On Virtual, Augmented, and Mixed Reality for Socially Assistive Robotics

by

Thomas Roy Groechel

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(COMPUTER SCIENCE)

May 2023

Acknowledgements

My Ph.D. was made possible by a large set of people who provided support throughout (hence the large section).

I would first like to thank my advisor Maja Matarić for her guidance and support throughout my graduate studies. One thing I tell all new potential students of the Interaction Lab is that Maja cares deeply about her students and exudes an unwavering consistency in her support as I've seen throughout the years. Maja endured my first ever research paper (including one of the most memorable/convoluted research games I attempted to describe) all the way through to this dissertation. I know I can always rely on her for advice and undoubtedly know choosing her lab five years ago was the right decision (as helped by praise sung by her former Ph.D. student and my undergraduate advisor Prof. Chad Jenkins). She is someone who I admire and who I now sometimes sound like when I tell researchers they should go out and pursue meaningful research outside of the lab environment.

I would also like to thank the members of my Dissertation Committee – Prof. Stefanos Nikolaidis, Prof. Mohammad Soleymani, Prof. Jesse Thomason, and Prof. Stephen Aguilar – for their advice and mentorship. Stefanos both taught me the core of HRI and sat with me multiple times to outline data analysis, providing the foundation of all of my subsequent analyses. Mohammad provided unique and insightful feedback within my Qualification Exam and Thesis Proposal that shaped the modelling direction of the remainder of my work. Jesse may have joined USC only 2 years ago but has already had an admirable impact on USC graduate student life as a whole. Jesse jumped right on my committee, providing feedback both in my dissertation as well as grant

writing. Finally, Stephen met with me multiple times, advising me on ideas founded in educational research and brainstormed many of the unique approaches found in my final research projects.

I would like to also thank my fellow #phd_postdoc lab members. In my first year, Naomi Fitter and Liz Cha gave me advice crucial to the success of my time at USC and a slide deck I referred to all the way to my final year. This was further helped by Matt Rueben who also provided a large amount of positive advice. Jessica Lupanow was helpful in our fun conversations, lab logistic buddies, and is the main reason my first paper was even readable. Chris Birmingham was a great work partner throughout, from writing a grant in our first year to talking over future plans. Lauren Klein is the most consistently positive person to whom I'm grateful for our many conversations and whose work was a great model to follow. I also wanted to thank Nathan Dennler for conversations about research, Mina Kian for her hospitality and conversation, and Amy O'Connell for taking on research projects with me. And finally, I want to thank Zhonghao Shi who was my first undergraduate mentee and a current Ph.D. student in the lab. Although he has impressed me with his work, he has impressed me more so in his kindness, willingness to help others, gratitude, and ability to highlight those around him.

I have been lucky to have the opportunity to work with and mentor high school, undergraduate, and master's students throughout my Ph.D. These students are first and/or co-authors on papers, won numerous research awards, and consistently made it to my 8am team meetings. Undergraduate students include Zhonghao Shi, Roxanna Pakkar, İpek Göktan, Karen Ly, Chloe Kuo, Julia Cordero, Adam Wathieu, Haemin Jenny Lee, Nisha Chatwani, Radhika Agrawal, Kartik Mahajan, Roddur Dasgupta, Ryan Stevenson, Adnan Karim, Daniel Ramirez, Rachel Channell, Evelyn Miguel Vargas, Charles Gary, and Dara Macareno. Master's students include Massimiliano Nigro and Karen Berba. High school students include Annika Modi, Jacob Zhi, İpek Göktan, Mena Hassan, Ashley Perez, and Bryan Pyo as a part of the USC SHINE program. I further want to recognize published, undergraduate first-authored works that are not included in this dissertation. The first authors and respective works are Adam Wathieu (Wathieu et al. 2022) and Julia Cordero (Cordero et al. 2022) – Best Paper Award finalist at Ro-MAN 2022 (4 nominated out of 256).

I would like to also thank the National Science Foundation who funded the research I worked on in the lab via the Expedition Grant for "Socially Assistive Robotics: An Expedition in Computing", NRI 2.0 grant for "Communicate, Share, Adapt: A Mixed Reality Framework for Facilitating Robot Integration and Customization" with Prof. David Krum (co-PI) NSF IIS-1925083, and NRI grant for "Socially Aware, Expressive, and Personalized Mobile Remote Presence: Co-Robots as Gateways to Access to K-12 In-School Education" with Prof. Gigi Ragusa (co-PI) NSF IIS-1528121. I would also like to thank USC Viterbi not only for the funding of my teaching assistanships but also (and more so) for funding the large majority of my high school, undergraduate, and master's students.

I would like to thank my family and their unending support. Countless calls to my parents – Michelle and Tom – helped push me through my graduate studies. The calls to my mom always led to me feeling better after and to my dad where we had great conversation inevitably ending in advice (in case anyone in the lab was wondering where I get it from). I also wanted to thank my 5 siblings in David, Katie, Mikey, Sarah, and Danny knowing that I can always rely on them.

Finally, I want to thank my wife EJ. There really are no words to describe how much support she has given me throughout these past five years. From moving to Los Angeles to listening to everything I have to say, she is the best person I could ever ask to be with. Also she and I would be disappointed if I didn't thank our cats Zeus and Athena who we adopted during the Ph.D. They provide needed unconditional support (as long as you feed Zeus on time).

Left) Zeus; **Middle)** EJ; and **Right)** Athena

Table of Contents

List of Tables

List of Figures

Abstract

Human-robot interaction (HRI) has seen significant growth as robotics research continues to advance. Subfields within HRI, such as socially assistive robotics (SAR) and virtual, augmented, and mixed reality for human-robot interaction (VAM-HRI) have also seen growth and impact in various domains. SAR provides non-physical assistance in areas such as mental healthcare, assisting older adults, and education, while VAM-HRI uses 3D virtual imagery to enhance human-robot interactions. However, those subfields developed separately and have not been integrated with each other. SAR is primarily focused on social domains and interactions, while VAM-HRI is primarily focused on completing domain-specific tasks efficiently and accurately.

This dissertation aims to *bridge the gap between SAR and VAM-HRI, providing a unifying framework for using VAM in SAR and demonstrating how to effectively leverage VAM for SAR problem domains*. By integrating VAM-HRI and SAR, we aim to improve the effectiveness and efficiency of SAR systems and enable new forms of interaction between humans and robots.

In this dissertation, an in-depth background on SAR and VAM-HRI is provided. The dissertation discusses the potential benefits and challenges of using VAM in SAR, and presents a unifying framework to leverage VAM for SAR, under the model-view-controller (MVC) software architecture paradigm. The dissertation explores every aspect of the MVC model, validating each through user studies and design considerations. The dissertation focuses on user state modeling, leveraging VAM for SAR, synthesizing reliable multimodal augmented reality data to support student kinesthetic curiosity, AR usability metrics, and learning human-robot proximal preferences. Additionally, it outlines how VAM can be used to expand the expressivity of SAR by creating designs for AR visualizations for both social and functional robot expressions, and providing design

recommendations for maximizing functional and social expressivity of AR robot gestures with different contextual factors. Lastly, this dissertation explores kinesthetic interaction paradigms for increasing human-robot flexibility of the controller, leveraging AR to create a wide range of interactions between the robot and users, including kinesthetic, direct interactions, and multi-party interactions.

Chapter 1

Introduction

This chapter provides an introduction to the growing fields of socially assistive robotics (SAR) and virtual, augmented, and mixed reality for human-robot interaction (VAM-HRI). The defining problem and goal of this dissertation is motivated by the need of a unifying framework to better leverage VAM for SAR. The chapter concludes with an outline of the rest of the dissertation and a list of primary and secondary contributions of the dissertation.

1.1 Growth and Impact of SAR and VAM-HRI

SAR and VAM-HRI have both seen recent, significant growth, having impact on a range of domains. SAR has been applied in many social contexts, while VAM-HRI has been used in a variety of functional HRI applications. Both of these fields have contributed significant work toward better understanding human-robot interaction (HRI).

1.1.1 Socially Assistive Robotics (SAR)

SAR, originally named by Feil-Seifer and Matarić (2005), is the combination of socially interactive robotics and assistive robotics. Matarić and Scassellati (2016) defines "social" interactions as any non-physical HRI (e.g., dialogue), and assistive robotics as HRIs where the goal is to help users

complete tasks that are difficult for them to do on their own, often in populations with additional needs (e.g., older adults and those with disabilities). Therefore, SAR is focused on providing meaningful assistance to users through non-physical HRI. Some examples of SAR include work in mental healthcare (Rabbitt et al. 2015), assisting older adults (Abdi et al. 2018), and education (Clabaugh et al. 2020) (Fig. 1.1). An in-depth background of SAR is given in Sec. 2.2.

Figure 1.1: Socially assistive robot system setup used for improving social and cognitive skills of students on the autism spectrum, evaluated in month-long in-home deployments (Shi et al. 2022; Clabaugh et al. 2019; Clabaugh et al. 2020; Jain et al. 2020a; Pakkar et al. 2019).

1.1.2 Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)

VAM-HRI is a relatively new area of research – originating in the 1980s (Kim et al. 1987) but seeing rapid growth starting in 2018 (Fig. 1.2) – that uses 3D virtual imagery to enhance HRIs. This field has seen rapid growth in recent years, which may be attributed to the increasing availability of VAM technology, such as commercial augmented and virtual reality (VR) devices (Walker et al. 2022), as well as the VAM-HRI workshops that have been held to bring together researchers and practitioners in this field (Williams et al. 2018a; Williams et al. 2020a; Williams et al. 2020b; Rosen et al. 2021; Chang et al. 2022). VAM-HRI has been applied in a variety of domains, including manufacturing (Nee and Ong 2013), training (Bric et al. 2016), construction (Ravi et al. 2021),

and research (Ikeda and Szafir 2022; Zea and Hanebeck 2021). These applications have largely focused on functional task metrics, such as time to complete a task or accuracy of task completion. An in-depth background on VAM-HRI in provided in Sec. 2.3.

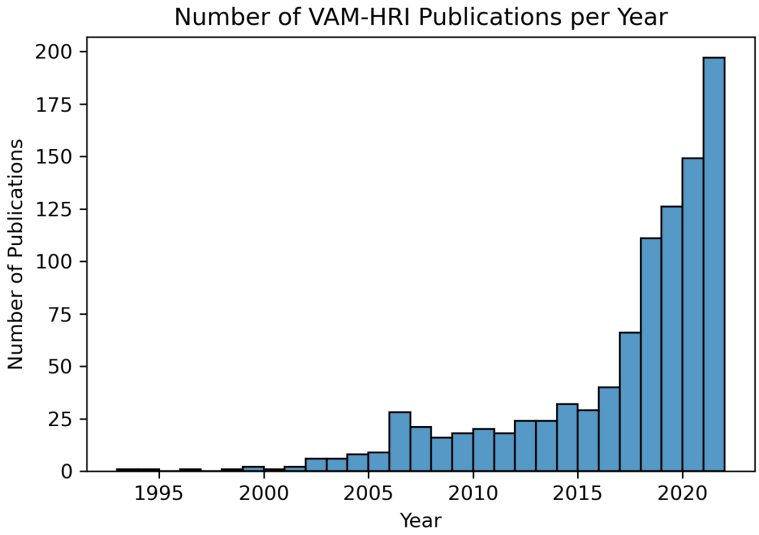


Figure 1.2: Growth of VAM-HRI publications over time. Histogram binned by year of publication of top 1,000 results from Google scholar with the search term '`("virtual reality" OR "augmented reality" OR "mixed reality") AND ("robot" OR "human-robot interaction")`'. The search was conducted September, 2022.

1.2 Motivation and Problem Statement

SAR has demonstrated the potential to have a significant impact across many domains, as it has the ability to assist and benefit a wide range of users in meaningful ways. The growth of SAR has highlighted the need for reliable data collection methods and the potential benefits of new interaction paradigms for different contexts. Similarly, VAM-HRI has shown great potential in traditional robot domains, enabling new interaction paradigms and modeling methods. VAM-HRI systems have been shown to be effective at improving the performance of human-robot tasks as measured by task efficiency and accuracy.

While the fields of SAR and VAM-HRI are both within robotics, they have developed separately and have not yet been integrated with each other. SAR is primarily focused on social

domains and interactions, while VAM-HRI is primarily focused on completing domain-specific tasks efficiently and accurately. *This dissertation aims to bridge the gap between SAR and VAM-HRI, provide a unifying framework for using VAM technology in SAR, and demonstrate how to effectively leverage VAM for SAR applications*. By integrating VAM-HRI and SAR, the aim is to improve the effectiveness and efficiency of SAR systems and enable new forms of interaction between humans and robots.

1.3 Contributions

The main contribution of this dissertation is to **define and demonstrate how VAM can be leveraged in SAR under the model-view-controller (MVC) paradigm** (Krasner and Pope 1988), presented as a framework for VAM-HRI in general, and then for SAR specifically, demonstrating how augmented reality (AR) can be used to reliably model user state, expand robot expressivity, and design kinesthetic interactions. This dissertation contributes to the field of HRI, where research in VAM-HRI and SAR has been growing rapidly and independently, this dissertation creates a bridge between the two fields so they can benefit each other.

The following are the primary contributions of this dissertation:

1. *A framework for VAM-HRI* that classifies research on 3D virtual imagery and VAM technology for HRI under the MVC paradigm (Krasner and Pope 1988). The framework focuses on VAM technology for improving the robot's internal model to better understand the user's state, increasing the expressivity of the robot's external expression, and enhancing the flexibility of the robot's controller for kinesthetic HRI.

2. *User state modeling leveraging VAM for SAR* that involves synthesizing reliable multimodal AR data to support student kinesthetic curiosity and AR usability metrics.

3. *VAM to expand SAR expressivity* that involves creating designs for AR visualizations for both social and functional robot expressions. The dissertation also provides design recommendations for maximizing functional and social expressivity of AR robot gestures with different contextual factors.

4. *Kinesthetic interaction paradigms for increasing human-robot flexibility of controller* that leverages AR to create a wider range of interactions between the robot and users, including kinesthetic, direct interactions, and multi-party interactions.

The following are secondary contributions of this dissertation:

1. *User studies* demonstrating the effectiveness of each technical approach in the work. These include studies with $3^{rd} - 5^{th}$ grade students, older adults (age 55+), and convenience populations (i.e., college students).

2. *Implemented and validated open-source software systems*. All code for studies throughout this dissertation are open-sourced and publicly available. These include:

    - `MoveToCode` – custom-made, extendable visual programming language (VPL) designed to encourage learning programming through movement alongside an embodied autonomous robot tutoring agent. The agent models a user's kinesthetic curiosity state in order to learn a personalized helping action policies for different users of the VPL
    - `PoseToCode` – converts user pose landmarks from webcam feeds to coding blocks using a convolutional neural network; performant to work on Chromebooks in schools
    - `NRI-SVTE` – robot capability visualization with a system that learns user-robot proxemic preferences in-situ using active transfer learning for representative sampling
    - `KuriAugmentedRealityArms` – AR arms for a robot to increase social expressivity

1.4 Outline

The remainder of this document is organized as follows:

- Chapter 2 defines key terms and provides background on the MVC software architecture paradigm as well as relevant existing work in SAR and VAM-HRI.
- Chapter 3 introduces the Tool for Organizing Key Characteristics of VAM-HRI Systems (TOKCS), the main unifying framework of the dissertation. TOKCS is based on the MVC software architecture paradigm and aims to provide a comprehensive understanding of VAM-HRI systems and how they can be leveraged for SAR.
- Chapter 4 describes three user modelling approaches with validation studies using VAM in SAR contexts. These include student kinesthetic curiosity and AR behavioral data for usability.
- Chapter 5 presents two designs for creating AR (AR) robot visualizations, along with validating studies on their effectiveness. Design considerations for using AR appendages to maximize both social and functional expressivity in robots are also given.
- Chapter 6 outlines two different AR, kinesthetic interactions to increase student interest in coding with a robot tutor–a pair coding activity and a pose-based coding activity–and provides design guidelines for both.
- Chapter 7 provides a summary and concluding statements as well as potential open problems and extensions to the dissertation.

Nota bene: This dissertation includes contributions from multiple researchers, specifically undergraduate students supervised and mentored by the dissertation author. The contributors are named in chapters and sections of the dissertation that cover their work. Specifically, a "**Contributors**" box is included at the beginning of each chapter or section that includes contributions by other researchers.

Chapter 2

Background and Related Work

This chapter presents relevant background information and related work. It first discusses the MVC software architecture paradigm – a key concept used throughout the dissertation. The chapter then describes SAR as a subfield, including the benefits of physically embodied robots. The field of VAM-HRI is also discussed, including the study of virtual embodiment. This chapter aims to provide context for the subsequent chapters, which will focus on how to leverage VAM technology to enhance SAR. The two fields of SAR and VAM-HRI exist in parallel; this dissertation aims to bridge the gap between them by describing how VAM can be used to enhance SAR.

2.1 Model-View-Controller (MVC) Software Architecture Paradigm

This dissertation presents using the MVC paradigm (Krasner and Pope 1988) to leverage VAM for SAR. As seen in Fig. 2.1, MVC is a software architecture pattern that separates an application into three main components: the model, the view, and the controller.

The *model* represents the data and the back end logic of the software application. It is responsible for maintaining the state of the application and handling interactions with the data store. The *view* is the user interface of the application. It presents the data to the user and allows the user

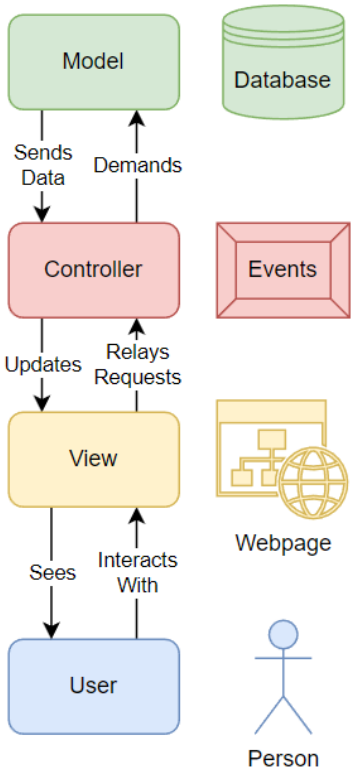


Figure 2.1: MVC (Krasner and Pope 1988) software architecture paradigm where the user interacts with the view, which relays these events to the controller, which then demands data from the model. The model sends the data back to the controller, which updates the view for the user to see. Some descriptions of MVC describe direct communication between the user and controller, with both descriptions being functionally equivalent.

to interact with the application. The *controller* is the bridge between the model and the view. It receives user input via events from the view, interacts with the model to update the state of the application, and then updates the view to reflect the changes in the model.

MVC is a widely used design pattern in software development, particularly in web applications. It allows for the separation of concerns, making it easier to develop and maintain the application. It also makes it easier to test individual components of the application independently. In this chapter, the background for each component of MVC with regards to SAR and VAM-HRI is described in detail to form the basis of the framework described in Chap. 3.

2.2 SAR

SAR is a relatively young yet rapidly growing subfield of HRI, initially defined by Feil-Seifer and Matarić (2005). In SAR, robots assist people in a variety of non-physical tasks, such as providing emotional support or aiding with learning. The physical robot embodiment enhances these interactions, leading to better, measurable outcomes. Advances in machine learning have led to improved approaches to modeling user state and a better understanding of how robots can express themselves, resulting in more meaningful interactions for a variety of user populations.

2.2.1 Social + Assistive + Robotics

Feil-Seifer and Matarić (2005) and Matarić and Scassellati (2016) defined SAR as in the intersection of *socially interactive robotics* and *assistive robotics*.

Socially interactive robotics was first introduced by Fong et al. (2003) as a subcategory of HRI, distinguishing between the main task of the robot was to form some interaction with the human as opposed to teleoperation research. They further broke down work into social interaction principles used by the robot (e.g., speech, gestures) as well as focusing on the human's perception of the robot in an interaction. This definition, however, does not properly encompass socially interactive robotics given the counter examples of teleoperated robots sending social signals back with their operators (Goodrich et al. 2013). Matarić and Scassellati (2016) instead categorizes HRIs into categories of *physical* and *social* interaction. This better defines the social component of SAR as anything pertaining to non-physical interaction.

A first definition of *assistive robotics*, on the other hand, is not clear-cut. Miller (1998) provides a possible starting point for defining it as using robotics to help those with disabilities. This does not fully encompass assistance though as people without disabilities are often assisted by robots in both social (Clabaugh and Matarić 2019) and physical (Mohebbi 2020) settings. Assistive robotics is therefor referred to as HRIs where the goal of the robot is to help users in tasks difficult for them to complete on their own, often in populations with additional needs (e.g., older adults and those

with disabilities). Matarić and Scassellati (2016) points to the large body of work predominantly focused on physical assistance with an increase in work looking to socially assist users. This is where SAR lies, the overlap of non-physical assistance for users predominantly in meaningful, need-based populations. The interactions with these populations encompass a variety of domains as further described in Sec. 2.2.5.

Finally the *robotics* components of SAR are encompassed by the question "why not use a cheaper and easier to deploy virtual agent?" One of the primary reasons for using a robot instead of a virtual assistant is the physical embodiment of the robot, which has been shown to lead to better outcomes (e.g., task time, learning gains, socio-emotional measures) across a variety of interactions (Deng et al. 2019) as further described in Sec. 2.2.2. A robot can also persist in the physical environment in a way that a virtual character cannot, allowing it to be present during a wider range of activities while persisting as a social agent. In addition, a robot can physically interact with the real world, further expanding the potential for interactions (e.g., tangible user interfaces (Law et al. 2019) and physically-based games (Andriella et al. 2019)). Similar to embodiment, these physical, embodied interactions have been shown to be beneficial to users especially in educational contexts (Cutica et al. 2014).

2.2.2 Physical Embodiment

Socially assistive robots do not necessarily need a physical embodiment to perform their tasks, which are, by definition, non-physical. However, Wainer et al. (2006) demonstrated the positive roles of physical embodiment as measurable outcomes of a HRI. The study of benefits of embodiment predates robotics having roots in social sciences (Varela et al. 1992), philosophy (Hendriks-Jansen 1996), and neuroscience (Zeman 2006). Further, Deng et al. (2019) surveyed a large body of research showing that a physical embodiment can lead to better outcomes, such as increased learning gains and positive socio-emotional measures. Thus the benefits of physically embodied robots is well studied and documented.

Additionally, a physical robot embodiment needs to be designed for its specific tasks and goals. The design of physical robot embodiments has a large body of work with outlining metaphors and their effects on users. This dissertation, however, primarily addresses the general role of physical embodiment and not the specifics of these design elements. The reader is directed to Dennler et al. (2022) for a comprehensive overview of physical robot design metaphors.

2.2.3 Model - User Hidden State Estimation

A large body of work in SAR focuses on modelling the user's state, particularly their latent (i.e., not directly observable) socio-emotional state and performance state during an interaction. This state information is then used by the robot to inform its action policy (e.g., **if** $(engagement < threshold)$ **then** $\rightarrow$ **do** $REENGAGEMENT\_BEHAVIOR$). Multimodal reasoning has been used to infer hidden user state such as knowledge (Schodde et al. 2017), engagement (Rich et al. 2010; Salam and Chetouani 2015a; Celiktutan et al. 2017; Jain et al. 2020b), curiosity (Ayub et al. 2022), and valence/arousal (Kulic and Croft 2007).

There is an emphasis on personalizing the user models (Clabaugh and Matarić 2019), often using machine learning techniques such as reinforcement learning (Kaelbling et al. 1996; Gordon et al. 2016; Clabaugh et al. 2019), transfer learning (Weiss et al. 2016; Spaulding and Shen 2021), and domain-adaptation (Daumé III 2009; Shi et al. 2022). These techniques aim to address the common issue of SAR datasets being relatively small compared to traditional machine learning datasets (e.g., vision (Ferraro et al. 2015; Janai et al. 2020; Mogadala et al. 2021), proximal preferences (Samarakoon et al. 2022), and natural language (Ferraro et al. 2015; Alyafeai et al. 2020; Khurana et al. 2022)). The small dataset problem is further compounded by the diversity and high variability of populations in SAR datasets (e.g., those with autism spectrum disorder (Lord et al. 2018)). In order to model phenomena, data for these data sets must be collected within SAR interactions. The sensor data collection of many SAR interactions can further complicate the modelling trained on these datasets. As exemplified by Williams et al. (2018b), data collection can be robot egocentric – from the robot's point of view, such as using an onboard camera – or allocentric –

from an external view, such as using a camera in the environment. With only robot-centric and external sensors, the user can easily move out of a sensor's field of view or a sensor may be moved out of place.

2.2.4 View - Physical Robot Expression

Expressivity in HRI refers to the robot's ability to use its modalities to non-verbally communicate the robot's intentions or its internal state (Charisi et al. 2019). Higher levels of expressiveness have been shown to increase trust, disclosure, and companionship with a robot (Martelaro et al. 2016). Expressivity can be conveyed with dynamic actuators (e.g., motors) as well as static ones (e.g., screens, LEDs) (Balit et al. 2018). HRI research into gesture has explored head and arm gestures, but many nonhumanoid robots partially or completely lack those features, resulting in low social expressivity (Cha et al. 2018).

A key component to social interaction is a user's perception of the robot during the interaction (Cha et al. 2015) which can be further broken down into social and functional components. *Social perception* is the user's belief that a robot is capable of participating in the interaction as a social actor that is an interactive, autonomous, and adaptable agent. *Functional perception* is the user's belief that a robot is able to accurately, and often quickly, perform a task's goal not related directly to the human-robot social relationship. Humans are social creatures and while tasks, by definition, have functional outcomes, the two categories, social and functional, are not mutually exclusive. Social and functional instead fall within two parallel spectra within any given task. Any interaction during a task has some combination of functional to social goals and benefits. Therefore defining social and functional tasks is similar to the difficulties found in defining "high-level" and "low-level" tasks in computer science and robotics, where the definitions are context- and domain-specific.

2.2.5 Controller - Domains and Interactions

SAR is a field of research with numerous applications in various domains, including healthcare (Rabbitt et al. 2015; Tapus et al. 2009; Matarić et al. 2007), older adult care (Vandemeulebroucke et al. 2018; Abdi et al. 2018), early childhood/infant development (Jeong et al. 2015; Klein et al. 2019), and education (Clabaugh et al. 2020; Chen et al. 2020; Shi et al. 2022; Short et al. 2014). In these contexts, SAR technologies are used to provide assistance and support to individuals with additional needs through meaningful interactions and assistance (as discussed in Sec. 2.2). These interactions can be one-on-one or involve multiple people (i.e., multi-party interactions (Birmingham et al. 2020; Short and Mataric 2017; Salam and Chetouani 2015b)), and the use of a physical robot embodiment can further enhance the shared experience of all people present in the same space.

One key area of focus in SAR research is education, where robots are used as personal tutors (Clabaugh et al. 2020; Chen et al. 2020; Shi et al. 2022; Short et al. 2014) or as part of the learning content itself (Bravo et al. 2017). These research efforts have demonstrated the benefits of using a one-on-one tutor, as it can personalize the learning experience for each student and provide additional help outside of the traditional classroom setting (Clabaugh and Matarić 2019).

Socially assistive robots, especially in the education domain (Clabaugh and Matarić 2019), are often designed for seated interactions for a variety of reasons. These may include the fact that many of the settings in which these robots are used, such as healthcare and education, often involve seated interactions, as well as the fact that seated interactions may be more cost-effective and easier to design and implement from a technical standpoint. Seated interactions may also be more traditional or familiar in certain domains, and may be more easily accepted by users. The emphasis on seated learning interactions creates an opportunity gap to research kinesthetic learning contexts (i.e., embodied learning (Macedonia 2019)) that include the well-documented benefits of the physical embodiment of the SAR tutor (Deng et al. 2019).

2.3 VAM-HRI

VAM-HRI is a nascent area of research that uses 3D virtual imagery to enhance HRIs. Fig. 1.2 shows the rapid growth of this field, which may be driven by the widespread availability of VAM technology such as commercial augmented and VR devices (Walker et al. 2022) in tandem with the the first VAM-HRI workshop (Williams et al. 2018a). One of the key challenges in VAM-HRI is to design and develop systems that can effectively support natural and intuitive interactions between humans and robots. This requires a deep understanding of how humans perceive, understand, and interact with VAM environments. To this end, VAM-HRI has many similarities to SAR, but there is very little substantiated overlapping work between the two areas of research. This presents an opportunity to leverage VAM technology to improve the effectiveness of SAR.

2.3.1 Milgram's Reality-Virtuality Continuum

The Milgram Reality-Virtuality Continuum (Milgram et al. 1995a) classifies environments and interfaces with respect to how much virtual and/or real content they contain (Fig. 2.2). On one end of the spectrum lies reality, which is any interface that does not use any virtual content and makes use of only real objects and imagery. The opposite end of the spectrum is VR, which would be an interface that consists of pure virtual content without any integration of the real world (for example, a simulated world presented in VR). Between these two extremes is mixed reality (MR), which captures all interfaces that incorporate a portion of both reality and virtuality in their design. There are two sub-classes of MR: (1) AR where virtual objects are integrated into the real world; and (2) augmented virtuality (AV) where real objects are inserted within virtual environments.

AR interfaces in VAM-HRI often communicate the state and/or intentions of a physical robot. For example, the battery levels of a robot can be displayed with a virtual object that hovers over a physical robot, or a robot's planned trajectory can be drawn on the floor with a virtual line to indicate the robot's future movement intentions.

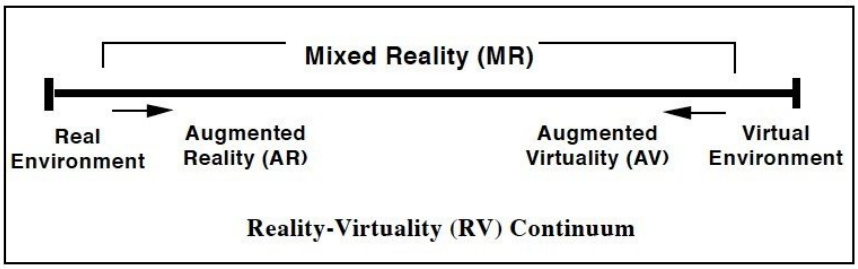


Figure 2.2: Milgram's Reality-Virtuality Continuum (Milgram et al. 1995a) - The continuum of interactions within only the physical reality (left) to fully VR (right). MR is the full continuum of combining any virtual and physical reality elements. There are two sub-classes of MR: (1) AR where virtual objects are integrated into the real world; and (2) AV where real objects are inserted within virtual environments.

VR interfaces are often used to provide simulated environments where human users can interact with virtual robots. In these virtual settings user interactions with robots can be monitored and evaluated without risk of physical harm for either robot or human. Additionally, the virtual robot models can be easily and quickly altered to allow for rapid prototyping of both robot and interface design. Without the need for physical hardware, robots can be added to any virtual scene without the typical costs associated with physical robots.

Virtual environments can also be used to teleoperate and/or supervise physical robots in the physical world. In cases like these, 3D data collected by the physical robot about its surrounding environment is integrated within virtual settings to create **AV** interfaces. Cyber-physical interfaces and virtual control rooms are two common VAM-HRI AV methods of enhancing remote robot operators ability by increasing situational awareness of their robot's state and location while mitigating the limitations of virtual interfaces such as cyber sickness (Lipton et al. 2017).

2.3.2 Virtual Embodiment

Embodiment in VAM-HRI can be thought of as existing on a spectrum, with one end representing interactions with a fully physical robot and the other end representing interactions with a fully virtual robot avatar. It is worth noting that these two ends of the spectrum are not mutually exclusive,

as a physical robot can be augmented with virtual 3D imagery, which is often the goal of VAM-HRI (Williams et al. 2020c; Walker et al. 2018; Walker et al. 2022). The benefits of physical robot embodiment are discussed in Sec. 2.2.2. In contrast, this section will focus on the embodiment effects of fully virtual 3D avatars in AR and VR.

Virtually embodied avatars are digital representations of a person in AR/VR environments. These avatars can be used to enhance social interactions, communication, and collaboration in virtual spaces. One benefit of virtually embodied avatars is that they can provide a sense of presence and co-presence, allowing people to feel as if they are physically present with others even when they are not in the same location (Wu et al. 2021; Murugan et al. 2021; Pakanen et al. 2022). This can be particularly useful in situations where face-to-face interactions are not possible due to physical distance or other constraints (e.g., COVID-19 pandemic (Asadzadeh et al. 2021)).

Virtually embodied avatars can also be used to facilitate social and communication skills training, such as in the treatment of social anxiety (Horigome et al. 2020) or autism spectrum disorder (Mosher and Carreon 2021). In these contexts, avatars can provide a safe and controlled environment for practicing social interactions and communication skills. Additionally, virtually embodied avatars can be used in AR/VR educational settings to create immersive and interactive learning experiences. For example, students can use avatars to collaborate on projects (Pidel and Ackermann 2020), participate in virtual field trips (Vellingiri et al. 2022), or engage in other interactive learning activities (Papanastasiou et al. 2019).

Overall, virtually embodied avatars offer a range of potential benefits in both personal and professional contexts, and their use is likely to continue to grow as AR/VR technologies continue to advance. As VAM-HRI grows, the field continues to study the potential benefits and challenges of mixed-embodiment, which refers to situations where individuals are partially physically and partially virtually embodied.

2.3.3 Model - User Hidden State Estimation

One key aspect of VAM-HRI is the ability of AR and VR devices to collect data that can be used to model user state, including pose and latent state such as intent. In VAM-HRI, data collection is typically performed using sensors that are built into AR and VR devices, such as head-mounted displays (HMDs) or tablets. These sensors can include cameras, depth sensors, inertial measurement units (IMUs), and other sensors that are used to track the user's head and hand movements, as well as their gaze direction and other features of their environment.

These data can be used to model the user's pose, the position and orientation of their body in space. This information is crucial for HRI, as it allows robots to understand the user's physical location and movement, and to respond appropriately. For example, if a user is pointing at a specific object, the robot can use their pose data to understand that they are trying to direct its attention to that object (Puljiz et al. 2021). Researchers further use these pose data to learn more fluid interaction via learning from demonstration (Wang and Belardinelli 2022).

In addition to modeling the user's pose, AR and VR devices can also be used to model latent state, such as intent. Latent state refers to unobserved or hidden variables that may influence a user's behavior. For example, if a user is utilizing an AR or VR system to interact with a robot, the robot may be able to infer their intent based on their gaze direction, hand gestures, and other subtle cues (Rosen et al. 2020; Higgins et al. 2022). This information can be used to guide the robot's behavior and to improve the overall effectiveness of the HRI system.

A key benefit VAM data collection is the consistency of these data given the user is either almost always wearing or holding the VAM technology. Thus systems can make assumptions about the data (e.g., user is at position $\{X,Y\}$ because HMD is at position $\{X,Y\}$)) with users unlikely to go "out-of-frame."

2.3.4 View - Virtual Robot Expression

VAM technology can be used independently of the robot's embodiment and can effectively communicate complex signals through 3D virtual imagery (Walker et al. 2022). This imagery has been

shown to increase the ease of robot programming (Gadre et al. 2019), remote teleoperation (Rosen et al. 2018; LeMasurier et al. 2022; Zea and Hanebeck 2021), human intent estimation (Rosen et al. 2019), and human-robot teaming tasks (Walker et al. 2018).

Visualizations of robot signals can improve human-robot communication by providing complex information in a simplified and accessible manner (Walker et al. 2022). However, survey analysis across the Miligram virtuality continuum has shown that early work in social MR for robots was limited (Holz et al. 2009), with examples including virtual expressive faces on Roomba vacuum cleaners (Young et al. 2007a) and virtual avatars on TurtleBot mobile robots (Dragone et al. 2006). To the best of the knowledge of the author, it wasn't until 2019 that virtual overlays focused on social outcomes were pursued further (as described in Sec. 5.1). Work outside this dissertation and after 2019, primarily focused on the use of deictic gesturing in AR-enhanced social robots, including the creation of different AR appendages and visualizations for robots which often involve trade-offs between social and functional designs (Hamilton et al. 2020; Hamilton et al. 2021; Tran et al. 2021). Chapter 5 builds upon and is intertwined with these works, as recently demonstrated by Brown et al. (2022).

2.3.5 Controller - Domains and Interactions

The field of VAM-HRI largely consists of human-robot teaming (Walker et al. 2019; Walker et al. 2018; Rosen et al. 2019; Chang et al. 2022; Walker et al. 2022), robot programming/debugging interfaces (Ikeda and Szafir 2022; Zea and Hanebeck 2021), and teleoperation (Zhang et al. 2018; Hedayati et al. 2018; Lipton et al. 2018). These span a large variety of domains including manufacturing (Nee and Ong 2013), healthcare (Viglialoro et al. 2021), training (Bric et al. 2016), construction (Ravi et al. 2021), robotics education (Villanueva et al. 2021), and research labs (Ikeda and Szafir 2022; Zea and Hanebeck 2021).

VAM-HRI interaction paradigms are largely dictated by the VAM hardware deployed in a given scenario. While hardware used for VAM can vary widely, there are certain types of hardware that

are commonly used in VAM-HRI. The most common, which enable experiences along the Reality-Virtuality Continuum, include: HMDs, projectors, displays, and peripherals.

HMDs: VR, MR, and AR all commonly use HMDs. HMDs allow for a full VR experience, visually immersing the user in a completely virtual environment. HMDs also allow for AV, such as in Wadgaonkar et al. (2021), where the user is in a virtual setting but the virtual robot being manipulated is also moving in the real world. Some HMDs are strictly AR headsets, where virtual images are rendered on top of the real world view of the user.

Projectors: Onboard projectors can provide a way for the robot to display virtual objects or information. Alternately, static projectors allow an area to contain AR elements. Images might be projected onto an object, on the floor, or onto a robot (Han et al. 2022; Gillen et al. 2012; Bolano et al. 2019).

Displays: This category of hardware ranges from handheld smartphones or tablets to room-size displays. Two-dimensional and three-dimensional monitors fall somewhere in between this range. Some of these exist in a single location, while mobile displays can be carried by a person or moved by a robot. A cave automated virtual environment (or CAVE) immerses the user in VR using 3 to 6 walls to partially or fully enclose the space. An AR display might include a real-time camera with overlaid virtual graphics, while a VR display contains completely virtual graphics. Displays can be an especially effective way to conduct user studies without investing in expensive hardware, for example by showing recorded videos to participants on Amazon Mechanical Turk (Mott et al. 2021).

Peripherals. Peripheral devices allow for a richer interaction within VAM. Leap Motion hand tracking can be combined with a headset such as the HTC Vive (as in (Mara et al. 2021)) to provide recording and playback of motions and commands. Some controllers are handheld and can be used individually or in tandem, giving the user a modality for both gesturing and selecting with the use of buttons on the device.

While the field of VAM-HRI has focused on a variety of domains and interaction paradigms, it is primarily focused on functional goals rather than social bi-directional interaction with the

robot. This dissertation proposes a framework for leveraging the benefits of VAM for SAR. By incorporating VAM technologies into SAR, it may be possible to enhance the social interaction and communication between humans and robots, as well as improve the capabilities of the robots in assisting with tasks.

Further surveys of VAM-HRI can be found in Walker et al. (2022) and Suzuki et al. (2022).

2.4 Summary

Socially interactive robotics (SIR) is a subfield of HRI that involves the use of robots to interact with humans in a social manner, using techniques such as speech and gestures. SIR has been defined more broadly as any non-physical interaction between humans and robots, while assistive robotics involves the use of robots to help users complete tasks that are difficult for them to complete on their own, often in populations with additional needs such as older adults and people with disabilities. There is overlap between SIR and assistive robotics in the area of SAR, which involves the use of robots to provide non-physical assistance to users in need-based populations. SAR interactions benefit largely from the robot's physical embodiment, persistent presence, and real-world physical interaction. SAR has numerous applications in domains such as healthcare, older adult care, early childhood development, and education.

VAM-HRI is a field of research that uses 3D virtual imagery to enhance HRIs. The goal of VAM-HRI is to design and develop systems that can support natural and intuitive interactions between humans and robots. VAM-HRI has many similarities to SAR, but there is limited overlap between the two fields. VAM-HRI applications include human-robot teaming, robot programming and debugging interfaces, and teleoperation, and they span a variety of domains including manufacturing, healthcare, training, construction, robotics education, and research labs. VAM-HRI hardware includes HMDs, projectors, displays, and peripherals. These devices can be used to create experiences along the Reality-Virtuality Continuum, ranging from fully virtual to fully real.

