.

Scale	Cronbach's α	95% Confidence Interval	
		Lower Bound	Upper Bound
Pragmatic quality	0.9012	0.8693	0.9331
Hedonic quality	0.8658	0.8212	0.9104

Table 4.12: UEQ-S questionnaire: Cronbach's α and 95% confidence intervals for the individual scales.

4.8.1 Pragmatic Experience

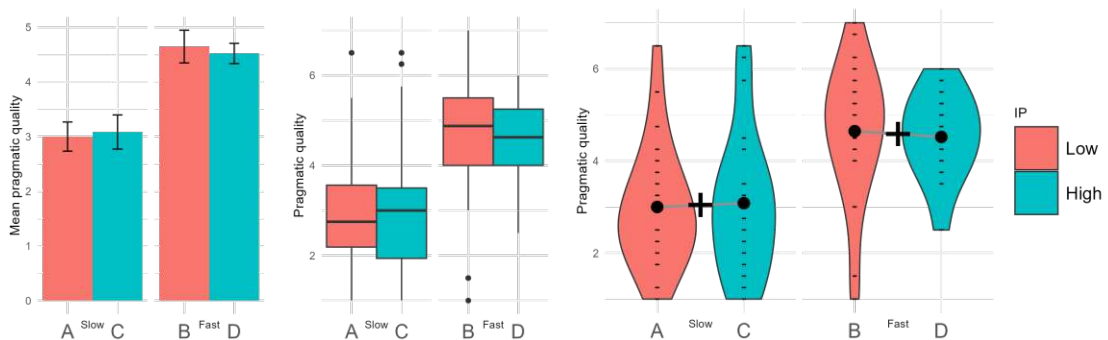


Figure 4.12: Overview of results for pragmatic quality: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	3.00	1.31	2.19	2.75	3.56	1.38	24
B	Low	Fast	4.65	1.46	4.00	4.88	5.50	1.50	24
C	High	Slow	3.08	1.53	1.94	3.00	3.50	1.56	24
D	High	Fast	4.52	0.91	4.00	4.62	5.25	1.25	24
A+B	Low	All	3.82	1.60	2.50	4.00	5.25	2.75	48
C+D	High	All	3.80	1.44	2.50	4.00	5.00	2.50	48
A+C	All	Slow	3.04	1.41	2.00	2.88	3.50	1.50	48
B+D	All	Fast	4.58	1.21	4.00	4.75	5.25	1.25	48

Table 4.13: Pragmatic quality: Summary for individual and combined test cases

Participants had no different user experience regarding the pragmatic quality, when compared between low ($Md = 4$, $IQR = 2.75$) and high ($Md = 4$, $IQR = 2.5$) initial proficiency configured test cases, which also was supported by the lack of a statistically significant main effect for initial proficiency ($F = 0.13$, $p = 0.73$, $\eta^2 = 0.002$) (please refer to table 4.13 and figure 4.12 for details). On the contrary, participants reported a better pragmatic user experience with fast ($Md = 4.75$, $IQR = 1.25$) learning rate, rather than with slow ($Md = 2.88$, $IQR = 1.5$) learning rate configured test cases. A measured statistically significant main effect for learning rate supports this finding ($F = 70.7$, $p < 0.001$, $\eta^2 = 0.51$). Finally, no there has been measured no statistically significant interaction effect for IP * LR ($F = 0.54$, $p = 0.47$, $\eta^2 = 0.008$).

4.8.2 Hedonic Experience

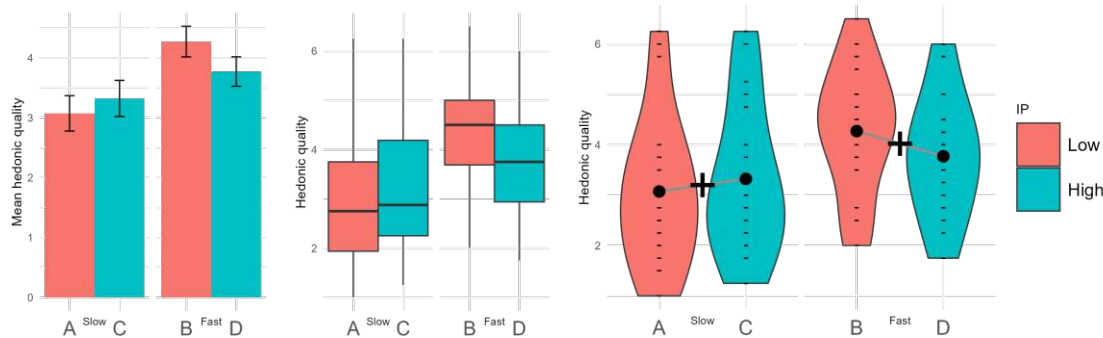


Figure 4.13: Overview of results for hedonic quality: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

Similar to the results of the pragmatic user experience, participants also did not have a different experience regarding the hedonic quality, when compared between low ($Md =$

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	3.07	1.45	1.94	2.75	3.75	1.81	24
B	Low	Fast	4.27	1.25	3.69	4.50	5.00	1.31	24
C	High	Slow	3.32	1.47	2.25	2.88	4.19	1.94	24
D	High	Fast	3.77	1.20	2.94	3.75	4.50	1.56	24
A+B	Low	All	3.67	1.47	2.50	3.75	4.56	2.06	48
C+D	High	All	3.55	1.35	2.50	3.38	4.50	2.00	48
A+C	All	Slow	3.20	1.45	2.25	2.75	3.81	1.56	48
B+D	All	Fast	4.02	1.24	3.19	4.12	4.81	1.62	48

Table 4.14: Hedonic quality: Summary for individual and combined test cases

3.75, $IQR = 2.06$) and high ($Md = 3.38$, $IQR = 2$) initial proficiency configured test cases (please refer to table 4.14 and figure 4.13), which also was supported by the lack of a statistically significant main effect for initial proficiency ($F = 1.22$, $p = 0.27$, $\eta^2 = 0.02$). Also similar to the results of the pragmatic user experience, participants again reported a better hedonic user experience with fast ($Md = 4.12$, $IQR = 1.62$) learning rate, rather than with slow ($Md = 2.75$, $IQR = 1.56$) learning rate configured test cases. A measured statistically significant main effect for learning rate supports this finding ($F = 29.42$, $p < 0.001$, $\eta^2 = 0.3$). Differently than with the results of the pragmatic user experience, there has been measured a statistically significant interaction effect for IP * LR ($F = 4.63$, $p < 0.05$, $\eta^2 = 0.06$). The following sets of test cases exhibited a statistically significant difference within the follow-up contrast tests: D-A ($p < 0.05$), C-B ($p < 0.001$) and B-A ($p < 0.001$). Thus, these results together with the visual differences of the test cases within the violin plot of figure 4.13 suggest that especially the configurations of test case B had a special influence on participants' hedonic experiences.

4.9 Teaching Motivation

Case	IP	LR	M	SD	Q1	Med	Q3	IQR	n
A	Low	Slow	3.33	1.90	2.00	3.00	4.25	2.25	24
B	Low	Fast	5.21	1.67	4.00	6.00	6.25	2.25	24
C	High	Slow	3.67	1.83	2.00	3.00	5.00	3.00	24
D	High	Fast	4.50	1.67	3.00	5.00	6.00	3.00	24
A+B	Low	All	4.27	2.01	3.00	4.00	6.00	3.00	48
C+D	High	All	4.08	1.78	3.00	4.00	6.00	3.00	48
A+C	All	Slow	3.50	1.86	2.00	3.00	5.00	3.00	48
B+D	All	Fast	4.85	1.69	4.00	5.00	6.00	2.00	48

Table 4.15: Self-rated teaching motivation: Summary for individual and combined test cases

The summarized data in table 4.15 suggests that participants self-rated their motivation to continue teaching not differently between low ($Md = 4$, $IQR = 3$) and high ($Md = 4$, $IQR = 3$) initial proficiency configured test cases, which also is supported by the lack of statistically significant main effect for initial proficiency ($F = 0.84$, $p = 0.36$, $\eta^2 = 0.01$). Other than with initial proficiency, participants self-rated their motivation to continue teaching had been higher with fast ($Md = 5$, $IQR = 2$) than with slow ($Md = 3$, $IQR = 3$) learning rate configured test cases, which was also supported by the statistically significant main effect for learning rate ($F = 30.04$, $p < 0.001$, $\eta^2 = 0.3$). Finally, there has been measured a statistically significant interaction effect for IP * LR ($F = 4.26$, $p < 0.05$, $\eta^2 = 0.06$). The follow-up contrast tests showed interesting findings, wherein the following pairs of test cases there has been measured a statistically significant between-group difference: D-A ($p < 0.01$), C-B ($p < 0.001$), and B-A ($p < 0.001$). Together with the comparison of the visual differences within the plots in figure 4.14, the results suggest that participants were self-rating their motivation to continue teaching the highest when the robot exposed a low initial proficiency at first while then being fast learning in the succeeding task.

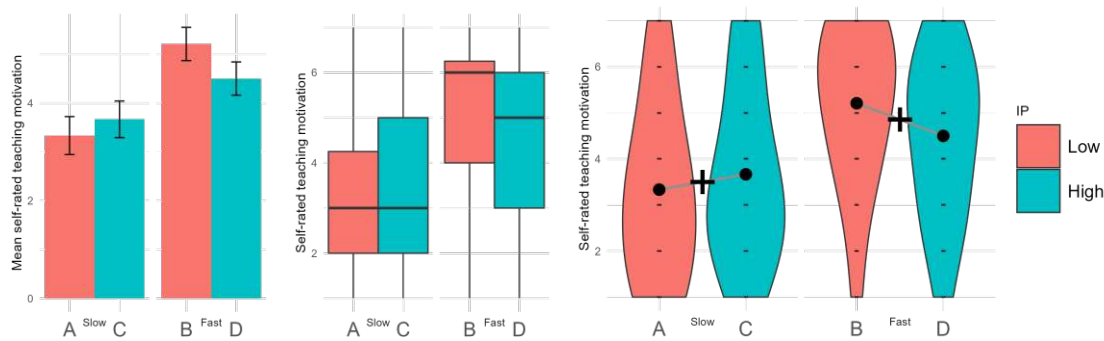


Figure 4.14: Overview of results for self-rated teaching motivation: A bar chart showing mean values with standard errors, a boxplot, and a violin plot with additional marks for individual mean values (dot symbol) and combined mean values (plus symbol). Each of the plots is grouped by the two independent variables initial proficiency (IP) and learning rate (LR).

