

Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations

KAYLA MATHEUS, REBECCA RAMNAUTH, and BRIAN SCASSELLATI, Computer Science, Yale University, New Haven, Connecticut, USA
NICOLE SALOMONS, Computer Science, Imperial College London, London, United Kingdom

In the past two decades, the field of social robotics has undergone significant growth, witnessing a surge in long-term human–robot interaction (HRI) studies. This review paper provides an in-depth analysis of 120 long-term HRI studies conducted between 2003 and 2023, spanning 7 major domains including education, entertainment, and physical and mental health. We define "long-term" as studies deploying social robots with the same users for more than three sessions across 3 consecutive days, aiming to employ a comprehensive approach and identify trends in this dynamic field. Our analysis explores various aspects of these studies, from participant demographics to the characteristics of the HRI and engagement measures. The findings reveal promising trends, such as diverse age group representation, a strong focus on real-world contexts, and autonomous robot operation. We also identify gaps, notably the limited representation of studies involving teenagers and those studying workplace settings. By presenting this overview, we aim to empower the HRI community to address challenges, refine methodologies, and foster innovation in the domain of long-term HRI.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Computer systems organization** → **Robotics**; • **General and reference** → *Surveys and overviews*;

Additional Key Words and Phrases: social robots, human-robot interaction, long-term robotics, multi-session robotics

ACM Reference format:
Kayla Matheus, Rebecca Ramnauth, Brian Scassellati, and Nicole Salomons. 2025. Long-Term Interactions with Social Robots: Trends, Insights, and Recommendations. *ACM Trans. Hum.-Robot Interact.* 14, 3, Article 55 (June 2025), 42 pages.
https://doi.org/10.1145/3729539

This work is supported by the National Science Foundation under grant numbers 1928448, 1955653, and 2106690. R. Ramnauth is supported by the NSF Graduate Research Fellowship and the NASEM Ford Predoctoral Fellowship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF and NASEM.
Authors' Contact Information: Kayla Matheus (corresponding author), Computer Science, Yale University, New Haven, Connecticut, USA; e-mail: kayla.matheus@yale.edu; Rebecca Ramnauth, Computer Science, Yale University, New Haven, Connecticut, USA; e-mail: rebecca.ramnauth@yale.edu; Brian Scassellati, Computer Science, Yale University, New Haven, Connecticut, USA; e-mail: brian.scassellati@yale.edu; Nicole Salomons, Computer Science, Imperial College London, London, United Kingdom; e-mail: n.salomons@imperial.ac.uk.



This work is licensed under Creative Commons Attribution International 4.0.

© 2025 Copyright held by the owner/author(s).
ACM 2573-9522/2025/6-ART55
https://doi.org/10.1145/3729539

1 Introduction

In the past two decades, the field of social robotics has experienced remarkable growth and, with it, a proliferation of studies focused on long-term **human–robot interaction (HRI)** over the course of multiple days, weeks, and months. A *social robot* possesses features and capabilities that enable it to interact with humans in ways that resemble social interactions between people [14]. Such robots can exhibit behaviors such as recognizing and expressing emotions [127], understanding and generating natural language [65], adapting to different social contexts [123], and even demonstrating a degree of empathy [82]. For many, the goal of social robotic study is to support a future where robots are not just present in individual lab sessions but are integrated *longitudinally* into daily lives.

By our accounts, the number of HRI research papers examining longitudinal social robot use has tripled from pre-2013 to 2023 (Section 4). Bajones et al. [7] have highlighted this transformation by noting that whereas the "burning question in HRI studies" was once "how many participants do I need?" it is now "how long should my study run for?" Such a shift reflects the field's progression beyond initial experimentation in social robotics and toward a deeper understanding of how robots can be effectively deployed in real-world applications. In a future where robots are integrated into our homes, schools, offices, and medical facilities, it becomes increasingly essential to research the dynamics of long-term interactions spanning days, weeks, months, and even years.

From a research perspective, the study of repeated HRIs offers several unique advantages compared with single sessions. For one, many types of robotic interactions require longitudinal engagement in order to have a positive effect. Robots that help the user develop a new skill or behavior change fall into this category, such as tutoring, training, and therapeutic robots. Additionally, as with human–human relationships, the relationship between humans and robots develops temporally. Answering questions surrounding how to create healthy and robust HRI relationships, how to personalize them, and what types work best in which settings are important benefits of long-term study. Finally, factors such as usage patterns, drop-off rates, and strategies for sustaining engagement [8, 160, 165] are crucial to understanding the effective integration of robots into daily life and society. Only through long-term deployments can researchers explore how users adopt robots and what methods impact successful adoption. A deeper discussion of the benefits of long-term HRI study is presented in Section 2.2.