This dissertation aims to develop a framework for leveraging VAM for SAR using the MVC paradigm (Chap. 3). The MVC paradigm is a software design pattern that separates the representation of information from the user's interaction with it. Each component of MVC is used to enhance SAR and validated through user studies. Chap. 4 focuses on user modeling, Chap. 5 focuses on robot expressions (i.e., the "view" in MVC), and Chap. 6 focuses on interaction paradigms (i.e., the "controller" in MVC). The main focus of each chapter is on its respective component of MVC, with the other components being used to validate the results. For example, the user modeling done in Chap. 4 uses different robot expressions (i.e., view) with different interaction paradigms (i.g., controller). However, the main focus of the work in Chap. 4 is primarily on user modelling.

Chapter 3

MVC: A Tool for Organizing Key Characteristics of VAM-HRI Systems (TOKCS)

This chapter presents TOKCS, a Tool for Organizing Key Characteristics of VAM-HRI Systems. TOKCS is based on the Interaction Cube framework and adopts a MVC approach to VAM-HRI, focusing on designer intent and user perception of virtual object anchor locations and manipulability. This chapter explains how TOKCS can be used to enhance understanding of VAM for SAR.

3.1 The Interaction Cube: an Existing VAM-HRI Framework

Contributors: Chapter 3 is based on Groechel et al. (2022a) written with co-first author Michael E. Walker. Additional authors of the published work include Christine T. Chang, Eric Rosen, and Jessica Zosa Forde.

The Interaction Cube (Williams et al. 2019a) uses three dimensions to characterize VAM-HRI work: the 2D Plane of Interaction to represent interactive design elements and the 1D Reality-Virtuality Continuum from Milgram (Milgram et al. 1995a) to characterize the environment.

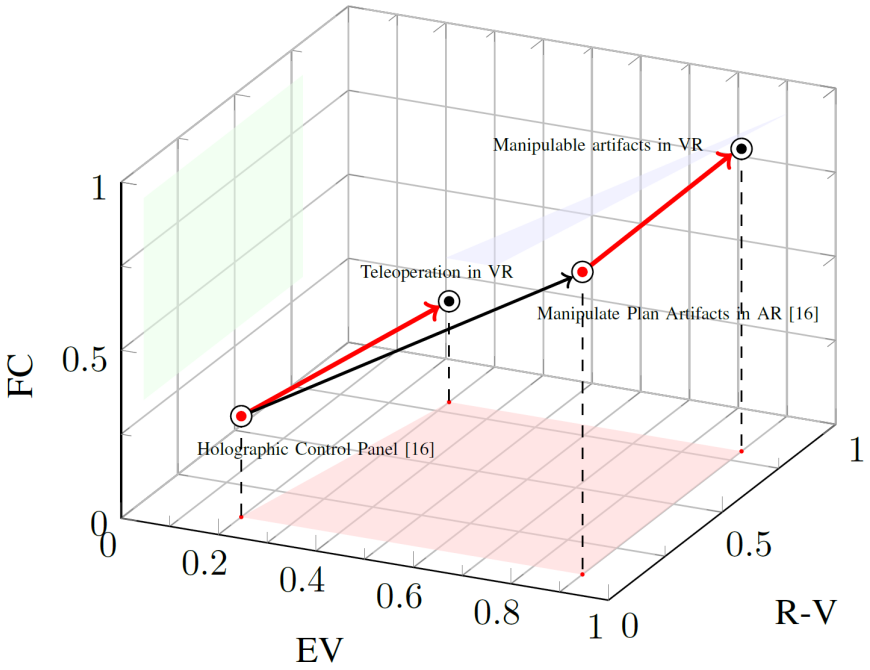


Figure 3.1: The Reality-Virtuality Interaction Cube used to visually categorize MR interaction design elements (MRIDE)s according to their Flexibility of Control (FC), Expressivity of View (EV), and where they lie upon the Reality-Virtuality Continuum (RV). Reality is indicated as 0 and Virtuality as 1.

3.1.1 Interaction Design Elements: Enhancing View and Control

Two of the three dimensions of the Interaction Cube (Fig. 3.1) are defined by the *Plane of Interaction*, which captures both (1) the opportunities to view into the robot's internal model, and (2) the degree of control the human has over the internal model. These two levels of interactivity (termed the **expressivity of view (EV)** and **flexibility of controller (FC)**, respectively) are the conceptual pillars for characterizing interactivity within the Interaction Cube, and any components that contribute or impact either EV or FC are called *interaction design elements*. This is similar to the MVC design pattern. However, in this case the 2D placement on the Interaction Plane depends on a vector whose direction results from the impact a design element has on EV and the impact a design element has on FC. The magnitude of the vector is scaled by the complexity of the robot's internal model. According to Williams et al. (2019a), "while it is likely infeasible to explicitly determine the position of a technology on this plane, it is nevertheless instructive to consider the

formal relationship between interaction design elements and the position of a technology on this plane."

3.1.2 Mixed-Reality Interaction Design Elements: Anchoring and Artifacts

The Interaction Cube categorizes the study of VAM virtual objects as MRIDEs (mixed-reality interaction design elements), which fall into one of three categories:

- *User-Anchored Interface Elements*: Objects attached to user view. This is similar to traditional GUI elements that are anchored to the user's camera coordinate frame and do not change along with the user's field of view. These elements may also be referred to as part of a user's heads up display as popularized by video games and movies.
- *Environment-Anchored Interface Elements*: Objects anchored to the environment or robot. For example, virtual arms that can be anchored to a robot (Groechel et al. 2019) or virtual objects that can be anchored to the physical environment.
- *Virtual Artifacts*: Objects that can be manipulated by humans or robots or may move "under their own ostensible volition" (Williams et al. 2019a). For example, virtual indicators of robot position, such as arrows, can move on their own within the environment.

3.1.3 The Reality-Virtuality Continuum

The Reality-Virtuality Continuum (Milgram et al. 1995a), as outlined in Sec. 2.3.1 and Fig. 2.2, is a scale that spans from physical reality, where there is no virtual imagery, to VR, where the user is completely immersed in virtual imagery. MRIDEs fall on this continuum along the third axis of the Reality-Virtuality Interaction Cube. MR, which is any combination of physical and VR, falls between these extremes. Within MR, there are two subcategories: AR and AV. AR involves the blending of virtual imagery with the user's present physical environment, while AV represents

physical world state information in 3D virtual imagery from a remote physical entity, such as a robot.

3.2 MVC: The TOKCS Framework

A key insight of this work is the addition of key characteristics of VAM-HRI not covered by the Interaction Cube to create TOKCS. These include VAM-HRI system hardware, research that seeks to increase a robot's model of the world around it, and additional granularity to *mixed-reality* interaction design elements (MRIDEs). *TOKCS defines a robot's* ***internal complexity of model*** *(Sec. 3.2.1) and, in conjunction with discretizing EV and FC from the Interaction Cube (Sec. 3.1.1), outlines VAM-HRI under the MVC paradigm (Krasner and Pope 1988).* The characteristics are part of TOKCS which is then applied to the 4^{th} VAM-HRI workshop's papers in Sec. 3.3. The application of TOKCS to the workshop informs the insights and future work recommendations outlined in Sec. 7.2.

3.2.1 Defining the Robot Internal Complexity of Model

The Interaction Cube emphasizes the increased expressivitiy of view and flexibility of controller aspects of projected visual objects having on the robot's underlying model. This fails to explore, however, the sensing capabilities and data afforded by VAM technologies (e.g., ARHMD). The framework can be expanded by including the technologies' ability to aid the robot's internal model of the world - namely increasing the robot's internal **complexity of model (CM)**. The robot's internal CM benefits from data typically difficult to gather (e.g., eye-gaze) as well as the technology affording data assumptions (e.g., a headset with various sensors being anchored to the user's head). These data aid in a robot's model of the *environment* and/or model of the *user*.

Environment - Data from the VAM technology further increases the robot's understanding of an environment. An example is provided in Fig. 3.2. Given a mobile robot with 2D SLAM, a 3D map from an ARHMD's SLAM can be transformed into the robot's coordinate frame. The map

can then be used for more accurate navigation. In another situation, a mobile phone camera can help with object recognition both in front and behind the robot.

User - Data from VAM technology further increase the robot's understanding of the user. For example, a robot can better infer a user's intent to choose an object by using ARHMD eye-gaze (Rosen et al. 2020). Data gathered from motion sensors can be used both for functional purposes (e.g., where is the human in relation to the robot) as well as to infer affective human state such as student curiosity (Groechel et al. 2021).

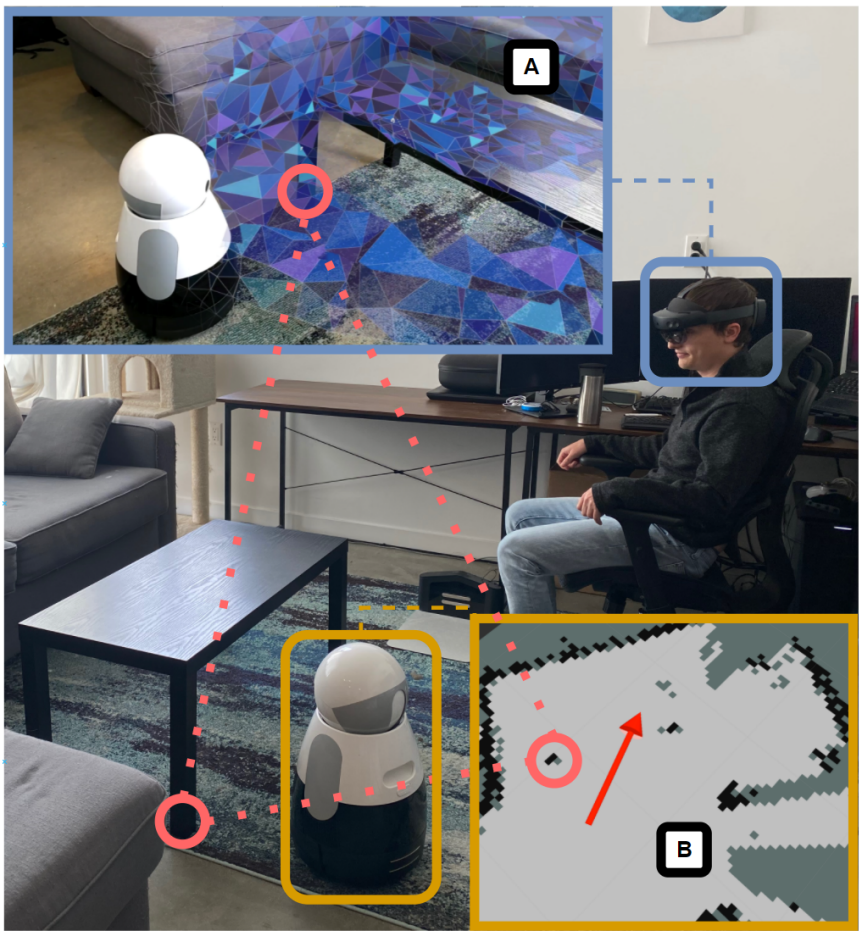


Figure 3.2: Demonstrates a navigation situation where the robot 2D SLAM map (B) benefits from the 3D SLAM map from the ARHMD (A). The robot only maps the two front table legs (bottom left) as it is only equipped with a 2D lidar. The robot, however, is too tall to move past the table so it will collide if it does not use the 3D map from the ARHMD. A combined SLAM map would be created from feature matching such as the table legs (circles).

3.2.2 User-Perceived Anchor Locations and Manipulability

The mixed reality interaction design element (MRIDE) categorizations of user-anchored interface elements, environment-anchored interface elements, and virtual artifacts (described in Sec. 3.1.2) are not mutually exclusive and lack necessary granularity. For example, a virtual artifact can be user-anchored, such as a movable user-anchored element or an environment-anchored object that moves on its own. Granularity can also be added to benefit MRIDE classifications such as distinguishing between robot- and environment-anchored objects.

To this end, two important distinctions can be added to expand the current framework. First, two characteristics are applied: **Anchor Location {User, Robot, Environment}** and **Perceived Manipulability {User, Robot, None}**. Second, MRIDEs are distinguished based on the intended user perception of the virtual object (i.e., where does the user perceive the anchor to be and who can/does move a virtual object).

The first distinction allows for multiple labels within each characteristic, such as objects that are manipulable by both the robot and the user. Visuals for path planning (e.g., (LeMasurier et al. 2021)) further highlight the benefits of these granular distinctions. A planned robot pose visualized within the environment could be argued as both robot- and environment-anchored since the same trajectory can be defined within the robot's local frame of reference or within a global frame of reference.

The latter distinction is important when characterizing Anchor Location as any object can be translated into the environment's coordinate frame. This translation may mathematically correct but the intended perception is important to the goals of studying a virtual object's effect on the user in the interaction. For example, the granularity of Anchor Location combined with intended user perception allows for labeling virtual objects intended to be perceived as part of the robot, such as added virtual robot appendages (Tran et al. 2020; Groechel et al. 2019). These virtual arms were specifically designed to be perceived as part of the robot to study their impact on the robot's functional and social expressivity, respectively. Labeling the study of virtual arms as anchored to the "environment" or "user" does not aid in surveying trends among research projects.

Expanding on this idea, a key property of virtual object manipulation is the user's action attribution of the manipulation (i.e., does the user perceive that they moved the object, the robot moved the object, or the object moved on its own). Perceived Manipulability is the action attribution, the perception the user has of the manipulation. For an object that the user manipulates (e.g., grasps), the Perceived Manipulability is the user. Virtual objects "manipulated" by the robotic system, however, are not necessarily physically manipulated by the robot nor perceived as such. In such a case, the virtual object may be scripted to move on its own to give the illusion of robot manipulation yet may fail in its illusion. This leads to a disconnect between the object moving and the robot's action of moving the object. When researching social robotics, this may have significant consequences on a user's perception of the robot (e.g., the robot's social presence). Therefore, to alleviate this complication, TOKCS is applied from the intended user perception of the designed system (i.e., if the system attempts an illusion of robot manipulation of a virtual object, it is classified under *Perceived Manipulability: Robot*).

Lastly, these MRIDE labels are only applied to virtual objects and are not tied to classifying VAM-HRI research under model, view, and control described in Sec. 3.1.1 and 3.2.1. VAM-HRI studies a variety of modalities provided by VAM technologies. HMD data use for improving a robot's SLAM, for example, still categorizes as increasing the robot's internal complexity of model but is not applicable under Anchor Location or Perceived Manipulability. Thus, these MRIDEs characteristics are designed for and only applied to virtual objects within VAM-HRI.

3.2.3 Classifying VAM Hardware Within TOKCS

Due to hardware technology making significant advances every year, labeling the specific technology (e.g., HoloLens 2) is important when classifying hardware within TOKCS. The hardware is not always described within VAM-HRI work, but is integral to understanding an interaction, as described in Sec. 2.3.5. These hardware technologies include HMDs, projectors, displays, and peripherals. These hardware technologies then fall under these categories for TOKCS.

3.2.4 Enumerating VAM Software Within TOKCS

There are a variety of software applications for facilitating 3D environments for VAM-HRI research. The most popular platforms like Unity3D support a wide variety of VR and MR hardware like those outlined in Section 3.2.3, and offer packages for networking with robot networks like ROS (Quigley et al. 2009) servers and rendering robot sensor data. ROS also offers a robot simulator, Gazebo (Koenig and Howard 2004), that directly interfaces with ROS applications and which has been used for VAM-HRI research. Other additional software generally relevant to HRI research is also included here, such as tracking AR tags to detect object poses using TagUp (Barentine et al. 2021). Software is not a direct part of the interaction as hardware, but relevant software are reported for a holistic understanding of what resources the VAM-HRI community uses to develop their applications.

3.2.5 Framework Limitations

The TOKCS framework was designed to capture and classify the key characteristics of VAM-HRI systems. However, the framework may become incomplete as advancements in both VAM-HRI research and VAM technology capabilities lead to new key characteristics differentiating VAM-HRI systems of the future. As field of VAM-HRI advances, the classification framework will need to grow as well.

3.3 Demonstrating TOKCS

TOKCS characterizes VAM-HRI systems with the following:

Anchor Location {User, Env, Robot} – where is the intended user perception of the virtual object's coordinate frame anchor (Sec 3.2.2);

Perceived Manipulability {User, Robot, None} – the intended user perception of "who" is able to or is currently manipulating the virtual object (Sec. 3.2.2);

Increases Expressivity of View (EV) $\{0,1\}$ – VAM technology is used to more explicitly show a robot's internal model such as using virtual objects to visualize robot sensors (Sec. 3.1.1);
Increases Flexibility of Controller (FC) $\{0,1\}$ – using VAM technology to add control modality to a robot (Sec. 3.1.1);
Increases Complexity of Model (CM) $\{0,1\}$ – using VAM technology to help the robot's understanding of the environment and/or the interaction (Sec. 3.2.1);
Milgram Continuum $\{$AR, AV, VR$\}$ – classification of which form of virtuality is being used (Sec. 3.1.3);
Hardware Description – which VAM technology is used (Sec. 3.2.3);
Software Description – which VAM software is used (Sec. 3.2.4)

TOKCS is applied to papers from the 4^{th} International Workshop on VAM-HRI to understand the ways in which researchers have been developing new technologies that leverage VAM. This was the most recent VAM-HRI workshop at the time of TOKCS creation. The ten papers and their categorization within the TOKCS are summarized in Table 3.1.

The ten papers cover a variety of contributions. In most cases, the presented system focused its improvements on a specific dimension of the TOKCS; five of the ten papers developed improvements within a single dimension. The two that contributed expansions along all three dimensions leveraged AR/VR in a domain that had previously not utilized AR/VR. For example, Higgins et al. (2021) developed a method for training grounded-language models in VR, instead of with real world robots. Ikeda and Szafir (2021) leverages AR-headsets for robotic debugging, where previous methods had used 2D screens. Four of the ten papers increased expressivity of view (EV), four increased the flexibility of the controller (FC), and three improved upon the robot internal complexity of model (CM). Of these 10 papers, half can be described as VR, three are AV, and two are AR. The majority of the presented methods are anchored at the environment level. Two methods' anchor is located at the robot and two are located at the user. If a perceived manipulable is available, it is typically available at the user level.

Table 3.1: Summary of TOKCS. Up arrow symbols (↑) indicate that the work increases the functionality within this aspect of TOKCS. Blank entries indicate that the contributions of the respective paper for this aspect are consistent with prior work. Column acronyms and abbreviations: AL → Anchor Location; PM → Perceived Manipulability; EV → Expressivity of View; FC → Flexibility of Controller; CM → Complexity of Model; MC → Milgram Continuum (Milgram et al. 1995a)

Paper	AL	PM	EV	FC	CM	MC	Software	Hardware
Boateng and Zhang (2021)	Robot, Env		↑			AR	Unity	Hololens video recordings via MTurk
Ikeda and Szafir (2021)	Env	User	↑	↑	↑	AR	Unity	Hololens
LeMasurier et al. (2021)	Env, Robot	User		↑		AV	Unity, ROSNET, ROS	HTC Vive
Puljiz et al. (2021)					↑	AV	Unity	Hololens
Wadgao- nkar et. al [20]	Env, Robot		↑			AV	Unity	HTC VIVE
Barentine et al. (2021)	Env			↑		VR	Unity, TagUp	Oculus Quest VR headset & controllers
Higgins et al. (2021)	User	User	↑	↑	↑	VR	Unity, ROS#, ROS, Gazebo	SteamVR headset
Mara et al. (2021)	Env	Robot, User				VR	Unity	HTC VIVE Pro Eye & Leap Motion
Mimnaugh et al. (2021)						VR	Unity	Oculus Rift S
Mott et al. (2021)	Env, User		↑			VR	Unity	MTurk Web Video of VR

A broad range of utilized hardware and software was also observed. Unity (Haas 2014) was overwhelmingly popular among papers as the 3D game engine of choice; nine of the ten papers explicitly mention Unity3D. The most popular HMD mentioned was the Hololens (Garon et al. 2016), which was used in three of the papers.

3.4 TOKCS for SAR

TOKCS, constructed largely out of the MVC paradigm, is a framework that can be applied to HRI, including the subfield of SAR. The defining characteristic of SAR as a subfield of HRI is its focus

on *social and assistive* interactions. Therefore, TOKCS can be particularly useful for designing and implementing social actions within assistive contexts in SAR.

The remainder of this dissertation explores how TOKCS can be applied to SAR. In particular, it discusses and demonstrates how to improve the robot's internal complexity of user models, increase the social and functional expressivity of robots, and develop new interaction paradigms for human-robot flexibility of controller within assistive kinesthetic contexts of STEM education. This will provide a comprehensive understanding of how TOKCS can be used to define, design, and implement effective and engaging interactions between humans and robots in SAR.

Chapter 4

Model - Expanding Internal Complexity of User Models for SAR

This chapter discusses the use of AR modalities for modeling user data. It highlights the importance of understanding the relationship between these data and the user's state. Through the examination of different use cases such as modelling student kinesthetic curiosity and AR usability, it illustrates that AR-based approaches have great potential for capturing user data and personalizing interactions in a range of contexts.

4.1 Student Kinesthetic Curiosity

Contributors: Section 4.1 is based on Groechel et al. (2021). Additional authors of the published work include Roxanna Pakkar, Roddur Dasgupta, Chloe Kuo, Haemin Jenny Lee, Julia Cordero, Kartik Mahajan, and Maja J. Matarić.

Surveys of SAR tutors reports that SAR tutoring has shown great promise with various learners, especially when using multiple interaction modalities. The work to date has largely focused on one-on-one learning companions for children with an emphasis on personalizing the learning interaction (Clabaugh and Matarić 2019). This is often characterized by Bloom's two sigma problem (Bloom 1984) where students performed two standard deviations better when tutored one-on-one

compared to traditional one-to-many lecture contexts. To personalize toward individual student needs, SAR tutor interactions have adopted interfaces (e.g., tablets) that increase the observability of student actions. The data from those interfaces are used to pursue multimodal reasoning about hidden student state such as a student's knowledge (Schodde et al. 2017), affect (Spaulding and Breazeal 2019), or engagement levels (Jain et al. 2020b). The emphasis on seated learning interactions creates a need to explore kinesthetic learning contexts (i.e., embodied learning (Macedonia 2019)) that include the well-documented benefits of the physical embodiment of the SAR tutor (Deng et al. 2019).

Embodied learning can be effectively explored in VAM-HRI settings. Consequently, this area of research has grown in recent years (Williams et al. 2020d), focusing on design (Williams et al. 2020c), teleoperation (Lipton et al. 2018), and signalling challenges (Walker et al. 2018) (Groechel et al. 2019). Many of the studied interactions employ augmented reality head-mounted displays (ARHMDs) that generate rich multimodal data minimizing or removing the need for external sensing.

This work explores using such rich, multimodal data for personalizing student interactions in an embodied learning context. This work creates synergies between SAR tutors, VAM-HRI, and embodied learning through the design and implementation of MoveToCode (M2C), an open-source, AR programming platform that interfaces with a robot tutor (Groechel et al. 2020) as seen in Fig. 4.1. This work introduces a real-time, multimodal measure of **student kinesthetic curiosity** (**KC^S**) and analyze how a curious robot tutor's actions impact KC^S during an interaction.

4.1.1 Technical Approach

The key insight of this work is combining the components of embodied learning (e.g., movement) and student curiosity (e.g., seeking new information) into a single measure of *kinesthetic curiosity* (KC^S) that is personalized for each student. To better understand how the design of KC^S can be leveraged for learning experiences with a robot, robot action policy is designed that uses KC^S_t.

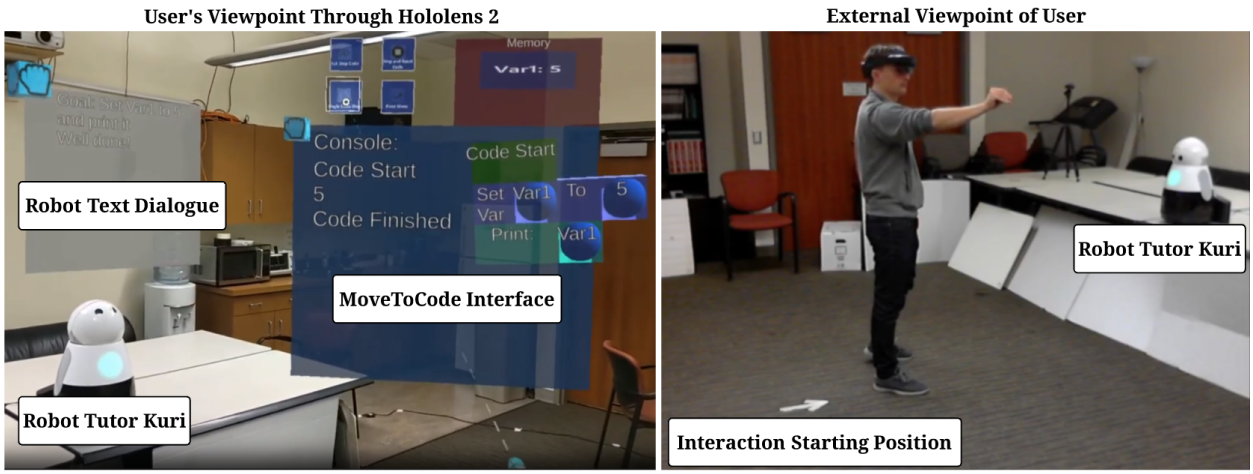


Figure 4.1: M2C Interaction; participants attempted to solve coding exercises involving 3D code blocks alongside the robot tutor, Kuri.

With the hypotheses given in Sect. 4.1.2, participants completed AR coding exercises with a curious SAR tutor described further in Sect. 4.1.2.

Measuring and Personalizing Kinesthetic Curiosity

To inform a robot's action policy, our real-time measures used a sliding-window approach to measure KC_t^S, a student's kinesthetic curiosity at a given time:

$$movement_t^S = \sum_{n=t-tw+1}^{t} dist(head\_pose_n, head\_pose_{n-1}) \tag{4.1}$$

$$curiosity_t^S = \sum_{n=t-tw}^{t} [ISA_n^S \neq NULL] \tag{4.2}$$

$$KC_t^S = w_0 * \frac{movement_t^S - \overline{movement^S}}{\sigma_{movement^S}} + w_1 * \frac{curiosity_t^S - \overline{curiosity^S}}{\sigma_{curiosity^S}} \tag{4.3}$$

where $movement_t^S$ (4.1) is measured with accumulated head pose change over a sliding time window tw, and $curiosity_t^S$ (4.2) is measured as the sum of information seeking actions (ISAs) over tw. ISAs are defined relative to the domain and action space of the learner. Specifically for this work, ISAs included snapping code blocks (Fig. 4.2), unsnapping code blocks, pressing interaction menu buttons (Fig. 4.1), and creating new code blocks. KC_t^S (4.3) assumes an underlying Gaussian distribution for $movement^S$ and $curiosity^S$ for all instances of time from 0 to t. This measure is a weighted combination of $movement_t^S$ deviation and $curiosity_t^S$ deviation from their respective mean

(i.e., z-normalization (Mariéthoz and Bengio 2005)). The resulting normalized scores are therefore personalized to each student at time t.

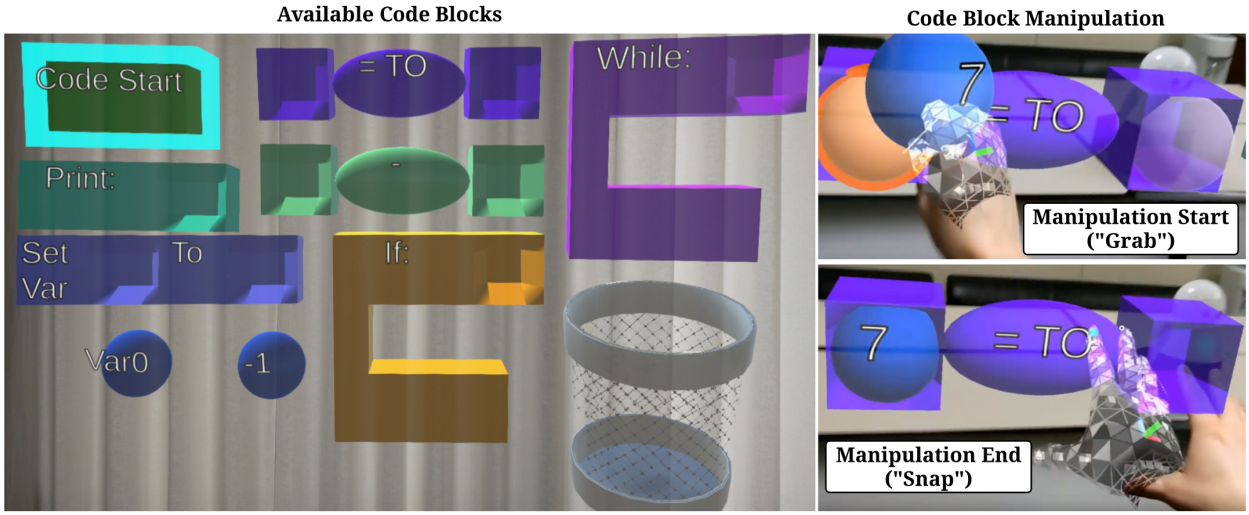


Figure 4.2: Available M2C code blocks (left) as seen by the participant through the Hololens 2. Code block manipulation (right) with a participant grabbing the block and letting it go to snap code blocks together.

For post-interaction analysis, a student's KC^S and a robot's KC^R values for an entire interaction consisted of the following:

$$KC^S = (movement^S, curiosity^S) \tag{4.4}$$

$$KC^R = (movement^R, curiosity^R) \tag{4.5}$$

where $movement^S$ is the student's total movement throughout the interaction, and $curiosity^S$ is the total number of ISAs throughout the interaction, allowing for normalization across interactions of varying lengths. Using the same set of actions to measure KC^S, a robot tutor's KC^R can be evaluated for both the real-time measure KC_t^R and the post-interaction analysis of KC^R.

***KC* Robot Tutor Action Policy**

The robot's adjustable rule-based policy, shown in Table 4.1, was designed using a changeable threshold T^{KC} and time-window tw that triggered robot actions based on the time since last action $tsla^R$. For example, a lower T^{KC} causes the robot to take more information-seeking actions in

order to motivate the user to do the same. The rule-based policy was designed to support a data collection for informing future data-driven policies.

Table 4.1: Robot Action Policy

Action	**Activation State**
Exercise Goal Dialogue	New Exercise Start
Virtual ISA	$KC_t^S < T^{KC}$ and $tsla^R \geq tw$
Positive Physical Affect	$KC_t^S \geq T^{KC}$ and $tsla^R \geq tw$ or Dialogue
Congratulatory Dialogue	Correct Answer
Scaffolding Dialogue	Incorrect Answer and Scaffolding Dialogue Left
Encouraging Dialogue	Incorrect Answer and ¬Scaffolding Dialogue Left

4.1.2 User Study

A single-session within-subjects experiment was conducted with approval by our university IRB (UP-17-00226) to evaluate a curious robot tutor policy (Table 4.1) with differing thresholds (T^{KC}). Ten participants (3F,7M) were recruited from the University of Southern California student population, with an age range of 19-27 ($\bar{x} = 22.8, \sigma = 2.9$). P8 experienced two separate operating system crashes (at 285.92 s and 341.44 s); having not experienced both experimental conditions, P8 was therefore removed from the behavioral data analysis.

Participants wore the ARHMD (Microsoft Hololens 2) and attempted to complete programming exercises with the help of a robot tutor, a Mayfield Robotics Kuri (Fig. 4.1). Kuri used the action policy described in Table 4.1 with tw empirically set to 20 seconds. The independent variable in the study was the robot KC threshold level using $T_{high}^{KC} = 0.5$ and $T_{low}^{KC} = -0.5$. The experiment lasted 20 minutes with T^{KC} substituted at 10 minutes. ISAs (see Eq. 4.2) in this experiment included snapping code blocks, unsnapping code blocks, pressing interaction menu buttons, and creating new code blocks.

User Study Hypotheses

A user study was performed (described in Sect. 4.1.2) to evaluate the following hypotheses regarding KC^S with equal weights ($w_0 = w_1 = 0.5$):

H1: KC^S data fulfill the stationarity assumption needed for z-normalization.

H2: Robot virtual information seeking actions (i.e., curious actions) will positively affect student KC_t^S.

H3: A more curios robot (i.e., lower T^{KC}) will encourage a higher KC^S when compared to a less curious robot (i.e., higher T^{KC}).

For **H1**, KC_t^S assumes an underlying Gaussian distribution over the time series which implies that the time series data are stationary. For **H2**, this work examined if robot virtual information seeking action (i.e., curious actions) could positively affect KC_t^S to better inform future robot action selection policies. For **H3**, this work tested the differences in conditions to examine longer interaction effects of a more or less curious robot.

4.1.3 Results and Analysis

KC_t^S depends on the time series for $movement_t^S$ and $curiosity_t^S$ to be stationary as they are modeled with the underlying assumption of a constant mean and variance needed for z-normalization. An Augmented Dickey-Fuller test was performed on each participants' $movement_t^S$ and $curiosity_t^S$ measures over the interaction to test for stationarity. With the exceptions of $movement_t^{P1}(p = .017, DF_\tau = -3.248)$ and $curiosity_t^{P2}(p = .012, DF_\tau = -3.378)$, all tests reported a significance of $p < .01$, supporting **H1**.

To analyze the effect of robot virtual information seeking actions (ISAs), the difference of measure (M) from time t to time $t + tw$ ($\Delta M_{t,t+tw}$) was calculated for all robot ISAs at time t. The robot totaled 170 ISAs with $\Delta M_{t,t+tw}$ distributions tested for normality (Fig. 4.3). A two-sided, single sample t-test was performed against a mean of 0. Significant results were found for all measures: $\Delta movement_{t,t+tw}^S(m)$ $(t = 4.51, p < .001, \bar{x} = 0.495, d = 0.35)$; $\Delta curiosity_{t,t+tw}^S(ISA)$ $(t = 4.637, p < .001, \bar{x} = 0.776, d = 0.36)$; $\Delta z\left(movement_{t,t+tw}^S\right)$ $(t = 4.623, p < .001, \bar{x} = 0.477, d = 0.36)$; $\Delta z\left(curiosity_{t,t+tw}^S\right)$ $(t = 5.087, p < .001, \bar{x} = 0.452, d = 0.39)$; $\Delta KC_{t,t+tw}^S$ $(t = 6.764, p <$

$.001, \bar{x} = 0.464, d = 0.52$). These findings demonstrate a positive short term effect of robot virtual ISAs on KC_t^S, supporting **H2**.

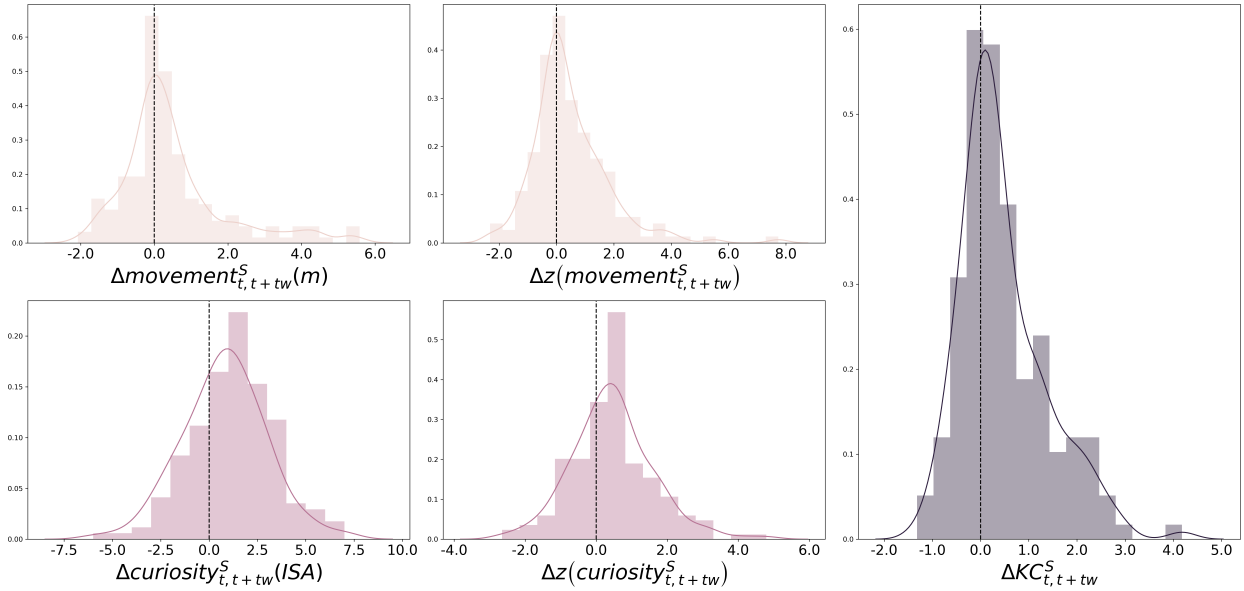


Figure 4.3: Action distributions for all differences in measures where the robot took an ISA at time t to the score at time $t+tw$ ($\Delta M_{t,t+tw}$). Zero line is plotted to show distribution shifts.

Differences in total $movement^S$ and $curiosity^{S,R}$ measures were analyzed between the robot action policy threshold conditions of $T^{KC}_{high} = 0.5$ and $T^{KC}_{low} = -0.5$ shown in Fig. 4.4. A Wilcoxon signed-rank test indicated a significant effect for $curiosity^R(ISA)$ ($\tilde{x}_{high} = 15, \tilde{x}_{low} = 3, W = 0, p = .008$). No significant effect was found for $movement^S(m)$ ($\tilde{x}_{high} = 52.2, \tilde{x}_{low} = 39.93, W = 21, p = .859$) or $curiosity^S(ISA)$ ($\tilde{x}_{high} = 90, \tilde{x}_{low} = 130, W = 11, p = .172$). These findings support the T^{KC} thresholds chosen but do not support a difference in KC^S between conditions posited by **H3**.

4.2 AR Behavioral Data for Usability

Contributors: Section 4.2 is based on Mahajan et al. (2020) written with co-first author Kartik Mahajan. Additional authors of the published work include Roxanna Pakkar, Julia Cordero, Haemin Jenny Lee, Maja J. Matarić.

SAR tutoring has been extensively explored with a variety of users and usability studies (Papadopoulos et al. 2020). Advances in VAM-HRI have enabled kinesthetic AR environments – as

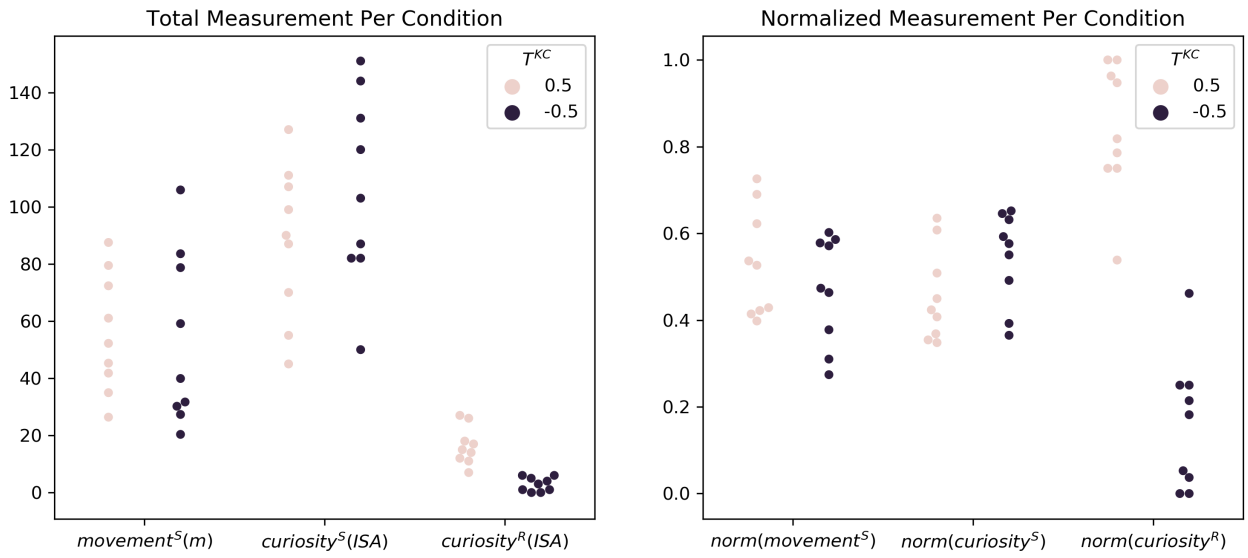


Figure 4.4: Measures between the more curious ($T^{KC} = -0.5$) and less curious ($T^{KC} = 0.5$) robot conditions. The left depicts the total score for each measure for all participants. The right depicts normalized measures for each participant between conditions.

described in Sec. 4.1 – where students move around and physically interact with coding blocks alongside a SAR tutor. Effective evaluation of usability of these nascent 3D interfaces requires a re-evaluation of common usability metrics employed in other tutoring environments.