4.10 Post-Hoc Explorative Analysis on Robot Success

As mentioned in section 2.3, Hedlund et al. found that the success or failure of the robot significantly affected their participants' perceptions of it. Additionally, the subset of the SE-HRI scale used in this study is concerned about how participants rate themselves in terms of their self-efficacy by asking whether or not they could imagine being able to teach the robot certain tasks. This rating of self-efficacy may also change when participants think that they successfully taught the robot.

However, since the design of the study was mainly to evaluate how different configurations of the robot's initial proficiency and learning rate affected the participants, the actual success was only indirectly tested in section 4.4. Nonetheless, an attempt to get more specific results in terms of robot success and failure has been made as described below.

It showed that a robot with a fast learning rate always had been successful at the end of the respective user tests. But, this was not always the case with a slow learning robot. In order to be able to compare results with those mentioned in the literature, additional post-hoc statistical tests were made, in which data has been isolated for slow learning robots (essentially cutting away half of the test data), while also being grouped into successful, i.e. robot proficiency reached 100%, and unsuccessful, i.e. robot proficiency was $< 100\%$, test runs. The two measures, teaching motivation and self-efficacy were considered for this special analysis, because of the above-mentioned reasons.

4.10.1 Teaching Motivation

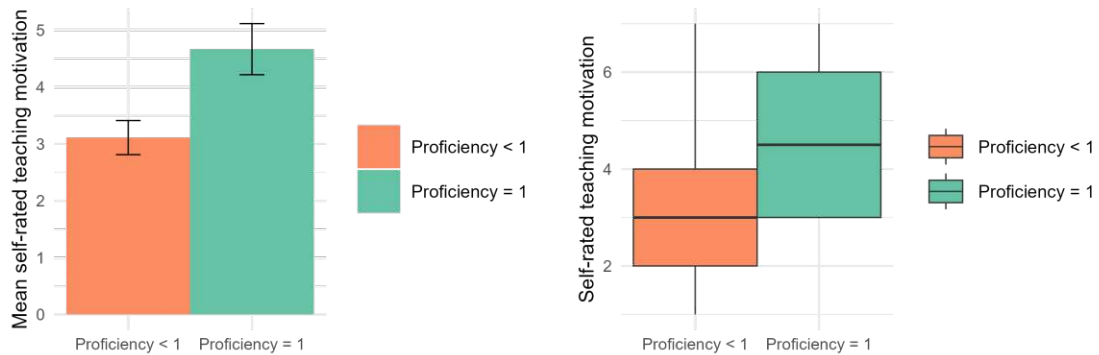


Figure 4.15: Overview of results for motivation isolated for slow learning rate configured test cases grouped by successful (proficiency = 1) and unsuccessful (proficiency < 1) test runs: A bar chart showing mean values with standard errors and a boxplot.

Group	M	SD	Q1	Med	Q3	IQR	n
Proficiency < 1	3.11	1.80	2.00	3.00	4.00	2.00	36
Proficiency = 1	4.67	1.56	3.00	4.50	6.00	3.00	12

Table 4.16: Self-rated teaching motivation: Summary for slow learning rate configured test cases grouped by successful (proficiency = 1) and unsuccessful (proficiency < 1) test runs.

The data in table 4.16 and plots in figure 4.15 show a significant gap in participants' self-rated motivation scores, depending on whether the robot successfully completed the task ($Md = 4.5$, $IQR = 3$) or failed ($Md = 3$, $IQR = 2$) to do so before they decided to discontinue the teaching process. A statistically significant effect on participants' self-rated teaching motivation ($t = 2.67$, $p = 0.01$, 95% CI [2.73, 0.38]) supports this

finding. The effect size was measured to be large (Cohens'd = 0.89) when comparing the group means.

4.10.2 Teaching Self-Efficacy

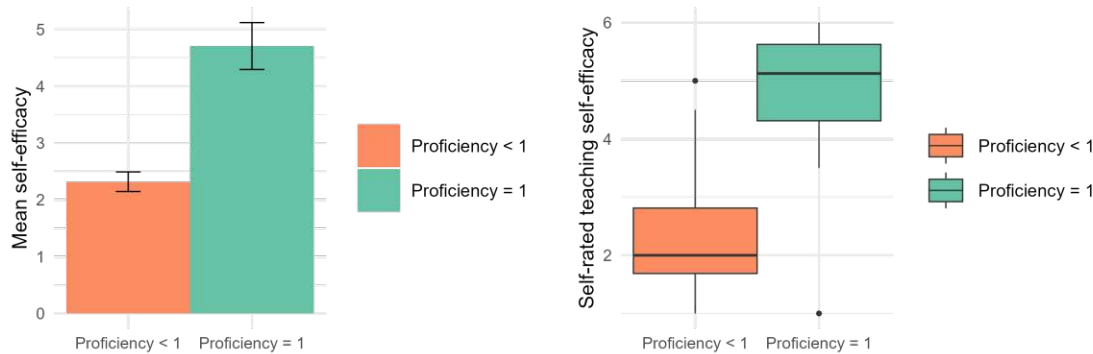


Figure 4.16: Overview of results for self-efficacy isolated for slow learning rate configured test cases grouped by successful (proficiency = 1) and unsuccessful (proficiency < 1) test runs: A bar chart showing mean values with standard errors and a boxplot.

Group	M	SD	Q1	Med	Q3	IQR	n
Proficiency < 1	2.32	1.04	1.69	2.00	2.81	1.12	36
Proficiency = 1	4.71	1.43	4.31	5.12	5.62	1.31	12

Table 4.17: Self-rated teaching self-efficacy: Summary for slow learning rate configured test cases grouped by successful (proficiency = 1) and unsuccessful (proficiency < 1) test runs.

The data in table 4.17 and plots in figure 4.16 show a significant gap in participants' ratings in terms of self-efficacy, depending on whether the robot successfully completed the task ($Md = 5.12$, $IQR = 1.31$) or failed ($Md = 2$, $IQR = 1.12$) to do so when they decided to discontinue the teaching process. A statistically significant effect on participants' self-efficacy scores ($t = 6.27$, $p < 0.001$, 95% CI [3.16, 1.62]) supports this finding. The effect size was measured to be huge (Cohens'd = 2.09) when comparing the group means.

4.11 Qualitative Results

This section presents insights into relevant parts of the participants' responses, examining first the final part of the robot assessment questionnaire, which asked why participants stopped teaching the robot, and second the debriefing interview, which evaluated certain aspects of teaching motivation. Please refer to 3.1.8 for details. Common themes of both parts were extracted throughout the data analysis phase. These themes further

were aligned with the independent variables of the experiment and formed the next subsections: 4.11.1 *Influence of Initial Proficiency* and 4.11.2 *Influence of Learning Rate*. All the findings which were still relevant, but would not fit into those two groups were put into the final subsection 4.11.3 *Other Findings*.

Participants reported that their expectations to be able to teach the robot were set low when the robot was not able to draw the rectangle in the initial task and vice versa. Some participants were surprised at how well the robot has been learning in test case B, where the robot was configured to show low initial proficiency and to be fast learning. Additionally, it has been mentioned by a few participants that they do not see any connection between the initial and the pick-and-place task, subsequently often expressing that the initial proficiency had no impact on their motivation to teach the robot. With a subset of these participants, there were found conflicting quantitative measurements on their self-rated motivation scores from the robot assessment questionnaire.

A major part of the participants brought up that any kind of visual learning progress of the robot increases their motivation to continue teaching. Further, when the robot is slow learning, some participants thought that they needed to change their teaching approach or strategy, that the robot may be dysfunctional, or that there might be a problem with the robot's motivation to learn. Further, a small number of participants attributed non-existent properties to the robot in certain test cases, e.g. different movement patterns in different test cases. Finally, some participants did not see a significant difference between the test cases, while mentioning that the time span to test each configuration was too short for them.

4.11.1 Influence of Initial Proficiency

More than half of the participants brought up that either when the robot drew the rectangle poorly (test cases A and B) or when the robot drew the rectangle perfectly (C and D), it changed their expectations towards how difficult or easy teaching the pick-and-place task would be.

“The last [C] one, yeah, I expected to be faster because I saw that the rectangle was perfect and I thought, okay, then you also do the task easier like the previous one, but, it took, I think, the longest, or the first [A] one took the longest, so, it was not that fast as I expected it.” — P1

“So I thought that in the last two tasks (test-cases) [D, C] they did it right, (meaning) the previous (drawing) task, and so I thought that the robot is better in learning things, but it wasn't.” — P5

“When I saw the robot drawing a rectangle that was not a rectangle at all, I already had the preconception that, oh, he might be a little bit dumber or something. It definitely skewed my interpretation.” — P17

As the interview questions were mainly concerned about motivational factors, P8 and P11 talked about how the initial proficiency did set the expectations for the learning efficiency of the robot on the pick-and-place task, but that it basically would not directly influence the motivation of the participant for teaching the robot.

“It definitely had like a feeling of like, okay, this one will be harder or easier than previous one. I said with the first [D] one, especially because the first one drew it correctly, I thought, okay, maybe that’s just how it goes. But then for example, on the second [A] one did not draw it correctly, I was already like, oh, this one won’t (learn as easy). [...] That sort of set my expectations, but I wouldn’t say I was less motivated to teach it just because it showed that it can (draw the rectangle) [...]” — P8

“It wasn’t relevant to my motivation, but it was relevant for my expectation how quickly it will learn.” — P11

P13 might have missed that also the robot in the third test case (D) was able to draw the rectangle shape without errors, but still mentioned that the robot drawing the rectangle perfectly in the last test was setting expectations for how well it could do things besides rectangle drawing.

“He (the robot) was slow, to put them in the right order. I don’t know if he needed more time, but he was the only one who could draw the rectangle in the previous (drawing) task, so I thought maybe he’s better, not only at rectangle drawing [C].” — P13

The quantitative measurement of self-rated teaching motivation for P18 showed that there was no big difference between low and high initial proficiency configured robots for this participant, but when asked about it in the interview, it was mentioned that it may have influenced the participants’ patience with the robot.

“[...] the performance of the previous task could lead to, I think, increased or decreased patience with the robot, like, the number of iterations, I tried to teach them.” — P18

Participants P19, P23, and P24 mentioned that they were especially surprised with the robot when it first showed that it had low initial proficiency, potentially also setting the expectation for it to learn the pick-and-place task low, but then learning fast afterward (test case B). Not only did it set expectations for the robot’s learning efficiency, but for these three participants there has also been measured a higher self-rated motivation score for this test case.