While the benefits of long-term HRI study are many, their pursuit often translates to researchers devoting more time, resources, and effort to build and deploy such systems [121]. Long-term studies involve multiple sessions and can often require deployments in uncontrolled, real-world environments. Consequently, the robots used in such studies must be more robust and sophisticated with advanced hardware and software components, often requiring full autonomy and a consideration of how the robot should behave at all hours of the day in multiple contexts. Studies investigating research topics such as personalization and maintaining engagement often require even deeper design and development for adaptive features. Ensuring privacy and managing data collection in a live environment also becomes increasingly intricate, and recruiting participants for extended periods can be challenging. A deeper discussion of the benefits of long-term HRI study is presented in Section 2.3.

Given the multitude of challenges associated with valuable long-term HRI studies, it becomes particularly important for researchers to learn from prior similar efforts. As long-term HRI progresses, it becomes essential to avoid unnecessary replication of work, identify gaps in knowledge, and pinpoint opportunities for advancement. In this spirit, we provide a comprehensive review of long-term HRI studies conducted over the past 20 years as a valuable resource for researchers. Our work fills a notable gap in the literature, as the last in-depth review of long-term social robots

[78] was conducted more than a decade ago in 2012. Building upon this prior work, our study presents a thorough analysis of 120 long-term studies spanning 7 major domains, such as education, entertainment, physical health, and mental health. Our analysis covers the period from 2003 until the time of analysis (April 2023), providing a comprehensive overview of the progress made in the field. To ensure inclusiveness and capture relevant trends, we have defined *long-term* in our corpus as studies deploying social robots with the same users for three or more sessions across 3 distinct days (e.g., one session per day for 3 days, once a week for 3 weeks). By adopting this criterion, we aim to adopt an inclusive methodology and identify patterns and insights that can contribute to a deeper understanding of long-term HRI. However, we also discuss alternative approaches to the *long-term* definition in Section 2.1.

In our analysis, we explore patterns over time and across domains for a number of study design aspects, including the longitudinal characteristics of the studies, the types and number of participants involved, study locations, defining elements of HRI, and the types of results and engagement measures employed. Figure 1 shows four different long-term social robotic interaction studies from our corpus, which vary widely across the characteristics we analyze. For instance, the studies vary in dyadic ((a) and (d)) interactions to triadic (c) and group interactions (b). The ages of participants in our studies ranged from children ((b), (c)) to the elderly (d). Some studies were conducted in the wild, such as in the home (c) or schools (b). Others ((a) and (d)) were done in laboratory environments. Lastly, they span a variety of domains, including education ((a) and (b)), social skills training (c), and exercising (d). By sharing insights and patterns on the design of such studies, we seek to empower the HRI community to collectively address challenges, refine methodologies, and foster innovation in the field of long-term HRI. We conclude this review with a discussion of the primary gaps and opportunities for future research in the domain of long-term HRI, as well as provide researchers with a number of study guidelines.

2 Background

In the following sections, we delve into a working definition of *long-term* within the context of HRI, explore the benefits and challenges inherent in conducting extended HRI research, and provide an overview of prior reviews in the realm of long-term HRI studies.

2.1 What Is Long-Term HRI?

As this review covers the topic of long-term HRI, the first question one may ask is: what does "long-term" signify in the context of social robotics? Could *any* repeated interaction with a robot be considered long-term, or is there an unspoken rule that a study must meet in order to be considered part of this category? Multiple perspectives are available in the HRI literature with no single agreed-upon approach. Many studies are organized around distinct sessions between users and robots, so one logical approach is to consider a minimum number of sessions. Kory-Westlund et al. take this approach for inclusion in their review of long-term, **socially interactive agents (SIAs)** [71], utilizing five sessions as their cutoff (while also acknowledging the arbitrary nature of this choice). By this measure, five sessions with an adaptive robotic tutor of Ramachandran et al. [116] would be considered long-term, but three sessions of Donnermann et al. [31] or four sessions of Jones and Castellano [60], also with adaptive robotic tutors, would not. Positing that there is a considerable difference between four and five sessions as a cut-off may not be productive. Furthermore, if five sessions were to occur on the same day, rather than spread out over multiple days, the "long-term" claim becomes more challenging. Relatedly, how should we consider a series of five 3-minute sessions over a week in comparison to five 30-minute sessions? While the number of sessions may be an easy yardstick, this alone does not tell the whole interaction story.