Usability has been studied in various tutoring environments using subjective and objective metrics. Subjective metrics include interview summaries (Malik et al. 2016) (Pino et al. 2015) and various survey tools, typically using Likert scales (Keizer et al. 2019), such as the commonly used System Usability Scale (SUS), a 0-100 scale. Objective metrics include user performance, manipulation time, and gaze (Feingold-Polak et al. 2018)(Caleb-Solly et al. 2018). SUS is used for evaluating both on-line and in-person tutoring, while objective metrics are more commonly used in on-line tutoring. In SAR tutors, observability is typically limited; this challenge is even greater in kinesthetic environments, where students move around.

This work applied objective usability metrics commonly used in seated 2D tutoring environments to data from a kinesthetic AR environment pilot study ($n = 9$) and validated those metrics with post-interaction SUS survey data. Three different usability metrics were studied: 1) student performance via problem-solving policies; 2) object manipulation time; and 3) gaze concentration.

The metrics were recorded over a 20 minute interaction as described in Sec. 4.1.2 involving a SAR tutor guiding a student through 7 coding exercises via an AR visual programming language M2C (Groechel et al. 2020). The strength of each usability metric was compared to subjective survey-based scores measured with the System Usability Scale (SUS). The results show that usability scores were correlated with the gaze metric but not with the manipulation time or performance metrics. The findings provide interesting implications for the design and evaluation of kinesthetic AR robot tutoring environments.

4.2.1 Technical Approach

Since VAM-HRI for SAR tutoring is a nascent area, to understand usability metrics for kinesthetic AR tutoring contexts, usability studies of programming tutors (Piech et al. 2015) and web interfaces (Wang et al. 2019) were reviewed, and chose three commonly used reliable metrics: user performance, manipulation time, and gaze concentration. These metrics were then adapated based on the context of our study.

Student Performance via Problem-Solving Policies: a participant's performance was measured by counting the number of good and bad policies created during a time frame. Introduced by Piech et al. (2015), a *policy* is defined as any group of two or more code blocks. For example, if the participant was tasked with adding integer block 1 and integer block 2, a correct solution would include combining the two integer blocks with an addition block. A *Good* Policy (*GP*) for this task includes combining the integer block 1 with the addition block. A *Bad* Policy (*BP*) includes some other, incorrect step(s), such as combining the integer block 3 and the addition block.

Manipulation Time: *MT* is defined as the amount of time it takes a participant to grab a coding block and *snap* it to another block as can be seen in Fig. 4.2. *Snapping* was defined as the action grabbing a code block, dragging it to be in contact with another code block, and then releasing the currently held block to snap it to the contacted block. A successful manipulation event was logged from the time when an object was first grabbed (t_{grab}) to when an object was *snapped* to another object (t_{snap}).

Gaze Concentration: *GC* is defined as the amount of time a participant looked a 2D (x,y) pixel (i.e., cell) within the *interaction space* over a rolling time window tw_{GC}. The *interaction space*, shown in Fig. 4.2, included the M2C interface and the physical robot tutor. The interaction space, measured in meters, was a 4m x 2.25m grid, totalling 3,600 cells measuring 0.05m x 0.05m each. A cell's score increased by 0.01 every frame the participant looked at that cell during tw_{GC}. The maximum cell score was capped at 1.

4.2.2 User Study

The dataset used in this work was from the within subjects ($n = 9$) study performed with an AR visual programming language M2C described in Sec. 4.1.2. In this study (Fig. 4.2) a SAR tutor aimed to increase a student's *kinesthetic curiosity* (*KC*), a metric involving the multimodal measure of a student's movement and curiosity. In the interaction, students combined *coding blocks* (e.g., if-blocks, print-blocks) by grabbing, dragging, and snapping blocks together in order to solve 7 beginner-level coding exercises. The acts of grabbing, dragging, and snapping blocks are part of the **manipulation time** metric, described in Sec. 4.2.1. Preset coding blocks were available to the participant at the beginning of each exercise. Tasks focused on building syntactic skills for integer addition, variable creation, and if-statements. The study and its results are under review for publication elsewhere.

The dataset includes 9 (2F,7M) of the 10 participants who were University of Southern California students with age range 19-27 ($\bar{x} = 22.8, \sigma = 2.9$). Participant 8 had two operating system crashes and was discarded from analysis. Behavioral data were collected at 0.02 sec intervals (50Hz) totaling 180 min of time series data yielding 540,000 rows. After the interaction, participants completed a ten-question questionnaire designed to measure individual SUS ratings.

This work examines the participants' objective and subjective metrics of usability. Specifically, it considers SUS survey results and the logged behavioral data of policy-evaluation, object manipulation time, and gaze concentration, described in the next section.

4.2.3 Results and Analysis

Data Analysis

All statistics were distributed between 0 and 1 using MinMax Scaling from Python's *sklearn* package (v0.24.2):

$$X_{scaled} = \sigma_x * (X_{max} - X_{min}) + X_{min} \tag{4.6}$$

where X_{scaled} is the new value for a data point in column X. To reduce skew of manipulation time (MT) results, a max MT of 10 seconds was empirically chosen, leaving 95.1% of the data. Any times over 10 seconds were adjusted to 10 seconds.

Not all participants completed all exercises; P5 failed to complete exercise 3 and P1,2,6,9 failed to complete exercise 6. This resulted in differently sized datasets for the different participants.

Post-interaction SUS scores are calculated for all participants based on a 10-question survey, as shown in Fig. 4.5 ($\bar{x} = 53.06, \tilde{x} = 55.0, \sigma^x = 17.2, CV = 32.8\%$).

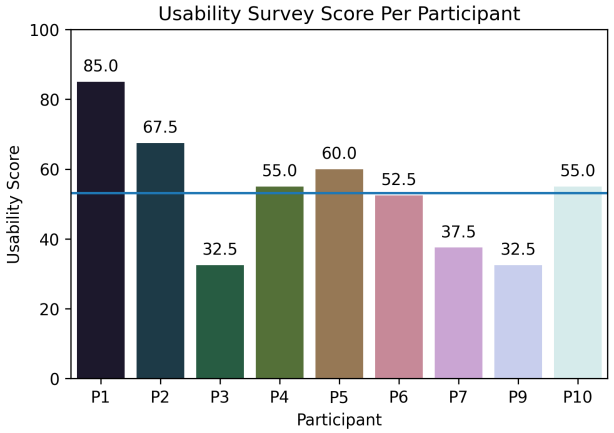


Figure 4.5: SUS rating (0-100) of the M2C interaction. The line indicates the median rating ($\tilde{x}$).

Unimodal Metric Analysis

The variance and correlation of each metric to the SUS score is also calculated (Fig. 4.5). A Levene's test was used to identify significant ($p < .05$) variance of each metric across participants, to signify the metric may distinguish different user behavior. Spearman's r correlation tests are used to validate the significance ($p < .05$) of each metric. Correlation of variance (CV) of each metric for unitless comparisons between metrics relative to their dispersion is also reported.

Policy-Evaluation Results

To evaluate user performance, total Good Policies GP ($\bar{x} = 24.125, \sigma = 11.243, CV = 46.6\%$) and total Bad Policies BP ($\bar{x} = 3.625, \sigma = 3.24, CV = 89.2\%$) were recorded per participant and per exercise (Fig. 4.6). A Pearson's r test showed that there was no significant correlation between the total GP and BP ($r_p(9) = 0.581, p = .100$). A Levene's test indicated unequal variances per participant over exercises for total GP ($F = 6.72, p = .001$) yet no significant variance for total BP ($F = 0.993, p = .439$). This supports that GP may be effective in helping to differentiate user behavior, whereas BP may not be, due to its consistency across all participants. A Spearman's r correlation indicated no significant relationship between total GP and SUS score ($r_s(9) = -0.369, p = .327$), indicating total GP is not indicative of SUS score when observed unimodally. A Spearman's r correlation indicates no significant relationship between total BP and SUS score ($r_s(9) = -0.340, p = .370$), supporting that total BP is not indicative of SUS score.

These findings show that neither of the user performance metrics (GP or BP) alone were significant indicators of usability.

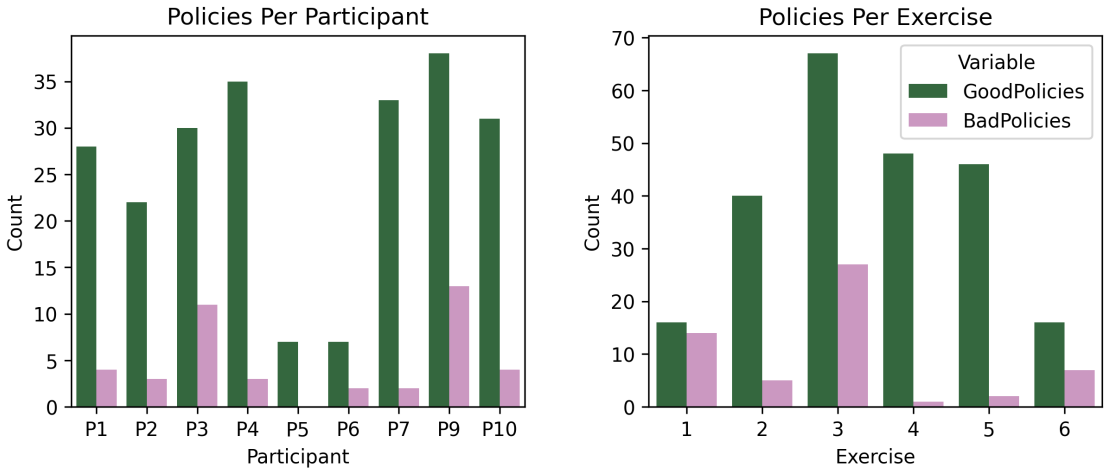


Figure 4.6: Good Policies (GP) and Bad Policies(BP) viewed per participant and per interaction. Exercise 7 was free-play so there are no GP or BP recorded for it.

Manipulation Time Results

To evaluate Manipulation Time MT, average MT per participant ($\bar{x} = 2.93s, \sigma^x = 0.76s, CV = 25.9\%$) and per exercise ($\bar{x} = 2.80s, \sigma^x = 0.57s, CV = 20.3\%$) were recorded, as shown in Fig. 4.7.

A Levene's test indicated a significant variance among participant's average MT per exercise ($W = 5.94, p < .0001$), indicating MT is able to differentiate user behavior. A Spearman's r correlation indicated no significant correlation between average MT and SUS score ($r_s(9) = 0.353, p = .351$), indicating that average MT is not indicative of SUS score.

As an additional MT metric, σ^{MT} ($\bar{x} = 3.51, \sigma^x = 0.816, CV = 23.2\%$) was calculated. A Spearman's r correlation also showed no significant correlation between σ^{MT} and SUS score ($r_s(9) = -0.417, p = .263$), indicating that average σ^{MT} was not indicative of the SUS score.

These findings indicate that neither of the Manipulation Time (MT) metrics alone was a significant indicator of usability.

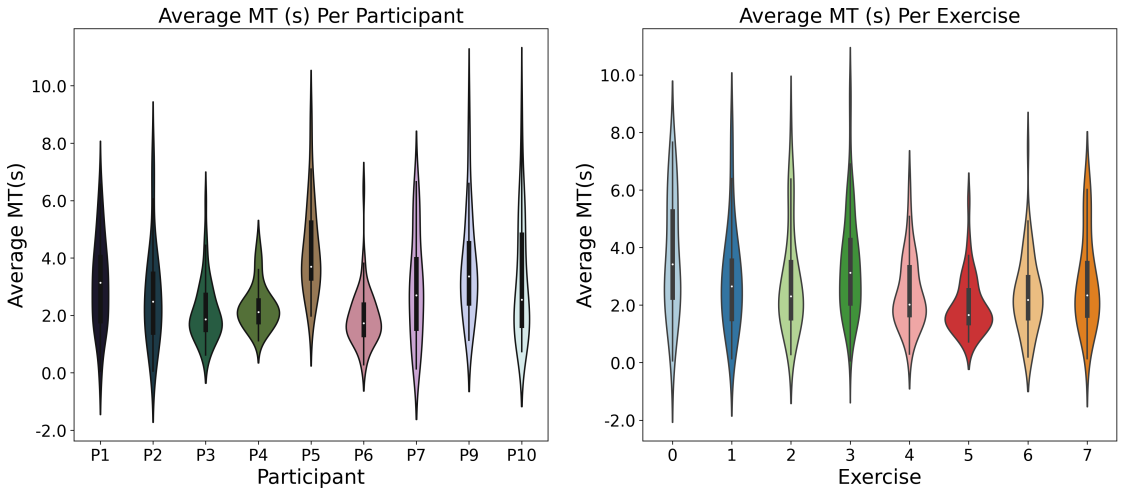


Figure 4.7: Average Manipulation Time (MT) per exercise and per participant. MT records time between when a student chooses a coding block (e.g. if-statement, integer block) and snaps it to another component. Refer to Sec. 4.2.1 for more detail.

Eye Gaze Concentration Results

To analyze eye gaze concentration GC, tw_{GC} was empirically set to 10 seconds. High intensity cell reads (HR) were defined as any cell with a value of 0.9 or higher, because 0.9 is over two standard deviations ($\sigma^{GC} = 0.306$) away from the average ($\bar{x} = 0.181$).

To evaluate HR, the total HR per participant was recorded ($\bar{x} = 153.44, \sigma^x = 23.733, CV = 15.4\%$) as shown in Fig. 4.8. A Levene's test showed a significant difference in variance among HR per time-step ($F = 38.1, p < .0001$), indicating that HR may distinguish user behavior. A

Spearman's r correlation also indicates a significant relationship between total HR and the SUS score ($r_s(9) = 0.77, p = .014$), supporting that total HR is indicative of SUS score.

As an additional metric of GC, σ^{GC} ($\bar{x} = 23.42, \sigma = 14.22, CV = 15.4\%$) was calculated based on 2D coordinates of gaze to represent how a participant's gaze traveled over a window. A Spearman's r correlation showed no significant relationship between σ^{GC} and the SUS score ($r_s(9) = 0.235, p = .542$), indicating that σ^{GC} is not indicative of SUS score.

These findings indicate that HR was a significant metric for usability, whereas σ^{GC} was not.

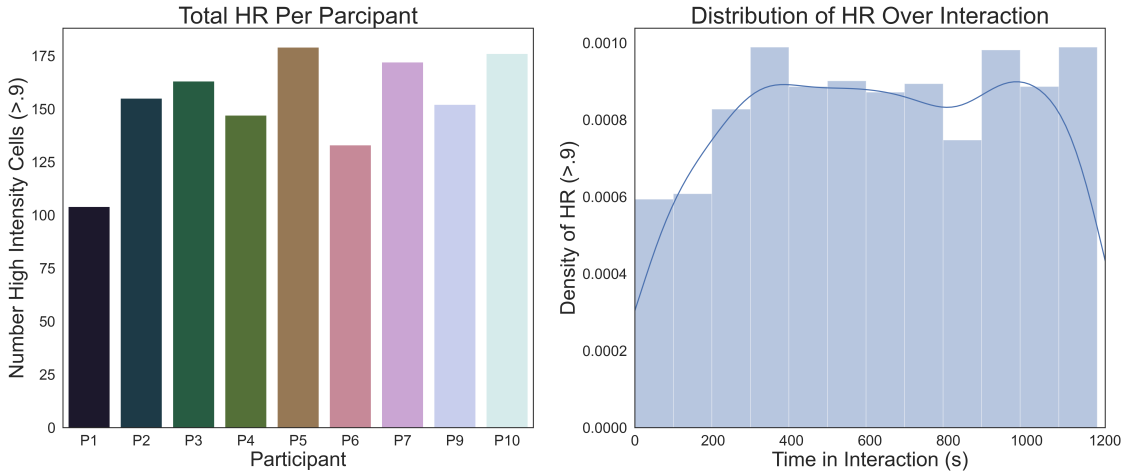


Figure 4.8: Total High Intensity Cell Reads (score > 0.9) HR per participant recorded over a rolling time window $tw_{GC} = 10$. Cells are defined as any 2D pixel in the interaction space. For more details, see Sec. 4.2.1.

4.3 Discussion and Summary

This chapter discussed the use of AR modalities for modeling user data. The sections presented in this chapter highlight the importance of understanding the relationship between these data and the user's state. Through the examination of the collected data, AR-based approaches show great potential for accurately capturing user data and personalizing interactions in a range of contexts. Overall, this chapter highlights the need for more research to better understand how to utilize AR modalities for modeling user data and the relationship of this data to the user's state.

First the chapter addressed the gap in understanding how to personalize interactions in kinesthetic (i.e., embodied) learning contexts in SAR tutoring. A multimodal measure of student kinesthetic curiosity (KC^S) was proposed that combines a student's movement and curiosity measures into a single, personalized measure. The study results indicate that the robot tutor was able to successfully use KC^S to personalize its action policy, positively affecting short-term KC^S. However, no significant results were found for longer state changes for each student. The AR visual programming language (M2C) created for this work has been made open-source. This study aims to inform future online features and measures for AR HRIs, shedding light on the importance of multimodal personalization measures in kinesthetic learning contexts.

Additionally, this research addresses the need to re-evaluate usability metrics for kinesthetic AR environments in SAR tutoring. Gaze was found as a key usability metric in this context, but not manipulation time or performance metrics. This contributes to a deeper understanding of how to evaluate and improve the usability of kinesthetic AR robot tutoring environments.

Chapter 5

View - Increasing SAR Social and Functional Expressivity of View

This chapter discusses using AR to increase a socially assistive robot's social and functional expressivity of view. AR animations of appendages can enhance social expressivity, and improving existing functional designs can accommodate individual user preferences. Design considerations for increasing both social and functional expressivity are also discussed, taking into account contextual factors and trade-offs.

5.1 AR Arms to Increase Robot Social Expressivity

Contributors: Section 5.1 is based on Groechel et al. (2019). Additional authors of the published work include Zhonghao Shi, Roxanna Pakkar, and Maja J. Matarić.

As described in Sec. 2.2, SAR have been shown to have positive impacts in a variety of domains, from stroke rehabilitation (Matarić et al. 2007) to tutoring (Clabaugh et al. 2017). Such robots typically have low expressivity due to physical, safety, and cost constraints. *Expressivity* in HRI refers to the robot's ability to use its modalities to non-verbally communicate the robot's intentions or its internal state (Charisi et al. 2019). Higher levels of expressiveness have been shown to increase trust, disclosure, and companionship with a robot (Martelaro et al. 2016). Expressivity can be conveyed with dynamic actuators (e.g., motors) as well as static ones (e.g., screens, LEDs)

(Balit et al. 2018). HRI research into gesture has explored head and arm gestures, but many non-humanoid robots partially or completely lack those features, resulting in low social expressivity (Cha et al. 2018).

Social expressivity refers to expressions related to communication of affect or emotion. In social and socially assistive robotics, social expressivity has been used for interactions such as expressing the robot's emotional state through human-like facial expressions (Chen et al. 2018; Meghdari et al. 2016; Kkedzierski et al. 2013), gestures (Cha et al. 2018), and poses (Bretan et al. 2015). In contrast, *functional expressivity* refers to the robot's ability to communicate its functional capabilities (e.g., using turn signals to show driving direction). Research into robot expressiveness has explored insights from animation, design, and cognitive psychology (Charisi et al. 2019).

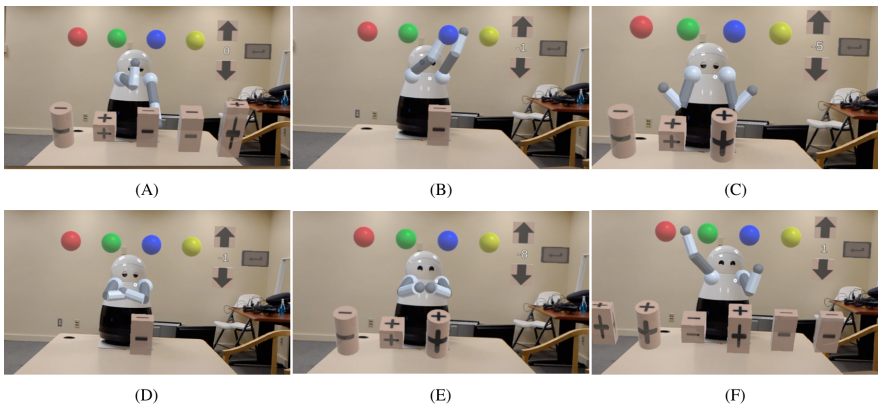


Figure 5.1: This dissertation explored how AR robot extensions can enhance low-expressivity robots by adding social gestures. Six AR gestures were developed: (A) facepalm, (B) cheer, (C) shoulder shrug, (D) arm cross, (E) clap, and (F) wave dance.

The importance of expressivity and the mechanical, cost, and safety constraints of physical robots call for exploring new modalities of expression, such as through the use of AR. As defined in Sec. 3.1.3, *AR* refers to virtual objects projected onto the real world while respecting physical reality and reacting to it.

Using virtual modalities for robots has led to the emerging field of VAM-HRI. VAM-HRI has already had advances in the functional expression of a robot (Walker et al. 2018; Williams et al. 2019b) but has not yet explored social expressiveness. Introducing such expressiveness into VAM-HRI allows us to leverage the positive aspects of physical robots–embodiment and physical

affordances (Deng et al. 2019)–as well as the positive aspects of AR–overcoming cost, safety, and physical constraints. As outlined in Sec. 1.3, *this section aims to synergize the combined benefits of the two fields by creating AR, socially expressive arms for low social expressivity robots (Fig. 5.1).*

This section describes the design, implementation, and validation of AR arms for a low-expressivity physical robot. A user study was conducted where participants completed AR mathematics task with a robot. This new and exploratory work in VAM-HRI did not test specific hypotheses; empirical data were collected and analyzed to inform future work. The results demonstrate a higher degree of perceived robot emotion, helpfulness, and physical presence by users who experienced the AR arms on the robot compared to those who did not. Participants who reported a higher physical presence also reported higher measures of robot social presence, perceived ease of use, usefulness, and had a more positive attitude toward using the robot with AR. The results from the study also demonstrate consistent ratings of gesture valence and identification of gesture intent.

5.1.1 Technical Approach

To study AR arms, a mobile robot with very low expressivieness was chosen: the Mayfield Robotics Kuri (Fig. 5.2), formerly a commercial product. Kuri is 50 cm tall, has 8 DoF (3 in the base, 3 between the base and head, and 1 in each of the two eyelids), and is equipped with an array of 4 microphones, dual speakers, lidar, chest light, and a camera behind the left eye. While very well engineered, Kuri is an ideal platform for the exploration of VAM-HRI in general, and AR gestures in particular, because of its lack of arms.

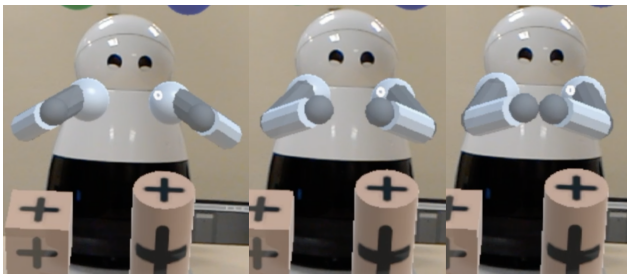


Figure 5.2: Keyframes for Kuri's clapping animation.

Implementation

The Microsoft Hololens ARHMD was used, which is equipped with a 30° x 17.5° field of view (FOV), an IMU, a long-range depth sensor, an HD video camera, four microphones, and an ambient light sensor. The AR extension prototypes were developed in Unity3D, a popular game engine. For communication between Kuri and the Hololens, the open-source ROS# library was used (Bischoff 2018). The full arm and experiment implementation is open-source and available at `https://github.com/interaction-lab/KuriAugmentedRealityArmsPublic`.

A balanced set of positive gestures (a dancing cheer, a clap, and a wave-like dance) and negative gestures (a facepalm, a shoulder shrug, and crossing the arms) were developed, as shown in Fig. 5.1. The Unity built-in interpolated keyframe animator was used to develop each gesture animation and a simple inverse kinematic solver to speed up the development of each keyframe.

Gesture Design

In designing the virtual gestures for the robot, inspiration was taken from social constructs of collaboration, such as pointing to indicate desire (Tomasello 2010), and emblems and metaphoric gestures (McNeill 1992), such clapping to show approval. The inclusion of such gestures goes beyond the current use of audio and dance feedback found in SAR systems (Leyzberg et al. 2012; Clabaugh et al. 2017; Scassellati et al. 2018).

Work in HRI has explored Disney animation principles (Thomas et al. 1995), typically either in simulation or with physically constrained robots (Takayama et al. 2011; Gielniak and Thomaz 2012; Ribeiro and Paiva 2012). This work explored a subset of Disney principles–squash and stretch, exaggeration, and staging–in the context of AR arm gestures. Each principle was considered for its benefit over physical world constraints. Squash and stretch gives flexibility and life to animations bringing life to robots that are rigid. Exaggeration has been shown to aid robot social communication (Gielniak and Thomaz 2012). Staging was considered for its role in referencing objects using arms to allow for joint attention.

The animated gestures were accompanied by physical body expressions to make the arms appear integrated with Kuri. For positive gestures, Kuri performed a built-in happy gesture (named

"gotit") that involved the robot's head moving up and emitting a happy, rising tone. For negative gestures, Kuri performed a built-in sad gesture (named "sad") that involved the robot's head moving down and being silent.

5.1.2 User Study

A single-session experiment consisting of two parts was conducted with approval by university IRB (UP-16-00603). The first part was a two-condition between-subjects experiment to test the AR arms. All participants wore the ARHMD and interacted with both physical and virtual objects as part of a math puzzle game. The independent variable was whether participants had arms on their Kuri robot (Experiment condition) or not (Control condition). Subjective measures included perceived physical presence, social presence, ease of use, helpfulness, and usefulness from Heerink et al. (2010), adapted for the AR robot. Task efficiency was objectively measured using completion time as is standard in VAM-HRI (Walker et al. 2018). After the first part of the experiment was completed, a survey of 7-point Likert scale questions abd a semi-structured interview were administered.

The second part of the experiment involved all participants in a single condition. The participants were shown a video of each of the six AR arm gestures and asked to rate each gesture's valence on a decimal scale from very negative (-1.00) to very positive (+1.00), as in prior work (Marmpena et al. 2018), and to describe verbally, in written form, what each gesture conveyed.

Part 1: AR Mathematics Puzzles

Participants wore the Hololens and were seated across from Kuri (Fig. 5.3) with a set of 20 colored physical blocks on the table in front of them. The blocks were numbered 1-9. The block shapes were: cylinder, cube, cuboid, wide cuboid, and long cuboid. The block colors were: red, green, blue, and yellow. The participants' view from the Hololens can be seen in Fig. 5.4, with labels for all objects pertinent to solving the mathematics puzzle. The view included cream-colored blocks in the same variety of shapes, labeled with either a plus (+) or minus (-) sign. Participants

were asked to solve an addition/subtraction equation based on information provided on the physical and virtual blocks, and virtually input the numeric answer.

Figure 5.3: Participant wearing the Hololens across from Kuri (left). Two sides of a single physical cuboid block (right).

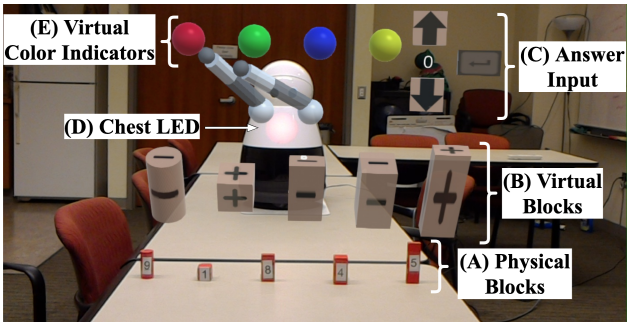


Figure 5.4: View when clicking a virtual block. Kuri is displaying red on its chest and pointing to the red sphere to indicate the virtual clicked block to the corresponding physical block color. From left to right, the blocks read: 9, 1, 8, 4, 5.

Participants were shown anywhere from 1 to 8 cream-colored **virtual blocks (B)** for each puzzle. To discover the hidden virtual block color, participants clicked on a virtual block (by moving the Hololens cursor over it and pressing a hand-held clicker); in response, Kuri's **chest LED (D)** lit up in the hidden block color. In the Experiment condition, Kuri also used AR arms to point to the **virtual color indicator (E)** of that color.

Once the color was so indicated, the participants selected a **physical block (A)** with the same shape and color. The number displayed on the physical block was part of the math equation. The + or - on the virtual block indicated whether the number should be added or subtracted. Once all virtual-to-physical correspondences for the blocks were found, participants added or subtracted the numbers into a running sum (initialized at 0), calculated the final answer, and input it into the virtual **answer input (C)**.

At the start of the session, participants were guided through the process in a scripted tutorial (Fig. 5.4). They were told to click on the virtual cylinder on the far left. Once clicked, Kuri lit up its chest LED in red and pointed at the red virtual ball-shaped color indicator. Participants then grabbed the red cylinder with the number 9 on it. This process was repeated for all the blocks. The resulting sum was: {(-9), (+1), (-8), (-4), (+5)} = -15; it was input into the virtual answer.

Kuri used social gestures in response to participants' answers in both conditions. For a correct answer, Kuri performed the positive physical gesture ("gotit"); for an incorrect answer, Kuri performed the negative physical gesture ("sad").

In the Experiment condition, Kuri also used the positive and negative AR arm gestures (Fig. 5.1) synchronized with the positive and negative physical gesture, respectively. The physical and AR gestures were combined, as opposed to using AR gestures only, based on feedback received from a pilot study. Participants in the pilot study indicated that gestures with both the body and AR arms (as opposed to AR arm gestures only) created a more integrated and natural robot appearance.

After the tutorial, participants attempted to solve a series of up to seven puzzles of increasing difficulty within a time limit of 10 minutes. When participants successfully input the correct answer to a puzzle, they advanced to the next puzzle. If the time limit was exceeded or all puzzles were completed, the system halted. Participants were then asked to do a survey and a semi-structured interview described in Section 5.1.2.

Ensuring Gesture Presentation Consistency

The puzzle task was designed to mitigate inconsistencies across participants. The first mitigation method addressed gesture randomness and diversity. The Experiment condition used gestures from the set of positive (PG = {cheer, clap, wave dance}) and negative (NG = {facepalm, shoulder shrug, arm cross}) gestures. To preserve balance, the system first chose gestures randomly without replacement for each set, thereby guaranteeing that each gesture was shown, assuming at least 3 correct and 3 incorrect answers. Once all gestures from a group were shown, the gestures were chosen randomly, with replacement. The Control condition did not require methods for ensuring

gesture diversity since Kuri used a single way of communicating correct answers and incorrect answers.

Steps were also taken to avoid only positive gestures being shown for users who had all correct answers. First, all participants were shown an incorrect answer and gesture during the tutorial. Second, some puzzles had a single physical block with two numbers on it (Fig. 5.3). In those cases, participants were told that the puzzle could have two answers. If their first answer was incorrect, they were told to turn the block over and use the number on the other side. Puzzles 3-7 all had this feature. Regardless of the participant's initial guess for these puzzles, they were told told they were incorrect and then shown a negative gesture. If the initial guess was one of the two possible answers, it was removed from the possible answers. After the initial guess, guesses were said to be correct if they were in the remaining set of correct answers. This consistency method ensured that each participant saw all of the negative gestures.

Part 2: Gesture Annotation

All participants were shown a video of Kuri using the arm gestures, as seen in Fig. 5.1 and can be found at `https://youtu.be/Ff08E9hvvYM`. The video was recorded through the Hololens camera, giving the same view as seen by participants in the Experiment condition of the math puzzles. After the participants watched all gestures once, they were given the ability to rewind and re-watch gestures as they responded to a survey. The gesture order of presentation was initially randomly generated and then presented in that same order to all participants. In total, the second part of the experiment took 5-10 minutes.

Measures and Analysis

A combination of objective and subjective measures was used to characterize the difference between the conditions.

Task Efficiency was defined as the total time taken to complete each puzzle. Users that did not complete all puzzles within the 10 minute time limit were noted. The post-study 7-point Likert scale questions used 4 subjective measures, adapted from Heerink et al. (2010) to evaluate the use of ARHMD with Kuri. The measures were: Total Social Presence, Attitude Towards

Technology, Perceived Ease of Use, and Perceived Usefulness. ***Total Social Presence*** measured the extent the robot was perceived as a social being (10 items, Cronbach's $\alpha = .89$). ***Attitude Towards Technology*** measured how good or interesting the idea of using AR with the robot was (3 items, Cronbach's $\alpha = .97$). ***Perceived Ease of Use*** measured how easy the robot with AR was to use (5 items, Cronbach's $\alpha = .73$). ***Perceived Usefulness*** measured how useful or helpful the robot with AR seemed (3 items, Cronbach's $\alpha = .81$).

Participants rated the robot's physical (0.00) to virtual (1.00) teammate presence to a granularity of two decimal points (e.g., 0.34) and were able to see and check the exact value they input. This measure was used to gauge where Kuri was perceived as a teammate on the Miligram virtuality continuum (Milgram et al. 1995b).

Qualitative coding was performed on the responses to the post-study semi-structured interviews, to assess how emotional and helpful Kuri seemed to the participants. Participants from the Experiment condition were also asked how "attached" the arms felt on Kuri; this question was coded for only those participants (Table 5.1). To construct codes and counts, one research assistant coded for: "How emotional was Kuri?" and "How helpful was Kuri?" without looking at the data from the Experiment condition. Another assistant coded for: "Do the arms seem to be a part of Kuri?" for participants in the Experiment condition. Codes were constructed by reading through interview transcripts and finding ordinal themes. Example quotes for each code are shown in Table 5.1.

For the gesture annotation, a similar approach to Marmpena et al. (2018) was used: users annotated each robot gesture on a slider from very negative (-1.00) to very positive (+1.00), in order to measure valence. The slider granularity was to two decimal points (e.g. -0.73) and participants were able to see the precise decimal value they selected.

To test annotator repeatability and ability to distinguish gestures, an inter-rater reliability test was conducted. This was used to measure the repeatability of choosing a single person from a generalized population to rate each gesture. To measure inter-rater reliability, intraclass correlation was used with a 2-way random effect model for a single participant against all participants (referred

to as "Raters") among the six gestures (referred to as "Subjects") to find a measure for absolute agreement among participants. Eq. 5.1 was used where k denotes the number of repeated samples, MS_R is the mean square for rows, MS_E is the mean square error, MS_C is the mean square for columns, and n is the number of items tested (Koo and Li 2016). Analysis consisted of using k = 1 as this evaluates the reliability of agreement when choosing a single rater to rate the gestures against all other raters. The *icc* function was used from the *irr* package of R (v3.6.0, `https://cran.r-project.org/`) with parameters "twoway", "agreement", "single". According to Koo et. al (Koo and Li 2016), poor values < 0.5, moderate values < 0.7, good values < 0.9, and excellent values ≥ 0.9.

$$ICC(2,k) = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \tag{5.1}$$

Each gesture also had an open-ended text box where users were asked: "Please describe what you believe gesture X conveys to you" where 'X' referred to the gesture number. These textual data were later coded by a research assistant (Table 5.2). Codes were constructed as the most common and relevant words for each gesture. Example quotes for each code are also included in Table 5.2.

Participants

A total of 34 participants were recruited and randomly assigned to one of two groups: Control (5F, 12M) and Experimental (8F, 9M). Participants were University of Southern California students with an age range of 18-28 ($M = 22.3, SD = 2.5$).

5.1.3 Results and Analysis

Arms Vs. No Arms Condition

For the math puzzles, the performance metric was analyzed but saw no statistically significant effect between conditions. An independent-samples t-test was conducted to compare ***Task Efficiency*** between the two experiment conditions. There was not significant difference in scores for

arms ($M = 83.0, SD = 33.0$) and no arms ($M = 77.8, SD = 23.7$) conditions ($t(16), p = .54$). There were an equal number of participants (6) in each group who timed out at 10 minutes.

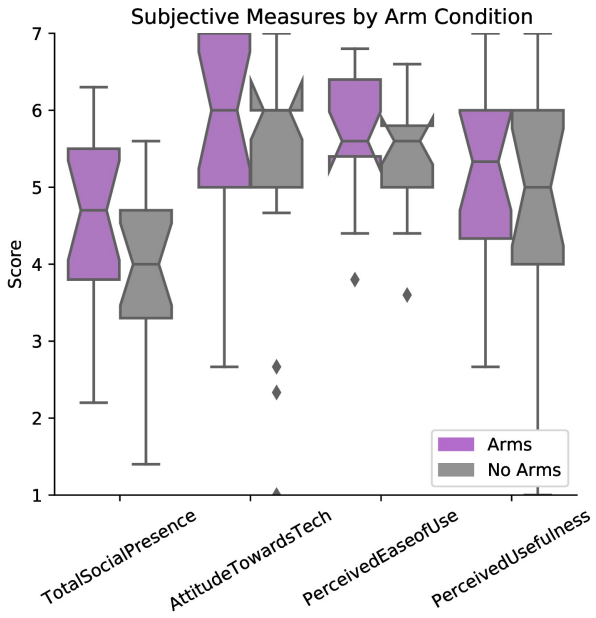


Figure 5.5: No statistical significance found for subjective measures. Boxes indicate 25% (Bot), 50% (Mid), and 75% (Top) percentiles. Notches indicate the 95% confidence interval about the median calculated with bootstrapping 1,000 particles (Efron and Tibshirani 1986). Thus notches can extend over the percentiles and give a "flipped" appearance (e.g., {Attitude, NoArms}).

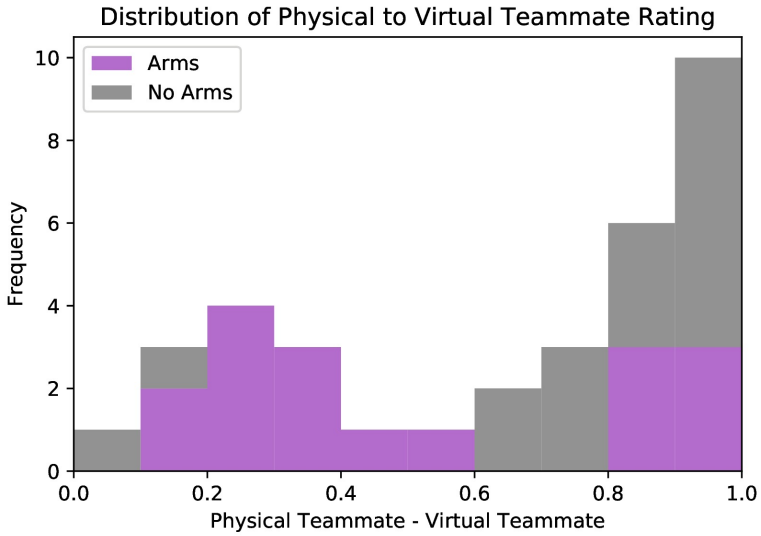


Figure 5.6: Stacked histogram with clustering to the left and right of 0.5 rating showing the distribution of virtual to phsyical teammate perception.

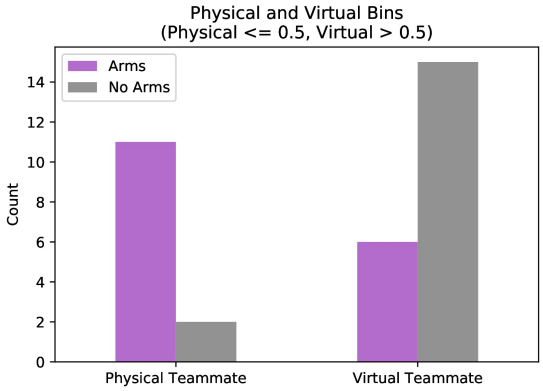


Figure 5.7: Participants in the Experiment condition were more likely to rate the AR robot as physical as opposed to virtual.