“then there was one robot who drew a ‘sanduhr’-shape (hourglass), instead of a square [...] but this robot learned to do the ordering of the cubes. [...] I expected it to be stupid, simply said, and then I was surprised that it was able to order the cubes.” — P19

“But I really liked this robot that was dumb in the beginning and then learned and like surprisingly good [B], because it still elevates some kind of feeling. Even though you know that probably it’s because of the system that is behind, but still makes you feel nice.” — P23

“Well, when I saw that, okay, this guy couldn’t accomplish previous tasks, probably it will not accomplish the next one. And it was just a fast thought. So, and then when it managed, I was like, ‘Yay, good, okay, such a nice improvement’.” — P23

“Yeah, absolutely, because my expectations what the robot can do, and for the last [B] one, for instance, the performance of the previous (drawing) task was poor again, and then the robot learned really fast from what I’ve shown to the robot, and then I was happy. [...] But yeah, it (initial proficiency) has a strong effect of what I could expect from the robot afterwards.” — P24

“[...] the drawing task was quite poor, and therefore my expectations were quite low, and then I was totally surprised, by the performance of the robot, and so, yeah, I was surprised and therefore, and then happy, and I thought, it was super fast to teach this robot.” — P24

One participant mentioned that there was uncertainty about how the rectangle drawing would be in connection to the pick-and-place task.

“As I mentioned, I didn’t see any correlation because the first [B] one was like, this sand-clock (hourglass). And I think he was the fastest, if I remember correctly. And the last [D] one did the rectangle quite nicely and this one was also quite fast. So that’s why I got a bit confused whether or not this first task had an effect on the efficiency of the robot.” — P22

“The first few times I wasn’t aware that there could be a connection, but when I learned that it could be a connection I draw my attention to it whether there is some sort of good rectangle - fast learner, bad rectangle - slow learner.” — P22

While also expressing uncertainty about the connection between the two tasks, P15 said that the robot in test case B was the most preferred one, which also seems to be reflected within the self-rated teaching motivation score compared to the remaining test cases (A, C, D).

“I tried to see a correlation between the rectangle drawing thing and the robot, but I wasn’t able to see it, so for example, as far as I remember the best robot was the one who drew like the set rectangle, like with the house of Nikola (referring to hourglass shape the robot drew in test case B) [...]” — P15

One other common theme regarding initial proficiency has been that some participants expressed that they did not see any connection between the drawing task and the pick-and-place task, so in a sense, they thought that the drawing task did not matter much for teaching the robot.

“Yeah, I didn’t know if I have to do this too, so I didn’t know what this had to do with my tasks.” — P07

“Well, first I thought, of course, that it (initial proficiency) said something about how well they would perform in the next (pick-and-place) task, but I think it didn’t. I think there was no connection [...] (between) the rectangle task and the cube task.” — P14

“No, no. Frankly I did not see the the why they were showing me what they did as a previous task. I did not understand that this was the link, in the last [B] one I thought, oh, so maybe that is why they are showing this to me. I did not get that they are sort of displaying their capabilities, or, what they might be an outlook of their ability to learn. I don’t know if that was the thought behind it but I did not get that connection. I thought this was this was sort of a relic how the test was designed that there were other tests that I am supposed to teach them and that was just cut or something.” — P20

Finally, when comparing the participants’ self-rated teaching motivation scores from the robot assessment questionnaire with the interview responses, when asked about whether or not initial proficiency affected their teaching motivation, some conflicting statements occurred. For example, P1 mentioned that initial proficiency did affect teaching motivation, when the participant also rated the teaching motivation with the same score compared between low and high initial proficiency configured test cases, i.e. same scores for low initial proficiency test cases A and C, and same scores for high initial proficiency test cases B and D.

“Yes, I would say, so, for the first [A] robot, it wasn’t that much, of how much it affected me, but I would say, with the iterations the better the robot got for the second [B] and third [D] run, it definitely affected me because I had this feeling of, okay, the robot can draw a rectangle, so he can also put the boxes in the correct order and, yeah, I would say that affected it a lot.” — P1

Vice versa, participants P3 and P21 responded that initial proficiency did not affect their motivation, while they systematically rated their teaching motivation lower within the questionnaire for low initial proficiency than with high initial proficiency configured test cases.

“No, it was confusing because it was a completely different task and the pattern was also different, that he drew [...] but the rectangle was also a little bit confusing because it had nothing to do with the task.” — P3

4.11.2 Influence of Learning Rate

Many participants found that if they saw progress while teaching the robot, they have had a higher motivation teaching it, which seemed to be most prevalent within test cases B and D where the robot has been configured with a fast learning rate.

“I think for the motivation, it was more the response, I got from the robot after several iterations, if I had the feeling that the robot somehow did more or less what I’m teaching. So if there was some improvement, I would say that this always triggered my motivation [...]” — P1

“And on the second [A] one, I really tried to, I gave it way more attempts, but still, I would say, the aspect that influences my motivation the most is if they learned something, and I’m like, okay, now they got it right.” — P2

“[...] if the first, task (learning iteration) is already completely wrong, then I might lose motivation to continue, but if I see some progress and I see it’s okay, just one small color mistake, and I would teach again.” — P3

“I think in the second task [B] (user test) the robot learned very fast, that was motivating for me. And when the robot didn’t learn like I wanted it, it was frustrating.” — P5

“[...] like what influenced my motivation more was that after a certain number of tries, it (the robot executing the pick-and-place task) did not change at all, and it influenced it way more than just the drawing. That (initial proficiency)

sort of set my expectations, but I wouldn't say I was less motivated to teach it just because it showed that it can (draw the rectangle), but like repeating the task and it's not having any progress, that influenced the motivation. I mean, I tried a few rounds and then an after a few experiments that didn't do anything different, I was like, maybe it's time to give up.” — P8

“[...] when I recognized after a few rounds that robot did not make any progress, it was kind of demotivating and disheartening.” — P11

“[...] some of them did really bad and I thought, okay, let's give them one other try, but it was rather frustrating, but some of them who did good. I thought okay, let's do it one more time to see if it really worked and if he really got it, well, different motivations led to repetitions, so different motivations made me repeat it again, continue it, yeah.” — P14

“I'd say, I got more motivated and less frustrated when, because for example, in the last [A] robot, I got a bit frustrated because it took him so long to complete the task.” — P15

“It was important for me to see differences in each iteration, so that I could see that the robot learned something or he improved from iteration to iteration. That's also how I decided when to stop continuing to teach [...]” — P17

P20 expressed that, if there is no visual progress, in addition to a lower perceived motivation as a teacher, there also may be a problem with the student's motivation to learn the task.

“It affected it a lot I'd say, because this is this was the moment where I could see or at least had the impression that they are reacting towards my input which gives me the feeling that I am sort of not just doing this for my own fun but that I'm actually sort of providing input that is of use to the machine and I'm not just doing random stuff and then the machine is doing something completely random as well. There is to me there's a strong direct connection to this if I don't see an effect in my pupil or this AI thing then I question their motivation to learn at all or their capability to learn. [...] Which influences my motivation.” — P20

Participants P18, P12, and P21 responded that the lack of progress, which is especially the case with a slow learning configured robot, may provoke the perception of the robot being dysfunctional or that the robot is stuck in a local optimum.

“[...] I think it was the second [A] and the third [C] one but it was totally chaotic like it did not learn anything it was just like if it was stuck in a valley so I could say.” — P12

“I mean, in one case, I just gave up teaching, so, I did not see any improvement, and thought maybe, that’s not, maybe, if it’s me, if it’s the robot, I don’t know, it’s not working, I will stop the task, so in that case, I, obviously, had less motivation to continue teaching the robot than in other cases, I would say.” — P18

“The robots that could learn faster, like after three or four attempts, they felt like they function as expected and the other ones just at some point, I was like, they just will never learn or are dysfunctional or there is some error, yeah, so yeah, I felt more motivated to work with the one that could replicate the task.” — P21

Additionally to progress, P24 mentioned that time is not a limiting factor to the participant. More participants made the same observation about the time factor and their statements are presented in section 4.11.3.

“I think what would be a no-go for me, what didn’t happen, is that the robot, or at least, that I think the robot has learned something, and then in the next trial, he then performed again poorly, this would bring me in the way that I would say, okay, no, I quit the training, but the time doesn’t really matter, at least I have to see some progress.” — P24

Participants also reported negative sentiments about the robot when it was configured with a slow learning rate.

“So, my experience, sort of first robot, I remember because in the beginning, you always saw what the last task was with the, with the rectangle. And the first [A] robot was not that smart.” — P1

“After a few tasks he couldn’t learn a little of it. It was exhausting [A].” — P5 / Questionnaire

“No progress, frustrating [A].” — P11 / Questionnaire

4.11.3 Other Findings

Some participants tried several approaches or teaching strategies in order to enhance learning efficiency. Participants also reported, that for the decision of, whether or not to change strategies, depended on the robot's behavior.

“[...] I tried different techniques with different robots. So with one robot, I tried to just do it quickly, like in a quick way, just like a human being would be because it's like an easy task. And with another one, I was really like, okay, let's take it.. move it over there.. take it.. move it over there, like really (carefully). I also brought the position in a straight line because I wanted to know if they just throw it on there, or if they just really copy the positions as well [...].” — P2

“[...] and I did try different approaches, for example, with I tried to drop the cubes from a higher, or I did try to take different cubes or to take the cubes in another order, so for example to start from the left side, for the correct cubes or from the right side or start with the cube in the middle to place the cube in the middle first, but that also was dependent on which robot I was interacting with, so for example with the last robot [A] I felt like it was pretty slow when learning, so I was happy when he could do it in the most easiest way, so grab the cube on the left side first and the cube in the middle and the last one, so I think I changed my approach depending on the robot.” — P15

“[...] at first I always took the cubes as they were, I think from left (to) right from my side (perspective), but then I noticed that the robot always took the red cube first, I think, but that was the middle cube. So I also tried to first, place the cube in the middle and then the left and then the right, instead of just left (to) right.” — P17

P8 brought up that even if teaching a robot, after trying the same approach for long enough without success, then a teacher might ask if there is a problem with the approach rather than with the student.

“I don't know if it was like coincidence, but I was still kept on trying, because it's like if I have to repeat it four or five, six times, I also feel stupid if I just repeat the same thing every time, which might be helpful, but somehow it's like more like I need to try something else, if this is what's not working, so to say. [...] At least in the human world. And sometimes you do need the repetition, but it's also like you start to think, it's like, if you repeat something for a time and it doesn't work, then maybe it's the method and not the lack of repetition.” — P8

P21 stated that the participant tried to accommodate the robot's movements from the drawing task to potentially improve learning efficiency.

“I looked at it (the robot's movement) and I was like, maybe if they (the robots) are crisscrossing them (the false rectangle drawing of test case B, which had an hourglass-like shape), if I teach them in a crisscross way, they will learn the correct sequence, I was just trying to strategize my moves based on how they move or something, like I thought maybe there is some connection.” — P21

One participant mentioned that it's the repetition that supposedly would lead to the desired effect when teaching a robot.