No significant effect among each metric was observed, as seen in Fig. 5.5. Mann-Whitney tests indicated no significant increases in ***Total Social Presence*** between arms $(Mdn = 4.7)$ and no arms $(Mdn = 4.0)$ conditions $(U = 115.5, p = .16)$, ***Attitude Towards Technology*** between arms $(Mdn = 6.0)$ and no arms $(Mdn = 6.0)$ conditions $(U = 117.5, p = .18)$, ***Perceived Ease of Use*** between arms $(Mdn = 5.6)$ and no arms $(Mdn = 5.6)$ conditions $(U = 117.0, p = .17)$, and ***Perceived Usefulness*** between arms $(Mdn = 5.33)$ and no arms $(Mdn = 5.0)$ conditions $(U = 138.0, p = .41)$. Qualitative coding for interviews can be found in Table 5.1. An explanation of the qualitative coding used for the interviews is found in Section 5.1.2.

Most participants answered towards the ends of the physical-to-virtual teammate scale, with very few near the middle (Fig. 5.6). Consequently, the participants were divided into two groups: "Physical Teammate" (ratings ≤ 0.5, $n = 13$) and "Virtual Teammate" (ratings > 0.5, $n = 21$) (Fig. 5.7) and performed a Chi-Square Independence test. A significant interaction was found $(\chi^2(1), p = .002)$. Participants in the Experiment condition, who experienced the arms, were more likely (64.7%) to rate Kuri as a physical teammate than participants in the Control condition, who did not experience the arms (11.8%). Next, post-hoc analyses were performed on subjective measures with the physical and virtual teammate binned groups.

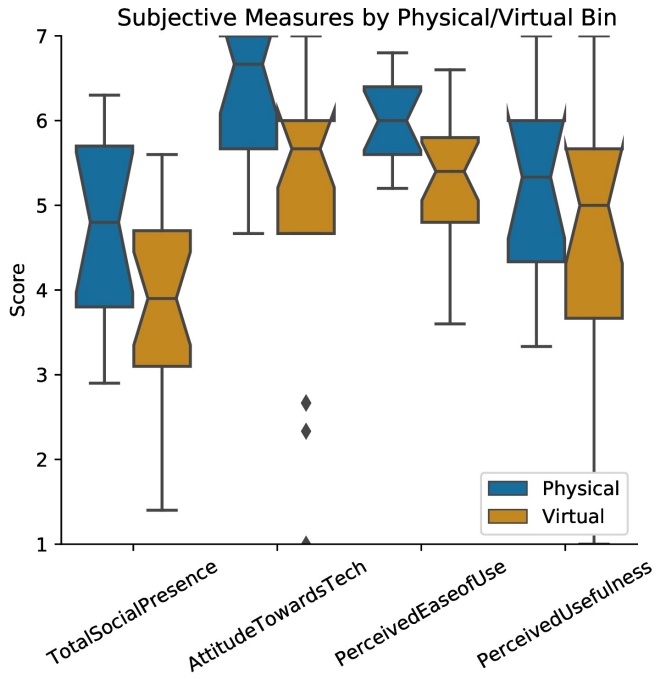


Figure 5.8: Significant increases for the first 3 measures with a marginally significant increase for measure 4. See Fig. 5.5 for notch box-plot explanation.

Physical Vs. Virtual Teammate

Survey data were analyzed with regard to the two bins and saw significant effects among metrics (Fig. 5.8). Mann-Whitney tests indicated a significant increase in ***Total Social Presence*** between physical ($Mdn = 4.8$) and virtual ($Mdn = 3.9$) groups ($U = 86.5, p = .04, \eta^2 = .092$), ***Attitude Towards Technology*** between physical ($Mdn = 6.7$) and virtual ($Mdn = 5.7$) groups ($U = 73.0, p = .01, \eta^2 = .149$), and ***Perceived Ease of Use*** between physical ($Mdn = 6.0$) and virtual ($Mdn = 5.4$) groups ($U = 79.0, p = .02, \eta^2 = .122$). Marginal significant increase was found for ***Perceived Usefulness*** between physical ($Mdn = 5.3$) and virtual ($Mdn = 5.0$) groups ($U = 93.5, p = .07, \eta^2 = .068$).

Gesture Validation

The data from gesture annotation were analyzed in order to validate participants' ability to distinguish the valence of gestures and consistency in interpreting gestures. As seen in Fig. 5.9 and Table 5.3, participants could distinguish the valence (negativity to positivity) of the gestures. The two-way, agreement intraclass correlation for a single rater, described in Section 5.1.2, resulted

Table 5.1: Qualitative Interview Coding

Code	No Arms	Arms	Quote
Not Emotional	7	5	"I didn't feel any emotion from the robot"
Close to Emotional	9	7	"Like not so emotional because the task was not based on the emotion"
Emotional	1	4	"It can talk and tell different emotions when I answer questions differently"
Very Emotional	0	1	"When it went like *crosses arms* it was like 'come on you're not helping me here.' And when her *acts out cheering*, yeah I would say very"
Not Helpful	6	2	"No"
Somewhat Helpful	2	3	"Sort of, yeah"
Helpful	9	12	"I like the way it had the visual feedback when I get right or wrong, and I just feel like it could reinforce it."
Are Arms a Part of Kuri?	-	Arms Count	Quote
No	-	2	"They seemed pretty detached"
Somewhat	-	4	"When it was pointing things it did seem like it a little bit"
Mostly	-	3	"I would say 60 percent", "8/10"
Yes	-	8	"What gave me the most information was her arms"

Table 5.2: Gesture Description Qualitative Code Counts

Gesture	**Code : Count**	**Example Quote**
G1: Facepalm	Disappointment :10, Frustration: 4, Facepalm: 3	"Facepalm, the robot is frustrated/disappointed"
G2: Cheer	Happy: 8, Celebration: 7, Cheer: 6	"That you got the answer correct and the robot is cheering you on"
G3: Shrug	Don't Know Answer: 11, Shrug: 5, Confuse: 4	"Shrugging, he doesn't know what the person is doing or is disappointed in the false guess"
G4: Arm Cross	Angry: 8, Disappointment: 7, Arm Cross: 4	"Crossing arms. *'Really??'* mild exasperation or judgment."
G5: Clap	Happy: 13, Clapping: 7, Excited: 4	"It's a very happy, innocent clap. I like the way its eyes squint, gives it a real feeling of joy."
G6: Wave Dance	Happy: 11, Celebrate: 4, Good Job: 4	"Celebration dance, good job!"

in a score $ICC(A,1) = 0.77$ with 95% confidence interval 0.55-0.95, and $F(5,190) = 125$, $p < 0.001$, which constitutes moderate to good reliability. Qualitative data are summarized in Table 5.2. Explanation for coding these data can be found in Section 5.1.2.

Table 5.3: Valence Rating Percentiles by Gesture.

$P_{\%}$	G1	G2	G3	G4	G5	G6
P_{25}	-0.85	0.72	-0.59	-0.87	0.64	0.55
P_{50}	-0.69	0.88	-0.30	-0.58	0.84	0.76
P_{75}	-0.41	1.00	-0.09	-0.36	1.00	1.00

5.2 Multidimensional Analysis of Functional AR Robot Capability Visualizations

Contributors: Section 5.2 is based on Groechel et al. (2022b) written with co-first authors Allison O'Connell and Massimiliano Nigro. Maja J. Matarić is also an author of the published work.

Whether it is a young student interacting with a SAR tutor in school or a trained roboticist debugging a system being created, the ability for a user to accurately estimate a robot's capabilities is critical for effective HRI. Within HRI research, *perceived robot capability* is defined as a user's perception of the robot's true capabilities (Cha et al. 2015). Under- and over-perception refer to a mismatch between the user's perceptions about the robot and the robot's true capabilities.

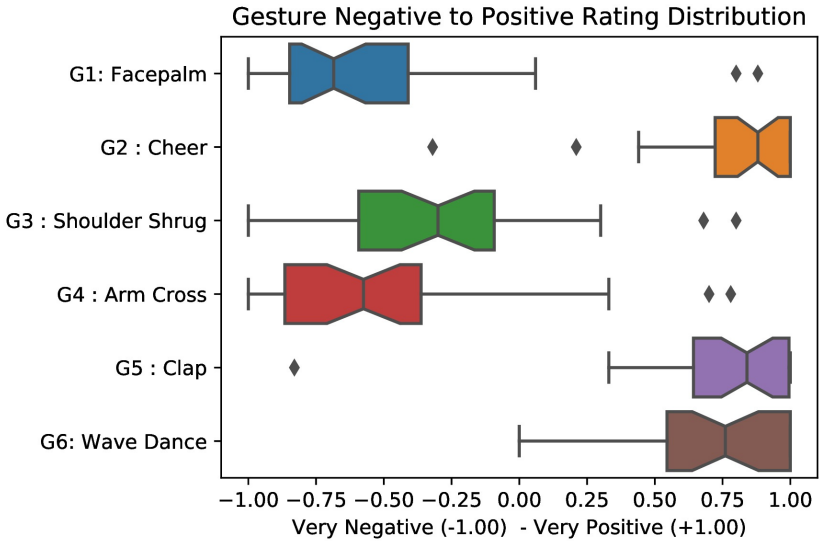


Figure 5.9: Distribution on ability to differentiate gesture valence.

Over-perception occurs when a user expects the robot to do something it is not capable of, while under-perception occurs when the user misses an interaction capability the robot possesses.

Visualizing robot capabilities as explicitly as possible aids capability signalling. Widely used software such as RViz (Hershberger et al. 2015) displays robot sensors (e.g., cameras) and reasoning capabilities (e.g., mapping and navigation waypoints).

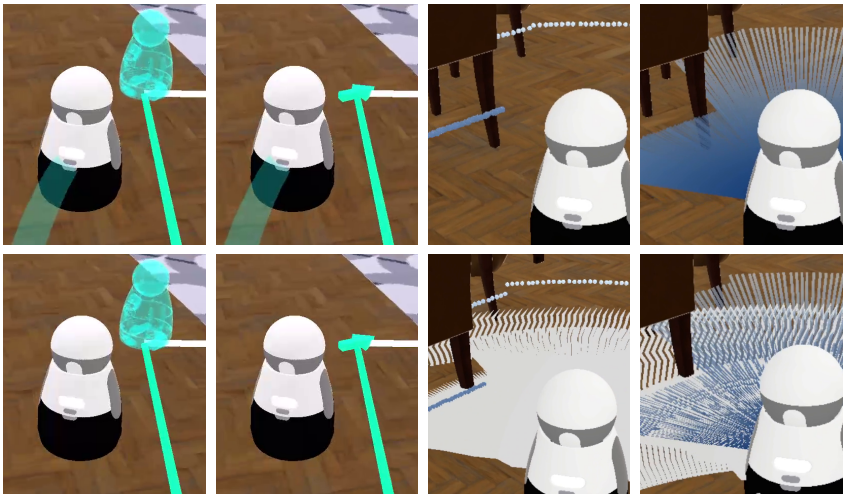


Figure 5.10: Combinations of Virtual Design Elements (VDEs) for navigation visualization (left) and LiDAR visualization (right). Details of each signal design are found in Sec. 5.2.1.

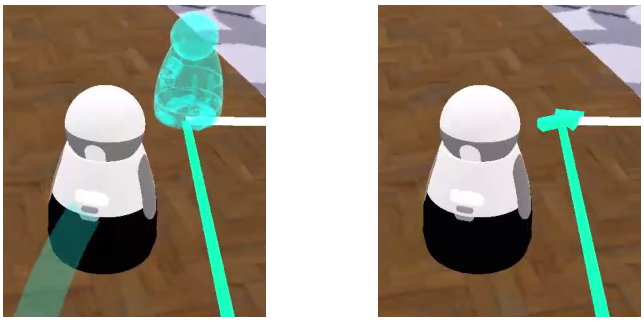
Many of these VAM-HRI signal designs, however, either explore a distinct set of visualizations (e.g., (Walker et al. 2018)) or only pose a single signal design for comparison against a non-VAM interface (e.g., RViz). To address these shortcomings, a set of salient Virtual Design Elements (VDEs (Walker et al. 2022)) were created for a set of core robot capabilities: **navigation**, **light detection and ranging (LiDAR)**, **camera**, **face detection**, **audio localization**, and **natural language processing (NLU)**. Two independent VDEs were created for each capability and validated the pairwise designs on Amazon Mechanical Turk (AMT) in a video-based study (n=150). AMT is a standard tool used in VAM-HRI research for initial signal design studies (Groechel et al. 2022a). Four videos of each signal displaying all combinations of the independent VDEs were shown to participants and evaluated for clarity and visual appeal. The results determine the clearest and most visually appealing design choices while also highlighting possible interaction effects of intra-signal VDEs.

5.2.1 Technical Approach

To create each signal's Virtual Design Element (VDE, (Walker et al. 2022)), inspiration was taken from prior visualization research and existing visualization software (e.g., RViz (Hershberger et al. 2015)). The Unity 3D game engine v2021.2.7f1 was used to create the visualizations for this work and have made them open-source and available at `https://github.com/interaction-lab/NRI-SVTE`. A video of all signal VDE combinations can be found at `https://youtu.be/Xw2_kHyN-xA`.

Navigation

Visualizations were created for the robot's ability to plan and execute paths in the environment.



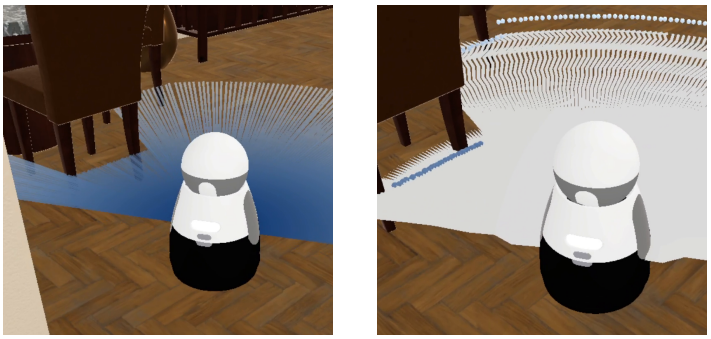
(a) Robot Outline + Trail (b) Arrow + No Trail

Figure 5.11: Navigation visualizations.

Robot Outlines / Arrows: Two different sets of navigation waypoints were designed: ghost robot outlines and arrows. The ghost robot outlines, inspired by (LeMasurier et al. 2022), place semi-transparent robots at each waypoint, oriented in the goal direction of the waypoint. The outlines were scaled down so as to reduce visual clutter and allow the robot to fully cover the outline. The 3D arrows were inspired by RViz (Hershberger et al. 2015). The differences between the two were expected to be in the salience of direction (arrows $>$ outline), environment occlusion (arrow $>$ outline), and visual appeal (outline $>$ arrow). **Trail / No Trail**: For a sense of momentum and direction, a trail was added to the visualization, but it occludes some of the environment.

LiDAR

Visualizations were created of the robot's ability to use LiDAR sensors to detect nearby objects.



(a) Lines + No Pulse (b) Points + Pulse

Figure 5.12: LiDAR visualizations.

Lines / Points: The robot measures the distance to objects in the environment with a 180 LiDAR sensors spaced 1° apart. 3D lines and points were used to indicate the distance returned by each of the sensors. The lines (see Fig. 5.12a) occupied the same position as the laser beam emitted by each sensor and were displayed as opaque and colored, with a gradient from dark blue at the source to white at the maximum distance measurable by the sensor. The points, inspired by RViz (Hershberger et al. 2015), were opaque 3D spheres. The color of each sphere was modulated to indicate the distance from the robot according to the same gradient scale used on the lines. The expected differences were in the salience of distance from the robot (lines $>$ points), indication of lasers (lines $>$ points), environment occlusion (points $>$ lines) and visual appeal (lines $>$ points).
Pulse / No Pulse: In order to give the impression that the sensors emitted lasers to measure distance, fixed length white lines were visualized emitting at a fixed interval from each sensor. The lines occluded part of the environment and do not continue to align with the sensors as the robot moves, visible in Fig. 5.12b.

Camera

Visualizations were created of the robot's ability to to collect visual information through a camera located in its left eye.

In-Scene / HUD: Input from the robot's camera was visualized in one of two locations depending on the condition. In the in-scene condition, the input was visible in a square window that hovered above the robot (see Fig. 5.13b). As the video rotated around the robot to show the scene

(a) HUD + No Pyramid (b) In-Scene + Pyramid

Figure 5.13: Camera visualizations.

from multiple viewpoints, the window rotated to always face the participant. In the HUD (heads-up display) condition, the input was visible in a fixed square on the screen. In both conditions, a grey frame was placed around the window to increase salience. These visualizations were inspired by the difference in "User-anchored" and "Robot-anchored" MR design elements (Groechel et al. 2022a; Walker et al. 2022). **Pyramid / No Pyramid**: A 3D translucent pyramid (see Fig. 5.13b) visualized the portion of the scene that was visible to the robot's camera, inspired by camera depictions in game engines. The pyramid was fixed to the robot's eye containing the camera and rotated with the robot's head to correspond with the camera input at all points in the video. The expected differences were in the salience of the portion of the environment visible to the robot (pyramid $>$ no pyramid), environment occlusion (no pyramid $>$ pyramid), and how appealing it is to look at (pyramid $>$ no pyramid).

Face Detection

Visualizations were created of the robot's ability to detect faces from its camera input.

Box / Face Mesh: Two different sets of face markers were designed: green boxes and triangle meshes. The boxes (see Fig. 5.14a), inspired by OpenCV (Bradski and Kaehler 2008) face detection software, were green square boxes fixed around the faces of each person in the robot's camera's field of view. Translucent triangle meshes, inspired by face detection software such as Google's MediaPipe (Lugaresi et al. 2019), were placed on each character model in the same way as the boxes. Both face markers were sized to fit each character model and rotated with the robot's head to remain oriented toward the robot's camera. **In-Scene / Not In-Scene**: The boxes and meshes

(a) Box + In-Scene (b) Mesh + Not In-Scene

Figure 5.14: Face detection visualizations.

were always visible in the camera input on the HUD. However, in the in-scene condition, the face markers were also visualized as objects fixed to the faces of the people in the scene. In the **not in-scene** condition, the markers were only visible in the HUD window but not in the environment.

The expected differences were in the salience of faces (in-scene $>$ not in-scene) and environment occlusion (in-scene $>$ not in-scene), inspired by the difference in "User-anchored" and "Environment-anchored" MR design elements (Groechel et al. 2022a; Walker et al. 2022).

Audio Localization

Visualizations were created of the robot's ability to to estimate the positions of sources sound in the environment.

(a) Cones + Small (b) Spheres + Large

Figure 5.15: Audio localization visualizations.

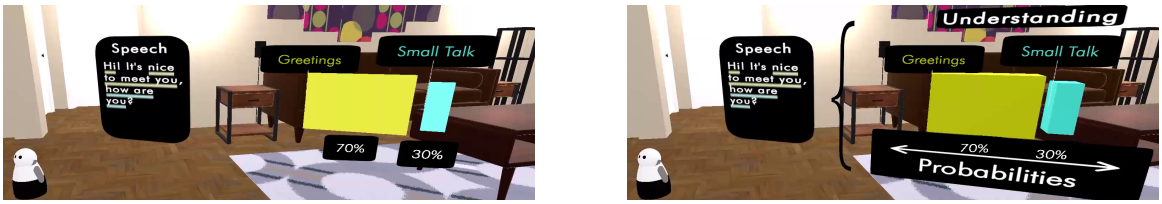
Spheres / Cones: 3D objects of two distinct shapes (spheres, cones) were overlayed around the robot and increased in size and color gradient in relation to the loudness input received from

the directional microphones. If no sound was perceived, the objects remained hidden. The cones and the spheres were inspired by (Kataoka et al. n.d.) and (Kose et al. 2018; Lopez-Rincon and Starostenko 2019), respectively. **Small / Large**: Small/Large visualizations differed by how much the 3D objects were increasing in size with respect to the microphone input. This characteristic explored the trade-off between robot visibility and visual indication of loudness.

Natural Language Understanding (NLU)

Visualizations were created of the robot's ability to to analyze spoken language and make predictions about the user's meaning. The main elements of the visualizations were:

- A speech bubble containing the speech understood by the robot in real time, with keyword highlighting (Ashktorab et al. 2019) as soon as the intents were extracted by NLU;
- A horizontal bar diagram, indicating the confidence for each intent understood in the sentence;
- Graphical elements (e.g., labels, curly brace) providing additional contextual information about the elements in the visualization.



(a) 2D + Sparse (b) 3D + Cluttered

Figure 5.16: NLU visualizations.

Cluttered / Sparse: The effect of the decluttering principle (Ajani et al. 2021) on the visualization was explored. The cluttered version presented all the graphical elements, providing more information to the user but cluttering the visualization. The sparse version removed the graphical elements and provided less information about the relationship between the elements in the visualization. **3D / 2D**: In the 3D/2D versions, the bar chart was displayed in its 3D or its 2D version.

5.2.2 User Study

To increase reproducibility, the study methods are also included in the open-source repository wiki `https://github.com/interaction-lab/NRI-SVTE/wiki`.

Participants

Participants were recruited through AMT. To determine the study size, (VanVoorhis, Morgan, et al. 2007) was followed: $50 + 8 * m$ where m is the number of independent variables (2 for each of the 6 signals, thus $m = 12$), resulting in $50 + 8 * 12 = 146$ participants. Four more participants were added in the case of incomplete data.

Inclusion criteria for the study were:

- At least 18 years of age;
- In the United States or US Minor Outlying Islands;
- Number of AMT HITs approved > 1000;
- AMT HIT approval rate $\geq 99\%$.

The 150 participants who completed the survey identified in open-ended questions as: Gender Identity – Cis Woman : 1, Female: 52, Male: 95, and left blank: 2; Race – African American : 3, American : 1, Asian : 8, Black : 9, Caucasian : 13, European : 1, Hispanic : 3, Latina : 1, Middle Eastern : 1, Native American : 1, White : 101, and left blank : 8; Age – Mdn : 35, $\bar{X}$: 37.22, σ : 9.74, and Range(20,71).

Procedure and Measures

This study was approved by the University's Institutional Review Board (IRB #UP-20-00030). Each participant first consented to the study, confirmed that their audio worked, and filled out a set of demographic questions. A within-subjects study design was used in which participants were shown four ten-second videos displaying each combination of conditions for the six signal visualizations described in Sec. 5.2.1. Both the signal set order and the order of videos within each

signal set were randomized, with the exception that the camera visualization was always shown immediately before the face detection visualization, as the latter depended on the former.

Quantitative and qualitative data were collected to measure the participants' opinions of the visualizations and their understanding of the robot based on the videos (found at `https://youtu.be/Xw2_kHyN-xA`), shown from the perspective of someone wearing an AR headset. The camera moved to different points of view as the robot moved to demonstrate a capability. Camera views were replicated exactly within each signal across conditions.

Quantitative Data: Participants were not told what capability of the robot was visualized in the videos when they first watched them. For each video, they rated the clarity and visual appeal via two 7-point Likert items ("The video shows the capability in a clear way", "The video is visually appealing") from "Strongly Disagree" to "Strongly Agree." The items were adapted from (Svalina et al. 2021) and (Amini et al. 2018), respectively.

Qualitative Data: After watching and rating the four videos, participants were asked the following three open-ended questions: "What did you like about the visualizations you scored higher?" "What did you dislike about the visualizations you scored lower?" and "What capability of the robot do you think is visualized in the videos above?". Once they had answered these three questions, the signal name and a short paragraph describing the signal (similar to those found in Sec. 5.2.1) were revealed to the participant. They were then asked one additional open-ended question, "How could the visualizations above be improved to better illustrate the robot's ability to perform [signal name]?"

Participants were required to watch all four videos and answer all questions about them before moving on to the next signal. The process was repeated until all six signal sets were completed. Upon completion, each participant received an Amazon gift card worth US$6.25.

Analysis

Both quantitative and qualitative analyses were performed. Survey data were treated as ordinal non-parametric tests were used (Schrum et al. 2020). To compare signal clarity and visual appeal, Wilcoxon signed-rank tests were used with Holm's corrected p values (Abdi 2010) and α levels

$<.05$*, $<.1$**, and $<.001$***. The brute force common language effect size (CLES) proposed by Vargha and Delaney (2000) was calculated; mean and standard deviation calculations are not appropriate for ordinal data (Schrum et al. 2020). CLES is the proportion of paired samples (s_{G0}, s_{G1}) where s_{G0} is higher than s_{G1}. To confirm the signals portrayed the intended capability, open response question "What does this visualization portray?" was annoted for with the criteria set in Sec. 5.2.1. For qualitative analysis, all participant open-ended answers were read marking down themes and associated quotes to each theme. All 150 participants were included in the analysis given the strict criteria described in Sec. 5.2.2.

5.2.3 Results and Analysis

Navigation

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are given in Table 5.4. A total of 102 participants (68%) correctly identified the robot's navigation capability.

Table 5.4: Navigation results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
RobotOutline + Trail / Arrow + No Trail	0.614	$<.001$***
RobotOutline + No Trail / Arrow + No Trail	0.589	$<.001$***
RobotOutline + Trail / Arrow + Trail	0.568	.015*
Arrow + Trail / Arrow + No Trail	0.549	.312
RobotOutline + No Trail / Arrow + Trail	0.543	.449
RobotOutline + Trail / RobotOutline + No Trail	0.526	.58
Visual Appeal		
RobotOutline + No Trail / Arrow + No Trail	0.589	.001**
RobotOutline + Trail / Arrow + No Trail	0.598	.003**
RobotOutline + Trail / Arrow + Trail	0.565	.178
RobotOutline + No Trail / Arrow + Trail	0.555	.25
Arrow + Trail / Arrow + No Trail	0.538	.427
RobotOutline + Trail / RobotOutline + No Trail	0.51	.999

Qualitative Analysis

Likes/Dislikes– Participants reported liking the arrows for their simplicity, clarity of direction, and lack of clutter, reporting the opposite for the robot outlines. Alternatively, other participants enjoyed the robot outlines citing them as more clear and visually "exciting" or "cool", reporting the arrows as "boring". A similar trend emerged for robot trails where participants liked the clarity of the trail (e.g., **P12**: "I like the trial in which it left behind to show where it was coming from and how it was moving with the way points") while those opposed disliked the clutter (e.g., **P67**: "I didn't like the blue line that followed him showing him actually doing it. It was overkill").

Suggested improvements– A common theme of dynamic objects and colors were suggested (e.g., **P12** : "as the robot goes through the path, the way point that the robot travel should fade or change colors"). Another suggested theme were requests for indication of robot planning (e.g., **P96**: "Before the robot moves, there should be a small thought bubble with a travel plan") as well as a map (e.g., **P100**: "Show a map").

Other unique factors– Participants described the visualization videos as "smoother" than others (e.g., **P1**: "[It was] less busy and the camera less nauseating"). Participants also reported confusion as to why the robot did not follow the straight line path (e.g., **P138**: "Why did it 'walk' in a rounded fashion when the lines were straight?").

LiDAR

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are shown in Table 5.5. A total of 62 participants ($\approx 41.3\%$) correctly identified the robot's LiDAR capability.

Qualitative Analysis

Likes/Dislikes– Rather than commenting on the individual points or lines for each sensor, participants emphasized the overall shape they formed when combined. For example, participants perceived the points as forming a "line" to denote the boundary around the area that the robot could sense, whereas they described the lines as forming a "cone" or "fan" around the front of the robot. Some preferred the points because they left the area around the robot visible (e.g., **P131**:

Table 5.5: LiDAR results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Line + No Pulse / Point + No Pulse	0.658	<.001***
Point + Pulse / Point + No Pulse	0.598	<.001***
Line + No Pulse / Line + Pulse	0.586	<.001***
Line + No Pulse / Point + Pulse	0.56	.036*
Line + Pulse / Point + No Pulse	0.563	.098
Point + Pulse / Line + Pulse	0.528	.422
Visual Appeal		
Line + No Pulse / Line + Pulse	0.662	<.001***
Line + No Pulse / Point + Pulse	0.622	<.001***
Point + No Pulse / Line + Pulse	0.595	<.001***
Line + No Pulse / Point + No Pulse	0.584	.003**
Point + No Pulse / Point + Pulse	0.55	.064
Point + Pulse / Line + Pulse	0.541	.134

"The outline of the sensor area was clean and simple") while others found the points too simplistic to convey the relevant information (e.g., **P73**: "Just having a line is confusing and doesn't really tell you much."). Participants described the pulsing videos as "confusing," "laggy," and "glitchy" (e.g., **P144**: "It looks like a glitch, as if the robot is creating ice or something when it scoots along.") although some understood the pulsing to be a more accurate representation of how the data are collected.

Suggested improvements– Participants suggested somehow indicating the objects that had been detected (e.g., **P119** : "When approaching an object, the object should be shown in red for clarification."). Another theme among the suggestions was some indication of the laser light returning to the robot after reflecting off of an object (e.g., **P51**: "Maybe if the line bounced back like a radar").

Other unique factors– Participants suggested that these visualizations could be enhanced by including more textual information about what the robot is doing (e.g., **P60**: "Maybe include some sample distances that were sensed for the viewer to see more clearly what is going (for those numerically inclined individuals)").

Camera

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are shown in Table 5.6. A total of 85 participants ($\approx 56.7\%$) correctly identified the camera capability.

Table 5.6: Camera results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Pyramid + HUD / No Pyramid + InScene	0.735	<.001***
Pyramid + HUD / No Pyramid + HUD	0.704	<.001***
Pyramid + InScene / No Pyramid + InScene	0.71	<.001***
Pyramid + InScene / No Pyramid + HUD	0.679	<.001***
No Pyramid + HUD / No Pyramid + InScene	0.54	.22
Pyramid + HUD / Pyramid + InScene	0.517	.999
Visual Appeal		
Pyramid + InScene / No Pyramid + InScene	0.621	<.001***
Pyramid + HUD / No Pyramid + InScene	0.608	<.001***
Pyramid + InScene / No Pyramid + HUD	0.597	.002**
Pyramid + HUD / No Pyramid + HUD	0.583	.005**
No Pyramid + HUD / No Pyramid + InScene	0.529	.999
Pyramid + InScene / Pyramid + HUD	0.518	.999

Qualitative Analysis

Likes/Dislikes– Participants reported finding the pyramid helpful and correctly interpreted it as the portion of the scene visible to the robot (e.g., **P90**: "I liked that the visualization is able to show the field of vision for the robot and gave us a view of what the robot is seeing without blocking our vision."). Participants disliked that the pyramid distorted or obstructed the scene (e.g., **P72**: "I didn't like the way the colors muted when showing the boundaries." **P56**: "The extra visual goodies kind of got in the way").

Participants preferences for the HUD or in-scene placement of the camera input were often related to the relative size of the displays. They found that the larger HUD display was easier to see (e.g., **P51**: "I liked that the inset picture was big enough to see well") but also obstructed the scene. They found the smaller in-scene window harder to see but less intrusive (e.g., **P84**: "I found

the smaller window showing the family less intrusive"), although some still found the in-scene placement obstructive and misleading (e.g., **P12**: "The robots perspective being above its head felt like it blocked out the people, as well made it look like the robot was thinking of something").

Suggested improvements– Participants suggested several ways to make the pyramid less intrusive in the scene without decreasing the salience of the visualization. They involved finding other ways to maintain the outline of the camera frame without distorting the scene with the translucent material (e.g., **P59**: "Instead of having the lens view be transparent have it just show an outline or bounding box", **P72**: "May a border line where the boundaries would be, but keep the colors and visuals not as distorted").

Other unique factors– As observed in the other signals, participants assumed the robot could perform functions beyond what the real robot was capable of, and suggested the visualization could better indicate what the robot is not doing in addition to what it is. In this example, participants were unsure if the robot was merely "seeing" the people in real time, or if the robot was taking pictures or recording videos of the people pictured (e.g., **P120**: "I think that placing a small red light bulb that turns on when the robot is recording or capturing images").

Face Detection

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are in Table 5.7. A total of 109 participants ($\approx 72.7\%$) correctly identified the robot's face detection capability.

Qualitative

Likes/Dislikes– Participants overwhelmingly viewed visualizations with green boxes around faces more favorably than those featuring triangle face meshes. Commonly cited reasons included visual salience (e.g., **P51**: "The green boxes on the faces are easy to see"), and more obvious meaning (e.g., **P63**: "The screen mapping of the face structure wasn't very clear and was a bit creepy"). The face meshes were often associated with mistrust of the robot, although both markers were described by some participants as unsettling (e.g.,**P15**: "I didn't like the green boxes around

Table 5.7: Face detection results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Box + InScene / Mesh + InScene	0.756	<.001***
Box + InScene / Mesh + Not InScene	0.747	<.001***
Box + Not InScene / Mesh + InScene	0.738	<.001***
Box + Not InScene / Mesh + Not InScene	0.729	<.001***
Box + InScene / Box + Not InScene	0.524	.935
Mesh + Not InScene / Mesh + InScene	0.51	.999
Visual Appeal		
Box + Not InScene / Mesh + InScene	0.614	<.001***
Box + Not InScene / Mesh + Not InScene	0.585	.002**
Box + InScene / Mesh + InScene	0.584	.039*
Box + InScene / Mesh + Not InScene	0.558	.319
Box + Not InScene / Box + InScene	0.519	.999
Mesh + Not InScene / Mesh + InScene	0.529	.999

the heads. It felt like they were a target", **P134**: "The crosshairs on the faces wasn't necessary and looked weird and technologically scary").

Participants who preferred having the markers in the scene in addition to on the HUD reported it as a more obvious indication of what the robot was doing (e.g., **P14**: "I liked the green boxes on both the floating frame and the actual scene, as it was a good way to reference exactly what the robot was seeing and interpreting"), while others found the in-scene markers unnecessary (e.g.,**P56**: "It was clear what the robot was doing, without the green frames getting in the way in the main physical scene").

Suggested improvements– Participants suggested changing the visualizations to indicate that the robot could differentiate between people, by numbering the boxes, using different colored boxes for the different faces, or labeling the boxes with the names if the robot recognizes specific individuals. These suggestions are evidence of a larger trend of participants seeking to clarify the extent of the robot's capabilities through the visualizations (e.g., can the robot recognize human faces, or simply detect them?).

Other unique factors– Participants had mixed suggestions on what color the box should be. For some, the green boxes were familiar and recognizable, while others found the green ”creepy” and suggested using a more neutral color.

Audio Localization

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are shown in Table 5.8. A total of 58 participants ($\approx 58.7\%$) correctly identified the robot's audio localization capability.

Table 5.8: Audio localization results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Sphere + Small / Sphere + Large	0.523	.191
Cone + Large / Sphere + Large	0.532	.468
Cone + Large / Cone + Small	0.513	.999
Cone + Large / Sphere + Small	0.51	.999
Cone + Small / Sphere + Large	0.519	.999
Sphere + Small / Cone + Small	0.502	.999
Visual Appeal		
Cone + Large / Sphere + Large	0.554	.009**
Cone + Large / Sphere + Small	0.544	.219
Cone + Small / Sphere + Large	0.536	.338
Cone + Large / Cone + Small	0.52	.999
Cone + Small / Sphere + Small	0.524	.999
Sphere + Small / Sphere + Large	0.512	.999

Qualitative Likes/Dislikes– Participants who preferred the cones described them as cleaner (e.g., **P81**: “I liked the smaller circles visuals more because it looked cleaner”) and less obstructive than the spheres (e.g.,**P101**: “The bubbles were too large and surrounded the robot”). Participants who preferred the spheres found them more salient (e.g., **P64**: “The bigger bubble made it clear where the robot was perceiving the noise from,”). Participants varied widely in their interpretations of the shapes, referring to the cones as ”dots” and ”arrows” and the to the spheres as ”bubbles” or a ”ring” formed around the robot as they overlapped. In some cases, these impressions contributed to the participants' interpretations of the visualizations (e.g., **P43** interpreted the cones as arrows

and expected them to point at the source of the audio: “I didn’t find the arrows very helpful as it did not accurately point to the source”).

Suggested improvements– The most common suggestions included adding a compass or one or more arrows that point to the source of the sound to more clearly indicate that the robot is interested in the direction/location of the sound source, and displaying some kind of icon near the robot to more clearly signal that the robot can detect sound, such as an ear or microphone. Participants also suggested changing the sound source, currently a boom box animated to move to different locations around the robot in the environment, to something a robot is more likely to encounter in a home (e.g.,**P3**: “I think it can be improved by showing sound moving from something more realistic, maybe like a toy train going around a toy track with the robot in the middle”).

Other unique factors– Participants were generally unclear about where the noise was coming from (i.e., some assumed the music was coming from the robot rather than the boom box, or that the robot was controlling the boom box).

Natural Language Understanding (NLU)

Quantitative Analysis

Survey results for Wilcoxon signed-rank test are shown in Table 5.9. A total of 112 participants ($\approx 74.7\%$) correctly identified the robot’s NLU capability.

Qualitative Analysis

Likes/Dislikes– In their open-ended responses, participants expressed preferring cluttered and sparse displays in approximately equal measure, which is consistent with analysis of the quantitative ratings. Participants who preferred cluttered displays valued the additional information and context provided by the additional axes and labels (e.g., **P56**: “The word ”probabilities” helped to make clear what the robot was showing.”). Conversely, participants who preferred sparse displays described non cluttered displays as ”cleaner” and less confusing to look at (e.g.,**P16**: “They were less cluttered with labels than the ones that had the understanding and probability labels”, **P102**: “The ones I scored lower had too much going on and made it a bit confusing”). Participants cited

Table 5.9: NLU results sorted by p_{cor}.

Combination	CLES	p_{cor}
Clarity		
Cluttered + 3D / Sparse + 3D	0.536	.092
Cluttered + 3D / Sparse + Flat	0.546	.245
Cluttered + 3D / Cluttered + Flat	0.52	.749
Cluttered + Flat / Sparse + Flat	0.527	.999
Cluttered + Flat / Sparse + 3D	0.517	.999
Sparse + 3D / Sparse + Flat	0.51	.999
Visual Appeal		
Cluttered + Flat / Sparse + Flat	0.526	.325
Cluttered + Flat / Cluttered + 3D	0.518	.783
Cluttered + Flat / Sparse + 3D	0.532	.999
Cluttered + 3D / Sparse + Flat	0.509	.999
Sparse + Flat / Sparse + 3D	0.507	.999
Cluttered + 3D / Sparse + 3D	0.517	.999

mainly aesthetic reasons for their ratings of the 3D and 2D graphs (e.g.**P14**: "I liked the bar graphs that were 3d, since they fit the overall aesthetic of the video better").

Suggested improvements– Participants indicated that the videos should be longer and show more interaction. Specifically, they formed the assumption that the robot talked back to the human, and were interested in observing how the robot would form a response based on the information it collected (e.g., **P12**: "Showing a response from the robot would help understand how it decides what response it takes"). They also suggested displaying more information about the robot's "thought processes" (e.g., **P14**: "It would be helpful to know which specific words are given more weight than others in determining which meaning. I also have to imagine some words would be classified into both categories, which would be helpful to be able to see visually").

Other unique factors– Participants hypothesized about other kinds of speech the robot understands and how the robot might act on its interpretations. Their suggestions involved testing the robot's knowledge or showing how it functions in varied situations (e.g., **P10**: "Could there be a graph for the robots understanding of body language aside from verbal communication?", **P122**: "Ask more question to test its knowledge base").

5.3 Maximizing AR Appendage Social and Functional Expressivity

Contributors: Section 5.3 is based on Goktan et al. (2022) written with co-first authors İpek Göktan and Karen Ly. Maja J. Matarić also is an author of the published work.

As described in Sec. 5.1 and Sec. 5.2, research has shown that gestures can be designed to increase users' functional or social perception of robots. *Functional perception* is the user's belief that a robot is able to accurately communicate its functional capabilities and intentions (e.g., the user can easily choose the target from a list of objects) (Groechel et al. 2019), (Hamilton et al. 2020). *Social perception* is the user's belief that a robot is capable of participating in the interaction as a social actor that is an interactive, autonomous, and adaptable agent that performs anthropomorphic actions (Jackson and Williams 2021).