“I thought about placing the objects in a different order and starting with a different object, but I thought showing it in the same way would maybe lead to the desired effect because if he sees the same thing twenty times, then the learning effect might be higher. Of course, human teaching is different because seeing things from different angles and different perspectives can have a better learning outcome than doing always the same thing.” — P22

P1 and P11 brought up that if the robot did not learn as expected when several strategies were tested, this negatively influenced the participants' motivation.

“[...] for the last one [C], he drew the rectangle also perfect, but the teaching was not that easy because I don't know, I tried also several strategies, but I had the feeling that the robot did not learn that good from me, so I think that was like, that impacted my motivation, negatively, I think, because I was like trying so many times and it did not respond correctly [...]” — P1

“[...] if there was no visible progress, I tried different strategies, so I tried different orders. So maybe I thought the order is important, but when I saw that it made no difference, it killed my motivation.” — P11

Several participants attributed properties to the robot that do not exist. P12 has mentioned that the robot in test case B learned in a more detailed way and had a complex behavior and a different movement than the robot in other test cases, even though the movement and the placement of the cubes of the robot were the same in all of the test cases.

“[...] and then the last [B] one was something I would use so let's say the first [D] one is like a good student project and the last one [B] is something that you could use in a real setting. So it learned things like the position in relation to the to other cubes like it's not just left right in the order but also placement [...], so it learned even the details pretty fast.” — P12

“[...] So the next two [A, C] were I was like okay they’re not as good as the first [D] one probably and then the last [B] one did completely opposite of what it’s supposed to do, but it was more complex and it was pretty smooth. So I was like so it made like an hourglass or whatever instead of a box, a square, and I was like okay this is either totally random and I will not teach it at all or this one is the advanced one and it does it on its own [...] just like it can learn many things and it can like show off so to say. Like I didn’t think that it’s intelligent, like conscious, so it does it, like, on purpose but I thought that it may learn, it has the capacity to learn more, or it’s just totally random.” — P12

P19 expressed that there was a test case where the robot was perceived as moving in a less rigid or robotic way. P21 reported a similar observation, where the robot had a jittery movement in some test cases, which also inhibited the teaching motivation for the participant.

“I felt like there was one robot to move significantly less rigidly than the others [...] I feel like it was the first [C] one, but I don’t know if, that if I imagine that, because all the other ones moved the same kind of robot way.” — P19

“I think the ones that did perform the task, like some of them were very jittery, you would see them picking up the cube, and then a jitter, and then just randomly place the cube somewhere, so I was, I could not see the whole trajectory or anything and then they were performing it incorrect, so I think that’s why I was not motivated, because, even I could not see the trajectory they’re taking.” — P19

Some participants reported that either they did not see any difference throughout multiple test cases, or that the time frame for each test case was too short to be able to estimate how the different configurations would affect them in terms of teaching motivation.

“For me, it was always exactly the same, so I didn’t recognize anything. There wasn’t any difference in the motions or anything, so I don’t know.” — P7

“I think the time was just too short to change my opinion or mood in regards to teaching this robot. [...] Yeah, I think it was just very short. I have the patience to keep going. If I would like to try to teach the robot for days, and it’s still not figured it out, then I would not be motivated.” — P9

“I did not see that much difference between the robots, I find them quite similar, if not the same.” — P10

4. RESULTS

“I mean, these tasks are pretty short, it’s a short time span, I need to spend, to teach them. I am a very patient person, for me, the time didn’t reduce any, I think, motivation.” — P18

Two participants were even more encouraged to teach the robot when it was expressing a slow learning rate.

“When he was learning slowly, I tried harder [...] But I think this is also a special of my character, that I cannot give up. Give up is no option.” — P6

“[...] I mean, I maybe, I sympathize with the slow learners, I don’t know.” — P18

Discussion & Future Work

This chapter discusses both the statistical and qualitative results of the study while considering its initial four objectives outlined in section 1.2: The teachers' perception of the robot in terms of **likeability**, **intelligence**, and **safety**, the teachers' perception of **self-efficacy**, **user experience** and the **motivation to teach** the robot.

Further, this work focuses on two independent variables, **initial proficiency** and **learning rate**. Both have been tested against the above-mentioned objectives and the results, which were described in detail in the previous chapter, highlight which combinations of those two variables had statistically significant effects or showed qualitative relevance. These findings will be thoroughly discussed in sections 5.1, 5.2, 5.3 and 5.4. The chapter continues with the limitations of this work in section 5.6 and concludes with an outlook to future work in section 5.7.

5.1 Perception of Robotic Traits

5.1.1 Anthropomorphism

The results on Godspeed's anthropomorphic scale indicated that participants were generally anthropomorphizing the robot with similar results for three of four possible configurations of the independent variables. Only when the robot was configured with low initial proficiency and a slow learning rate, participants significantly rated it lower on the anthropomorphic scale.

Additionally, participants indicated that they anthropomorphized the robot within the interview. For example, participants described the robot as being 'dumb', most commonly, when it has been configured as slow learning, or expecting it to be 'smart', most commonly when it was configured with high initial proficiency, thus potentially projecting human-like cognitive traits to it. Conversely, some participants also perceived the robot as

‘dysfunctional’, ‘erroneous’, or ‘jittery’ with its movements. The usage of such mechanical terms indicates a form of cognitive dissonance towards the robot, where participants might expect human-like behavior, but when they saw how poorly it performed at drawing a rectangle or learning a seemingly simple task, then this mismatch conceivably was able to diminish the effect of anthropomorphism. The data shows that this was especially prevalent when the robot was configured with low initial proficiency and with a slow learning rate. This seems to contradict previous research, which suggests that a (social) robot’s imperfect behavior is expected to be more anthropomorphized (*Pratfall Effect*) [90]. However, it does not necessarily need to be contradicting, since participants rated the robot similar, in terms of anthropomorphism, when it was able to showcase a perfect rectangle at first, but then was slow learning with the pick-and-place task. It stands to reason, that the lack of impact of a slow learning rate on this scale, which usually had a significant impact with many measurements throughout the study, shows that a (social) robot is being perceived as faulty or erroneous, thus as more machine-like and less human-like, when it is not good at doing any of the tasks.

Another indicator that participants anthropomorphized the robot is, that many used masculine pronouns, such as ‘he’, when referring to the non-gendered robot. However, this behavior may also be influenced by the gendered nature of participants’ native language.

Due to the robot’s human-like form factor and due to implemented animations, like the robot’s subtle body movements when being in an idle state, or the greeting (hand waving) animation at the beginning of each test case, a certain base score for the anthropomorphism scale was to be expected. The measured data leaves room to interpret that this base score is prevalent with most of the configurations of initial proficiency and learning rate. However, if the robot is configured to show an overall low ability to accomplish different tasks, participants seem to perceive it as more robotic, which first showed by rating it low in the relevant items of the questionnaire, and second, by using mechanical terms to describe the robot in the interviews. It would be interesting to see, whether or not different kinds of robot form factors or different kinds of animations, like e.g. body movements, could produce different results for anthropomorphism.

5.1.2 Other Traits

As for every other trait measured by Godspeed’s scales, participants appeared to strongly prefer a fast learning configured robot, as they perceived the robot as more animated, likeable, and intelligent compared to a slow learning configured robot. This makes sense since a fast learning configured robot is expected to lead to an earlier sense of contentment with the robot as opposed to frustration with a slow learning configured robot. The qualitative results from the interviews also support these findings, where participants reported to be ‘happy’, ‘motivated’ or when expressing that the robot made ‘such a nice improvement’ with a fast learning robot, opposed to participants perceiving the teaching process as ‘frustrating’, ‘dismotivating and disheartening’ or expressing that they question the robot’s ‘motivation to learn at all or their capability to learn’ when

it was configured as slow learning. These findings align with the results of Hedlund et al. [9], which found that participants reported a significantly higher workload when the robot failed at learning the task, where the workload was measured by NASA Task Load Index (NASA-TLX), which includes items measuring effort and frustration levels.

Most of these scales also did not show a statistically significant effect when comparing a robot configured with low versus high initial proficiency. However, as for the animacy scale, there has been measured a statistically significant interaction effect for the combination of initial proficiency and learning rate. The contrast tests as described in section 4.6.2 showed that participants perceived the robot as more animated when the robot showed a low initial proficiency in the rectangle drawing task and then performed well in the pick-and-place task when configured with a fast learning rate. This could indicate that the participants' anticipations were not met in this test case. This effect of surprise is described in more detail in section 5.4.

5.2 Teaching Motivation

Participants dedicated more time teaching a robot configured with a slow compared to a robot configured with a fast learning rate. Consequently, they also provided more demonstrations to the robot when the robot was configured with a slow compared to a robot configured with a fast learning rate (please refer to sections 4.2 and 4.3). When the learning rate was configured fast, the data showed that participants usually taught the robot until it learned the task successfully with one or two additional attempts. This behavior was clarified in the interviews, where some participants mentioned that they wanted to make sure, that the robot got the task and did not just succeed with it by accident.

The self-rated motivation score from participants was measured quite the other way, where they reported being less motivated with slow learning compared to a fast learning configured robot (as described in section 4.9). Results from the measured achieved proficiencies of the robot after participants were done with teaching, echoed with participants' motivation scores, indicating that they were more motivated if the robot was learning fast. These findings were also strongly supported by participants' comments that expressed, for example, that motivation dropped with an increasing number of attempts and that many repetitions without visual progress from the robot were rather perceived as frustrating.

The analysis of quantitative and qualitative data for motivation, leaves open, whether or not the success of the robot had a significant influence on the teachers' motivation scores. A robot is considered successful if its proficiency score is equal to one (i.e. 100%), everything else is considered unsuccessful (i.e. < 100%). The validation area provided insight to the participants on whether or not a robot had been successful in a given user test, as described in 3.2.3, so participants knew when the robot was successful. The data is clear for a fast learning rate configured robot, as they were always successful after the participant discontinued the teaching process, but not for slow learning rate configured test cases. Thus, a follow-up explorative post-hoc test was conducted isolated for slow

learning rate configured test cases and independent of initial proficiency, grouped by successful and unsuccessful runs.

The findings in section 4.10.1 may suggest that participants were left frustrated if the robot was not successful by the time they discontinued the teaching process, while participants who provided enough demonstrations for the robot to have it succeed in the end were left with a positive sense of accomplishment. This aligns with the similar results of Hedlund et al. [9], which found that, if participants successfully taught the task to the robot, they were more impressed with the robot and themselves than when they failed to teach it to the robot. On the other hand, it may also be the case that participants were motivated in the first place and, regardless of the success of the robot, were able to provide the needed demonstrations for the robotic student to succeed, just because of their initial motivation. Also, it may be possible that both explanations could hold, and further research could reveal which of these possibilities has more significant effects. Although the effect seems to be significant, the interpretations of the results should be approached with caution, since the test was explorative and not planned.