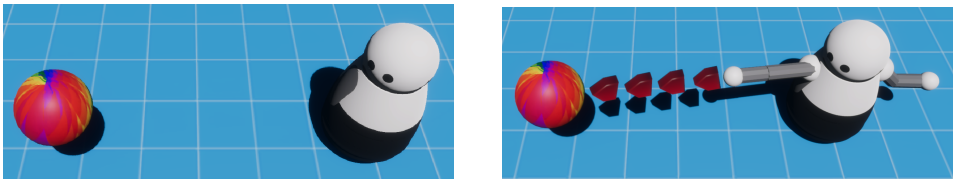
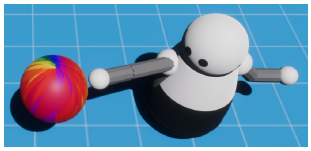


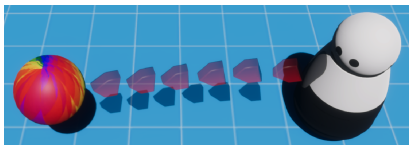
Figure 5.17: Kuri indicating a target object for the user to attend to. **Left**: Kuri without AR additions. **Right**: Kuri gestures at a target object using combined anthropomorphic AR appendages (arms) and non-anthropomorphic gestures (arrows).

A virtual gesture performed by a robot can be either anthropomorphic or non-anthropomorphic. Anthropomorphic gestures, such as those in Fig. 5.18a, are human-like gestures (Erel et al. 2018), (Wadgaonkar et al. 2021) that support HRI (Hamilton et al. 2020). Non-anthropomorphic gestures are abstract gestures (e.g., virtual arrows) that aim to communicate goals of actions efficiently and accurately (Erel et al. 2018), (Hamilton et al. 2020). While non-anthropomorphic gestures can be 2D (e.g., a flat dashed line) or 3D (e.g., arrow line in Fig. 5.18b), proposed recommendations suggest using 3D non-anthropomorphic gestures because they situate gestures in real interaction space and communicate added information such as directionality and depth (e.g., via shading/shadows). Further, utilizing 3D non-anthropomorphic gestures may increase social perceptions of the

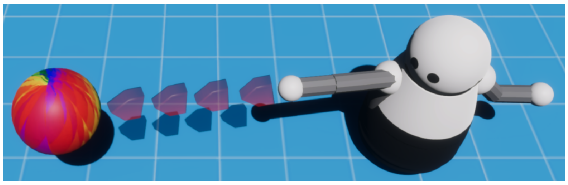
robot as users may associate the non-anthropormorphic signal as part of the robot. Additionally, anthropomorphic and non-anthropomorphic gestures can be used in tandem, as shown in Fig. 5.17, where the robot points at an object and an arrow connects its virtual arm to the object.



(a) Anthropomorphic



(b) Non-Anthropomorphic



(c) Anthropomorphic and Non-Anthropomorphic

Figure 5.18: Kuri robot gestures toward the target sphere by (a) pointing with projected AR arms, (b) using an arrow line, and (c) pointing with projected AR arms and an arrow line. Compared to Fig. 5.17, the arrow line is transparent thus obfuscating less of the background.

While a variety of virtual robot deictic gestures–gestures that point toward objects or areas in order to redirect user's attention–have been implemented to improve VAM-HRI, both anthropomorphic and non-anthropomorphic gesture systems have been found to involve tradeoffs between functional task performance and user social perception of robots (Hamilton et al. 2020), creating a need for a unified approach that considers both functional and social attributes.

Toward that goal, a set of design recommendations and techniques is proposed that combine anthropomorphic and non-anthropomorphic virtual gestures based on the placement of the robot, user, and target object during an interaction. Design recommendations were compiled by analyzing the relevant literature for design considerations, and highlighting factors that were reported to significantly influence either social perception or task efficiency (Hamilton et al. 2020), (Walker et al. 2018), (Stogsdill et al. 2021), (Piwek 2009), (Tran et al. 2021). The findings are distilled into the following key factors: motivation of the interaction, a target's visibility, a target's salience, and a target's distance. To illustrate the recommendations, the example of a robot attempting to draw

attention to an object via an anthropomorphic deictic gesture is considered using the proposed recommendations to create AR gestures that consider both functional and social user perceptions.

5.3.1 Tradeoffs Between Influencing Social and Functional Perception

This work builds on past work on communicating robot intent, robot gestures, anthropomorphism, and expressivity in VAM-HRI, briefly summarized next.

Influencing Functional Perception

Functional perception is a key concept in AR. Developers often project AR on objects in the physical world in order to quickly and concisely communicate concepts. Walker et al. (2018) designed and evaluated various explicit and implicit AR interfaces and reported that certain designs (e.g., NavPoints) were able to significantly improve the communication of robot intent as well as objective task efficiency. Other studies have also shown that utilizing AR can increase user task speed and facilitate human-robot collaboration (Tran et al. 2021), (Bagchi, Marvel, et al. 2018), (Chandan et al. 2021), (Krenn et al. 2021). These findings suggest that AR visualizations can be used to accurately communicate robot intent within a desirable reaction time, and therefore boost the functional perception of a robot.

Influencing Social Perception

AR has also been used as a medium to enhance the expressivity of mobility-constrained robots (i.e., robots that are either unable to move or move slowly in the context of the interaction) and give them a prominent role as social actors within interactions (Groechel et al. 2019) (Young et al. 2007b). Additionally, past research suggests that implementing AR arms on a low-expressivity robot can increase its physical and social presence. Groechel et al. (2019) projected AR arms on a physical robot and found that participants were more likely to view the robot as a physical teammate. VAM-HRI has explored other techniques for using AR to boost expressivity, such as "robot expressionism through cartooning," which applies common comic and cartoon art (Young et al. 2007b). Social perception is also highly influenced by the user's action attribution of the manipulation (e.g., when an object moves during an interaction, the user attributes the movement to

the robot) (Groechel et al. 2022a). The intended anchor location perception being robot-anchored increases social perception.

Trade-Offs Between Functional and Social Perception

Although Walker et al. (2018) reported an improvement in the communication of robot intent and task efficiency, they also observed trade-offs between intent clarity and users' perception of the robot as a teammate. This trade-off between social and functional perception has been recognized in other work as well. Hamilton et al. (2020) compared ego-sensitive-gestures (e.g., a virtual arrow placed on a robot) and non-ego-sensitive allocentric gestures (e.g., a virtual arrow placed on a target), and found that using non-ego-sensitive gestures resulted in faster reaction time and greater accuracy, whereas ego-sensitive gestures resulted in higher social perception and likability (Hamilton et al. 2020).

5.3.2 Design Considerations and Recommendations

Research has shown that when discussing an object, people consider the context of the target object and the surrounding environment when making the decision between using deictic and non-deictic gestures to draw attention to the it (Stogsdill et al. 2021). Similar to how people take account of contextual factors when determining when and how to use gestures, four main design considerations are highlighted for the VAM-HRI community, based on a literature review (Hamilton et al. 2020), (Walker et al. 2018), (Stogsdill et al. 2021), (Piwek 2009), (Tran et al. 2021). From these design considerations, a set of proposed recommendations were developed. The recommendations were designed for mobility-constrained robots with an AR field of view. A summary of the recommendations is found in Table 5.10, which shows the proposed method of gesturing for all combinations of design considerations.

Motivation of the Interaction

The *motivation of the interaction* is the intended end goal or purpose of the interaction. An *interaction* is a single correspondence between a robot and human, and a *composite interaction* as sequence of multiple interactions. The motivation of the interaction is for the robot to communicate

Table 5.10: Recommended Gesture Type Based on Design Considerations: **"A"** is anthropomorphic gesture and **"NA"** is non-anthropomorphic gesture. **"NA*"** suggests a directional non-anthropomorphic gesture (e.g., vector with arrow pointing in the direction of the target). The ordering between **"A"** and **"NA"** suggests the prioritization between the two types.

	Functional Motivation, High Salience	**Functional Motivation, Low Salience**	**Social Motivation, High Salience**	**Social Motivation, Low Salience**	**Functional and Social Motivation, High Salience**	**Functional and Social Motivation, Low Salience**
Close to Target, Both in FOV	A	NA + A	A	A + NA	A	A + NA
Close to Target, Only one in FOV	NA	NA	A + NA	NA + A	NA + A	NA + A
Close to Target, Neither in FOV	NA*	NA*	A + NA*	A + NA*	NA*	NA*
Far from Target. Both in FOV	NA + A	NA + A	A + NA	NA + A	A + NA	NA + A
Far from Target Only one in FOV	NA	NA	NA + A	NA + A	NA + A	NA
Far from Target, Neither in FOV	NA*	NA*	NA* + A	NA* + A	NA*	NA*

information to the user in order to achieve a given goal. More specifically, motivation can have a functional component, social component, or both.

Motivation as a Functional Component - An interaction is motivated by a functional component when the objective is the overall task performance. The accuracy and efficiency of how information is conveyed is prioritized, placing the user's functional perception of the robot above the social perception.

Motivation as a Social Component - An interaction contains a social component when the intent of doing the action is for the robot to be perceived as more social. In other words, the user believes the robot has social agency within an interaction (Jackson and Williams 2021), potentially viewing it as acting more human-like than machine-like. Increasing social perception is an example of social motivation. Social perception is the user's belief that a robot is a social agent; the belief may not accurately reflect the robot's true role or capabilities within an interaction.

Proposed Recommendations

Functional Component - If the motivation for a HRI contains only a functional component (e.g., increasing functional perception), prior research suggests utilizing a non-anthropomorphic gesture to maximize the efficiency and accuracy of task performance (Hamilton et al. 2020), (Tran et al. 2021). Similarly, Krenn et al. (2021) found that robot deictic gestures can help the user interpret the next intended task. Given that the social presence of the robot is not prioritized, solely using non-anthropomorphic gestures will ensure clear and efficient communication (Hamilton et al. 2020).

Social Component - If the motivation of a HRI contains only a social component (e.g., increasing social perception), prior research suggests including anthropomorphic gestures (e.g., virtual arm for pointing). Deictic gesturing, for example, portrays the robot as a social agent that is an active part of the interaction (Hamilton et al. 2020). Using this type of gesture solidifies the robot as an agent and encourages the user to refer back to the robot, increasing the robot's salience within the interaction (Hamilton et al. 2020). While anthropomorphic gestures provide robots with the communication and social presence of human-like gesturing, they impose limitations in terms of communication speed and accuracy. Physically-realistic gesturing takes more time than displaying an arrow on the screen. This phenomenon increases social perception but hinders functional perception due to the reduction in the speed of communication.

Functional and Social Component - If the motivation consists both of functional and social components, then non-anthropomorphic and anthropomorphic gestures can be used in combination, as illustrated in Fig. 5.18c, but the combination is not always the best approach. For consistency, it may be best to use anthropomorphic gestures and consider including non-anthropomorphic gestures only when needed for speed or clarity. The goal of a composite interaction may not always be to maximize functional and social components. For example, if the interaction requires high functional capability, it may be advantageous to forgo anthropomorphic gesturing altogether.

Visibility

Visibility refers to whether the user can see the target object in their field of view (FOV) (Stogsdill et al. 2021). This category can be affected by virtual or physical occlusions. During an interaction, the user may see both the robot and target (AND), only the robot or target (XOR), or neither (NAND).

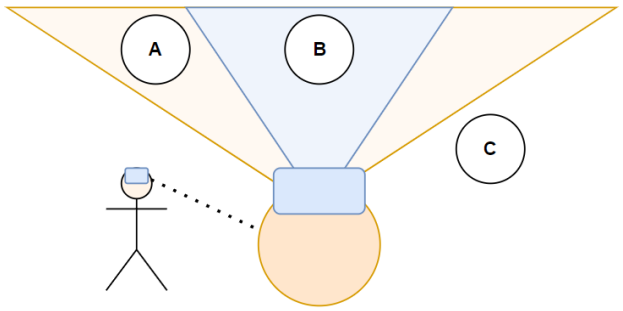


Figure 5.19: Top down view of a user wearing an ARHMD. Object **A** is in the user's field of view (FOV) but not the AR FOV. Object **B** is in both views. Object **C** is in neither.

This section assumes that an object within the FOV is within the boundaries of the AR FOV (e.g., object B in Fig. 5.19). AR FOV is a subset of human FOV that is limited by the scale in which the platform displaying AR components is able to capture the environment the user sees. A physical object that is in the user's FOV but outside of the AR FOV is considered to be out of frame. This includes the robot as long as the AR appendages are not visible within the AR FOV. Because the FOV of AR devices is dynamically changing and interaction-dependent, proposed recommendations specifically target the use of an AR FOV rather than a human FOV. Consequently, the recommendations for this category may not be applicable if a non-AR FOV is being used.

Both Robot and Target Object are Visible (AND) - The user can clearly view the robot and the target object in their FOV without having to adjust their gaze.

Visible Robot or Object but Not Both (XOR) - The user can only view the robot or only view the target object in their FOV. The user may need to adjust their gaze in order to see both targets at once or it may not be possible for them to be in the same FOV.

Neither Robot nor Target Object is Visible (NAND) - The user cannot see the robot or the target object because they are placed outside of the viewer's FOV.

Proposed Recommendation

Both Target and Robot in User's FOV (AND) - Humans are more likely to use deictic gestures (e.g., pointing) when the target is visible (Stogsdill et al. 2021). To maximize the robot's social perception, if both the target object and the robot are within the user's FOV, prior research suggests to use anthropomorphic gestures, as shown in Fig. 5.18a. The robot using anthropomorphic deictic gestures will improve the user's recall and human-robot rapport (Stogsdill et al. 2021).

Visible Robot or Object but Not Both (XOR) - If the target is outside of the user's FOV but the robot is within it, an anthropomorphic and non-anthropomorphic gesture (e.g., an arrow from the robot's pointing finger) should be considered to associate the robot to the target. This implementation increases both functional and social perception because anthropomorphic deictic gesturing draws attention and connects the robot's intentions with the specific task and functionality (Stogsdill et al. 2021). The non-anthropomorphic gesture contributes to the functional perception by allowing the user to accurately distinguish the location of the object. The same recommendation applies if user can only see the target.

In a scenario where only the target is visible, developers can choose whether or not to implement non-anthropomorphic gestures. If non-anthropomorphic gestures are present, the user can clearly identify the target without having to refer back to the robot, increasing functional perception. However, if only anthropomorphic gestures are used, the user will be compelled to find the robot. Looking back at the robot in order to identify its anthropomorphic gestures will make the robot a more prominent as an active agent in the interaction, promoting its social perception.

Neither in User's FOV (NAND) - If neither the target nor robot are in the user's FOV, the first consideration is what the motivation for the interaction is and whether it consists of a functional and/or social component. For instance, one might use a directional non-anthropomorphic gesture, such as a large arrow that points to the direction of the robot's location or an arrow as in Fig. 5.17 to redirect the user's FOV to the robot. However, if the motivation is social, using anthropomorphic deictic gesturing that encourages the user to find the robot first and then search for the object may be more effective for social perception.

Salience of Target

Salience refers to how noticeable a target is within its environment, a property analogous to accessibility (Piwek 2009). Gestures made toward an object can change its salience value.

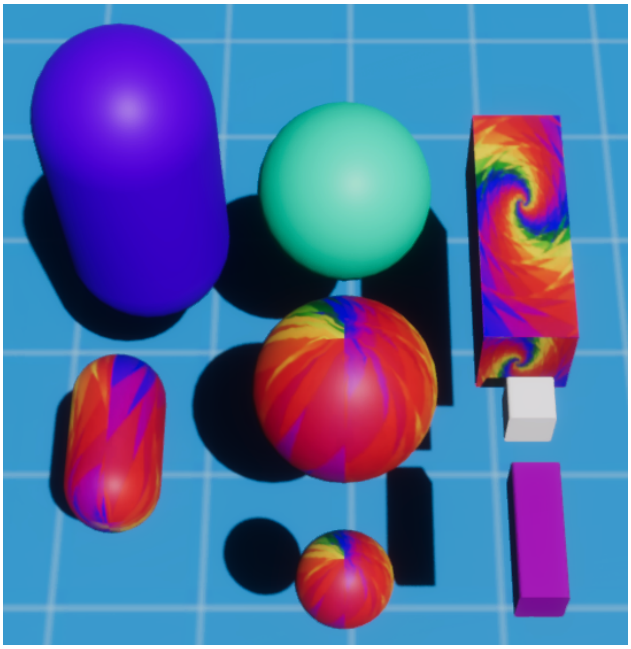


Figure 5.20: The target object's salience depends on the characteristics of the objects around it (e.g., color, scale, shape). The more the objects share similar characteristics, the lower the target object's salience.

When there are multiple objects in close proximity (e.g., Fig. 5.20), each may have attributes that distinguish it from others, (e.g., color, scale, shape) and increases its salience (Piwek 2009). Non-verbal robots are often unable to or highly constrained in their ability to indicate the location of a specific object while utilizing existing salience factors (Cha et al. 2018); they are typically

limited to gesturing toward the object to draw attention to it. Therefore, when it comes to nonverbal signaling, gesturing can increase the salience of an object, but anthropomorphic deictic gesturing often cannot distinguish the object if it is in close proximity to others.

Anthropomorphic deictic gestures can also increase the salience of surrounding objects by bringing attention to the overall area. This is known as implied spatial salience (Piwek 2009). It is necessary to consider salience when choosing to include anthropomorphic and non-anthropomorphic gestures because it is also determined by how distinguishable the target is from its neighbors.

Proposed Recommendation

Target is Far Away from Other Objects - When the target object in an interaction is not in close proximity to other objects, the user can clearly interpret the intent of an anthropomorphic gesture made toward that object (Piwek 2009). If there are extraneous factors that prevent the user from clearly identifying the object (e.g., distance, FOV), non-anthropomorphic gestures can be implemented to more accurately communicate the location of the target.

Target is in Close Proximity with Other Objects - The target has low salience if it is in close proximity with other objects (Piwek 2009). The user may find it difficult to distinguish between the objects in the direction the robot is pointing, and may require additional visual assistance. Therefore, non-anthropomorphic gestures, along with anthropomorphic gestures, can be used to clearly indicate the target. If the motivation of the interaction is functional, developers may avoid anthropomorphic gestures and simply use non-anthropomorphic gestures to quickly indicate the target.

Salience of AR Components:

While design recommendations emphasize the salience of the target object, the salience of the non-anthropomorphic gestures should also be considered. Factors such as color, size, and opacity may influence how the user perceives the robot's gesture. Salience of a non-anthropomorphic gesture affects how disruptive it is to the user, which determines how noticeable it is (Pärsch et al. 2019) and how quickly the user can take action after the gesture is produced. For instance,

developers generating an arrow to gesture towards a target object may choose to select an opaque color, as in Fig. 5.17, to make it more noticeable, or select a shade that is more transparent, as in Fig. 5.18c, to make it more cohesive with the robot and obstruct less of the background. Considering the opacity, scale, and color of non-anthropomorphic gestures is important because overlaying AR visualizations onto the real world may distract the user and obstruct their FOV (Pärsch et al. 2019). Therefore, non-anthropomorphic gestures that have higher opacity, large scale, or similar color to the background are undesirable.

Choosing the color of the non-anthropomorphic signal objects (e.g., arrows) can be specifically targeted for different use cases. A signal can be given the same color as the target object for better salience with that object. To avoid blending in with the background, a dynamic coloring system can be created by contrasting the color of the signal's background. A color can be generated as a pre-programmed map (e.g., complimentary colors), inverted, or changed to a secondary color choice. Finally, when aiming for the most social signal, developers should choose a color on the robot or average color of the robot to boost the signal attribution toward the robot. The same recommendation applies to choosing colors for adding appendages to the robot.

Distance

In an interaction composed of a user, target object, and robot, the distance model prioritizes the displacement between the robot and the object. Hamilton et al. (2020) reported that there was no significant evidence that the distance between a user and target influenced the social presence of a robot. The findings of their study suggest that there may be a correlation between the robot-target distance and the robot's social presence, which calls for further study. This work focuses on mobility-constrained robots; distance issues are more complex for highly mobile robots (e.g., drones) that can move quickly and efficiently toward a target.

Proposed Recommendation:

Target Close to Robot - If the target object is in close proximity to the robot, anthropomorphic gestures may be used to promote the robot's social presence. Using deictic gestures would enhance the robot's anthropomorphism because humans tend to use the same gesturing model when objects

are visible and in close proximity (Stogsdill et al. 2021). The close proximity will also reduce issues of salience (Hamilton et al. 2020). Solely using anthropomorphic gestures when the robot is close to the target boosts the social perception of the robot. However, if there are extraneous factors (e.g., salience, field of view) that prevent the user from clearly distinguishing the target object, non-anthropomorphic gestures may be helpful.

Target Far from Robot - If the target object is outside of a predetermined distance boundary (calculated for each interaction considering the specific robot and environment), it can be considered far from the robot. Since distance between the robot and target make it difficult to interpret the direction of deictic gestures, developers may use non-anthropomorphic gestures (Stogsdill et al. 2021) along with anthropomorphic gestures to indicate the location of a target object without sacrificing functional or social perception. If the interaction is motivated by a functional component, developers may use only non-anthropomorphic gestures in order to promote speed and accuracy (Hamilton et al. 2020).

5.4 Discussion and Summary

This chapter outlines how the design of nonverbal gestures and visual elements in AR for SAR can significantly impact user perception of a robot's expressivity. To make these elements clear and effective for all types of users, they must be adaptable to different preferences and backgrounds. Future work should focus on developing adaptable visualizations and exploring the effects of different gesture types on user perception in various scenarios and with different types of robots.

For social expressivity (Sec. 5.1), the arms vs. no arms conditions did not show significant differences for task efficiency or subjective measures. However, the two conditions were highly correlated with user perception of Kuri as a physical or virtual teammate. The results suggest that participants may have associated arms with physical tasks, such as picking up objects or pointing. The experiment condition also involved more overall movement, which may have conveyed a perception of physicality. The binned subjective results suggest that users preferred a physically

associated MR robot. This aligns with the importance of embodiment for social presence. Future work should explore other factors that increase physical presence and the potential for new gestures and actions in VAM-HRI.

For functional expressivity (Sec. 5.2), this work presents Virtual Design Elements (VDEs) for 6 robot signals and evaluates their designs for clarity and visual appeal using an AMT study. Some VDEs consistently scored higher than others, such as the pyramid for the camera and the box design for face detection. In other cases, different groups of users had different preferences for VDEs. Common themes in the qualitative responses included the salience of the visual and the inclusion of verbal and text-based information. To make VDEs clear for all types of users, they must be adaptable to different preferences and backgrounds. Future work should focus on developing these adaptable visualizations as a basis for signal design.

Finally in Sec. 5.3, this work presents design recommendations for nonverbal robot gesturing that consider both the social and functional aspects of the interaction. By taking into account factors such as motivation, visibility, salience, and distance, guidance is provided on selecting gestures that can maximize perception by the user. However, these recommendations are not exhaustive and require further investigation and empirical evaluation. Future work should explore the effects of different gesture types on user perception in various scenarios and with different types of robots.

Chapter 6

Controller - Designing Embodied, Flexible, and Extensible Interactions

This chapter explains two applications that utilize iteratively learned and validated design conventions and considerations for the interface between the model and view: the controller. The first application is the extension of M2C first described in Sec. 4.1, a visual programming language with a robot tutor that uses AR to increase young students' curiosity in coding. The second application, PoseToCode (P2C), is a kinesthetic web application that aims to increase students' curiosity in coding, deployed in schools. The chapter covers both existing and new design conventions and considerations for creating AR experiences with socially assistive robots.

6.1 MoveToCode: Iterative Design of an Embodied AR Visual Programming Language

Contributors: Section 6.1 is based on work drafted to be submitted. Additional authors on the drafted work include İpek Göktan, Karen Ly, Anna-Maria Velentza, and Maja J. Matarić.

As described in Sec. 4.1, M2C was developed to leverage VAM for SAR. M2C is an open-source, embodied (i.e., kinesthetic) learning visual programming language that aims to increase

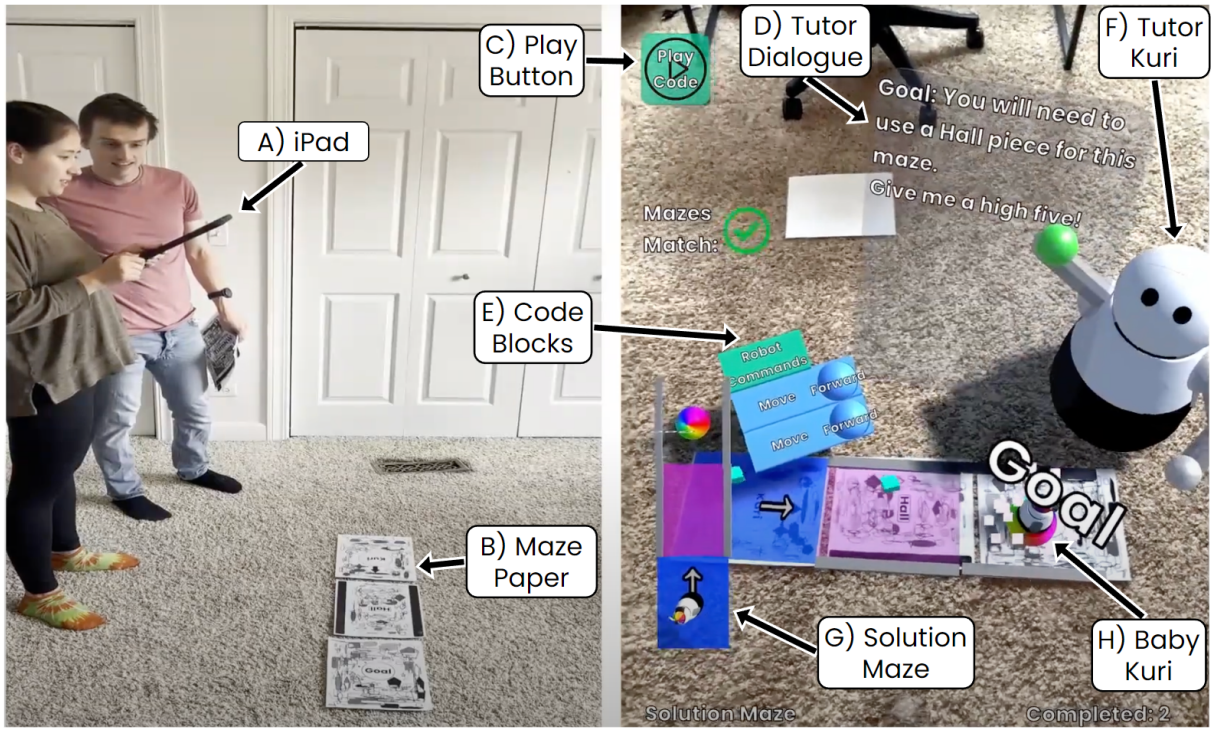


Figure 6.1: M2C pair-programming exercise. Left) External view of pair programmers. Right) M2C activity view. A) vertically held mobile tablet; B) tangible maze paper tacked by the tablet; C) code play button; D) virtual tutor dialogue; E) code blocks that control the miniature robot through the maze; F) autonomous, AR robot tutor Kuri posed for a high five; G) goal maze configuration; & H) miniature robot starting on the blue tile and programmed to reach the goal tile.

the curiosity of young students (ages 8-12) in programming. This section describes the extension and improvement of M2C while emphasizing learned and known design conventions.

The improved M2C application utilizes an AR autonomous robot tutor named Kuri which models the students' kinesthetic curiosity with habituation and responds to help promote their curiosity in programming. The design of M2C was informed by a series of pilot studies and was subsequently validated in local Los Angeles elementary classrooms ($n = 21$). Results from these final classroom deployments validated the design decisions when compared to the final classroom pilot ($n = 15$), showing an improvement in perceived robot helpfulness (median $+\Delta 1.25$ out of 5) and number of completed exercises (median $+\Delta 1$ out of a maximum of 11). Although no significant changes were found for pre-post student interest or intention to program later in life, students wrote more open-ended questions post-study revolving around topics of the robot, programming,

research, and if they were able to do the activity again later. Overall, this work demonstrates the potential of using VAM-HRI in a kinesthetic context for SAR tutors, and highlights the existing conventions and new design considerations for creating AR applications for SAR.

6.1.1 Technical Approach

M2C was designed through a series of university ethics board approved pilot studies, culminating in its final deployment as described in Sec. 6.1.2. These studies included transitioning M2C from expensive AR headsets to more affordable tablets, redesigning the exercise environment to support tangible pair-programming for 8-12 year old students (Fig. 6.2), and revising the actions and policy of the robot tutor. Throughout this process, known design conventions were blended with new considerations specific to M2C, a embodied learning AR activity with a robot tutor. This section highlights the most relevant aspects of the M2C design process and implementation. The application is open-source and available at `https://github.com/interaction-lab/MoveToCode`.

M2C Implementation and Exercises

M2C was originally designed (Groechel et al. 2021) for the Microsoft Hololens 2 - an ARHMD. The application was redesigned to work on any ARKit or ARCore supported phone or tablet (Nowacki and Woda 2020) as can be seen in Fig. 6.1. These phones and tablets are significantly cheaper (at the time of writing it costs ≈$3.5k USD for the Hololens 2 where a 9th Generation iPad costs ≈$330 USD) and are therefor much more feasible for real-world classroom deployments.

M2C was originally designed to teach traditional computing concepts such as print statements, math equations, if-statements, variables, and looping through console-based exercises. As a result, it was difficult for new learners to understand in one 20 minute session. Further this original design did not have any physical connection to the real world, which is a common benefit of tangible programming languages (Sapounidis and Demetriadis 2017).

To better suit the 8-12 year old target audience, the design of M2C was changed to involve programming a miniature robot through a maze created by the students using physical pieces of

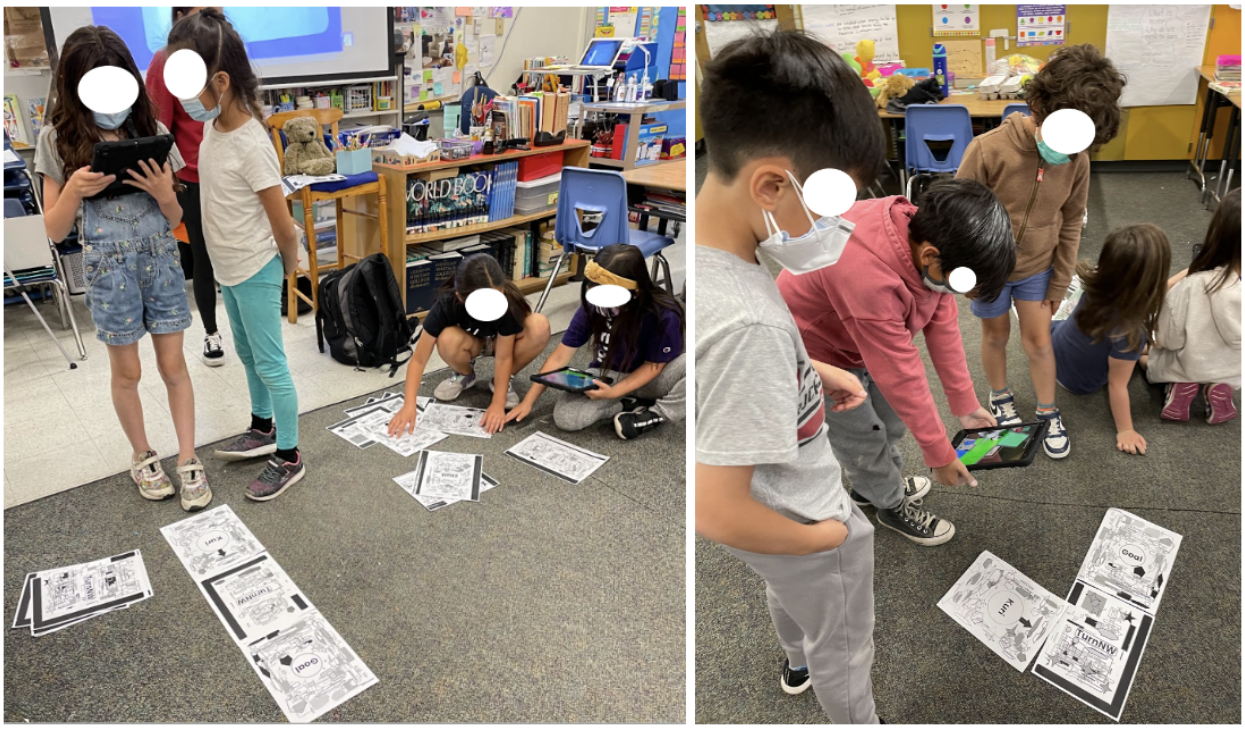


Figure 6.2: View of the final M2C pilot study with a local Los Angeles school of 8-12 year old students.

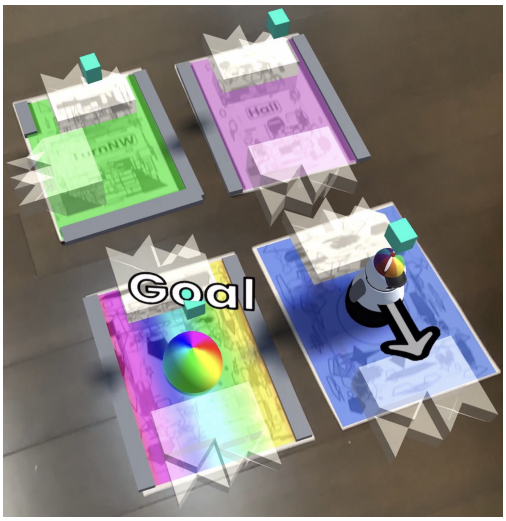


Figure 6.3: The four types of maze pieces include the turn piece, the hall piece, the goal piece, and the baby Kuri starting position piece.

paper. This approach has been shown to be effective in increasing computation thinking skills in a variety of systems and age groups (Kanellopoulou et al. 2021; Guenaga et al. 2021; Ternik et al. 2017). The miniature robot is called "baby Kuri" (Fig. 6.1.H) and there is also a larger, autonomous

robot tutor called “tutor Kuri” (Fig. 6.1.F). Both robots are 3D models of the Mayfield Kuri robot, with tutor Kuri also featuring socially expressive arms from Groechel et al. (2019). Each piece of paper (Fig. 6.3) corresponds to a type of maze piece that the mobile device tracks:

- `Turn` - piece with walls on 2 contiguous sides
- `Hall` - piece with walls on 2 parallel sides
- `Goal` - final navigation goal
- `Start` - where the baby Kuri starts

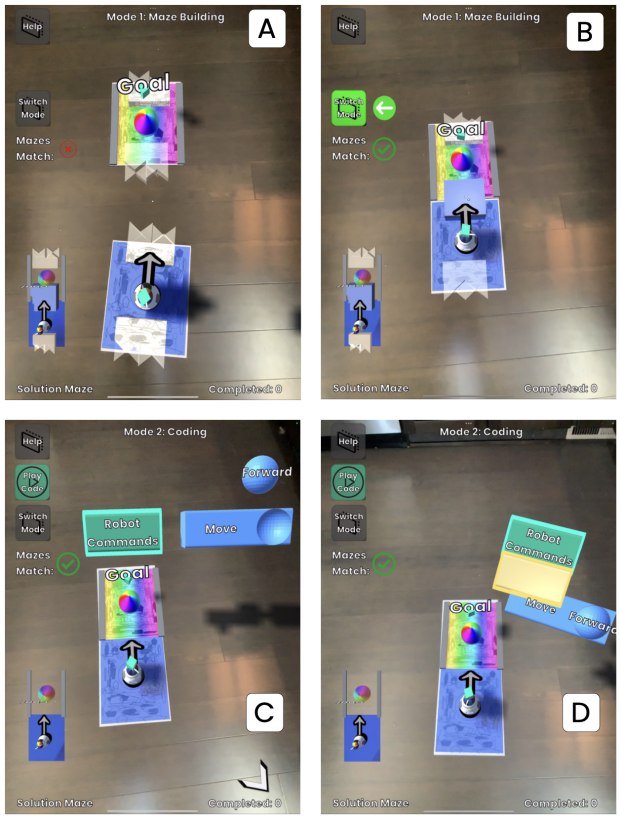


Figure 6.4: M2C exercises are split into two modes. In mode 1 (A & B) the user connects maze pieces to match a solution maze. In mode 2 (C & D) the user codes the baby Kuri to complete the maze.

M2C exercises are divided into two modes as shown in Fig. 6.4. In mode 1, students rearrange physical pieces of paper to create a maze that is identical to the “Solution Maze” (Fig. 6.1.G).

Maze pieces have connector pieces that are lined up and highlight when they are connected (Fig. 6.4.B). When the maze is identical, the student holding the iPad can press a "Lock Maze" button thus locking the connectors and transitioning the app to mode 2. In mode 2, the student uses 3D code blocks (Fig. 6.1.E) to program the baby Kuri robot to navigate the maze.

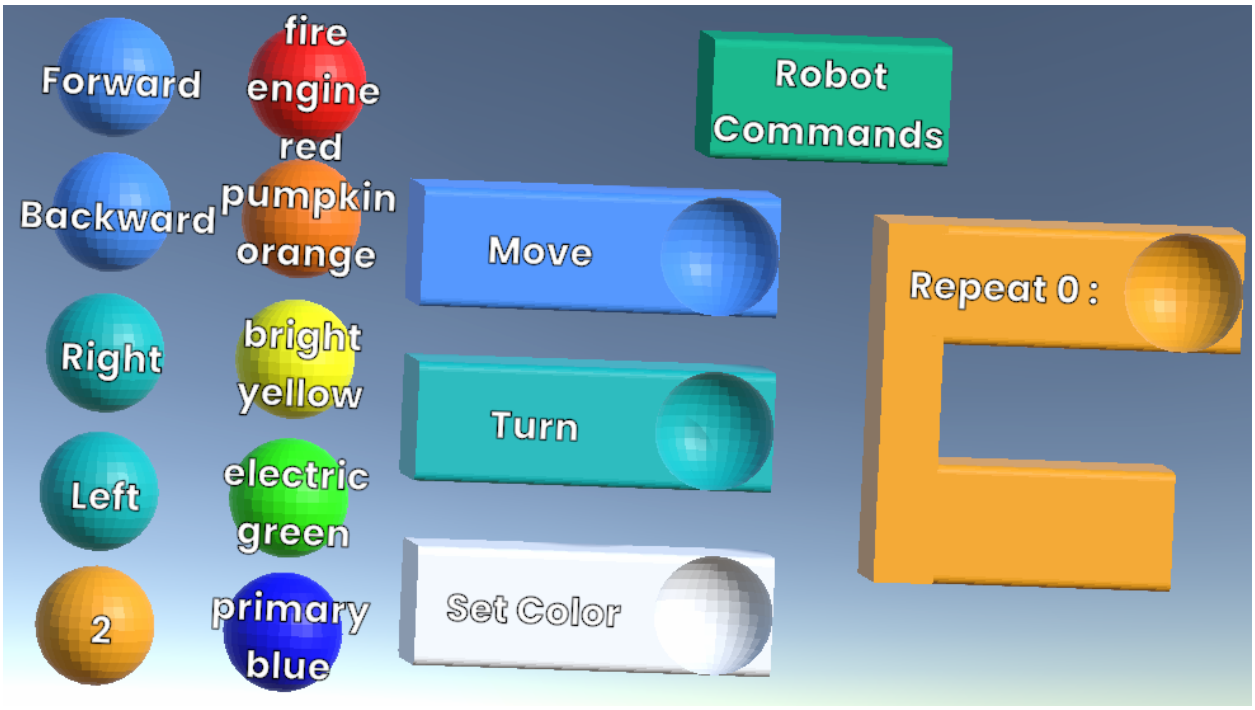


Figure 6.5: All code block types used for programming the baby Kuri through maze exercises.

The code blocks used in the exercises are shown in Fig. 6.5 and include:

- `Robot Commands` - starting point of the code
- `Move` - moves the robot according to given argument
- `Turn` - turns the robot according to given argument
- `Repeat` - repeats blocks inside a given argument times
- `Set Color` - sets baby Kuri color given an argument
- `Forward/Backward` - argument for `Move` block
- `Right/Left` - $\pm 90^{\circ}$ argument for `Turn` block

- `Integer` - number argument for `Repeat` block
- `Color` - argument for `Set Color`

This is the subset of M2C supported code blocks used for the maze exercises. The complete set of supported code blocks can be found at `https://github.com/interaction-lab/MoveToCode`.

M2C exercises cover computational thinking concepts such as sequencing, looping, and using different blocks to solve the same problem (e.g., `Move(Forward)` ≈ `Turn(Left)`→`Turn(Left)`→`Move(Backward)`). Ten exercises were created and ordered based on the time taken to complete them in a pilot study. All necessary code blocks for each exercise are provided. As the exercises become more complex, erroneous blocks are also added. These blocks are either 1) not possible to be part of a correct solution; or 2) `Set Color + Color` blocks, which allow students to change the color of baby Kuri but have no direct effect on the correct solution.

Kinesthetic Curiosity Habituation & Tutor Action Policy

Tutor Kuri's action selection was largely based upon a student's *kinesthetic curiosity* (KC^S) outlined in Sec. 4.1.1 and Eq. 4.3. Information seeking actions (ISA) ISA scores are defined relative to the domain and action space of the learner. Improving upon the original definition, a version was implemented using habituation saliency (Wu and Miao 2013) for a given human action defined as:

$$KDS = 1 - \frac{ActionCounts[A_t]}{max(ActionCounts)} \tag{6.1}$$

$$TS = min(\frac{(t - lastRecordedTime(A_t))^2}{60^2}, 1) \tag{6.2}$$

$$ISA_t^S = \begin{cases} 2 & \text{if } A_t \text{ has never been done} \\ 0 & \text{if } A_t \text{ is recorded in last } tw \\ KDS + TS & \text{otherwise} \end{cases} \tag{6.3}$$

where A_t is the human action taken at time t in seconds, $ActionCounts$ is a map of unique actions to the total number of times each action has been completed, KDS represents knowledge-driven saliency, and TS represents temporal saliency. KDS is discounted the more an action has been done normalized by the max number of times any action has been taken. TS grows at a quadratic rate by squaring the difference of time t and the last recorded time of A_t followed by normalization via squaring a max time constant of 60 seconds. TS maxes out at 1 denoting any action done ≥ 1 minute ago.

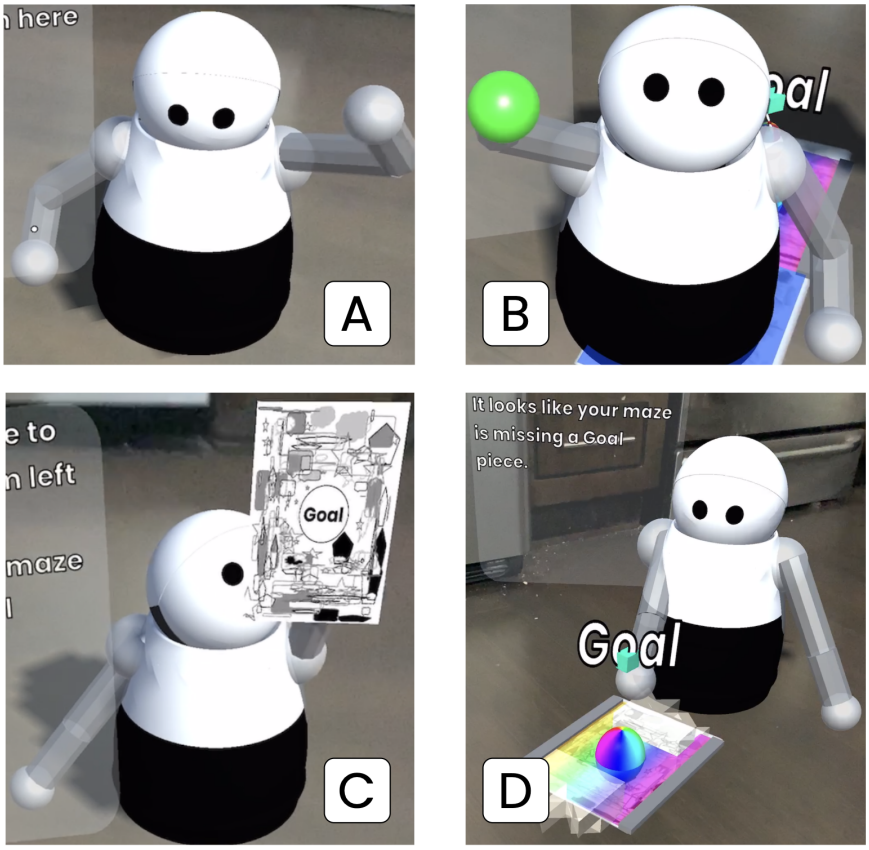


Figure 6.6: A subset of actions tutor Kuri performed. A) wave; B) high five; C) showing a type of missing paper; and D) moving to and pointing at a misaligned maze piece.

Shown in Fig. 6.6, tutor Kuri actions include the following with * indicating *context-dependent helpful actions*:

- Idle and look around
- Wave to user

- Interactive high-five
- Move out of the user's way
- Dialogue (Fig. 6.1.H)
  - Exercise goal
  - Congratulatory phrases
    * Encouragement phrases
    * Referencing a maze piece or code block

* Showing a type of maze paper not yet used

* Moving and pointing to a misaligned maze piece

* Moving and pointing to a misaligned code block

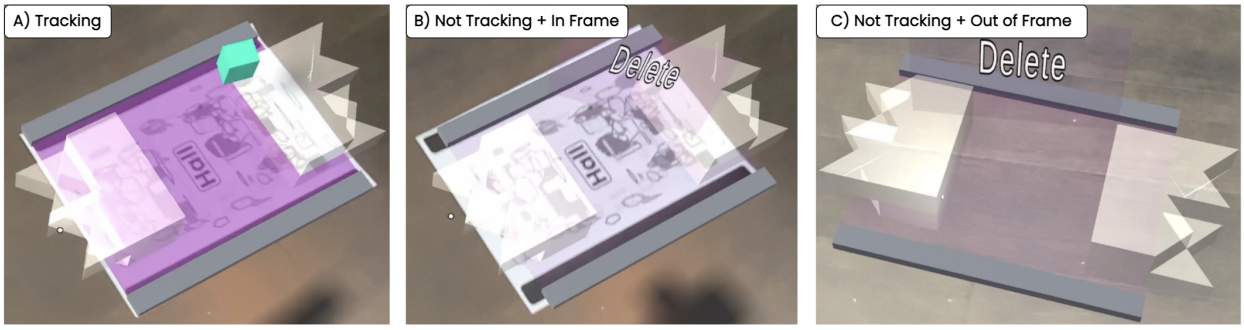


Figure 6.7: All possible states of tracking a piece of maze paper. A) virtual analog is overlaid with spinning tracking indicator cube; B) virtual analog persists having higher transparency, removing the spinning indicator cube, and adding a delete button; and C) virtual analog is the same as B when the paper is out of view of the mobile device.

The tutor Kuri action policy is designed to give *context-dependent helpful actions* whenever $KC_t^S < 0.5$ and the last time Kuri performed an action was $> tw$. The 0.5 threshold was chosen because it was shown to produce higher short-term KC_t^S scores compared to a lower threshold (Groechel et al. 2021). Tutor Kuri gives the exercise goal at the start of a new exercise and offers a high-five with accompanying congratulatory dialogue upon exercise completion. Tutor Kuri

moves out of the user's way whenever it is not performing an action and collides with virtual objects (often maze pieces) or is $< 0.75m$ from the user. The target position is calculated in the horizontal plane (i.e., $y = 0$ for a 3D $\{x, y, z\}$ vector where y is defined as up) as follows: A vector from the collided object to the user is calculated and normalized as $\hat{V_C}$. A vector $\hat{V_{avg}}$ is calculated as the vector between $\hat{V_C}$ and the user's unit forward vector. $\hat{V_{avg}}$ is added to the user's position to denote the target destination for tutor Kuri. Tutor Kuri only waves to the user when it first arrives and whenever it is about to leave as defined by the within-subjects study design conditions (Sec. 6.1.2).

Design Considerations Implemented from Pilot Studies

The M2C system was iteratively designed through a series of pilot studies, which allowed for adjustments based on user feedback and observations. The design process began with a pilot study ($n = 10$) involving college students and the original language headset design (Groechel et al. 2021). This was followed by a pilot study ($n = 5$) of Ph.D. and Master's students from the USC Interaction Lab, who tested a console-based M2C system adapted for tablets with the AR tutor Kuri added. The exercises were then redesigned to focus on programming the baby Kuri through a maze, using tangible pieces of paper to anchor the experience. A final pilot study was conducted with elementary school students ($n = 15$) in Los Angeles as part of an after school robotics club consisting of students ages 8-12, and the final design was a combination of lessons learned from these pilots and existing design conventions drawn from Google's AR User Experience Design Guidelines (Google 2022). These designs were then tested in a full study of two Los Angeles elementary school classrooms, as described in Sec. 6.1.2.

Existing Design Conventions

During the design process for M2C, a number of existing AR-specific design conventions were identified as useful for improving the user experience. These conventions, which have been effective in other AR applications, were chosen based on observations and feedback from pilot studies. Despite being known conventions, they were repeatedly emphasized during the design process due to their importance for improving the user experience and therefor emphasized within this work.

One important design consideration was the difficulty that users had in understanding the z-depth of virtual objects, or how far away an object was from the user. To address this issue, efficient, shader-based shadows were added to objects (Fig. 6.4), which helped to improve depth perception (Diaz et al. 2017). Another useful design convention was the use of context-aware arrows to reference off-screen objects of interest (Fig. 6.4.C), such as code blocks or the tutor Kuri. These arrows functioned as non-anthropomorphic deictic gestures, aimed at quickly drawing the user's attention to an object (Goktan et al. 2022; Brown et al. 2022). Finally, it was observed that users struggled with manual rotation of 3D objects, such as code blocks. To address this issue, a feature was implemented that locked the rotation of these objects to the user when manipulated. This feature, based on guidelines from Google's Augmented Reality User Experience Design Guidelines (Google 2022), helped to improve the usability of M2C.

New Design Considerations

During the pilot studies for M2C, new design problems and solutions emerged that focused on two specific categories: the use of physical pieces of paper and the 3D code blocks. It was also observed that holding mobile devices vertically allowed for prolonged use compared to holding them horizontally, leading to the final design of M2C being in a vertical format.

One key solution was the use of physical pieces of paper as a exercise medium (Fig. 6.3), which connected learning to the physical world similar to that of tangible programming languages (Sapounidis and Demetriadis 2017). The paper allowed for a defined role for a second student, eliminated the need for one student to hold the mobile device while the other watched, and provided a physical anchoring point for virtual content. The paper anchors served as spawning reference points for virtual objects, such as code blocks and virtual maze pieces, and naturally defined the AR play area for any physical environment. AR experiences need to account for many possible physical domains (Google 2022), from a large convention center to a cramped room. This means you need to define the play area accounting for these different areas and adjust the virtual content to avoid object clipping or unreachable objects. The physical pieces of paper naturally restrict the

play area as you can only place them in realistic locations. Finally, paper was chosen over custom-made or 3D printed objects to improve accessibility for real-world classroom deployments.

Tracking the pieces of paper, however, presented a new problem (Fig. 6.7). When the maze paper was tracked by the mobile device, a virtual analog was positioned exactly where the paper was. However, when the paper stopped being tracked, the virtual analog persisted, causing confusion for users. This persistence was necessary for mazes that required a large number of pieces, as the mobile device might not be able to track them all at once even if the physical pieces are all in the camera frame. To address the confusion, spinning tracking indicators were first added (Fig. 6.7.A) to the virtual analogs when the paper was being tracked, and made them disappear when tracking ended. Despite this solution, pilot users remained confused. To eliminate this confusion, the virtual analogs were turned primarily translucent ($\alpha \approx 11.7\%$) when compared to the tracking ($\alpha \approx 54.9\%$). The difference, although not directly measured, was evident by the lack of tracking questions asked by participants between the final elementary pilot study and the full study described in Sec. 6.1.2. Finally, a `delete` button was included that appeared when tracking ended, allowing the user to remove the virtual analog.

The 3D code blocks were designed using tangible and virtual programming language benefits. The benefits of tangible programming language interfaces, such as those used in AR settings, are well-established in the literature (Sapounidis and Demetriadis 2017; Jin et al. 2018; Hattori and Hirai 2019). These interfaces allow users to interact with code in a more tactile, spatial, and persistent manner, as the physical pieces can be placed in the environment without taking up space on the mobile device interface. In contrast, 2D block-based coding interfaces, like Scratch (Resnick et al. 2009), offer unlimited supplies of blocks and ease of rearrangement, but take up a large portion of mobile device screen real estate and are tied to the device reference frame. To address these issues, 3D code blocks were created for the M2C system that offer the benefits of both virtual blocks (e.g., ease of spawning, dynamic expansion, and repositioning) and tangibles (e.g., persistence in the 3D environment and spatial grounding).

Connecting these code blocks, however, presented a new challenge, as users struggled with z-depth manipulation. To solve this problem, a system was implemented that casts a ray from the user's grab point to the object, which was treated as part of the code block's hit box, effectively extending it along an infinite z plane relative to the user. This approach, along with design choices outlined in Section 6.1.1, helped to improve user interaction with the M2C system.

The last design considerations revolved around the robot tutor actions (listed in Sec. 6.1.1). Two specific actions learned from piloting were the need for a way to interact back with the robot (leading to the interactive high-five) and the need to prevent the robot from bothering the user too often. In the final classroom pilot, students were heard saying "get out of the way Kuri" or similar phrases. When interviewed, they mentioned wanting to "move Kuri out of the way", for Kuri to "go above the ceiling", or for Kuri to "go outside." This led to the creation of the "move out of the way of the user" action, as students did not want to be constantly pestered by the robot tutor in the same way they wouldn't want to be constantly pestered by a human tutor.

6.1.2 User Study

Hypotheses

H1: When comparing the final pilot with the final studies, design decisions described in Sec. 6.1.1 increase:

- **A:** ***Perceived Robot Helpfulness***
- **B:** number of completed exercises

H2: The presence of a virtual robot tutor, when compared to no virtual robot tutor, increases the amount of time the user looks at the tutor robot or tutor dialogue.

H3: When comparing post-interaction to pre-interaction, students inidicate an increase in:

- **A:** ***Interest in Programming***
- **B:** ***Future Intention to Program***

Recruitment and Participants

This study was approved by the University's Institutional Review Board (IRB #UP-20-00030). First a recruitment flyer was sent out to a list of local schools, both public and private, from the USC Viterbi School of Engineering K-12 STEM Center. Inclusion criteria for the study were students 7-13 years of age and proficient knowledge of the English language. Two teachers from two different schools responded and scheduled 1 hour study sessions.

Twenty-one students participated. Before each study, legal guardian consent and child assent were obtained for all 21 of these students. Students who did not have signed parental consent still participated in the M2C activity but no data were collected. These students were subsequently not included in any data analysis.

These students first filled out a demographic questionnaire prior to the activity. There were 8 students in the first school and 13 in the second. Fifteen identified their gender as male, 5 female, and 1 preferred not to specify. Their age ranged from 9-10 years old ($\bar{X} = 9.5, \sigma = 0.5$). Identified ethnicity of the students were of Hispanic origin : 13, Black/African American + Asian: 2, Black/African American + Hispanic origin: 1, Hispanic origin + White: 1, Black/African American: 1, Asian + Middle Eastern or North African: 1, and preferred not to specify: 2. Prior coding experience included Code.org: 9, Scratch + Code.org: 3, Scratch: 2, Scratch + Code + Roblox: 1, Robotics: 1, Other (not specified): 2, and None: 3.

Measurements

To evaluate the task and measure students' attitudes and curiosity toward programming and Kuri robot, both quantitative and qualitative measurements were used by giving the pre- and post-tests to submit.

Quantitative Questions – To measure students' curiosity, the established question generation task was used in which the students are prompted to ask as many questions about a topic, without providing answers (Harris 2012). The task has been used in relevant research, measuring kid's curiosity after interacting with a social robot(Gordon et al. 2015). Students were instructed to

write down as many questions as they could after the briefing section to avoid any biases from the task or the questionnaires pre-test- and then after the end of the task -post test-.

Qualitative Questionnaire – The questionnaire was constructed based on valid and reliable existing questionnaires, adapted for the research needs. It was also evaluated by two independent teachers to test it for being appropriate for the students' age needs. The pre-test questionnaire had two parts: first, the students' attitudes, separated into two major thematic areas 1) ***Interest in Programming*** construct which had 13 items and evaluated the students' interest in programming, based on (Gul et al. 2022) questionnaire, and 2) ***Future Intention to Program***, which had 3 items and evaluated students' intention and motivation to follow a future career in programming, subset of the STIMEY Horizon Project questionnaire (Velentza et al. 2020). Students evaluated them on a Likert scale from "Strongly Disagree" to "Strongly Agree". An additional four demographic questions were added to the pre-survey regarding students' age, ethnicity, gender identity, and prior experience with programming. The post-test questionnaire had the same questions with the pre-test plus one more thematic area, ***Perceived Robot Helpfulness*** with four items based on (Putnam et al. 2020) usability scale appropriate for children.

The overall questionnaire's validity was tested by a multidisciplinary group of engineers and psychologists with Lawshe's subject-matter expert rating methodology (SMEs). The Content Validity Ratio (CVR critical) of the questionnaire was acceptable for five experts at .99, with two-tailed p=.01 (Wilson et al. 2012).

Procedure

Students were first given a pre-survey which included demographic questions, interest in programming, and intention to program (described in detail in Sec. 6.1.2). For classroom 1, students were shown a video of M2C (`https://youtu.be/6CMuADWboD8`). An issue arose from this class not fully understanding the basic usage of M2C (i.e., the video was later cited as too quickly leading to students not even being able to assemble the first simple maze) so, as done for the pilot, a demo was given to the second class in place of the video. Students in the first class were shown basic demos after noticing they struggled with basic mechanics of M2C within the first 2

minutes. It is recognized this is a study confound between the classes and therefore pool all data as one dataset for analysis, not as separate classroom datasets. As this application is designed for a real world deployment, this work equates this class difference to different teachers giving varying levels of explanation and demonstrations.

After the video or demo, students were given five minutes to write any and all questions they had. During this time, the respective teacher assigned pairs of students to ensure students with consent and assent forms were paired. The second class had 13 students, leading to one group of 3. Working in dyads in a supporting environment enhances students' problem-solving, and difficulties management, while by communicating with their peers they increase their belief in their capabilities (Çakır et al. 2017), (Carvajal-Ayala and Avendaño-Franco 2021). Although collaboration can be challenging depending on students' social and cognitive skills(Laru et al. 2012), when working in a supportive environment provides engagement in STEM activities (Puvirajah et al. 2020) and the teachers were asked to pair the students, to match their individual characteristics.

Each group was given a tablet and maze papers. The 9^{th} generation Apple 10.2-inch iPads were used as Apple has 52% of the United States market share for tablets as of 2022 (GlobalStats 2022). Each group was assigned a sufficient work area in the classroom (e.g., Fig. 6.2).

The M2C activity, described in Sec. 6.1.1, was 20 minutes and 22 seconds broken down as follows:

- 6 seconds allowing the iPad to scan the room geometry
- 10 minutes 8 seconds condition A
- 10 minutes 8 seconds condition B

The **experimental conditions** were of the AR robot tutor Kuri model either being visible or not visible. In both conditions, the tutor Kuri dialogue box (Fig. 6.1.D) was still visible. The conditions were randomly counterbalanced having 5 groups starting by seeing tutor Kuri and the remaining 5 groups only seeing the tutor dialogue box. The action policy described in Sec. 6.1.1 operated the exact same in both conditions, with the only difference being the visibility of the AR

robot tutor body and arms. At the beginning of each condition, tutor Kuri would either wave and say hello or wave and say goodbye for 8 seconds, depending on the current state of tutor Kuri.

The application automatically closed itself upon completion. Although all students were instructed this would happen, every group restarted the application. They were instructed to stop with this second running of the application and returned to their seats. Students were then given a post-survey with the same interest in programming and intentions to program later in life surveys. Finally, students were asked to write down any questions they now had for 5 minutes.

Data Analysis

Both the final study group ($n = 21$) and the final pilot group ($n = 15$) were compared. In the pilot group, 8 identified their gender as female and 7 as female. Their age ranged from 8-12 years old ($\bar{X} = 9.7, \sigma = 1.1$). Identified ethnicity of the students was Asian + White : 3, White: 3, Hispanic origin + Asian: 2, Asian: 2, Hispanic origin + White: 1, Black/African American + Asian + White: 1, and preferred not to specify: 3. Prior coding experience included Scratch : 9, Scratch + BotBall (robotics): 4, and Scratch + Code.org : 2. This pilot group was given the same demo as the second classroom in the final study.

In order to evaluate the effectiveness of the design decisions outlined in Section 6.1.1, a comparison was conducted between the final pilot study with a sample of 15 students and the final classroom study with a sample of 21 students. Specifically, the students' scores for ***Perceived Robot Helpfulness*** were measured by conducting a two-sided Mann-Whitney U test (McKnight and Najab 2010). Additionally, behavioral data were collected and analyzed during the study. These data were collected at a rate of 50Hz, or every 0.02 seconds, and included the number of exercises reached by both the final groups (with 10 groups total) and pilot groups (with 7 groups total). Furthermore, the amount of time spent looking at the tutor robot (Kuri) or the tutor dialogue box within the final groups were compared, as shown in Figure 6.1.D. To analyze ***Interest in Programming*** and ***Future Intention to Program*** the post-interaction survey data were compared to the pre-interaction survey data via a two-sided Wilcoxon signed-rank test (Woolson 2007). Cliff's delta (δ) is reported for effect size (Macbeth et al. 2011).

Student written open-ended questions were qualitatily coded by first reading through all questions, creating categories, and then categorizing each question. Question categories included: 1) ***Robots*** – curious about the robot; 2) ***Programming*** – curious about programming; 3) ***Research*** – questions pertaining to the researchers; and 4) ***Repetition*** – asking about being able to do the activity again. A subset of questions qualified and counted for multiple categories (e.g., "How do you code robots?").

6.1.3 Results and Analysis

Perceived Robot Helpfulness

Robot helpfulness scores are compared between the final pilot ($n = 15$) to the final classroom studies ($n = 21$) with individual scores plotted in Fig. 6.8. Two-sided Mann-Whitney tests indicated a significant increase in ***Perceived Robot Helpfulness*** between the final classroom ($Mdn = 4.25$) and pilot ($Mdn = 3.0$) conditions ($U = 283.0, p < .001, \delta = .797$). This supports **H1.A** indicating design changes from the final pilot to the final classroom deployments described in Sec. 6.1.1.

Behavioral Data

A plot the number of exercises reached by each pair from the final pilot group ($Mdn = 5$) to the final study ($Mdn = 6$) groups can be seen in Fig. 6.9. The pilot group was not of sufficient sample size ($n = 7$ pairs) therefor no statistical tests were performed. The data may support **H1.B** but is not sufficient to be used as confirmatory evidence.

Also observed was the total amount of time spent looking at the tutor Kuri robot or the tutor dialogue in the visible robot and not visible robot conditions (Fig. 6.10). This is measured by casting a ray each time step from the center of the iPad with the first colliding object recorded. In the invisible robot tutor condition the collision boxes for the tutor Kuri remain active with only the meshes turned off. Ray collisions with these collisions boxes are still counted as to not favor the visible robot condition by merely having a larger target area. Seven groups looked at the robot or dialogue box when the robot mesh was visible and two groups (one from each class) looked more

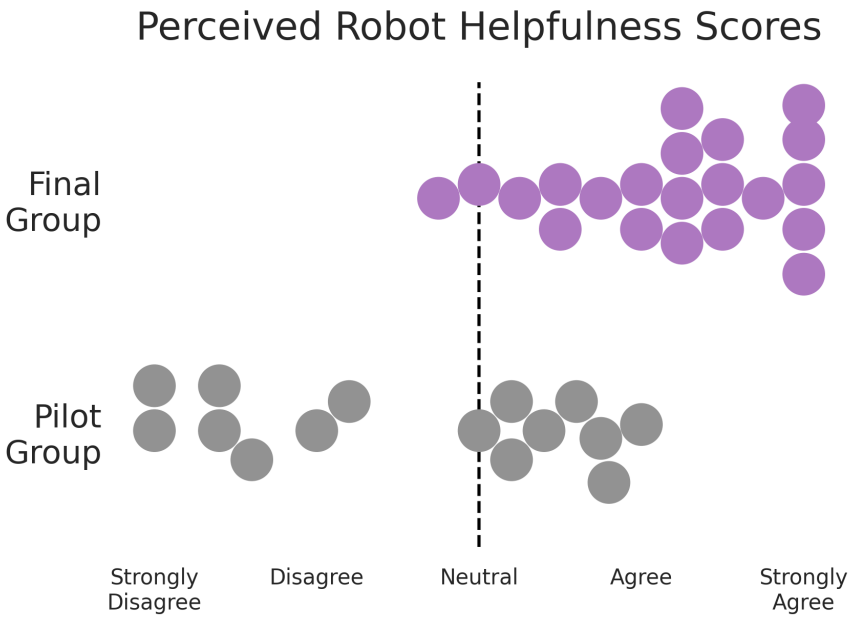


Figure 6.8: Perceived robot helpfulness of the final elementary pilot plotted with the final classroom studies. The scores are an average of 4 perceived robot helpfulness items described in Sec. 6.1.2.

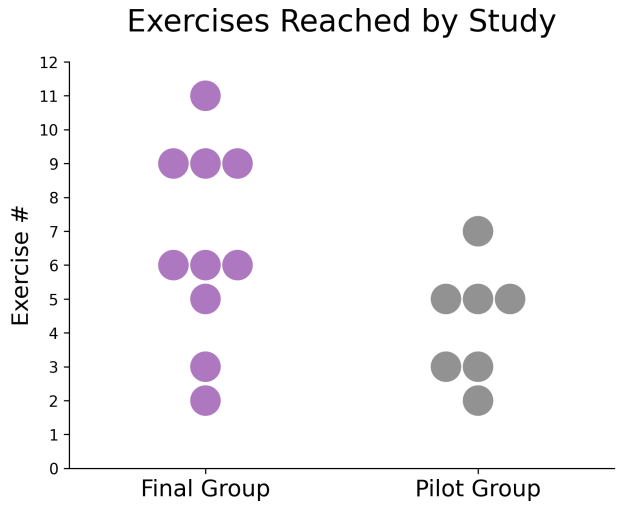


Figure 6.9: Number of exercises reached by each study group at the end of the exercises.

when the robot mesh was invisible. One group recorded less than 1 second of looking at tutor Kuri or the dialogue box. Again, there is not sufficient sample size to conduct statistical tests on these data. Thus the data may support **H2** but is not sufficient to be used as confirmatory evidence.

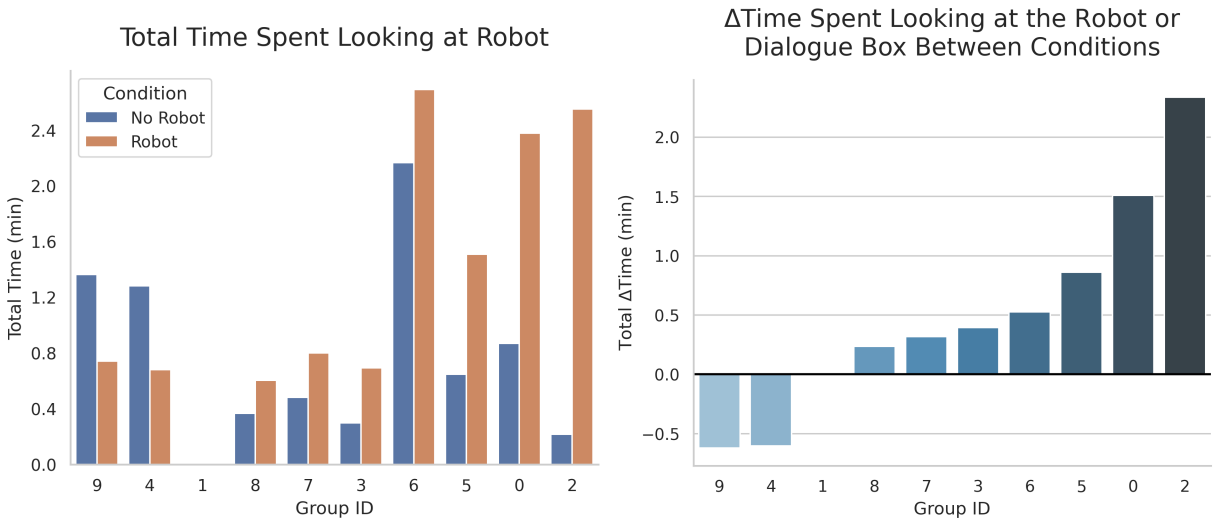


Figure 6.10: **Left**:Time spent looking at the robot or dialogue box between conditions sorted by difference in time of the Robot vs. No Robot conditions. **Right**: The sorted difference in time spent looking at the robot or dialogue box in the robot condition and the robot or dialogue box when the robot was not visible.

Interest and Future Intention to Program

Interest in programming and future intention to program were compared between pre- and post-interaction survey responses with no significant effect. Two-sided Wilcoxon signed-rank tests indicated no significant increases in ***Interest in Programming*** between post-interaction ($Mdn = 4.31$) and pre-interaction ($Mdn = 4.07$) surveys ($z = 128, p = .390$). A Wilcoxon signed-rank test indicated no significant increases in ***Future Intention to Program*** between post-interaction ($Mdn = 3.67$) and pre-interaction ($Mdn = 3.33$) surveys ($z = 49.5, p = .849$). Thus neither **H3.A** nor **H3.B** are supported.

Pre-post Student Questions

Question categories and counts are shown in Table 6.1. Nine out of the 21 students from both schools generated questions during the pre-test phase and the total number of written questions from all students was 22. During the post-test phase, both the number of students who wrote questions and the total number of questions increased, 15 students generated 36 questions.

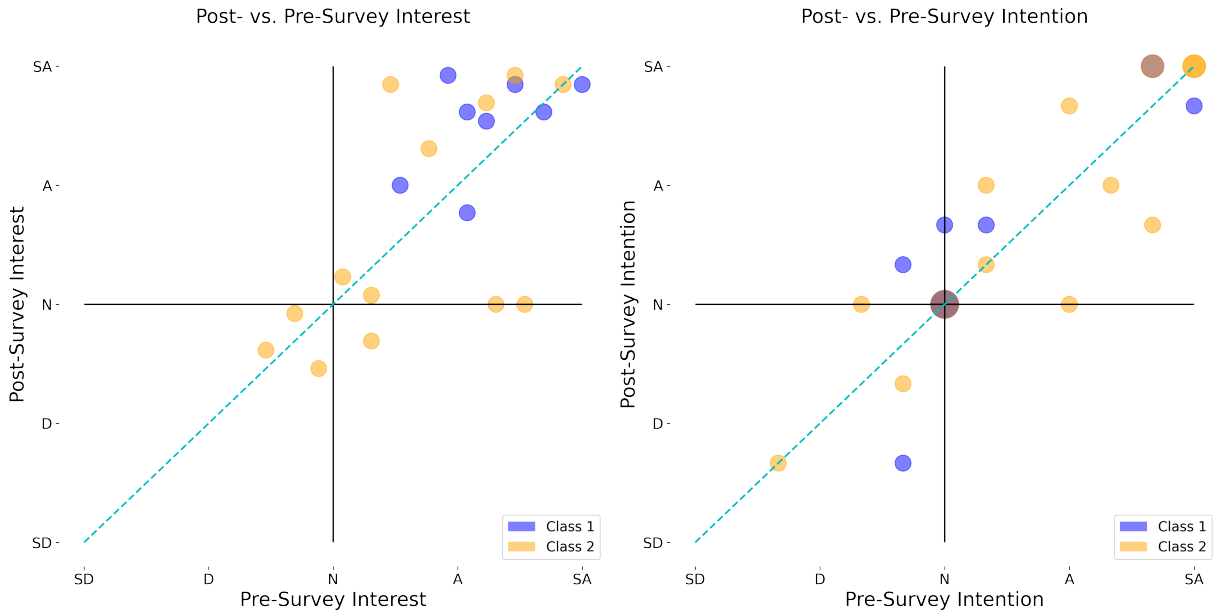


Figure 6.11: **Left**: Curiosity in programming. **Right**: Intention to pursue programming further. Axes of each graph are from "Strongly Disagree" to "Strongly Agree". Scores above the diagonal line indicate higher post scores when compared to pre. Dot size is relative to the number of score occurrences.

Category	**Pre Total**	**Post Total**	**% of Pre**	**% of Post**
Robot	9	12	40.9%	33.3%
Programming	6	20	27.3%	55.6%
Research	9	9	40.9%	25.0%
Repetition	1	11	4.5%	30.6%

Table 6.1: Students' question generation per category for the pre-interaction (22 total) and post-interaction (36 total) question writing sessions. The percentages are calculated relative to the total questions asked within that session (e.g., $\frac{9}{22} = 40.9\%$).

Example questions asked include the following:

- "How does the robot works?"
- "How did the robot move?"
- "Does the robot have emotions?"
- "How do you code robots?"
- "How old are you, do you code for a job?"

- "Do you like this career?"
- "Can you make more blocks and free play and make a block for you can code something for?"
- "Can we do more coding?"
- "When did you start coding?"
- "Can we expect more of this in the future?"
- "Will you have a different program if we see you again?"

6.2 PoseToCode: Design Considerations for a Pose-Based AR Input System

Contributors: Section 6.2 is based on Chatwani et al. (2022a) written with co-first authors Nisha Chatwani and Chloe Kuo. Maja J. Matarić is also an author of the published work.

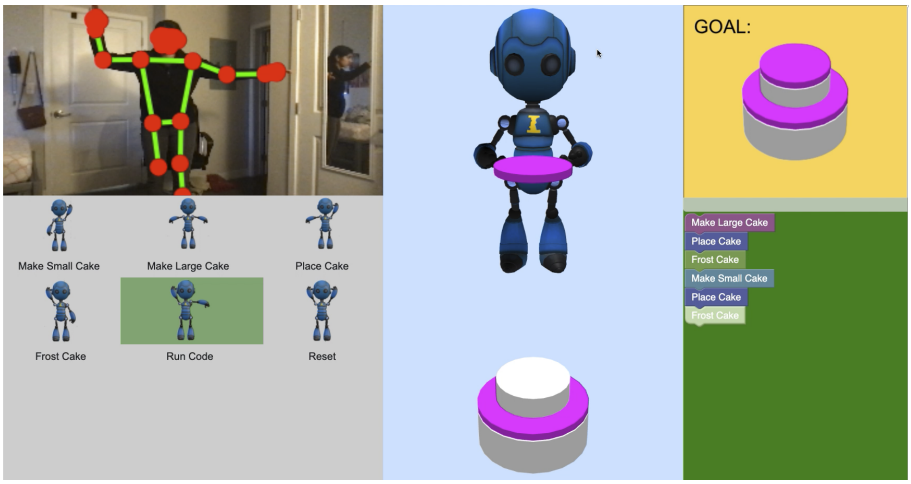


Figure 6.12: P2C Exercise 3: Build a Cake. The user (top left) must physically perform poses (bottom left) to create a sequence of codeblocks (bottom right), which, when executed, instruct the virtual robot (middle) to construct a cake (top right). The robot is displayed near the top of the screen because the following exercise involves programming objects underneath it.

While M2C allows for tablet-based interactions, it is not accessible for those without mobile devices. To address this gap, P2C was created which combines embodied learning with block-based programming to create a usable system with the end goal of increasing student curiosity and understanding of programming. P2C (Fig. 6.12) is an embodied learning block-based coding activity where students perform poses to create code blocks that guide a virtual robot through a series of exercises. A key feature of P2C was to efficiently run on low-end computers that are more accessible to schools.

Design considerations were created for P2C discovered in informal pilot testing. To validate P2C design, it was deployed in a local 5th grade classroom of 24 students. Usability was measured and compared to a traditional block-based programming activity, and analyzed the study results to develop future design considerations for making P2C more intuitive and easier to use. The results of the study support P2C as a usable design and identify improvements for future coding languages that integrate embodied learning and block-based programming. P2C is open-source and has a publicly accessible repository (Chatwani et al. 2022b) and demonstration (Chatwani et al. 2022c).

6.2.1 Technical Approach

P2C enables users to create and execute code by posing with their body. The P2C interface (Fig. 6.13) has the user's **video feed (A)** in the top left corner where the Mediapipe pose detection library (Lugaresi et al. 2019) draws lines showing landmarks on the user's body. Directly below the video feed, a grid shows a set of **progress bars (B)**, with images depicting the virtual robot performing a pose; each pose image is labelled to indicate what code block it will create. The **virtual robot (C)** is in the middle of the screen, and in the top right corner, there is an image of the **goal state (D)** indicating what the code blocks should produce to complete the exercise. Lastly, the Blockly workspace with the **code blocks (E)** created so far is on the right of the visual interface.

In P2C, the user's poses are recognized with Google Mediapipe pose detection software (Lugaresi et al. 2019) executing at 60Hz. At time t, the pose key points are run through a deep neural

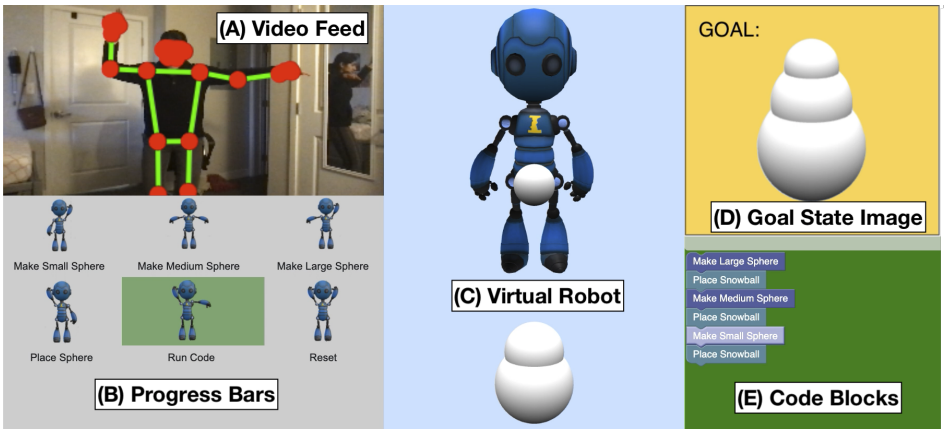


Figure 6.13: P2C Exercise (Level 2): Student (top left) posing to create code blocks (bottom right) that guide the virtual robot to build a snowman. A) Flipped video feed with the MediaPipe detected pose drawn. B) Grid of pose images and progress bars for each pose. C) Virtual robot that performs instructions from code blocks. D) Goal state image. E) Blockly workspace with crated code blocks.

network to map each arm to {HIGH, MED, LOW, NONE}. Based on the arm mappings, the corresponding pose progress bar (e.g., {Left: HIGH, Right: MED} $\rightarrow$ "Run Code") fills at a rate of $1.2 * \Delta(t, t-1)$ while all other progress bars decay at the rate of $0.8 * \Delta(t, t-1)$. When the user holds a pose for 4 seconds, a custom Blockly (Pasternak et al. 2017) code block that corresponds to that pose is created. Code blocks are instructions that control a virtual robot on the screen; they can be created, erased, and executed. If the user's executing code reaches the goal state, they are moved to the next exercise. Alternatively, if the code fails to reach the goal state, the user must continue attempting the same challenge until they succeed or their time spent on the activity reaches the 10-minute limit. The time limit was chosen based on pilot testing the activity and ensuring that it fit into the available classroom time.