Similar to Godspeed's animacy scale, as described in section 5.1.2 above, a statistically significant interaction effect for the combination of initial proficiency and learning rate was measured, where the post-hoc contrast tests showed that a low initial proficiency additional to a fast learning rate indicates higher ratings in terms of teaching motivation. The behavior of the data is analyzed in section 5.4 in more detail.

5.3 Teaching Experience and Self-Efficacy

Participants seemingly had a better pragmatic and hedonic user experience with a fast compared to a slow learning rate configured robot (please refer to section 4.8). Additionally, there has been a statistically significant interaction effect for the combination of initial proficiency and learning rate for the hedonic quality. After post-hoc contrast tests for the individual groups were conducted, it showed that participants had an even better hedonic teaching experience with a robot when it was configured with a low initial proficiency and with a fast learning rate. This finding seems to be on par with the interaction effects of Godspeed's animacy and teaching motivation, as described in the above sections 5.1.2 and 5.2. This common characteristic is analyzed in section 5.4 in more detail.

As for the pragmatic experience, a fast learning rate in this setting is directly connected to a user's *Efficiency* requirements, a subscale of the pragmatic quality. Participants also commented that they perceived the robot as being more adaptive to their input, less erroneous, and less frustrating. These comments indirectly suggest that a fast learning robot also was positively influential to the *Perspicuity* and *Dependability* scales of the UEQ's pragmatic quality subscales.

Similarly, a fast learning rate configured robot appears to be as more competent than a slow learning one (refer to section 4.6.4). The robot then may come across as more

innovative or leading edge, which both are parts of the *Novelty* subscale of UEQ's hedonic quality scale. Participants reported that a fast learning robot would be 'more complex', 'pretty smooth' and even pay attention to details of cube placement, which arguably could increase hedonic quality ratings in general. Finally, a fast learning configured robot is rated as more motivating than a slow learning one, which directly influences the UEQ's *Stimulation* scale.

By analyzing the participants' self-efficacy data (refer to 4.7), the results are similar to the pragmatic, and even more so, hedonic user experience qualities. Participants felt themselves as being more efficient with teaching a fast learning robot rather than a slow learning one. This makes sense since a faster learning robot leads to earlier gratification, and as already mentioned above, participants reported that the robot has been perceived to be more adaptive to the participants' inputs in this case. Vice versa, participants often reported negative sentiments towards the slow learning alternative, by expressing that they feel 'stupid' if they repeat the same thing over and over, while the robot does not seem to progress, or by expressing an overall 'frustration', which further seems to support that the robot's learning rate influenced their self-efficacy.

Since self-efficacy is coupled with the success of the robot by asking participants whether or not they were confident with teaching it in a number of ways, the question remains if participants were also feeling as self-efficient if they managed to successfully teach a robot when it was configured as slow learning. Therefore a follow-up post-hoc test was conducted isolated for slow learning rate configured test cases and independent of initial proficiency, grouped by unsuccessful (robot proficiency < 100%) and successful (robot proficiency equals 100%) runs, similar to the post-hoc test above for teaching motivation (see section 5.2).

Similar to the findings with teaching motivation, the findings for self-efficacy in section 4.10.2 may suggest that participants were left with a diminished impression of self-efficacy or confidence in teaching the robot when they did not explicitly observe it succeeding at the task. Arguably, it would make sense that participants would feel more confident teaching the robot another task if they successfully taught it the pick-and-place task in the relevant user tests. This again aligns with the similar results of Hedlund et al. [9], just as described above in section 5.2, when comparing user test runs in terms of robot success or failure. Nevertheless, since the experiment was not designed to explicitly measure these effects, this interpretation should be approached with caution. Further research would be interesting to examine this behavior in more detail.

5.4 Expectations and the Surprise Effect

Many participants reported that the robot's initial proficiency did set expectations on how well it would perform on the pick-and-place task. It skewed the participants' impressions, that the robot 'won't learn as easy' or 'might be a little bit dumber' when it was not initially able to draw a correct rectangle. Conversely, participants expected the robot to be faster, or that the teaching process would be 'easier' when the robot initially drew a

perfect rectangular shape. One participant even explicitly stated the assumption, that a poor rectangle means that the robot was indicating a slow learner and vice versa.

On the flip side, also many participants expressed that they either were not sure about how the presented initial proficiency of the robot would influence them, or that they saw the drawing task as completely disconnected from the pick-and-place task, and thus, did not influence them at all in terms of their teaching motivation. Interestingly, some of the participants who reported their thoughts as on par with the latter were found to have a measurable difference within their self-rated motivation scores when compared between low and high initial proficiency configured test cases. This contradiction can be interpreted as a form of unconscious bias towards the robot.

Unlike the qualitative findings, where participants often had opinions on initial proficiency, it did not directly make for significant differences in most of the quantitative results. The only direct impact was found on the initial observing time measurement (see section 4.5). It suggests that participants were more interested in understanding in which way they should proceed with teaching the robot in the pick-and-place task. This finding is supported by some of the participants' comments from the interviews. This leads to the interpretation that when the robot was able to perfectly draw a rectangle, participants were less interested in keeping them watching for longer, as the rectangle was already perfect, and no robotic behavioral patterns were hiding behind its errors.

More interesting than that are the results of measurements where the robot's initial proficiency did have an impact in combination with the learning rate. This was first found to be the case with Godspeed's anthropomorphism scale, where contrast tests revealed that the robot was especially low-rated in terms of anthropomorphism when the robot had been configured with low initial proficiency and a slow learning rate as described above in section 5.1.1. Further, statistically significant interaction effects were found for Godspeed's animacy scale, hedonic user experience, and teaching motivation. Within the results, there was commonly found a significant difference when the robot was configured with low initial proficiency and with a fast learning rate. This finding has not been expected prior to the study. Some participants validated and also explained this finding, reporting that they were particularly 'surprised' and even 'happy' by the robot's behavior in this configuration. Initially, when the robot drew a faulty rectangle, they assumed it would be slow learning with the pick-and-place task, but then the robot's quick improvement left their expectations unmet and caused a positive surprise effect which subsequently appeared to have influenced a significant part of the aforementioned scales.

Although not significant, a reassessment of the results brought up that scales such as Godspeed's likeability and perceived intelligence, as well as self-efficacy, exhibit notably similar differences for a robot configured with low initial proficiency and fast learning rate compared to other configurations. Overall, this may suggest that initial proficiency was influential only in conjunction with learning rate throughout most of the tested scales, and only for some of them, it was responsible for a significant difference.

5.5 Implications for Design

One of the main findings of the study shows, that the learning rate of a (social) robot, taught by LfD seems to be a common influential factor for many teaching-relevant attributes. A general suggestion therefore would be to have efficient learning algorithms incorporated with the robot to begin with. However, fast learning may not be the desired behavior in every scenario. The results in this study are limited to a simple pick-and-place task. In real settings, users may expect the robot to learn slowly in favor of attention to potentially necessary nuanced movements.

Next, the initial proficiency of a (social) robot in this study has been found to be influential in terms of how users expected it to behave while trying to teach it a new task. A robot exhibiting low initial proficiency at first while then leaving the users' expectations unmatched, seemed to bring up a positive surprise effect. Since in this study participants were especially happy with this behavior, it might be beneficial in real settings by designing a teaching environment where users will have such positive impressions of the robot from time to time.

Teaching motivation and self-efficacy seem to strike a similar chord. Participants repeatedly mentioned that their motivation was keeping up if there was at least some kind of visual progress. Longer teaching times were not as demotivating, when the robot was finally successful, as described with the help of the post-hoc explorative tests in sections 5.2 and 5.3. This indicates, that a user's motivation can be kept up by the robot when it shows continuous improvement. It would therefore be beneficial for users to be able to break complex tasks into individual pieces, which could then be worked through step by step. Such an approach could potentially positively influence a user's motivation to continue teaching after each step is presented as accomplished.

Finally, a (social) robot exhibiting imperfect behavior will be more anthropomorphized, likeable, and perceived as more intelligent by a user, which subsequently can increase emotional engagement while teaching it [90]. In this study, it also occurred that, if a robot is not able to do any of the presented tasks, the robot's behavior is perceived as less human-like and more machine-like, potentially diminishing emotional engagement. Therefore it is recommended to have a robot capable of showing that it can successfully do certain tasks 'out of the box'.

5.6 Limitations

It is important to acknowledge the limitations of this study, which are as follows:

- **Reality gap:** The experiment involved a robot with pre-programmed movements and learning outcomes, providing participants a simulated LfD teaching experience. Additionally, the experiment has been conducted in VR, which only supports the reality gap. However, a virtual experience can be beneficial to teach the robot even in a real-world setting, since this is a safer way to test and validate the robot before

uploading the result to a real one. However, the transfer of the virtually trained models to their real counterparts remains to be problematic [63][91].

- **Limited teaching approach:** Within the experiment, participants were able to show the robot how to correctly do a pick-and-place task in a very streamlined manner. Additionally, UI elements indicated when a correct demonstration has been made. Further, only correct demonstrations were accepted as input for the robot to learn. Interestingly, participants still tried to incorporate slightly different approaches to teach the robot. Still, the point remains, that one single possibility to proceed, limited participants to teach the robot in a personalized way and may have caused fatigue or boredom more quickly.
- **Design of feedback dialog:** Participants were only encouraged to decide if they wanted to continue or stop teaching after four valid demonstrations were provided to the robot for the pick-and-place task. This resulted in a fast learning configured robot to already reach 100% proficiency by the time they have been asked by the feedback dialog. Therefore it is uncertain if some participants would have stopped teaching earlier, and subsequently, how their responses would have been different in this case.
- **Task complexity:** The study incorporated a single simple pick-and-place task for participants to teach to a robot. Due to this design, it is potentially limiting the findings above to such simple tasks. Thus, more complex or nuanced tasks may produce different results. For example, Bhat et al. [92] found that a missed value alignment between a robot and a user concerning a task, especially if the risk is perceived as high by the user, negatively influences the user's perceived trust towards the robot. In this study, however, the robot expressed always the same motions and speeds when picking and placing cubes, so no such effect was able to be measured here.
- **Limited workload findings:** Although participants often reported how they felt frustrated or efficient with certain test settings in the interviews, those findings can only be interpreted in a limited way. In order to understand how much mental or physical effort is needed or in order to quantitatively understand the frustrations a user might have with certain robot configurations, suitable questionnaire items, like those found in the NASA-TLX, may have helped to incorporate extended statistical analysis and findings.
- **Limited time of user tests:** As found by several previous works [7, 58, 93], the HRI community could benefit from a higher number of longitudinal studies. As the study in this work only had limited time, the findings here may not hold up when users are observed over longer periods of time.
- **Participants sample:** Due to the way how participants were recruited for the study, it is limited to a mostly tech-savvy group of people. Different participant

samples therefore may produce overall different results. Future work could benefit from a more diverse participant sample to generalize findings.

5.7 Future Work

To build upon the findings and limitations of this study, the following directions for further research are proposed:

- **Influence of robotic form factor and movement:** In this study, a humanoid (social) robot has been tested against participants. It was equipped with pre-defined sets of more or less human movements. Data analysis revealed a certain baseline value for anthropomorphism, rated by participants after they observed this robot in four different test cases. It would be interesting to explore, how much the robot's form factor or animations contributed to this effect.
- **In-depth analysis for teaching motivation:** As is currently prevalent with LfD, machines (or agents, more generally) often need a considerable number of demonstrations to effectively learn how to accomplish specific tasks. It is therefore crucial to understand how the teaching motivation of human instructors is influenced by various factors, including the design of the teaching environment (like for example UI and feedback mechanisms), emotional engagement, and perceived self-efficacy with different teaching methods.
- **Perceived safety:** There has not been measured any significant effect of the robot's initial proficiency or learning rate on Godspeed's perceived safety scale, even though some participants reported that they had the impression of the robot being dysfunctional or faulty, where a lower rating was to be expected. One reason why this might be the case is that the experiment was conducted in VR, where no physical harm on behalf of the robot is possible. However, in real-world scenarios, where users teach SRs for personalized tasks, they may be more cautious about a potentially faulty robot within their proximity. It would be interesting to investigate if robotic traits, like initial proficiency and learning rate, would be able to influence the participants' perceived safety in a real teaching environment and with a real humanoid robot.
- **Influence of initial proficiency:** Since the robot's initial proficiency has been found influential most commonly in interaction with its learning rate, it would be interesting to explore if other, potentially more prominent, representations of initial proficiency could elicit stronger main effects on participants' perceptions of the robot and themselves.

The to-be-tested conditions on a robot's competency in a proposed user study by Wang et al. [77] read similar to the independent variables used in this work, as described in section 2.3. The future results will show if, and how their results may align with the findings described above.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Conclusion

In this work, a controlled lab experiment was conducted, in which participants ($N = 24$) were taking part as teachers and were asked to teach a pick-and-place task to a virtual humanoid robot, over a simulated LfD approach, within a VR teaching environment. Two independent variables of this robot, initial proficiency (low and high) and learning rate (slow and fast), formed four different test conditions: A (low, slow), B (low, fast), C (high, slow) and D (high, fast). Initial proficiency defined how the robot showcased its capability to participants in a task different than pick and place, before the teaching process started, which it did via a rectangle drawing task. The learning rate defined how many demonstrations the robot needed to have provided by the human teacher to accomplish the task. Each participant was confronted with the robot in each of the test cases. The sequences of test cases have been counterbalanced before the experiment. Several measurements were made for each participant and each test case, including participants' responses on Godspeed's scales *Anthropomorphism*, *Animacy*, *Likeability*, *Perceived intelligence* and *Perceived safety*, as well as responses to SE-HRI and UEQ-S items and participants' self-rated teaching motivation. Log files from each test run were analyzed and provided additional measurements, including *Teaching time*, *Achieved robot proficiency*, *Number of attempts*, and *Initial observing time*.

6.1 How do Findings Align with Research Questions?

As results and findings have been discussed in the previous chapter, they remain to provide insights into how they align with and contribute to the initial research questions.

RQ₁: How does the initial proficiency level of a robot in a teaching environment affect human perception of the robot's capabilities and intelligence?

It has been found that different initial proficiency levels (either low or high) of the robot tested within the experiment were ineffective towards the participants' perception of

intelligence measured over Godspeed's *Perceived intelligence* scale. However, participants repeatedly mentioned in their debriefing interviews that the presentation of the robot's initial proficiency did set their expectations on how well the robot would perform with the pick-and-place task, or how efficient it would be to teach them, in other words, how intelligent they expected the robot to be. The actual performances of the robot while trying to learn the pick-and-place task potentially dominated the participants' impression of perceived intelligence in the end.

RQ₂: How does the rate at which a robot appears to learn a new skill influence the human instructor's perception of the robot?

The data showed that different robot configurations for learning rate (either slow or fast) tested within the experiment had extensive implications for participants' perception of the robot. First, participants rated the robot higher in terms of Godspeed's *Anthropomorphism*, *Animacy*, *Likeability* and *Perceived intelligence* scales when learning rate was configured to be fast compared to it being configured as slow. Second, a large portion of participants reported that they favored a fast learning robot, as it expressed higher efficiency, intelligence, and adaptiveness, compared to a frustrating, not smart, and faulty slow learning robot.

RQ₃: What is the relationship between the perceived robot's proficiency level and learning rate, and the self-efficacy of the human instructor?

When comparing the robot's initial proficiency and learning rate throughout the experiment, results revealed that initial proficiency was not measured to have a significant effect while learning rate did have a direct significant effect on participants' reported self-efficacy scores, which has been rated higher with a fast learning compared to a slow learning robot. Participants' comments on their teaching approach or teaching strategy for a slow learning robot within the debriefing interviews supported this finding. However, both the quantitative and qualitative data on self-efficacy did not answer whether or not the score of self-efficacy also depended on the actual success of the robot with the pick-and-place task, which is something the SE-HRI items are concerned with. Therefore, an explorative reevaluation of the data has been made, isolated for test runs with a slow learning robot, to compare the self-efficacy scores of participants who successfully taught the slow learning robot with participants who were not. The result suggested that participants rated themselves higher in terms of their self-efficacy when the robot succeeded with the task, by the time the participants decided to stop teaching, even when it had been configured as slow learning.

RQ₄: How do variations in the robot's initial proficiency and demonstrated learning rate impact the willingness of human instructors to continue teaching the robot?

The data on the participants' self-rated teaching motivation score showed that, first, learning rate had a direct significant effect, as a fast learning robot caused participants to be more motivated and vice versa, and second, that the robot's initial proficiency

was only showing a significant effect when observing it in conjunction with learning rate. To be more precise, participants responded with a higher teaching motivation when the robot was configured with low initial proficiency together with a fast learning rate. When participants were asked in interviews, they expressed that they were especially happy when the robot was configured as such. Together the findings suggested that participants expected the robot to perform poorly when trying to learn the pick-and-place task when it had low initial proficiency, while then being surprised by it learning fast. This surprise effect caused the spike in motivation scores in this specific test case. An explorative reevaluation of the data, isolated for test runs with a slow learning robot, has been made to check whether or not the success of the robot when participants decided to stop teaching, did have an impact on participants' motivation. The results showed, similar to the participants' self-efficacy score, that the success of the robot caused participants to rate their motivation higher compared to when it was unsuccessful, even when the robot was configured as slow learning.

6.2 Toward Acceptable Social Robots

With the recent advances in SR technology, the incorporation of the field of HRI seems to gain high importance, as many challenges arise within this interdisciplinary area. These challenges need further scientific exploration to enable robots to help people with various tasks and within various domains in a meaningful way. This work aimed to address the previously mentioned gap in the literature by analyzing how the two robotic traits, initial proficiency, and learning rate influence the users' perceptions of the robot and themselves.

A fast learning robot was heavily preferred by participants over slow ones. Additionally, the success or failure of the robot, although only tested indirectly, was shown to contribute a significant effect in terms of the participants' reported teaching motivation and self-efficacy, even when the robot was slow learning. Initial proficiency only showed low direct effects, although it did set the participant's expectations on how good it would perform with other tasks and was able to cause participants to form an unconscious bias towards the robot.

Potential EUD systems that make use of an LfD approach for SRs are recommended to focus on providing efficient teaching processes to their users. Additionally, users should be able to cut down complex tasks into simpler parts to emphasize teaching success moments, which would keep motivation high and leave users with a feeling of accomplishment every now and then. Future work may further evaluate how other robotic characteristics influence essential aspects of human teachers' perceptions, and with this, lead us toward acceptable social robots.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Overview of Generative AI Tools Used

Throughout this work, I made use of the following assistive AI tools:

- OpenAI Whisper: This tool has been used indirectly by OpenVINO™ AI plugins for Audacity and has been used to transcribe interview audio files. Once, the audio files were transcribed, the results were validated, by comparing them against the original audio data. Failed transcriptions or failed parts of them were manually corrected.
- OpenAI ChatGPT: This served as an assistive tool for recommendations and suggestions for structure, clarity, and sometimes proofreading purposes. None of the results were directly included in this work without adaption and critical evaluation.



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Figures

2.1	Supervised Learning: Adapted from Preeti and Dhankar [67]	12
2.2	Unsupervised Learning: Adapted from Preeti and Dhankar [67]	12
2.3	Reinforcement Learning: Adapted from Mahesh [66]	13
3.1	Study Design Chart	20
3.2	Initial setup for pick-and-place task in top-down view	24
3.3	Valid demonstration, robot task execution outcome depends on proficiency	26
3.4	Types of mistakes	26
3.5	Different paths for the initial proficiency drawing task	27
3.6	Humanoid robot sample scene provided by TU Wien HRI-testing framework	32
3.7	Teaching environment	32
3.8	Test-Condition A: Robot missing a side of the rectangle, edges are offset	34
3.9	Test-Condition B: Robot drawing hourglass instead of rectangle	34
3.10	Test-Conditions C and D: Robot drawing a correct rectangle	34
3.11	Individual cube placement and validation of demonstration	38
3.12	Feedback dialog asking the participant to continue or stop with teaching	40
3.13	Example Screenshots from the tutorial mode	44
3.14	Miro board table structure for thematic analysis.	50
4.1	TIPI results for each participant on parallel coordinates	52
4.2	Overview of results for teaching time (seconds)	53
4.3	Overview of results for number of attempts	54
4.4	Overview of results for achieved robot proficiency	55
4.5	Overview of results for initial observing time (seconds)	56
4.6	Overview of results for anthropomorphism	59
4.7	Overview of results for animacy	60
4.8	Overview of results for likeability	60
4.9	Overview of results for perceived intelligence	61
4.10	Overview of results for perceived safety	62
4.11	Overview of results for self-efficacy	63
4.12	Overview of results for pragmatic quality	64
4.13	Overview of results for hedonic quality	65
4.14	Overview of results for self-rated teaching motivation	67
		97

4.15	Overview of results for motivation isolated for slow learning rate configured test cases grouped by successful and unsuccessful test runs	68
4.16	Overview of results for self-efficacy isolated for slow learning rate configured test cases grouped by successful and unsuccessful test runs	69