A full P2C activity is composed of a series of three challenges for the user to complete within 10 minutes. The process consists of creating code blocks by posing and then executing all of the created code blocks. To complete the first challenge (Fig. 6.14), the user must instruct the virtual robot to perform a dance routine of four or more dance moves. To complete the second challenge (Fig. 6.13), the user must instruct the robot to build a snowman. To complete the third challenge (Fig. 6.12), the user must instruct the robot to construct a frosted three-tiered cake. After all three

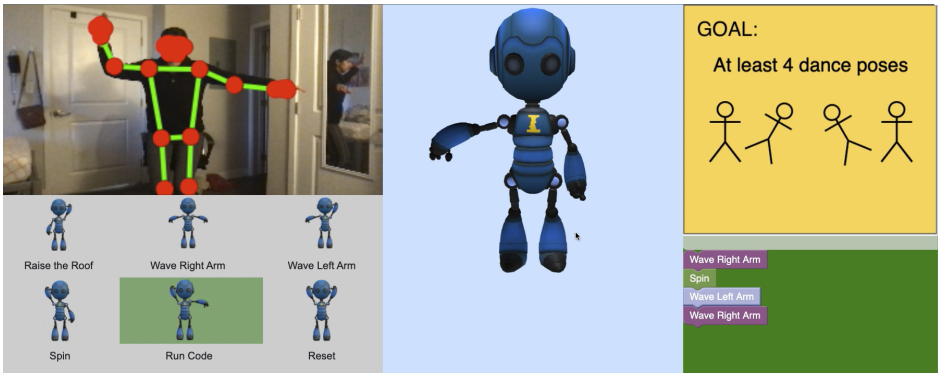


Figure 6.14: P2C Exercise 1: Choreographing a dance routine for the virtual robot.

challenges are completed, the user moves on to a freeplay mode challenge until the time limit is reached.

Pilot Study Insights About Design Considerations

An informal pilot test of P2C was conducted with two engineering students using the snowman building exercise. The following insights were gained about P2C design considerations:

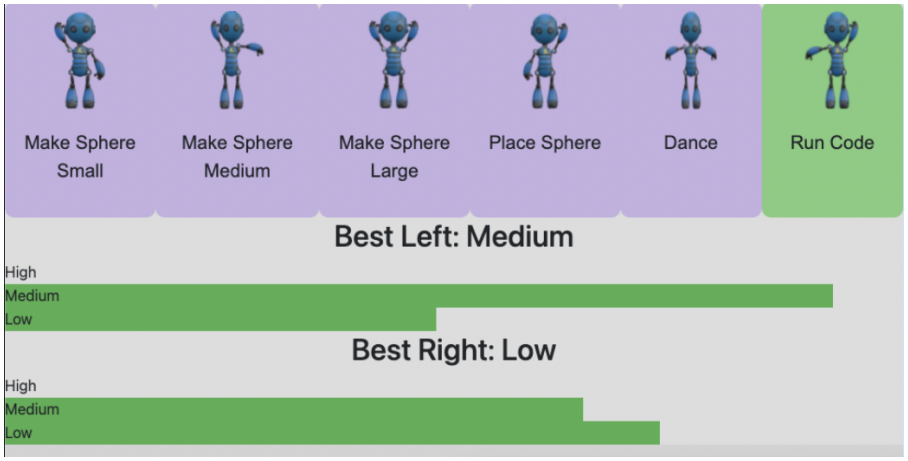
- *Accessibility across low-end computers*: Schools use a variety of laptops with a wide range of processing power and operating systems. Therefore, P2C was created to be operating-system agnostic, executing via the web at 60Hz on the currently most popular Chromebook with a webcam.

- *Accessibility at a distance*: Multiple interface elements are needed to accommodate students who stood away from the computer while completing the programming exercises.

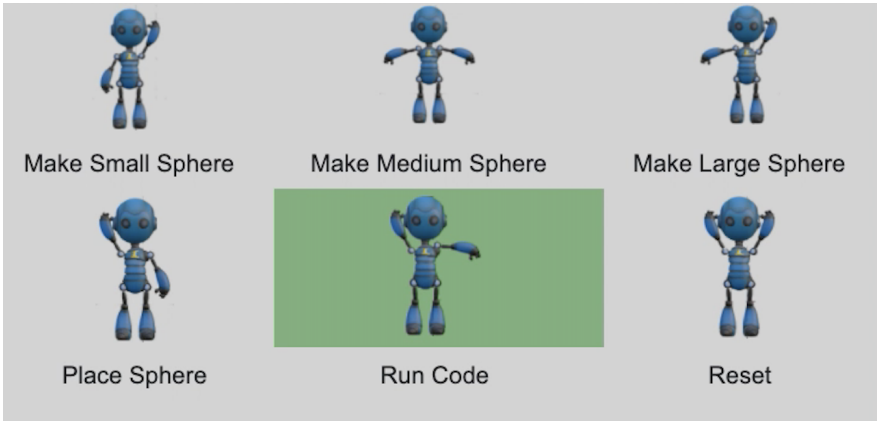
  - *Visibility*: Code blocks must have clear visibility, requiring them to be large, and therefore also enforcing having fewer on-screen components.

  - *Webcam as the only input source*: different interfaces were piloted (e.g., space bar pressing) but participants did not wish to step to and away from the computer, preferring to only use their body pose as input.

- *Real time input and feedback*: The webcam continually collects input from the user at 60Hz, resulting in the following considerations:



(a) Original pose key with individual bars for each arm.



(b) Updated pose key where pose bars fill up directly.

Figure 6.15: Original (a) and updated (b) pose key designs. The original design showed each individual arm state and a pose map. Participants found this difficult to map the arm states to each pose. The updated design directly filled up each respective pose.

– *Direct pose key* The original pose key (Fig. 6.15a) consisted of left and right arm states ({HIGH,MED,LOW,NONE}) with only the accumulated states shown to the user, not the state of the best current pose. When the progress bar corresponding to the state from each arm reached 100% completion, the mapped pose produced the corresponding code block. Many pilot users cited this as complicated as they needed to read the arm states and then the respective texted-based pose map. To address this, an updated direct pose key was developed (Fig. 6.15b) that combines the state of the arms into one pose state that corresponds directly to a progress bar.

– *Reactive and persistent pose bars*: To increase reactivity, the best pose progress bar at time t increases at the rate of $1.2 * \Delta(t, t-1)\%$ while all other bars decay at the rate of $0.8 * \Delta(t, t-1)\%$. The slower decay rate (0.8) compared to the growth rate (1.2) allows for a semi-persistent pose meter, because the progress bars grow faster than they shrink. For example, when a user is doing run code pose, if their left arm goes out of frame, the run code meter decays more slowly than it grows. Thus when the user brings their arm back into the frame, the progress of the run code pose is easily recovered.

– *Both arms down not used as input*: Users need to take time to think about their input. The most common pose while thinking was leaving both arms at their sides, leading to a {L=LOW, R=LOW} classification.

6.2.2 User Study

The insights from the pilot study were used to design the following full user study.

Procedure

A single-session within-subject study was conducted virtually over Zoom with a local 5th grade class in the context of a K-12 STEM outreach event. The study was approved by the University's Institutional Review Board (IRB #UP-20-01171).

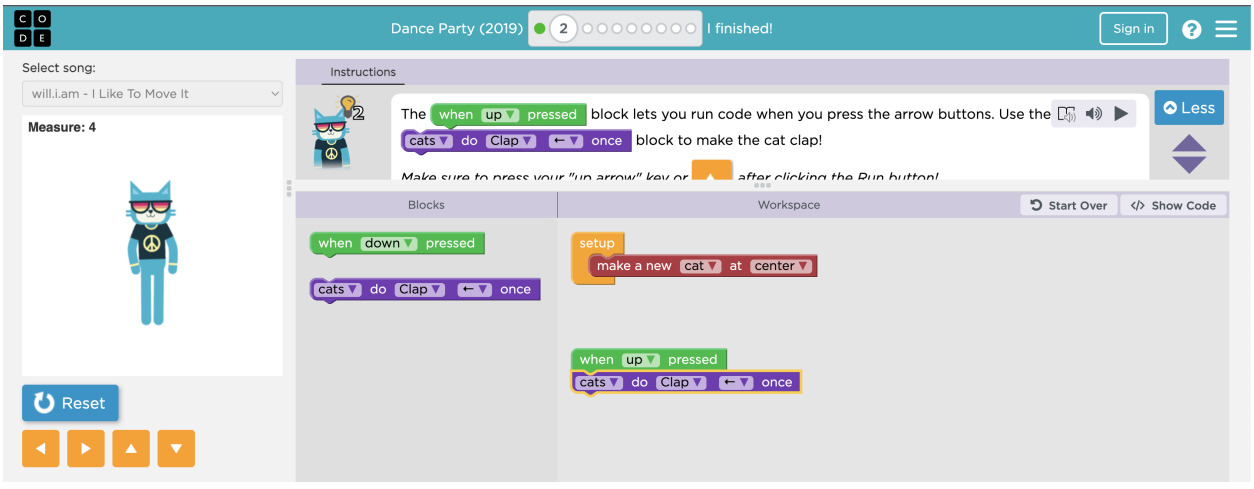


Figure 6.16: Code.org: Dance Party 2019 block programming activity that aims to teach basic coding concepts by guiding users to code a dance routine for a virtual character (Kalelioğlu 2015).

The student participants were randomly assigned to one of two conditions: 1) P2C programming activity first; and 2) Code.org (Kalelioğlu 2015) Dance Party 2019 activity (Fig. 6.16) first, a drag-and-drop block programming exercise giving students instructions to create a dance routine for a virtual character. Code.org's Dance Party 2019 Activity was chosen because, similarly to P2C, it teaches sequential code logic, and code blocks are used as instructions for a virtual character. Each student participant used a Chromebook to complete the activities.

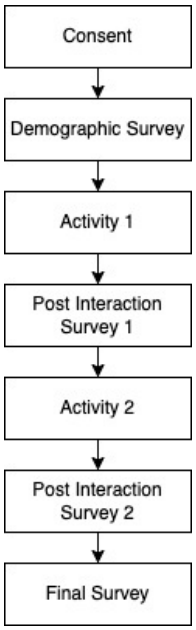


Figure 6.17: Diagram of the study procedure: pre-study survey, two coding activities, post-activity surveys after each activity, and an activity preference survey.

Student participants first completed a demographic survey. They then engaged in their first coding activity for up to 10 minutes, followed by a post-activity survey. The student participants then moved to their second activity, followed by a second post-activity survey. Finally, the student participants were given a final survey comparing the two activities. The student participants were allowed to end an activity early at any point, in which case they were automatically directed to the next step in the study procedure. A diagram of the study procedure is shown in Fig. 6.17.

This study was conducted virtually, given the COVID-19 pandemic. The virtual format of the study constrained the ability to provide clarifications and assistance to the student participants to only the Zoom chat function, where the students asked questions via this function. A single

teacher in the classroom had to physically move to each individual student participant to answer their questions. Conducting the study in person has the potential to make student participants' experience more enjoyable.

Participants

A local Los Angeles 5th grade class of 24 students (7 male, 16 female) was recruited. All student participants were volunteers and were given no form of compensation. Prior to the study, formal consent was obtained from each students' legal guardian and child assent was obtained from each student participant. A total of 10 participants (5 male, 5 female) completed all surveys and both programming activities. This paper reports on the data and analyses from those 10 participants. It is recognized this could lead to survivorship bias, but this was chosen to reduce possible ordering effects and to use all pairwise data.

In the pre-survey, student participants were asked what programming education platforms, if any, they had used in the past. All 10 indicated that they had previous technical experience with Scratch and two indicated that they also had past experience with Code.org. Additionally, student participants were asked to indicate their level of agreement with the statement "I want to learn more about computer programming." Of the 10 participants, 5 strongly agreed with the statement, and the other 5 agreed with the statement.

Data Collection

The pre-study surveys collected data on the students' prior exposure to programming. The post-activity surveys obtained system usability scores (SUS) (Bangor et al. 2009), aiming to assess the perceived activity difficulty for P2C and Code.org, and to capture the student participants' attitudes towards programming after each activity. The final post-study survey obtained each student's preferred activity between P2C and Code.org, as well as qualitative data via a write-in form on why they preferred one activity over the other.

P2C behavioral data were automatically collected in order to 1) find behavioral data correlations for usability; and 2) create a dataset to compare future iterations of P2C to. As the student participants performed the P2C activity, behavioral data were collected consisting of the time each

student took to complete the activity, the number of exercises each student successfully completed within the time limit, and the number of code blocks created throughout the activity.

6.2.3 Results and Analysis

This work evaluates the usability of P2C through participant surveys outlined in Section 6.2.2. Participant interviews were also analyzed for details about their experience.

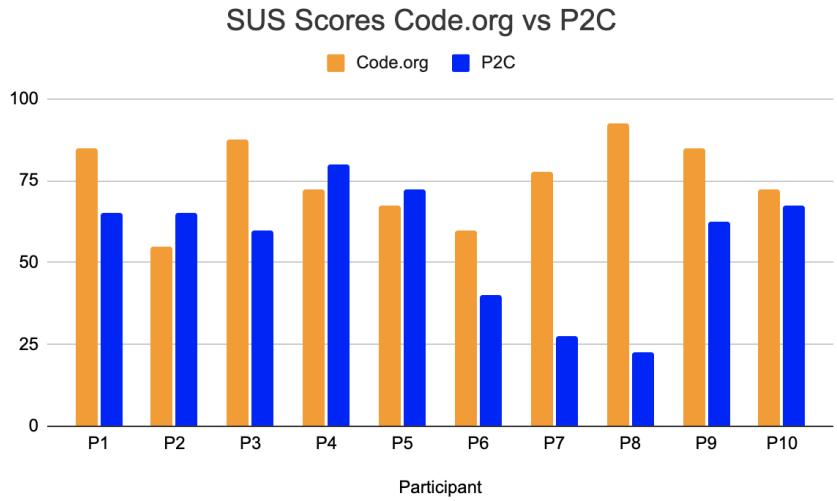


Figure 6.18: Bar graph comparing SUS scores for Code.org and P2C for all 10 student participants (Ok = 25-59, Good = 60-89, Excellent = 90-100 (Klug 2017)).

Quantitative Results

To explore design considerations about P2C, System Usability Scale (SUS) (Bangor et al. 2009) questionnaire was adapted for young students in order to compare usability compared to the Code.org activity. P2C yielded a median SUS score of 63.75, slightly below the SUS average of 68, and Code.org yielded a median of 75, above average. A SUS score between 68 and 89 indicates that the system has "good" or above average usability, and P2C's SUS score is within the 35-40th percentile (Klug 2017). The statistical power of the SUS scores generated are low since there are only 10 responses. Thus, this supports P2C as a usable system but leaves room for improvement. Fig. 6.18 shows the difference in SUS scores between Code.org and P2C. Mann-Whitney tests

indicated that P2C (*Mdn* = 4) is more difficult than Code.org (*Mdn* = 5) with conditions ($U = 20$, $p = .020$).

A Pearson's r correlation indicated no significant relationship between total time spent on the P2C activity and SUS score (rs(10) = -0.367). Similarly, Pearson's r correlation did not yield statistical significance between the number of code blocks created by a student participant during the P2C activity and SUS score (rs(10) = -0.252). Both post-activity surveys asked participants how much they agree with the statement "I want to learn more about computer programming." Mann-Whitney tests revealed an insignificant difference between the responses for P2C (*Mdn* = 4 (Agree)) and Code.org (*Mdn* = 5 (Strongly Agree)) conditions ($U = 41$, $p = .480$).

Qualitative Results

The qualitative results from the post-study survey showed that 5 of 10 student participants preferred P2C over Code.org. For the participants who started with P2C, 4 of 5 preferred P2C, and for the participants who started with Code.org, 4 out of 5 preferred Code.org. Two themes emerged from analyzing the free responses, as follows.

P2C is more active than than Code.org - Participants who preferred P2C found P2C to be a more active and engaging activity than Code.org. Three of five participants who preferred P2C over Code.org used the word *fun* to describe P2C in their reasoning for choosing their preference. Only one student out of five who preferred the Code.org activity over P2C described Code.org as *fun.* P3 indicated that they preferred P2C because they *get to be active and move around,* and similarly, another participant noted that with P2C, *you do more activity and more exercise* than Code.org.

Code.org is easier to use - Student participants who preferred Code.org found Code.org easier to use than P2C because it had fewer software bugs when compared to P2C. For example, P1 wrote about Code.org, *It's much more easy to use. (I like coding this way)* while P2 wrote, *in Code.org it's simple and fun and it's a good way to pass the time but P2C is super frustrating and got me annoyed because of glitched like when it highlights your screen green and then kicks you out.* Three participants wrote in the final survey that P2C was *frustrating* because it *glitched* often, making

Code.org more desirable. Additionally, two participants wrote that Code.org gave *more instruction* than P2C and provided more guidance on how to be successful in the activity.

6.3 Discussion and Summary

This chapter presented two applications that utilize iteratively learned and validated design conventions for the interface between the model and view, also known as the controller. The first application is an extension of Sec. 4.1 in M2C, a visual programming language that utilizes an AR robot tutor to increase young students' curiosity in coding. The second application, P2C, is a kinesthetic web application that aims to increase students' interest in coding and has been deployed in schools. This chapter discussed both existing and new design conventions and considerations for creating AR experiences with socially assistive robots. These applications demonstrate the potential for using AR in SAR tutors to promote students' curiosity and interest in coding.

The chapter first demonstrated the potential of using VAM in the field of SAR tutors through the development of M2C - an open-source, embodied learning visual programming language. The application utilizes an AR autonomous robot tutor named Kuri to model students' kinesthetic curiosity and promote their interest in programming. The design of M2C was informed by pilot studies and validated in local elementary classrooms, resulting in an improvement in perceived robot helpfulness and number of completed exercises. Although no significant changes were found in pre-post student interest or intention to program later in life, open-ended questions post-study indicate that students were more interested in the robot, programming, and research topics. This work highlights the potential of VAM-HRI in a kinesthetic context for SAR tutors and the design considerations for creating AR applications for SAR.

The chapter finally explored the design and usability of P2C, an embodied learning block-based programming language. Results showed that half of the 5th grade student participants preferred P2C over Code.org, citing it as more fun and engaging. However, some students reported technical issues and unclear instructions for P2C. The study highlights the need for better initial instruction

and technical improvements for P2C. Additionally, students helping one another during the study suggests potential for future development of collaborative programming activities. Limitations of the study include technical glitches and unclear instructions for P2C, as well as survey length and potential bias from student age and data collection methods. Future work should address these limitations through unit and integration tests, video tutorials, and individual student interviews.

Chapter 7

Summary and Conclusions

This chapter first summarizes the contributions of this dissertation on leveraging VAM for SAR. Trends in VAM-HRI are then discussed and it concludes by speculating on future directions for VAM for SAR within the MVC framework.

7.1 Contributions

The main contribution of this dissertation is to **define and demonstrate how VAM can be leveraged in SAR under the MVC paradigm**. The dissertation presents a framework for VAM-HRI in TOKCS which classifies research on 3D virtual imagery for HRI under the MVC paradigm. The framework focuses on VAM technology for improving the robot's internal model to better understand the user's state, increasing the expressivity of view of the robot's external expression, and enhancing the flexibility of controller for kinesthetic SAR.

The dissertation also provides several key contributions to the field of SAR. The first demonstrates synthesizing reliable multimodal AR data to support student kinesthetic curiosity and AR usability metrics. Additionally, the dissertation provides a contribution in expanding SAR expressivity with VAM by creating designs for AR visualizations for both social and functional robot expressions. The dissertation also provides design recommendations for maximizing functional

and social expressivity of AR robot gestures with different contextual factors. Finally, the dissertation presents research on kinesthetic interaction paradigms for increasing human-robot flexibility of controller that leverages AR to create a wider range of interactions between the robot and users.

7.2 Current Trends & the Future of VAM-HRI

In this dissertation, the 4^{th} VAM-HRI Workshop was used as a case study for MRIDE classification and categorization within the Reality Virtuality Interaction Cube (Sec. 3.3); however, the papers submitted to that workshop can also be used to exemplify and project current and future trends in the field of VAM-HRI. This growing sub-field of HRI is showing promise in enhancing many areas of HRI, from robot control (e.g., teleoperation and supervision interfaces) to collaborative robotics and improving teamwork with autonomous systems. This chapter will discuss some of the key insights gathered from this year's workshop that show how VAM-HRI is evolving and improving the field of HRI as whole.

7.2.1 Experimental Evaluation of VAM-HRI Systems

Research in HRI features user studies in the evaluation of robotic systems and their interfaces. It has been an ongoing challenge to adequately record and play back human interactions with a robot, and to answer questions such as: "Where was the user looking at X time?," "How close was the human positioned relative to the robot at Y moment?," "What were the user's joint values when using a new interface and how are the physical ergonomics evaluated?" VAM-HRI allows for recording, playback, and analysis of user interactions with virtual or real robots and objects in an experimental setting due to the ability of HMDs to record body/hand/head position/orientation and gaze direction from a seemingly limitless number of virtual cameras recording from different angles (Williams et al. 2020c). This is effectively exemplified in CoBot Studio (Mara et al. 2021) (see Fig. 7.1).

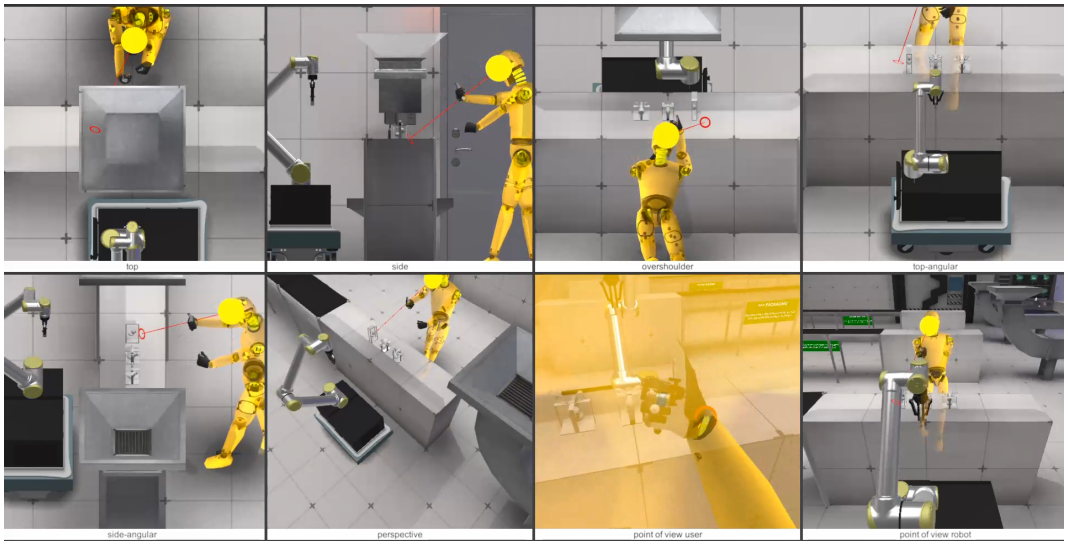


Figure 7.1: Advances in VAM-HRI research have enhanced the ability to precisely record, play back, and analyze human interactions with robots and other experimental stimuli in controlled user studies. This is exemplified in Mara et al. (Mara et al. 2021) CoBot Studio project where HRI user studies were conducted in a VR environment with numerous virtual cameras monitoring the experimental area from a multitude of angles. The cameras made use of the VR hardware to track body and head motion to record human postures and posture shifts, task-related human movements, gestures, and gaze behaviors, etc. Such techniques can benefit the field of HRI as a whole and allow for more complete and feature-rich data of human behavior.

Although VR interfaces have the aforementioned strengths for enhancing experimental evaluation, they have evaluation challenges as well. One is the use of online studies with crowdworkers (e.g., on Amazon Mechanical Turk). HRI in general has made prolific use of online user studies (especially during the COVID-19 pandemic) that take advantage of affordable and readily available participants. VAM-HRI draws upon 3D visualizations (as often seen in with HMD-based interfaces), which cannot be properly displayed to crowdworkers who lack HMDs and/or 3D monitors. Additionally, a strength of AR interfaces is that 3D data and visualizations can be rendered contextually in user's environments and are able to be observed from any angle desired by the user. VAM-HRI studies that utilize crowdworkers to evaluate VAM interfaces (e.g., Mott et al. (2021)) are restricted to online images and videos viewed by Mechanical Turkers on 2D monitors that restrict their viewpoint to that of pre-recorded videos that do not allow for a true VAM experience. It remains an open question if results from crowdsourced VAM-HRI studies provide comparable

results to in-person VAM-HRI studies since 3D VAM technology is inherently experienced differently than the 2D experiences found on crowdsourcing platforms. Regardless, crowdworkers still holds value in the early prototyping phases of VAM-HRI research when the initial formulation of object and interaction designs can be evaluated quickly and inexpensively.

7.2.2 VAM-HRI as an Interdisciplinary Field

HRI is an interdisciplinary field and VAM-HRI is as well. For example, the CoBot Studio project brought together roboticists, psychologists, AI experts, multi-modal communication researchers, VR developers, and professionals in interaction design and game design (Mara et al. 2021). As the VAM-HRI field grows, it will likely become increasingly common (and needed) to engage teams made up of members with a variety of experiences and skill sets all contributing to collaborative research.

Research in multi-robot systems is an underexplored domain of VAM-HRI research, in particular with regard to enhancing the complexity of model (CM). VAM technology can be formulated as another robot within a system, a robot with non-deterministic, non-directly controllable behavior but one with a data rich sensor suite. The frameworks and techniques of the adjacent field of multi-robot systems may be able to be modified or even directly applied when treating the user as an autonomous mobile sensor platform, akin to the user being treated as though they are another robot in the system. For example, spatial and semantic scene understanding are important perceptual capabilities for active robots (to navigate their environment) and passive VAM technologies (to localize the user's field of view).

Additionally, experimentation techniques seen in the field of general VR may aid in the administering of questionnaires and gathering participant feedback. Typical questionnaires administered by VAM-HRI researchers can be quite jarring for participants who experience extreme context shifts between virtual worlds (where a study took place) and the real world (where the feedback is gathered). This poses as a potential confounding factor for participants who no longer visually reference what they are evaluating and may incorrectly remember experimental stimuli they can

no longer see. The field of VR has similar challenges; some studies have started to provide in situ evaluations where questionnaires are posed to users within the virtual environments (Lin et al. 2019). In situ surveys are starting to be used in VAM-HRI as well. In the CoBot studio project, surveys were administered within the experiment's virtual setting, removing the confounding factors of: (1) reality-virtuality context shifts (having to leave the immersive virtual environment by taking off an HMD to take a mid-task survey); and (2) retrospective surveys provided well after exposure to experimental stimulus (Mara et al. 2021).

The cross-disciplinary trends and ideas from VR are not unidirectional; VAM-HRI is currently poised to inform and improve the field of VR in return. Enhancing immersion has been a primary goal of VR since its inception many decades ago. With the rise of mass-produced consumer HMDs, visual immersion has reached new heights. However, the challenge of providing *physical* immersion through the use of haptics has largely remained an open question: how can a user reach out and touch a dynamic character in a virtual world? Research in VAM-HRI has proposed a potential solution for dynamic haptics, where robots mimic the pose and movements of virtual dynamic objects. Work by Wadgaonkar et al. (2021) exemplifies the notion of VAM-HRI supporting the field of VR with robots acting as dynamic haptic devices and allowing users to touch characters in virtual worlds and further enhance immersion in VR settings.

7.2.3 Advancements in VAM-HRI

A strength of VAM-HRI is the ability to alter a robot's morphology with virtual imagery. This technique can take the form of body extensions where virtual appendages are added to a real robot, such as limbs (Groechel et al. 2019), or form transformations where the robot's entire morphology is altered, such as transforming a drone into a floating eye (Walker et al. 2018). Recent VAM-HRI developments have further expanded on the idea of changing a real robot's appearance through morphological alterations to include superficial alterations as well, where virtual imagery can be used to change a robot's cosmetic traits. Prior work has demonstrated that robot cosmetic alterations can communicate robot internal states (e.g., robotic system faults) (De Pace et al. 2018).

Although the interactions studied in HRI are typically focused on the end-user, a lesser studied category of interaction is that between robots and their developers and designers. Debugging robots is often challenging and tedious; robot faults and unexpected behavior are hard to understand and explain without parsing through command lines and error logs. To address this issue, prior work in VAM-HRI has used AR interfaces to enhance debugging capabilities (Collett and Macdonald 2010; Millard et al. 2018). Work by Ikeda and Szafir (2021) in VAM-HRI 2021 built upon those concepts by providing in situ AR visualizations of robot state and intentions, allowing users to better compare robot plans and actions during debugging. As AR hardware becomes increasingly intertwined with robot systems, debugging tools such as these will likely become more common, increasing efficiency and enjoyment of robot design.

Finally, VAM-HRI interfaces have been a popular topic of study within HRI for many years, and many standard methods of interacting with robots through MR and VR have emerged (e.g., AR waypoints for navigation and AR lines for displaying robot trajectory (Walker et al. 2018)). However, novel methods of interacting with robots are still being designed, such as persistent virtual shadows, aimed at tackling the issue of knowing a robot's location when out of the user's field of view. While prior solutions have tried using 2D top-down radars for showing robot locations (Walker et al. 2018), issues remained because the interfaces required repeated context shifts by the user to look at the physical surroundings and then to the radar. Solutions such as persistent virtual shadows circumvent this limitation by embedding robot location data into the user's environment, providing a natural method of displaying a robot's location. Creative advances will continue to emerge in the relatively nascent field of VAM-HRI, presenting an exciting new future for both VAM-HRI and HRI as a whole.

7.3 Future Direction of VAM for SAR

There is a significant opportunity for further development and exploration of VAM technology for SAR across all three MVC areas. This includes improving user modeling techniques and adapting

them to the kinesthetic domain, designing intuitive and effective interfaces for user interaction with the VAM environment, and exploring new approaches to controller design such as multi-embodied transferable agents and multiparty physical interactions.

Modelling user state in SAR interactions through VAM may be a straightforward process, as it utilizes existing methods of user modelling and replaces traditional sensing methods (e.g., external cameras) with VAM. However, adapting VAM to the kinesthetic domain, which involves movements and interactions with the environment, presents additional challenges. Previous research has shown that there are difficulties in using VAM for user modelling, such as determining intent through eye gaze (Rosen et al. 2020). For example, it can be challenging to differentiate between someone using their eyes to express intent versus using them to explore an area. To improve VAM for SAR modelling, it is crucial to identify and address these key changes in the domain.

AR anthropomorphic gestures are also being studied for robot expression, as was done in this dissertation. However, there is a lack of research on social cues through non-anthropomorphic and mixed gestures. While there is a significant amount of work in both academia and industry on creating and interacting with virtual characters, these characters are not limited to human forms and can still be anthropomorphized. A potential benefit of using non-anthropomorphic appendages and gestures is that they do not generate the same existing human expectations. For example, the arms described in Sec. 5.3 do not have fingers and are partially non-anthropomorphic, reducing the expectation for the user. This can be a key challenge when designing an agent, as an expectation mismatch can lead to poor interactions. For example, if a robot is able to talk but does not have a good dialogue system, this can lead to worse interactions than if the robot does not talk. In addition, there is also potential for users to customize their robots, as in the video game industry where users invest in digital artifacts to show off to others. Similar to how humans often wear clothes for form as well as function, customization allows users to personalize their robots and could lead to work in community building via robot customization (Fitter et al. 2020).

The interaction space for controllers is highly variable, and one possibility for exploration is the use of multi-embodied transferable agents. This refers to the ability to have a physical robot

that learns about an user during an interaction, such as in a classroom, and then the individual can "take the robot with them" through an AR application or on-screen analog. There are many potential benefits to this approach, including shared data stores and sensing capabilities. Additionally, leveraging the physical embodiment of a robot for multiparty interactions can be a valuable area of research. While screens tend to isolate users, physical robots can allow multiple people to interact in the physical world. AR expands the interaction space and allows individuals to share that space.

The potential for VAM in SAR is vast, and continued research and development in this area may greatly improve the user experience and enable new forms of interaction.

7.4 Final Words

This dissertation is a starting point for understanding how to use VAM for SAR. The hope is that this work will serve as a foundation for future research in VAM for SAR, toward pursing the many open research challenges in this area.

Bibliography

Wathieu, Adam, Thomas R Groechel, Haemin Jenny Lee, Chloe Kuo, and Maja J Matarić (2022). "RE: BT-Espresso: Improving Interpretability and Expressivity of Behavior Trees Learned from Robot Demonstrations". In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 11518–11524.

Cordero, Julia R, Thomas R Groechel, and Maja J Matarić (2022). "A Review and Recommendations on Reporting Recruitment and Compensation Information in HRI Research Papers". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 1627–1633.

Feil-Seifer, David and Maja J Matarić (2005). "Defining socially assistive robotics". In: *9th International Conference on Rehabilitation Robotics, 2005. ICORR 2005.* IEEE, pp. 465–468.

Matarić, Maja J and Brian Scassellati (2016). "Socially assistive robotics". In: *Springer handbook of robotics*, pp. 1973–1994.

Rabbitt, Sarah M, Alan E Kazdin, and Brian Scassellati (2015). "Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use". In: *Clinical psychology review* 35, pp. 35–46.

Abdi, Jordan, Ahmed Al-Hindawi, Tiffany Ng, and Marcela P Vizcaychipi (2018). "Scoping review on the use of socially assistive robot technology in elderly care". In: *BMJ open* 8.2, e018815.

Clabaugh, Caitlyn, Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Leena Mathur, Kartik Mahajan, Gisele Ragusa, and Maja Matarić (2020). "Month-long, in-home socially assistive robot for children with diverse needs". In: *International Symposium on Experimental Robotics*. Springer, pp. 608–618.

Shi, Zhonghao, Thomas R Groechel, Shomik Jain, Kourtney Chima, Ognjen Rudovic, and Maja J Matarić (2022). "Toward Personalized Affect-Aware Socially Assistive Robot Tutors for Long-Term Interventions with Children with Autism". In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.4, pp. 1–28.

Clabaugh, Caitlyn, Kartik Mahajan, Shomik Jain, Roxanna Pakkar, David Becerra, Zhonghao Shi, Eric Deng, Rhianna Lee, Gisele Ragusa, and Maja Matarić (2019). "Long-term personalization of an in-home socially assistive robot for children with autism spectrum disorders". In: *Frontiers in Robotics and AI* 6, p. 110.

Jain, Shomik, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J Matarić (2020a). "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders". In: *Science Robotics* 5.39, eaaz3791.

Pakkar, Roxanna, Caitlyn Clabaugh, Rhianna Lee, Eric Deng, and Maja J Mataricć (2019). "Designing a socially assistive robot for long-term in-home use for children with autism spectrum disorders". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 1–7.

Kim, Won, Frank Tendick, and LAWRENCEW Stark (1987). "Visual enhancements in pick-and-place tasks: Human operators controlling a simulated cylindrical manipulator". In: *IEEE Journal on Robotics and Automation* 3.5, pp. 418–425.

Walker, Michael, Thao Phung, Tathagata Chakraborti, Tom Williams, and Daniel Szafir (2022). "Virtual, Augmented, and Mixed Reality for Human-Robot Interaction: A Survey and Virtual Design Element Taxonomy". In: *arXiv preprint arXiv:2202.11249*.

Williams, Tom, Daniel Szafir, Tathagata Chakraborti, and Heni Ben Amor (2018a). "Virtual, augmented, and mixed reality for human-robot interaction". In: *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 403–404.

Williams, Tom, Daniel Szafir, Tathagata Chakraborti, and Elizabeth Phillips (2020a). "Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)". In: *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. HRI '19. Daegu, Republic of Korea: IEEE Press, pp. 671–672.

Williams, Tom, Daniel Szafir, Tathagata Chakraborti, Ong Soh Khim, Eric Rosen, Serena Booth, and Thomas Groechel (2020b). "Virtual, augmented, and mixed reality for human-robot interaction (vam-hri)". In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 663–664.

Rosen, Eric, Thomas Groechel, Michael E Walker, Christine T Chang, and Jessica Zosa Forde (2021). "Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)". In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 721–723.

Chang, Christine T, Eric Rosen, Thomas R Groechel, Michael Walker, and Jessica Zosa Forde (2022). "Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)". In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 1237–1240.

Nee, Andrew YC and Soh-Khim Ong (2013). "Virtual and augmented reality applications in manufacturing". In: *IFAC proceedings volumes* 46.9, pp. 15–26.

Bric, Justin D, Derek C Lumbard, Matthew J Frelich, and Jon C Gould (2016). "Current state of virtual reality simulation in robotic surgery training: a review". In: *Surgical endoscopy* 30.6, pp. 2169–2178.

Ravi, Kaushik Selva Dhanush, Jesús Medina Ibáñez, Daniel Mark Hall, et al. (2021). "Real-time Digital Twin of On-site Robotic Construction Processes in Mixed Reality". In: *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*. Vol. 38. IAARC Publications, pp. 451–458.

Ikeda, Bryce and Daniel Szafir (2022). "Advancing the design of visual debugging tools for roboticists". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 195–204.

Zea, Antonio and Uwe D Hanebeck (2021). "iviz: A ROS visualization app for mobile devices". In: *Software Impacts* 8, p. 100057.

Krasner, Glenn E and Stephen T Pope (1988). "A description of the model-view-controller user interface paradigm in the smalltalk-80 system". In: *Journal of object oriented programming* 1.3, pp. 26–49.

Fong, Terrence, Illah Nourbakhsh, and Kerstin Dautenhahn (2003). "A survey of socially interactive robots". In: *Robotics and autonomous systems* 42.3-4, pp. 143–166.

Goodrich, Michael A, Jacob W Crandall, and Emilia Barakova (2013). "Teleoperation and beyond for assistive humanoid robots". In: *Reviews of Human factors and ergonomics* 9.1, pp. 175–226.

Miller, David P (1998). "Assistive robotics: an overview". In: *Assistive technology and artificial intelligence*, pp. 126–136.

Clabaugh, Caitlyn and Maja Matarić (2019). "Escaping Oz: Autonomy in Socially Assistive Robotics". In: *Annual Review of Control, Robotics, and Autonomous Systems* 2, pp. 33–61.

Mohebbi, Abolfazl (2020). "Human-robot interaction in rehabilitation and assistance: a review". In: *Current Robotics Reports* 1.3, pp. 131–144.

Deng, Eric, Bilge Mutlu, Maja J Matarić, et al. (2019). "Embodiment in Socially Interactive Robots". In: *Foundations and Trends® in Robotics* 7.4, pp. 251–356.

Law, Matthew V, JiHyun Jeong, Amritansh Kwatra, Malte F Jung, and Guy Hoffman (2019). "Negotiating the creative space in human-robot collaborative design". In: *Proceedings of the 2019 on Designing Interactive Systems Conference*, pp. 645–657.

Andriella, Antonio, Alejandro Suárez-Hernández, Javier Segovia-Aguas, Carme Torras, and Guillem Alenya (2019). "Natural teaching of robot-assisted rearranging exercises for cognitive training". In: *International Conference on Social Robotics*. Springer, pp. 611–621.

Cutica, Ilaria, Francesco Ianı, and Monica Bucciarelli (2014). "Learning from text benefits from enactment". In: *Memory & Cognition* 42.7, pp. 1026–1037.

Wainer, Joshua, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric (2006). "The role of physical embodiment in human-robot interaction". In: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, pp. 117–122.

Varela, Francisco J., Evan T. Thompson, and Eleanor Rosch (1992). *The Embodied Mind: Cognitive Science and Human Experience*. New edition. The MIT Press.

Hendriks-Jansen, Horst (1996). *Catching Ourselves in the Act: Situated Activity, Interactive Emergence, Evolution, and Human Thought*. The MIT Press. DOI: 10.7551/mitpress/1748.001.0001.

Zeman, Adam (2006). "Wider than the sky The phenomenal gift of consciousness". In: *Journal of Clinical Investigation* 114. DOI: 10.1172/JCI23795.

Dennler, Nathaniel, Changxiao Ruan, Jessica Hadiwijoyo, Brenna Chen, Stefanos Nikolaidis, and Maja Matarić (2022). "Design Metaphors for Understanding User Expectations of Socially Interactive Robot Embodiments". In: *ACM Transactions on Human-Robot Interaction*.

Schodde, Thorsten, Kirsten Bergmann, and Stefan Kopp (2017). "Adaptive robot language tutoring based on Bayesian knowledge tracing and predictive decision-making". In: *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 128–136.

Rich, Charles, Brett Ponsler, Aaron Holroyd, and Candace L Sidner (2010). "Recognizing engagement in human-robot interaction". In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 375–382.

Salam, Hanan and Mohamed Chetouani (2015a). "A multi-level context-based modeling of engagement in human-robot interaction". In: *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. Vol. 3. IEEE, pp. 1–6.

Celiktutan, Oya, Efstratios Skordos, and Hatice Gunes (2017). "Multimodal human-human-robot interactions (mhhri) dataset for studying personality and engagement". In: *IEEE Transactions on Affective Computing* 10.4, pp. 484–497.

Jain, Shomik, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J. Matarić (2020b). "Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders". In: *Science Robotics* 5.39.