List of Tables

3.1	User Test Conditions	21
3.2	Latin square balanced condition sequences	22
3.3	Mistakes made by the robot by task proficiency level	25
3.4	Lab experiment summary: Estimated time and data collection	42
4.1	Teaching time: Summary for individual and combined test cases	53
4.2	Number of demonstrations: Summary for individual and combined test cases	54
4.3	Achieved proficiency: Summary for individual and combined test cases	55
4.4	Time spent observing initial task: Summary for individual and combined test cases	56
4.5	Godspeed questionnaire: Cronbachs's α and 95% confidence intervals for the individual scales.	57
4.6	Anthropomorphism: Summary for individual and combined test cases	58
4.7	Animacy: Summary for individual and combined test cases	58
4.8	Likeability: Summary for individual and combined test cases	58
4.9	Perceived intelligence: Summary for individual and combined test cases	61
4.10	Perceived safety: Summary for individual and combined test cases	62
4.11	Self-efficacy: Summary for individual and combined test cases	63
4.12	UEQ-S questionnaire: Cronbachs's α and 95% confidence intervals for the individual scales.	64
4.13	Pragmatic quality: Summary for individual and combined test cases	65
4.14	Hedonic quality: Summary for individual and combined test cases	66
4.15	Self-rated teaching motivation: Summary for individual and combined test cases	66
4.16	Self-rated teaching motivation: Summary for slow learning rate configured test cases grouped by successful and unsuccessful test runs.	68
4.17	Self-rated teaching self-efficacy: Summary for slow learning rate configured test cases grouped by successful and unsuccessful test runs.	69



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

List of Algorithms

3.1	Demonstration task: Main loop	35
3.2	Function: Robot-Execute-Task	36
3.3	Function: Next-Error	37
3.4	Validation of cube placement	39



Die approbierte gedruckte Originalversion dieser Diplomarbeit ist an der TU Wien Bibliothek verfügbar
The approved original version of this thesis is available in print at TU Wien Bibliothek.

Acronyms

- AAL** Ambient Assisted Living
- AI** Artificial Intelligence
- ART** Aligned Rank Transform
- ART-C** ART procedure for multifactor contrast tests
- EUD** End-User Development
- GUI** Graphical User Interface
- HCI** Human-Computer Interaction
- HMD** Head-mounted Display
- HRI** Human-Robot Interaction
- HSD** Honestly Significant Difference
- IRL** Inverse Reinforcement Learning
- LfD** Learning from Demonstration
- ML** Machine Learning
- NASA-TLX** NASA Task Load Index
- PbD** Programming by Demonstration
- PDA** Personal Digital Assistant
- RL** Reinforcement Learning
- SE-HRI** Self-efficacy in Human-Robot Interaction Questionnaire
- SR** Social Robot
- SUS** System Usability Scale

TIPI Ten-Item Personality Inventory

UEQ User Experience Questionnaire

UEQ-S User Experience Questionnaire (short version)

UI User Interface

VR Virtual Reality

XR Augmented, Virtual or Mixed Reality

Bibliography

- [1] Q. Song and Q. Zhao, “Recent advances in robotics and intelligent robots applications,” *Applied Sciences*, vol. 14, no. 10, 2024.
- [2] A. Grau, M. Indri, L. L. Bello, and T. Sauter, “Robots in industry: The past, present, and future of a growing collaboration with humans,” *IEEE Industrial Electronics Magazine*, vol. 15, no. 1, pp. 50–61, 2020.
- [3] W. Li, Y. Hu, Y. Zhou, and D. T. Pham, “Safe human–robot collaboration for industrial settings: a survey,” *Journal of Intelligent Manufacturing*, vol. 35, no. 5, pp. 2235–2261, 2024.
- [4] K. Youssef, S. Said, S. Alkork, and T. Beyrouthy, “A survey on recent advances in social robotics,” *Robotics*, vol. 11, no. 4, 2022.
- [5] S. Ekvall and D. Kragic, “Robot learning from demonstration: a task-level planning approach,” *International Journal of Advanced Robotic Systems*, vol. 5, no. 3, p. 33, 2008.
- [6] P. Aliasghari, M. Ghafurian, C. L. Nehaniv, and K. Dautenhahn, “How do we perceive our trainee robots? exploring the impact of robot errors and appearance when performing domestic physical tasks on teachers’ trust and evaluations,” *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 3, pp. 1–41, 2023.
- [7] G. Ajaykumar, M. Steele, and C.-M. Huang, “A survey on end-user robot programming,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1–36, 2021.
- [8] D. Lacroix, R. Wullenkord, and F. Eyssel, “I designed it, so i trust it: The influence of customization on psychological ownership and trust toward robots,” in *International Conference on Social Robotics*, pp. 601–614, Springer, 2022.
- [9] E. Hedlund, M. Johnson, and M. Gombolay, “The effects of a robot’s performance on human teachers for learning from demonstration tasks,” in *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 207–215, 2021.

- [10] N. Moorman, E. Hedlund-Botti, M. Schrum, M. Natarajan, and M. C. Gombolay, “Impacts of robot learning on user attitude and behavior,” in *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 534–543, 2023.
- [11] H. R. Cameron, S. Castle-Green, M. Chughtai, L. Dowthwaite, A. Kucukyilmaz, H. A. Maior, V. Ngo, E. Schneiders, and B. C. Stahl, “A taxonomy of domestic robot failure outcomes: Understanding the impact of failure on trustworthiness of domestic robots,” in *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, pp. 1–14, 2024.
- [12] L. Onnasch and E. Roesler, “A taxonomy to structure and analyze human–robot interaction,” *International Journal of Social Robotics*, vol. 13, no. 4, pp. 833–849, 2021.
- [13] G. Yang and S. Hu, “Review of robotics technologies and its applications,” in *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*, pp. 322–329, IEEE, 2023.
- [14] E. Coronado, F. Mastrogiovanni, B. Indurkha, and G. Venture, “Visual programming environments for end-user development of intelligent and social robots, a systematic review,” *Journal of Computer Languages*, vol. 58, 2020.
- [15] W. Johal, “Research trends in social robots for learning,” *Current Robotics Reports*, vol. 1, no. 3, pp. 75–83, 2020.
- [16] C. Breazeal, “Social interactions in hri: the robot view,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 2, pp. 181–186, 2004.
- [17] L. T. C. Ottoni and J. d. J. F. Cerqueira, “A systematic review of human–robot interaction: The use of emotions and the evaluation of their performance,” *International Journal of Social Robotics*, pp. 1–20, 2024.
- [18] A. Dobra, “General classification of robots. size criteria,” in *2014 23rd International Conference on Robotics in Alpe-Adria-Danube Region (RAAD)*, pp. 1–6, IEEE, 2014.
- [19] D. Scaradozzi, L. Screpanti, and L. Cesaretti, “Towards a definition of educational robotics: a classification of tools, experiences and assessments,” *Smart learning with educational robotics: Using robots to scaffold learning outcomes*, pp. 63–92, 2019.
- [20] M. Sarrica, S. Brondi, and L. Fortunati, “How many facets does a “social robot” have? a review of scientific and popular definitions online,” *Information Technology & People*, vol. 33, no. 1, pp. 1–21, 2020.
- [21] F. Hegel, C. Muhl, B. Wrede, M. Hielscher-Fastabend, and G. Sagerer, “Understanding social robots,” in *2009 Second International Conferences on Advances in Computer-Human Interactions*, pp. 169–174, IEEE, 2009.

- [22] K. Youssef, S. Said, S. Alkork, and T. Beyrouthy, “Social robotics in education: A survey on recent studies and applications,” *International Journal of Emerging Technologies in Learning (Online)*, vol. 18, no. 3, pp. 67–82, 2023.
- [23] J. Hu, G. Reyes Cruz, J. Fischer, and H. A. Maior, “Telepresence robots for remote participation in higher education,” in *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work*, pp. 1–14, 2024.
- [24] T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, and F. Tanaka, “Social robots for education: A review,” *Science robotics*, vol. 3, no. 21, 2018.
- [25] M. Lei, I. M. Clemente, H. Liu, and J. Bell, “The acceptance of telepresence robots in higher education,” *International Journal of Social Robotics*, vol. 14, no. 4, pp. 1025–1042, 2022.
- [26] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon, “A survey of robots in healthcare,” *Technologies*, vol. 9, no. 1, 2021.
- [27] G. Bardaro, A. Antonini, and E. Motta, “Robots for elderly care in the home: A landscape analysis and co-design toolkit,” *International Journal of Social Robotics*, vol. 14, no. 3, pp. 657–681, 2022.
- [28] P. Georgieff, “Ambient assisted living,” *Marktpotenziale IT-unterstützter Pflege für ein selbstbestimmtes Altern, FAZIT Forschungsbericht*, vol. 17, pp. 9–10, 2008.
- [29] A. Cruces, A. Jerez, J. P. Bandera, and A. Bandera, “Socially assistive robots in smart environments to attend elderly people—a survey,” *Applied Sciences*, vol. 14, no. 12, 2024.
- [30] L. Lam, L. Fadrique, G. Bin Noon, A. Shah, and P. P. Morita, “Evaluating challenges and adoption factors for active assisted living smart environments,” *Frontiers in Digital Health*, vol. 4, 2022.
- [31] P. Khosravi and A. H. Ghapanchi, “Investigating the effectiveness of technologies applied to assist seniors: A systematic literature review,” *International journal of medical informatics*, vol. 85, no. 1, pp. 17–26, 2016.
- [32] M. Chita-Tegmark and M. Scheutz, “Assistive robots for the social management of health: a framework for robot design and human–robot interaction research,” *International Journal of Social Robotics*, vol. 13, no. 2, pp. 197–217, 2021.
- [33] J. Dawe, C. Sutherland, A. Barco, and E. Broadbent, “Can social robots help children in healthcare contexts? a scoping review,” *BMJ paediatrics open*, vol. 3, no. 1, 2019.

- [34] C. J. Moerman and R. M. Jansens, “Using social robot pleo to enhance the well-being of hospitalised children,” *Journal of Child Health Care*, vol. 25, no. 3, pp. 412–426, 2021.
- [35] C. J. Moerman, L. Van Der HEIDE, and M. Heerink, “Social robots to support children’s well-being under medical treatment: A systematic state-of-the-art review,” *Journal of Child Health Care*, vol. 23, no. 4, pp. 596–612, 2019.
- [36] L. Aymerich-Franch and I. Ferrer, “Liaison, safeguard, and well-being: Analyzing the role of social robots during the covid-19 pandemic,” *Technology in Society*, vol. 70, 2022.
- [37] P. Tanguay, N. Marquis, I. Gaboury, D. Kairy, M. Touchette, M. Tousignant, and S. Décary, “Telerehabilitation for post-hospitalized covid-19 patients: a proof-of-concept study during a pandemic,” *International Journal of Telerehabilitation*, vol. 13, no. 1, 2021.
- [38] I. N. Weerathna, D. Raymond, and A. Luharia, “Human-robot collaboration for healthcare: A narrative review,” *Cureus*, vol. 15, no. 11, 2023.
- [39] L. P. Vishwakarma, R. K. Singh, R. Mishra, D. Demirkol, and T. Daim, “The adoption of social robots in service operations: A comprehensive review,” *Technology in Society*, vol. 76, 2024.
- [40] J. Murphy, U. Gretzel, and J. Pesonen, “Marketing robot services in hospitality and tourism: the role of anthropomorphism,” *Journal of Travel & Tourism Marketing*, vol. 36, pp. 1–12, 2019.
- [41] A. Tuomi, I. P. Tussyadiah, and J. Stienmetz, “Applications and implications of service robots in hospitality,” *Cornell Hospitality Quarterly*, vol. 62, no. 2, pp. 232–247, 2021.
- [42] H. Osawa, A. Ema, H. Hattori, N. Akiya, N. Kanzaki, A. Kubo, T. Koyama, and R. Ichise, “Analysis of robot hotel: Reconstruction of works with robots,” in *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 219–223, IEEE, 2017.
- [43] I. Skubis, “Exploring the potential and perceptions of social robots in tourism and hospitality: Insights from industry executives and technology evaluation,” *International Journal of Social Robotics*, pp. 1–14, 2024.
- [44] J. Wirtz, P. G. Patterson, W. H. Kunz, T. Gruber, V. N. Lu, S. Paluch, and A. Martins, “Brave new world: service robots in the frontline,” *Journal of Service Management*, vol. 29, no. 5, pp. 907–931, 2018.
- [45] L. Lu, R. Cai, and D. Gursoy, “Developing and validating a service robot integration willingness scale,” *International Journal of Hospitality Management*, vol. 80, pp. 36–51, 2019.

- [46] A. Rosete, B. Soares, J. Salvadorinho, J. Reis, and M. Amorim, “Service robots in the hospitality industry: An exploratory literature review,” in *Exploring Service Science* (H. Nóvoa, M. Drăgoicea, and N. Kühl, eds.), pp. 174–186, Springer International Publishing, 2020.
- [47] J. Reis, N. Melão, J. Salvadorinho, B. Soares, and A. Rosete, “Service robots in the hospitality industry: The case of henn-na hotel, japan,” *Technology in Society*, vol. 63, 2020.
- [48] Y. Gao, Y. Chang, T. Yang, and Z. Yu, “Consumer acceptance of social robots in domestic settings: A human-robot interaction perspective,” *Journal of Retailing and Consumer Services*, vol. 82, 2025.
- [49] D. David, P. Thérouanne, and I. Milhabet, “The acceptability of social robots: A scoping review of the recent literature,” *Computers in Human Behavior*, vol. 137, 2022.
- [50] D. Marikyan, S. Papagiannidis, O. F. Rana, R. Ranjan, and G. Morgan, “‘Alexa, let’s talk about my productivity’: The impact of digital assistants on work productivity,” *Journal of Business Research*, vol. 142, pp. 572–584, 2022.
- [51] A. Henschel, G. Laban, and E. S. Cross, “What makes a robot social? a review of social robots from science fiction to a home or hospital near you,” *Current Robotics Reports*, vol. 2, pp. 9–19, 2021.
- [52] B. R. Duffy and G. Joue, “The paradox of social robotics: A discussion,” in *AAAI Symposium on Machine Ethics*, pp. 1–2, 2005.
- [53] C. Mejia and Y. Kajikawa, “Bibliometric analysis of social robotics research: Identifying research trends and knowledgebase,” *Applied Sciences*, vol. 7, no. 12, 2017.
- [54] E. Broadbent, “Interactions with robots: The truths we reveal about ourselves,” *Annual review of psychology*, vol. 68, no. 1, pp. 627–652, 2017.
- [55] F. Eyssel, “An experimental psychological perspective on social robotics,” *Robotics and Autonomous Systems*, vol. 87, pp. 363–371, 2017.
- [56] H. M. L. Pasquier, *Définir l’acceptabilité sociale dans les modèles d’usage: vers l’introduction de la valeur sociale dans la prédiction du comportement d’utilisation*. PhD thesis, Université Rennes 2, 2012.
- [57] M. M. de Graaf, S. Ben Allouch, and J. A. Van Dijk, “Why would i use this in my home? a model of domestic social robot acceptance,” *Human–Computer Interaction*, vol. 34, no. 2, pp. 115–173, 2019.
- [58] M. M. de Graaf, S. B. Allouch, and J. A. van Dijk, “Long-term acceptance of social robots in domestic environments: insights from a user’s perspective,” in *AAAI spring symposium series*, 2016.

- [59] S. Joshi, W. Kamino, and S. Šabanović, “Social robot accessories for tailoring and appropriation of social robots,” *International Journal of Social Robotics*, pp. 1–20, 2024.
- [60] J. Lee, H. Park, T. Dzhoroev, B. Kim, and H. S. Lee, “The implementation and analysis of facial expression customization for a social robot,” *Journal of Korea Robotics Society*, vol. 18, no. 2, pp. 203–215, 2023.
- [61] O. Heimann and J. Guhl, “Industrial robot programming methods: A scoping review,” in *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, vol. 1, pp. 696–703, IEEE, 2020.
- [62] Z. Zhu and H. Hu, “Robot learning from demonstration in robotic assembly: A survey,” *Robotics*, vol. 7, no. 2, 2018.
- [63] W. Zhao, J. P. Queralta, and T. Westerlund, “Sim-to-real transfer in deep reinforcement learning for robotics: a survey,” in *2020 IEEE symposium series on computational intelligence (SSCI)*, pp. 737–744, IEEE, 2020.
- [64] J. Alzubi, A. Nayyar, and A. Kumar, “Machine learning from theory to algorithms: An overview,” *Journal of Physics: Conference Series*, vol. 1142, no. 1, 2018.
- [65] R. Muhamedyev, “Machine learning methods: An overview,” *Computer modelling & new technologies*, vol. 19, no. 6, pp. 14–29, 2015.
- [66] B. Mahesh, “Machine learning algorithms – a review,” *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, pp. 381–386, 2020.
- [67] D. K. S. Preeti and A. Dhankar, “A review on machine learning techniques,” *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, pp. 778–882, 2017.
- [68] J. Kober, J. A. Bagnell, and J. Peters, “Reinforcement learning in robotics: A survey,” *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238–1274, 2013.
- [69] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, “A survey of robot learning from demonstration,” *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.
- [70] S. Calinon, F. Guenter, and A. Billard, “On learning, representing, and generalizing a task in a humanoid robot,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 2, pp. 286–298, 2007.
- [71] A. Billard, S. Calinon, R. Dillmann, and S. Schaal, “Survey: Robot programming by demonstration,” *Springer handbook of robotics*, pp. 1371–1394, 2008.

- [72] E. Yigitbas, I. Jovanovikj, and G. Engels, “Simplifying robot programming using augmented reality and end-user development,” in *Human-Computer Interaction – INTERACT 2021* (C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, and K. Inkpen, eds.), pp. 631–651, Springer International Publishing, 2021.
- [73] F. Paternò, “End user development: Survey of an emerging field for empowering people,” *International Scholarly Research Notices*, vol. 2013, no. 1, 2013.
- [74] F. Paternò and V. Wulf, *New perspectives in end-user development*. Springer International Publishing, 2017.
- [75] J. Scholtz, “Theory and evaluation of human robot interactions,” in *36th Annual Hawaii International Conference on System Sciences*, IEEE, 2003.
- [76] Y. Duan, M. Andrychowicz, B. Stadie, O. Jonathan Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, “One-shot imitation learning,” *Advances in neural information processing systems*, vol. 30, 2017.
- [77] S. Wang, B. Scassellati, and T. Fitzgerald, “Toward measuring the effect of robot competency on human kinesthetic feedback in long-term task learning,” *TBD*, 2024.
- [78] D. Zielasko, B. Rehling, D. Clement, and G. Domes, “Carry-over effects ruin your (cybersickness) experiments and balancing conditions is not a solution,” in *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 1–5, 2024.
- [79] A. L. Edwards, “Balanced latin-square designs in psychological research,” *The American journal of psychology*, vol. 64, no. 4, pp. 598–603, 1951.
- [80] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, “Recent advances in robot learning from demonstration,” *Annual review of control, robotics, and autonomous systems*, vol. 3, no. 1, pp. 297–330, 2020.
- [81] F. Steinmetz, A. Wollschläger, and R. Weitschat, “Razer—a hri for visual task-level programming and intuitive skill parameterization,” *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1362–1369, 2018.
- [82] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, “A very brief measure of the big-five personality domains,” *Journal of Research in personality*, vol. 37, no. 6, pp. 504–528, 2003.
- [83] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, “Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots,” *International journal of social robotics*, vol. 1, pp. 71–81, 2009.

- [84] A. R.-V. D. Pütten and N. Bock, “Development and validation of the self-efficacy in human-robot-interaction scale (se-hri),” *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 7, no. 3, pp. 1–30, 2018.
- [85] M. Schrepp, A. Hinderks, and J. Thomaschewski, “Design and evaluation of a short version of the user experience questionnaire (ueq-s),” *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 103–108, 2017.
- [86] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.
- [87] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, “The aligned rank transform for nonparametric factorial analyses using only anova procedures,” in *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 143–146, 2011.
- [88] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock, “An aligned rank transform procedure for multifactor contrast tests,” in *The 34th annual ACM symposium on user interface software and technology*, pp. 754–768, 2021.
- [89] N. King, “Using templates in the thematic analysis of text,” *Essential guide to qualitative methods in organizational research*, vol. 256, 2004.
- [90] N. Mirnig, G. Stollnberger, M. Miksch, S. Stadler, M. Giuliani, and M. Tscheligi, “To err is robot: How humans assess and act toward an erroneous social robot,” *Frontiers in Robotics and AI*, vol. 4, 2017.
- [91] M. Sasaki, J. Muguro, F. Kitano, W. Njeri, and K. Matsushita, “Sim-real mapping of an image-based robot arm controller using deep reinforcement learning,” *Applied Sciences*, vol. 12, no. 20, 2022.
- [92] S. Bhat, J. B. Lyons, C. Shi, and X. J. Yang, “Value alignment and trust in human-robot interaction: Insights from simulation and user study,” in *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*, pp. 39–63, Springer, 2024.
- [93] I. Leite, C. Martinho, and A. Paiva, “Social robots for long-term interaction: a survey,” *International Journal of Social Robotics*, vol. 5, pp. 291–308, 2013.