Ayub, Ali, Marcus Scheunemann, Christoforos Mavrogiannis, Jimin Rhim, Kerstin Dautenhahn, Chrystopher L Nehaniv, Verena V Hafner, and Daniel Polani (2022). "Robot Curiosity in Human-Robot Interaction (RCHRI)". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 1231–1234.

Kulic, Dana and Elizabeth A Croft (2007). "Affective state estimation for human–robot interaction". In: *IEEE transactions on robotics* 23.5, pp. 991–1000.

Kaelbling, Leslie Pack, Michael L Littman, and Andrew W Moore (1996). "Reinforcement learning: A survey". In: *Journal of artificial intelligence research* 4, pp. 237–285.

Gordon, Goren, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal (2016). "Affective personalization of a social robot tutor for children's second language skills". In: *Thirtieth AAAI Conference on Artificial Intelligence*.

Weiss, Karl, Taghi M Khoshgoftaar, and DingDing Wang (2016). "A survey of transfer learning". In: *Journal of Big data* 3.1, pp. 1–40.

Spaulding, Samuel and Jocelyn Shen (2021). "Towards Transferrable Personalized Student Models in Educational Games". In: *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS'21)*.

Daumé III, Hal (2009). "Frustratingly easy domain adaptation". In: *arXiv preprint arXiv:0907.1815*.

Ferraro, Francis, Nasrin Mostafazadeh, Lucy Vanderwende, Jacob Devlin, Michel Galley, Margaret Mitchell, et al. (2015). "A survey of current datasets for vision and language research". In: *arXiv preprint arXiv:1506.06833*.

Janai, Joel, Fatma Güney, Aseem Behl, Andreas Geiger, et al. (2020). "Computer vision for autonomous vehicles: Problems, datasets and state of the art". In: *Foundations and Trends® in Computer Graphics and Vision* 12.1–3, pp. 1–308.

Mogadala, Aditya, Marimuthu Kalimuthu, and Dietrich Klakow (2021). "Trends in integration of vision and language research: A survey of tasks, datasets, and methods". In: *Journal of Artificial Intelligence Research* 71, pp. 1183–1317.

Samarakoon, SM Bhagya P, MA Viraj J Muthugala, and AG Buddhika P Jayasekara (2022). "A Review on Human–Robot Proxemics". In: *Electronics* 11.16, p. 2490.

Alyafeai, Zaid, Maged Saeed AlShaibani, and Irfan Ahmad (2020). "A survey on transfer learning in natural language processing". In: *arXiv preprint arXiv:2007.04239*.

Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh (2022). "Natural language processing: State of the art, current trends and challenges". In: *Multimedia Tools and Applications*, pp. 1–32.

Lord, Catherine, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele (2018). "Autism spectrum disorder". In: *The lancet* 392.10146, pp. 508–520.

Williams, Tom, Nhan Tran, Josh Rands, and Neil T Dantam (2018b). "Augmented, mixed, and virtual reality enabling of robot deixis". In: *International Conference on Virtual, Augmented and Mixed Reality*. Springer, pp. 257–275.

Charisi, Vicky, Selma Sabanovic, Serge Thill, Emilia Gomez, Keisuke Nakamura, and Randy Gomez (2019). "Expressivity for Sustained Human-Robot Interaction". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 675–676.

Martelaro, Nikolas, Victoria C Nneji, Wendy Ju, and Pamela Hinds (2016). "Tell me more: Designing hri to encourage more trust, disclosure, and companionship". In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, pp. 181–188.

Balit, Etienne, Dominique Vaufreydaz, and Patrick Reignier (2018). "PEAR: Prototyping Expressive Animated Robots-A framework for social robot prototyping". In: *HUCAPP 2018-2nd International Conference on Human Computer Interaction Theory and Applications*, p. 1.

Cha, Elizabeth, Yunkyung Kim, Terrence Fong, Maja J Matarić, et al. (2018). "A Survey of Nonverbal Signaling Methods for Non-Humanoid Robots". In: *Foundations and Trends® in Robotics* 6.4, pp. 211–323.

Cha, Elizabeth, Anca D Dragan, and Siddhartha S Srinivasa (2015). "Perceived robot capability". In: *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 541–548.

Tapus, Adriana, Cristian Tapus, and Maja J Mataric (2009). "The use of socially assistive robots in the design of intelligent cognitive therapies for people with dementia". In: *2009 IEEE international conference on rehabilitation robotics*. IEEE, pp. 924–929.

Matarić, Maja J, Jon Eriksson, David J Feil-Seifer, and Carolee J Winstein (2007). "Socially assistive robotics for post-stroke rehabilitation". In: *Journal of NeuroEngineering and Rehabilitation* 4.1, p. 5.

Vandemeulebroucke, Tijs, Bernadette Dierckx de Casterlé, and Chris Gastmans (2018). "How do older adults experience and perceive socially assistive robots in aged care: a systematic review of qualitative evidence". In: *Aging & mental health* 22.2, pp. 149–167.

Jeong, Sooyeon, Kristopher Dos Santos, Suzanne Graca, Brianna O'Connell, Laurel Anderson, Nicole Stenquist, Katie Fitzpatrick, Honey Goodenough, Deirdre Logan, Peter Weinstock, et al. (2015). "Designing a socially assistive robot for pediatric care". In: *Proceedings of the 14th international conference on interaction design and children*, pp. 387–390.

Klein, Lauren, Laurent Itti, Beth A Smith, Marcelo Rosales, Stefanos Nikolaidis, and Maja J Matarić (2019). "Surprise! predicting infant visual attention in a socially assistive robot contingent learning paradigm". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 1–7.

Chen, Huili, Hae Won Park, and Cynthia Breazeal (2020). "Teaching and learning with children: Impact of reciprocal peer learning with a social robot on children's learning and emotive engagement". In: *Computers & Education* 150, p. 103836.

Short, Elaine, Katelyn Swift-Spong, Jillian Greczek, Aditi Ramachandran, Alexandru Litoiu, Elena Corina Grigore, David Feil-Seifer, Samuel Shuster, Jin Joo Lee, Shaobo Huang, et al. (2014). "How to train your dragonbot: Socially assistive robots for teaching children about nutrition through play". In: *The 23rd IEEE international symposium on robot and human interactive communication*. IEEE, pp. 924–929.

Birmingham, Chris, Zijian Hu, Kartik Mahajan, Eli Reber, and Maja J Matarić (2020). "Can I Trust You? A User Study of Robot Mediation of a Support Group". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 8019–8026.

Short, Elaine and Maja J Mataric (2017). "Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 385–390.

Salam, Hanan and Mohamed Chetouani (2015b). "Engagement detection based on mutli-party cues for human robot interaction". In: *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 341–347.

Bravo, Flor A, Alejandra M González, and Enrique González (2017). "A review of intuitive robot programming environments for educational purposes". In: *2017 IEEE 3rd Colombian Conference on Automatic Control (CCAC)*. IEEE, pp. 1–6.

Macedonia, Manuela (2019). "Embodied Learning: Why at school the mind needs the body". In: *Frontiers in psychology* 10.

Milgram, Paul, Haruo Takemura, Akira Utsumi, and Fumio Kishino (1995a). "Augmented reality: A class of displays on the reality-virtuality continuum". In: *Telemanipulator and telepresence technologies*. Vol. 2351. Spie, pp. 282–292.

Lipton, Jeffrey I, Aidan J Fay, and Daniela Rus (2017). "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing". In: *IEEE Robotics and Automation Letters* 3.1, pp. 179–186.

Williams, Tom, Leanne Hirshfield, Nhan Tran, Trevor Grant, and Nicholas Woodward (2020c). "Using augmented reality to better study human-robot interaction". In: *International Conference on Human-Computer Interaction*. Springer, pp. 643–654.

Walker, Michael, Hooman Hedayati, Jennifer Lee, and Daniel Szafir (2018). "Communicating Robot Motion Intent with Augmented Reality". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pp. 316–324.

Wu, Yuanjie, Yu Wang, Sungchul Jung, Simon Hoermann, and Robert W Lindeman (2021). "Using a fully expressive avatar to collaborate in virtual reality: Evaluation of task performance, presence, and attraction". In: *Frontiers in Virtual Reality* 2, p. 641296.

Murugan, Amarnath, Rishi Vanukuru, and Jayesh Pillai (2021). "Towards Avatars for Remote Communication using Mobile Augmented Reality". In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, pp. 135–139.

Pakanen, Minna, Paula Alavesa, Niels van Berkel, Timo Koskela, and Timo Ojala (2022). ""Nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in AR and VR based telexistence systems". In: *Entertainment Computing* 40, p. 100457.

Asadzadeh, Afsoon, Taha Samad-Soltani, and Peyman Rezaei-Hachesu (2021). "Applications of virtual and augmented reality in infectious disease epidemics with a focus on the COVID-19 outbreak". In: *Informatics in medicine unlocked* 24, p. 100579.

Horigome, Toshiro, Shunya Kurokawa, Kyosuke Sawada, Shun Kudo, Kiko Shiga, Masaru Mimura, and Taishiro Kishimoto (2020). "Virtual reality exposure therapy for social anxiety disorder: A systematic review and meta-analysis". In: *Psychological Medicine* 50.15, pp. 2487–2497.

Mosher, Maggie A and Adam C Carreon (2021). "Teaching social skills to students with autism spectrum disorder through augmented, virtual and mixed reality". In: *Research in Learning Technology* 29.

Pidel, Catlin and Philipp Ackermann (2020). "Collaboration in virtual and augmented reality: a systematic overview". In: *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, pp. 141–156.

Vellingiri, Shanthi, Ryan P McMahan, Vinu Johnson, and Balakrishnan Prabhakaran (2022). "An augmented virtuality system facilitating learning through nature walk". In: *Multimedia Tools and Applications*, pp. 1–12.

Papanastasiou, George, Athanasios Drigas, Charalabos Skianis, Miltiadis Lytras, and Effrosyni Papanastasiou (2019). "Virtual and augmented reality effects on K-12, higher and tertiary education students' twenty-first century skills". In: *Virtual Reality* 23.4, pp. 425–436.

Puljiz, David, Bowen Zhou, Ke Ma, and Björn Hein (2021). "HAIR: Head-mounted AR Intention Recognition". In: *Proc. of the 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*.

Wang, Chao and Anna Belardinelli (2022). "Investigating explainable human-robot interaction with augmented reality". In.

Rosen, Eric, David Whitney, Michael Fishman, Daniel Ullman, and Stefanie Tellex (2020). "Mixed reality as a bidirectional communication interface for human-robot interaction". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 11431–11438.

Higgins, Padraig, Ryan Barron, and Cynthia Matuszek (2022). "Head Pose for Object Deixis in VR-Based Human-Robot Interaction". In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 610–617.

Gadre, Samir Yitzhak, Eric Rosen, Gary Chien, Elizabeth Phillips, Stefanie Tellex, and George Konidaris (2019). "End-user robot programming using mixed reality". In: *2019 International conference on robotics and automation (ICRA)*. IEEE, pp. 2707–2713.

Rosen, Eric, David Whitney, Elizabeth Phillips, Daniel Ullman, and Stefanie Tellex (2018). "Testing robot teleoperation using a virtual reality interface with ROS reality". In: *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, pp. 1–4.

LeMasurier, Gregory, Jordan Allspaw, Murphy Wonsick, James Tukpah, Taskin Padir, Holly Yanco, and Elizabeth Phillips (2022). "Designing a User Study for Comparing 2D and VR Human-in-the-Loop Robot Planning Interfaces". In.

Rosen, Eric, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex (2019). "Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays". In: *The International Journal of Robotics Research* 38.12. Publisher: SAGE Publications Ltd STM, pp. 1513–1526. DOI: 10.1177/0278364919842925.

Holz, Thomas, Mauro Dragone, and Gregory MP O'Hare (2009). "Where robots and virtual agents meet". In: *International Journal of Social Robotics* 1.1, pp. 83–93.

Young, James E, Min Xin, and Ehud Sharlin (2007a). "Robot expressionism through cartooning". In: *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 309–316.

Dragone, Mauro, Thomas Holz, and Gregory MP O'Hare (2006). "Mixing robotic realities". In: *Proceedings of the 11th international conference on Intelligent user interfaces*. ACM, pp. 261–263.

Hamilton, Jared, Nhan Tran, and Tom Williams (2020). "Tradeoffs Between Effectiveness and Social Perception When Using Mixed Reality to Supplement Gesturally Limited Robots". In: *International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction*. Vol. 3.

Hamilton, Jared, Thao Phung, Nhan Tran, and Tom Williams (2021). "What's The Point? Tradeoffs Between Effectiveness and Social Perception When Using Mixed Reality to Enhance Gesturally Limited Robots". In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '21. New York, NY, USA: Association for Computing Machinery, pp. 177–186. DOI: 10.1145/3434073.3444676.

Tran, Nhan, Trevor Grant, Thao Phung, Leanne Hirshfield, Christopher Wickens, and Tom Williams (2021). "Robot-Generated Mixed Reality Gestures Improve Human-Robot Interaction". In: *International Conference on Social Robotics*. Springer, pp. 768–773.

Brown, Landon, Jared Hamilton, Zhao Han, Albert Phan, Thao Phung, Eric Hansen, Nhan Tran, and Tom Williams (2022). "Best of Both Worlds? Combining Different Forms of Mixed Reality Deictic Gestures". In: *ACM Transactions on Human-Robot Interaction*.

Walker, Michael, Hooman Hedayati, and Daniel Szafir (2019). "Robot Teleoperation with Augmented Reality Virtual Surrogates". In: DOI: 10.1109/HRI.2019.8673306.

Zhang, Tianhao, Zoe McCarthy, Owen Jowl, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel (2018). "Deep imitation learning for complex manipulation tasks from virtual reality teleoperation". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 1–8.

Hedayati, Hooman, Michael Walker, and Daniel Szafir (2018). "Improving collocated robot teleoperation with augmented reality". In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pp. 78–86.

Lipton, Jeffrey I, Aidan J Fay, and Daniela Rus (2018). "Baxter's homunculus: Virtual reality spaces for teleoperation in manufacturing". In: *IEEE Robotics and Automation Letters* 3.1, pp. 179–186.

Viglialoro, Rosanna Maria, Sara Condino, Giuseppe Turini, Marina Carbone, Vincenzo Ferrari, and Marco Gesi (2021). "Augmented reality, mixed reality, and hybrid approach in healthcare simulation: a systematic review". In: *Applied Sciences* 11.5, p. 2338.

Villanueva, Ana M, Ziyi Liu, Zhengzhe Zhu, Xin Du, Joey Huang, Kylie A Peppler, and Karthik Ramani (2021). "Robotar: An augmented reality compatible teleconsulting robotics toolkit for augmented makerspace experiences". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Wadgaonkar, Chinmay P, Johannes Freischuetz, Akshaya Agrawal, and Heather Knight (2021). "Exploring Behavioral Anthropomorphism With Robots in Virtual Reality". In.

Han, Zhao, Jenna Parrillo, Alexander Wilkinson, Holly A Yanco, and Tom Williams (2022). "Projecting robot navigation paths: Hardware and software for projected ar". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 623–628.

Gillen, Matthew, Joseph Loyall, Kyle Usbeck, Kelly Hanlon, Andrew Scally, Joshua Sterling, Richard Newkirk, and Ralph Kohler (2012). "Beyond line-of-sight information dissemination for force protection". In: *MILITARY COMMUNICATIONS CONFERENCE, 2012-MILCOM 2012*. IEEE, pp. 1–6.

Bolano, Gabriele, Christian Juelg, Arne Roennau, and Ruediger Dillmann (2019). "Transparent robot behavior using augmented reality in close human-robot interaction". In: *2019 28th IEEE*

International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, pp. 1–7.

Mott, Terran, Thomas Williams, Hao Zhang, and Christopher Reardon (2021). "You Have Time to Explore Over Here!: Augmented Reality for Enhanced Situation Awareness in Human-Robot Collaborative Exploration". In.

Mara, Martina, Kathrin Meyer, Michael Heiml, Horst Pichler, Roland Haring, Brigitte Krenn, Stephanie Gross, Bernhard Reiterer, and Thomas Layer-Wagner (2021). "CoBot Studio VR: A Virtual Reality Game Environment for Transdisciplinary Research on Interpretability and Trust in Human-Robot Collaboration". In.

Suzuki, Ryo, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt (2022). "Augmented Reality and Robotics: A Survey and Taxonomy for AR-enhanced Human-Robot Interaction and Robotic Interfaces". In: *CHI Conference on Human Factors in Computing Systems*, pp. 1–33.

Groechel, Thomas, Michael Walker, Christine T Chang, Eric Rosen, and Jessica Forde (2022a). "A Tool for Organizing Key Characteristics of Virtual, Augmented, and Mixed Reality for Human-Robot Interaction Systems: Synthesizing VAM-HRI Trends and Takeaways". In: *IEEE Robotics & Automation Magazine*.

Williams, Tom, Daniel Szafir, and Tathagata Chakraborti (2019a). "The reality-virtuality interaction cube". In: *Proceedings of the 2nd International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

Groechel, Thomas, Zhonghao Shi, Roxanna Pakkar, and Maja J Matarić (2019). "Using socially expressive mixed reality arms for enhancing low-expressivity robots". In: *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 1–8.

Groechel, Thomas, Roxanna Pakkar, Roddur Dasgupta, Chloe Kuo, Haemin Lee, Julia Cordero, Kartik Mahajan, and Maja J Matarić (2021). "Kinesthetic curiosity: Towards personalized embodied learning with a robot tutor teaching programming in mixed reality". In: *International Symposium on Experimental Robotics*. Springer, pp. 245–252.

LeMasurier, Gregory, Jordan Allspaw, and Holly A Yanco (2021). "Semi-Autonomous Planning and Visualization in Virtual Reality". In: *arXiv preprint arXiv:2104.11827*.

Tran, Nhan, Kai Mizuno, Trevor Grant, Thao Phung, Leanne Hirshfield, and Tom Williams (2020). "Exploring mixed reality robot communication under different types of mental workload". In: *International Workshop on Virtual, Augmented, and Mixed Reality for Human-Robot Interaction*. Vol. 3.

Quigley, Morgan, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng (2009). "ROS: an open-source Robot Operating System". In: *ICRA workshop on open source software*. Vol. 3. 3.2. Kobe, Japan, p. 5.

Koenig, Nathan and Andrew Howard (2004). "Design and use paradigms for gazebo, an open-source multi-robot simulator". In: *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*. Vol. 3. IEEE, pp. 2149–2154.

Barentine, Christian Michael, Andrew McNay, Ryan Pfaffenbichler, Addyson Smith, Eric Rosen, and Elizabeth Phillips (2021). "Manipulation Assist for Teleoperation in VR". In: *Proc. of the 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*.

Boateng, Andrew and Yu Zhang (2021). "Virtual Shadow Rendering for Maintaining Situation Awareness in Proximal Human-Robot Teaming". In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '21 Companion. Boulder, CO, USA: Association for Computing Machinery, pp. 494–498.

Ikeda, Bryce and Daniel Szafir (2021). "An AR Debugging Tool for Robotics Programmers". In.

Higgins, Padraig, Gaoussou Youssouf Kebe, Adam Berlier, Kasra Darvish, Don Engel, Francis Ferraro, and Cynthia Matuszek (2021). "Towards Making Virtual Human-Robot Interaction a Reality". In: *Proc. of the 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*.

Mimnaugh, Katherine J, Markku Suomalainen, Israel Becerra, Eliezer Lozano, Rafael Murrieta, and Steven LaValle (2021). "Defining Preferred and Natural Robot Motions in Immersive Telepresence from a First-Person Perspective". In: *Proc. of the 3rd International Workshop on Virtual, Augmented, and Mixed-Reality for Human-Robot Interactions (VAM-HRI)*.

Haas, John K (2014). "A history of the unity game engine". In: *Diss. WORCESTER POLYTECHNIC INSTITUTE* 483, p. 484.

Garon, Mathieu, Pierre-Olivier Boulet, Jean-Philippe Doiron, Luc Beaulieu, and Jean-François Lalonde (2016). "Real-time high resolution 3D data on the HoloLens". In: *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*. IEEE, pp. 189–191.

Bloom, Benjamin S (1984). "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring". In: *Educational researcher* 13.6, pp. 4–16.

Spaulding, Samuel and Cynthia Breazeal (2019). "Frustratingly Easy Personalization for Real-time Affect Interpretation of Facial Expression". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp. 531–537.

Williams, Tom, Daniel Szafir, Tathagata Chakraborti, Ong Soh Khim, Eric Rosen, Serena Booth, and Thomas Groechel (2020d). "Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)". In: *Companion of the 2020 ACM/IEEE International Conference on*

Human-Robot Interaction. HRI '20. Cambridge, United Kingdom: Association for Computing Machinery, pp. 663–664. DOI: 10.1145/3371382.3374850.

Groechel, Thomas, Chloe Kuo, and Roddur Dasgupta (2020). *interaction-lab/MoveToCode: DOI Release*. Version v0.0.1. DOI: 10.5281/zenodo.3924514.

Mariéthoz, Johnny and Samy Bengio (2005). "A unified framework for score normalization techniques applied to text-independent speaker verification". In: *IEEE signal processing letters* 12.7, pp. 532–535.

Mahajan, Kartik, Thomas Groechel, Roxanna Pakkar, Julia Cordero, Haemin Lee, and Maja J Matarić (2020). "Adapting usability metrics for a socially assistive, kinesthetic, mixed reality robot tutoring environment". In: *International Conference on Social Robotics*. Springer, pp. 381–391.

Papadopoulos, Irena, Runa Lazzarino, Syed Miah, Tim Weaver, Bernadette Thomas, and Christina Koulouglioti (2020). "A systematic review of the literature regarding socially assistive robots in pre-tertiary education". In: *Computers & Education* 155, p. 103924. DOI: https://doi.org/10.1016/j.compedu.2020.103924.

Malik, Norjasween Abdul, Fazah Akhtar Hanapiah, Rabiatul Adawiah Abdul Rahman, and Hanafiah Yussof (2016). "Emergence of socially assistive robotics in rehabilitation for children with cerebral palsy: A review". In: *International Journal of Advanced Robotic Systems* 13.3, p. 135.

Pino, Maribel, Mélodie Boulay, François Jouen, and Anne Sophie Rigaud (2015). ""Are we ready for robots that care for us?" Attitudes and opinions of older adults toward socially assistive robots". In: *Frontiers in aging neuroscience* 7, p. 141.

Keizer, Richelle ACM Olde, Lex Van Velsen, Mathieu Moncharmont, Brigitte Riche, Nadir Ammour, Susanna Del Signore, Gianluca Zia, Hermie Hermens, and Aurèle N'Dja (2019). "Using socially assistive robots for monitoring and preventing frailty among older adults: a study on usability and user experience challenges". In: *Health and Technology* 9.4, pp. 595–605.

Feingold-Polak, Ronit, Avital Elishay, Yonat Shahar, Maayan Stein, Yael Edan, and Shelly Levy-Tzedek (2018). "Differences between young and old users when interacting with a humanoid robot: A qualitative usability study". In: *Paladyn, Journal of Behavioral Robotics* 9.1, pp. 183–192.

Caleb-Solly, Praminda, Sanja Dogramadzi, Claire AGJ Huijnen, and Herjan van den Heuvel (2018). "Exploiting ability for human adaptation to facilitate improved human-robot interaction and acceptance". In: *The Information Society* 34.3, pp. 153–165.

Piech, Chris, Mehran Sahami, Jonathan Huang, and Leonidas Guibas (2015). "Autonomously generating hints by inferring problem solving policies". In: *Proceedings of the second (2015) acm conference on learning scale*, pp. 195–204.

Wang, Jiahui, Pavlo Antonenko, Mehmet Celepkolu, Yerika Jimenez, Ethan Fieldman, and Ashley Fieldman (2019). "Exploring relationships between eye tracking and traditional usability testing data". In: *International Journal of Human–Computer Interaction* 35.6, pp. 483–494.

Clabaugh, Caitlyn, Konstantinos Tsiakas, and Maja Matarić (2017). "Predicting Preschool Mathematics Performance of Children with a Socially Assistive Robot Tutor". In: *Proceedings of the Synergies between Learning and Interaction Workshop IROS, Vancouver, BC, Canada*, pp. 24–28.

Chen, Chaona, Oliver GB Garrod, Jiayu Zhan, Jonas Beskow, Philippe G Schyns, and Rachael E Jack (2018). "Reverse engineering psychologically valid facial expressions of emotion into social robots". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, pp. 448–452.

Meghdari, Ali, Minoo Alemi, Ali Ghorbandaei Pour, and Alireza Taheri (2016). "Spontaneous human-robot emotional interaction through facial expressions". In: *International Conference on Social Robotics*. Springer, pp. 351–361.

Kkedzierski, Jan, Robert Muszyński, Carsten Zoll, Adam Oleksy, and Mirela Frontkiewicz (2013). "EMYS–emotive head of a social robot". In: *International Journal of Social Robotics* 5.2, pp. 237–249.

Bretan, Mason, Guy Hoffman, and Gil Weinberg (2015). "Emotionally expressive dynamic physical behaviors in robots". In: *International Journal of Human-Computer Studies* 78, pp. 1–16.

Williams, Tom, Daniel Szafir, Tathagata Chakraborti, and Elizabeth Phillips (2019b). "Virtual, Augmented, and Mixed Reality for Human-Robot Interaction (VAM-HRI)". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 671–672.

Bischoff, Dr. Martin (2018). *Ros#*. `https://github.com/siemens/ros-sharp`.

Tomasello, Michael (2010). *Origins of human communication*. MIT press.

McNeill, David (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago press.

Leyzberg, Daniel, Samuel Spaulding, Mariya Toneva, and Brian Scassellati (2012). "The physical presence of a robot tutor increases cognitive learning gains". In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34.

Scassellati, Brian, Laura Boccanfuso, Chien-Ming Huang, Marilena Mademtzi, Meiying Qin, Nicole Salomons, Pamela Ventola, and Frederick Shic (2018). "Improving social skills in children with ASD using a long-term, in-home social robot". In: *Science Robotics* 3.21, eaat7544.

Thomas, Frank, Ollie Johnston, and Frank Thomas (1995). *The illusion of life: Disney animation*. Hyperion New York.

Takayama, Leila, Doug Dooley, and Wendy Ju (2011). “Expressing thought”. In: *Proceedings of the 6th international conference on Human-robot interaction - HRI ’11*.

Gielniak, Michael J and Andrea L Thomaz (2012). “Enhancing interaction through exaggerated motion synthesis”. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI ’12*.

Ribeiro, Tiago and Ana Paiva (2012). “The illusion of robotic life”. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI ’12*.

Heerink, Marcel, Ben Kröse, Vanessa Evers, and Bob Wielinga (2010). “Assessing acceptance of assistive social agent technology by older adults: the almere model”. In: *International journal of social robotics* 2.4, pp. 361–375.

Marmpena, Mina, Angelica Lim, and Torbjørn S Dahl (2018). “How does the robot feel? Perception of valence and arousal in emotional body language”. In: *Paladyn, Journal of Behavioral Robotics* 9.1, pp. 168–182.

Milgram, Paul, Anu Rastogi, and Julius J Grodski (1995b). “Telerobotic control using augmented reality”. In: *Robot and Human Communication, 1995. RO-MAN’95 TOKYO, Proceedings., 4th IEEE International Workshop on*. IEEE, pp. 21–29.

Koo, Terry K and Mae Y Li (2016). “A guideline of selecting and reporting intraclass correlation coefficients for reliability research”. In: *Journal of chiropractic medicine* 15.2, pp. 155–163.

Efron, Bradley and Robert Tibshirani (1986). “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy”. In: *Statistical science*, pp. 54–75.

Groechel, Thomas R, Amy O’Connell, Massimiliano Nigro, and Maja J Matarić (2022b). “Reimagining RViz: Multidimensional Augmented Reality Robot Signal Design”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 1224–1231.

Hershberger, Dave, David Gossow, Josh Faust, and William Woodall (2015). *rviz*. `https://github.com/ros-visualization/rviz`.

Bradski, Gary and Adrian Kaehler (2008). *Learning OpenCV: Computer vision with the OpenCV library*. ” O’Reilly Media, Inc.”

Lugaresi, Camillo, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. (2019). “Mediapipe: A framework for building perception pipelines”. In: *arXiv preprint arXiv:1906.08172*.

Kataoka, Yuta, Wataru Teraoka, Yasuhiro Oikawa, and Yusuke Ikeda (n.d.). “Effect of handy microphone movement in Mixed Reality visualization system of sound intensity”. In: (), p. 8.

Kose, Ahmet, Aleksei Tepljakov, Sergei Astapov, Dirk Draheim, Eduard Petlenkov, and Kristina Vassiljeva (2018). "Towards a Synesthesia Laboratory: Real-time Localization and Visualization of a Sound Source for Virtual Reality Applications". In: *Journal of Communications Software and Systems* 14.1. Number: 1, pp. 112–120. DOI: 10.24138/jcomss.v14i1.410.

Lopez-Rincon, Omar and Oleg Starostenko (2019). "Music Visualization Based on Spherical Projection With Adjustable Metrics". In: *IEEE Access* 7, pp. 140344–140354.

Ashktorab, Zahra, Mohit Jain, Q. Vera Liao, and Justin D. Weisz (2019). "Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3290605.3300484.

Ajani, Kiran, Elsie Lee, Cindy Xiong, Cole Nussbaumer Knaflic, William Kemper, and Steven Franconeri (2021). "Declutter and focus: Empirically evaluating design guidelines for effective data communication". In: *IEEE Transactions on Visualization and Computer Graphics*.

VanVoorhis, CR Wilson, Betsy L Morgan, et al. (2007). "Understanding power and rules of thumb for determining sample sizes". In: *Tutorials in quantitative methods for psychology* 3.2, pp. 43–50.

Svalina, Ana, Jesenka Pibernik, Jurica Dolić, and Lidija Mandić (2021). "Data Visualizations for the Internet of Things Operational Dashboard". In: *2021 International Symposium ELMAR*. IEEE, pp. 91–96.

Amini, Fereshteh, Nathalie Henry Riche, Bongshin Lee, Jason Leboe-McGowan, and Pourang Irani (2018). "Hooked on data videos: assessing the effect of animation and pictographs on viewer engagement". In: *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*.

Schrum, Mariah L, Michael Johnson, Muyleng Ghuy, and Matthew C Gombolay (2020). "Four years in review: Statistical practices of Likert scales in human-robot interaction studies". In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 43–52.

Abdi, Hervé (2010). "Holm's sequential Bonferroni procedure". In: *Encyclopedia of research design* 1.8, pp. 1–8.

Vargha, András and Harold D Delaney (2000). "A critique and improvement of the CL common language effect size statistics of McGraw and Wong". In: *Journal of Educational and Behavioral Statistics* 25.2, pp. 101–132.

Goktan, Ipek, Karen Ly, Thomas R Groechel, and Maja J Mataric (2022). "Augmented Reality Appendages for Robots: Design Considerations and Recommendations for Maximizing Social and Functional Perception". In: *arXiv preprint arXiv:2205.06747*.

Jackson, Ryan Blake and Tom Williams (2021). "A theory of social agency for human-robot interaction". In: *Frontiers in Robotics and AI*, p. 267.

Erel, Hadas, Guy Hoffman, and Oren Zuckerman (2018). "Interpreting non-anthropomorphic robots' social gestures". In: *Proceedings of the 2018 ACM/IEEE international conference on Human-Robot Interaction, Chicago, Illinois. ACM*.

Stogsdill, Adam, Grace Clark, Aly Ranucci, Thao Phung, and Tom Williams (2021). "Is it Pointless? Modeling and Evaluation of Category Transitions of Spatial Gestures". In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 392–396.

Piwek, Paul (2009). "Salience in the generation of multimodal referring acts". In: *Proceedings of the 2009 international conference on Multimodal interfaces*, pp. 207–210.

Bagchi, Shelly, Jeremy A Marvel, et al. (2018). "Towards augmented reality interfaces for human-robot interaction in manufacturing environments". In: *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*.

Chandan, Kishan, Vidisha Kudalkar, Xiang Li, and Shiqi Zhang (2021). "ARROCH: Augmented reality for robots collaborating with a human". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 3787–3793.

Krenn, Brigitte, Tim Reinboth, Stephanie Gross, Christine Busch, Martina Mara, Kathrin Meyer, Michael Heiml, and Thomas Layer-Wagner (2021). "It's your turn!–A collaborative human-robot pick-and-place scenario in a virtual industrial setting". In: *arXiv preprint arXiv:2105.13838*.

Young, James E., Min Xin, and Ehud Sharlin (2007b). "Robot expressionism through cartooning". In: *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 309–316. DOI: 10.1145/1228716.1228758.

Pärsch, Nikolai, Clemens Harnischmacher, Martin Baumann, Arnd Engeln, and Lutz Krauß (2019). "Designing augmented reality navigation visualizations for the vehicle: a question of real world object coverage?" In: *International Conference on Human-Computer Interaction*. Springer, pp. 161–175.

Nowacki, Paweł and Marek Woda (2020). "Capabilities of arcore and arkit platforms for ar/vr applications". In: *International Conference on Dependability and Complex Systems*. Springer, pp. 358–370.

Sapounidis, Theodosios and Stavros Demetriadis (2017). "Educational robots driven by tangible programming languages: A review on the field". In: *International Conference EduRobotics 2016*. Springer, pp. 205–214.

Kanellopoulou, Ioanna, Pablo Garaizar, and Mariluz Guenaga (2021). "First Steps Towards Automatically Defining the Difficulty of Maze-Based Programming Challenges". In: *IEEE Access* 9, pp. 64211–64223.

Guenaga, Mariluz, Andoni Eguiluz, Pablo Garaizar, and Juanjo Gibaja (2021). "How do students develop computational thinking? Assessing early programmers in a maze-based online game". In: *Computer Science Education* 31.2, pp. 259–289.

Ternik, Žan, Anja Koron, Tine Koron, and Irena Nančovska Šerbec (2017). "Learning programming concepts through maze game in scratch". In: *European Conference on Games Based Learning, Academic Conferences International Limited, str*, pp. 661–670.

Wu, Qiong and Chunyan Miao (2013). "Curiosity: From psychology to computation". In: *ACM Computing Surveys (CSUR)* 46.2, pp. 1–26.

Google (2022). *Augmented reality design guidelines - google developers*.

Diaz, Catherine, Michael Walker, Danielle Albers Szafir, and Daniel Szafir (2017). "Designing for depth perceptions in augmented reality". In: *2017 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, pp. 111–122.

Jin, Qiao, Danli Wang, Xiaozhou Deng, Nan Zheng, and Steve Chiu (2018). "AR-Maze: a tangible programming tool for children based on AR technology". In: *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pp. 611–616.

Hattori, Keisuke and Tatsunori Hirai (2019). "An intuitive and educational programming tool with tangible blocks and AR". In: *ACM SIGGRAPH 2019 Posters*, pp. 1–2.

Resnick, Mitchel, John Maloney, Andrés Monroy-Hernández, Natalie Rusk, Evelyn Eastmond, Karen Brennan, Amon Millner, Eric Rosenbaum, Jay S Silver, Brian Silverman, et al. (2009). "Scratch: Programming for all." In: *Commun. Acm* 52.11, pp. 60–67.

Harris, Paul L (2012). *Trusting what you're told: How children learn from others*. Harvard University Press.

Gordon, Goren, Cynthia Breazeal, and Susan Engel (2015). "Can children catch curiosity from a social robot?" In: *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 91–98.

Gul, Demet, Ibrahim Cetin, and M Yasar Ozden (2022). "A scale for measuring middle school students' attitudes toward programming". In: *Computer Applications in Engineering Education* 30.1, pp. 251–258.

Velentza, Anna-Maria, Stavros Ioannidis, and Nikolaos Fachantidis (2020). "Service robot teaching assistant in school class-room". In: pp. 12115–12117.

Putnam, Cynthia, Melisa Puthenmadom, Marjorie Ann Cuerdo, Wanshu Wang, and Nathaniel Paul (2020). "Adaptation of the system usability scale for user testing with children". In: *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pp. 1–7.

Wilson, F Robert, Wei Pan, and Donald A Schumsky (2012). "Recalculation of the critical values for Lawshe's content validity ratio". In: *Measurement and evaluation in counseling and development* 45.3, pp. 197–210.

Çakır, Nur Akkuş, Arianna Gass, Aroutis Foster, and Frank J Lee (2017). "Development of a game-design workshop to promote young girls' interest towards computing through identity exploration". In: *Computers & Education* 108, pp. 115–130.

Carvajal-Ayala, Daisy Catalina and Ricardo Alonso Avendaño-Franco (2021). "Implementing Lesson Plans for Collaborative Learning with Children in an EFL Context." In: *GIST Education and Learning Research Journal* 22, pp. 199–226.

Laru, Jari, Sanna Järvelä, and Roy B Clariana (2012). "Supporting collaborative inquiry during a biology field trip with mobile peer-to-peer tools for learning: a case study with K-12 learners". In: *Interactive Learning Environments* 20.2, pp. 103–117.

Puvirajah, Anton, Geeta Verma, and Todd Campbell (2020). "Advancing Minoritized Learners' STEM Oriented Communication Competency Through a Science Center-Based Summer Program". In: *Journal of Museum Education* 45.4, pp. 437–449.

GlobalStats (2022). *Tablet vendor market share united states of america*.

McKnight, Patrick E and Julius Najab (2010). "Mann-Whitney U Test". In: *The Corsini encyclopedia of psychology*, pp. 1–1.

Woolson, Robert F (2007). "Wilcoxon signed-rank test". In: *Wiley encyclopedia of clinical trials*, pp. 1–3.

Macbeth, Guillermo, Eugenia Razumiejczyk, and Rubén Daniel Ledesma (2011). "Cliff's Delta Calculator: A non-parametric effect size program for two groups of observations". In: *Universitas Psychologica* 10.2, pp. 545–555.

Chatwani, Nisha, Chloe Kuo, Thomas R Groechel, and Maja J Mataric (2022a). "PoseToCode: Exploring Design Considerations toward a Usable Block-Based Programming and Embodied Learning System". In.

Chatwani, Nisha, Chloe Kuo, Groechel Thomas, Julia Cordero, and Radhika Agrawal (2022b). *PoseToCode*. `https://github.com/interaction-lab/PoseToCode`.

Chatwani, Nisha, Chloe Kuo, and Thomas R Groechel (2022c). *PoseToCode Demonstration*. `https://posetocode.web.app/tutorial.html`.

Pasternak, Erik, Rachel Fenichel, and Andrew N Marshall (2017). "Tips for creating a block language with blockly". In: *2017 IEEE blocks and beyond workshop (B&B)*. IEEE, pp. 21–24.

Kalelioğlu, Filiz (2015). "A new way of teaching programming skills to K-12 students: Code. org". In: *Computers in Human Behavior* 52, pp. 200–210.

Bangor, Aaron, Philip Kortum, and James Miller (2009). "Determining what individual SUS scores mean: Adding an adjective rating scale". In: *Journal of usability studies* 4.3, pp. 114–123.

Klug, Brandy (2017). "An overview of the system usability scale in library website and system usability testing". In: *Weave: Journal of Library User Experience* 1.6. DOI: 10.3998/weave.12535642.0001.602.

Lin, Lorraine, Aline Normoyle, Alexandra Adkins, Yu Sun, Andrew Robb, Yuting Ye, Massimiliano Di Luca, and Sophie Jörg (2019). "The effect of hand size and interaction modality on the virtual hand illusion". In: *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, pp. 510–518.

De Pace, Francesco, Federico Manuri, Andrea Sanna, and Davide Zappia (2018). "An augmented interface to display industrial robot faults". In: *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, pp. 403–421.

Collett, Toby Hartnoll Joshua and Bruce Alexander Macdonald (2010). "An augmented reality debugging system for mobile robot software engineers". In.

Millard, Alan G, Richard Redpath, Alistair M Jewers, Charlotte Arndt, Russell Joyce, James A Hilder, Liam J McDaid, and David M Halliday (2018). "ARDebug: an augmented reality tool for analysing and debugging swarm robotic systems". In: *Frontiers in Robotics and AI* 5, p. 87.

Fitter, Naomi T, Luke Rush, Elizabeth Cha, Thomas Groechel, Maja J Matarić, and Leila Takayama (2020). "Closeness is key over long distances: Effects of interpersonal closeness on telepresence experience". In: *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 499–507.