Fig. 1. Four long-term social robotic interaction studies from our corpus, which vary widely across the characteristics we analyze: (a) an investigation into the persistence of first impressions of a Furhat robot [107], (b) a Robovie robot that interacts with children in a classroom setting [62], (c) a robot that provides social skills training for children with Autism Spectrum Disorder (ASD) [131], and (d) a robot that provides exercising motivation to the elderly [34].

Perhaps a better approach would be to calculate the total interaction time with a robot over the study period. A few studies report this metric explicitly, such as Afyouni et al. [1] (averaging 2 hours 22 minutes of free-use interactions over 1 week) or Kidd and Breazeal [66] (averaging 50.6 days). However, most authors do not measure or calculate this metric, perhaps due to complexity or a lack of historical precedent. Total interaction time also fails to capture any information on time gaps between interactions. However, a more core challenge persists: a certain number of minutes or days of interaction may have different implications for robots within different domains and purposes. What is "long-term" for a tutoring robot may be different from that of a game-playing robot, based on interaction aspects such as the cognitive load of the activity or inherent repetition in the content.

Another approach to bench-marking a "long-term" definition is to instead use a measurement of the interaction having passed the *novelty effect* [78]. The novelty effect is generally characterized by a shift from initial excitement with a new experience to more stable and meaningful engagement. Capturing passing this effect means that researchers can be more sure that their results represent truly "long-term" social robot effects rather than shorter-term responses to novelty. For instance, Bajones et al. [7] argued that a claim of surpassing the novelty phase could be made with reasonable confidence only after a 3-week period, thus setting their study duration at 21 days. Prior work has attempted to establish when the novelty effect wears off [78, 160]. However, there is no fixed or universally agreed-upon timeframe given the numerous variables that can influence novelty

effects. For instance, the complexity of the interaction affordances on the robot may impact how inherently engaging it is and the diversity of features to explore. Alternatively, the setting of the robot (e.g., public vs. private) may impact the level of desire a user has to interact with the robot. Such variability ultimately places the responsibility on researchers to substantiate attenuation of the novelty effect in their specific use-case scenario by employing appropriate metrics and analysis, which relatively few authors in our corpus have done. Among those who have, Weiss et al. [160] noted the presence of novelty effects by observing a decline in attachment levels with a Vector robot after just 2 weeks out of a 30-week study. In contrast, Bodala et al. [13] found no discernible evidence of novelty effects even after five weekly sessions involving a mindfulness training robot. This result presents a conundrum: Was there an absence of novelty from the start, or was 5 weeks not yet enough time to overcome them?

With no one hard and fast rule, we adopt broad strokes in this review with regard to the definition of *long-term. We thus consider studies as having more than 3 sessions across 3 separate days, with the same user, as applicable to our corpus of* ***long-term*** *social robotics studies.* In this way, *long-term* studies are inherently multi-session, but not all multi-session studies are *long-term* (for instance, studies with two sessions). While we do not aim to set a strict definition for "long-term" HRI from our analysis, we do seek to illuminate how the field to date has considered long-term characteristics for HRI studies. In Section 4.1, we share how our corpus spans a wide range of temporal characteristics (sessions, days, interaction time, etc.) and offer ways researchers can consider these elements for future studies.

2.2 Benefits of Long-Term HRI Research

Long-term HRI studies require significant resources and effort. In this section, we highlight three high-level reasons why researchers may want to consider investing in conducting long-term HRI studies.

2.2.1 Unique Interaction Types. A number of robot interaction types and benefits are only possible with long-term deployments. Certain robotic systems designed to teach new skills, such as tutoring, physical training, or cognitive training necessitate multiple sessions over an extended period to effectively support skill development (e.g., [26, 31, 34, 53, 80, 117, 151, 157, 161]). Assessing the effectiveness of these programs also requires longitudinal measurements to demonstrate the user's skill acquisition over time. Therapeutic robots, such as those supporting individuals with mental health challenges [133] or **Autism Spectrum Disorder (ASD)** [67], benefit from a long-term study in similar ways. Successful therapeutic engagements are often repetitive in nature [90], and long-term interaction allows researchers to observe the effects of sustained engagement on the well-being of individuals receiving therapy.

2.2.2 User Perceptions and Relationships. Another benefit from long-term study is the ability to explore how a user's perceptions of a robot and their relationship with the robot may change over time. Kory-Westlund et al. discuss the notion of moving "beyond interaction to relationship" with long-term HRI [71]; just as relationships between humans can enhance learning, health, and social outcomes, building and sustaining deeper relationships with robots can do the same. For example, Paetzel et al. [107] found that while the perceived competence of a robot remained steady over multiple repeated sessions, perceived threat and discomfort fluctuated. Jeong et al. [58] found that a home-based social robot was able to build rapport and working alliance while supporting the mental well-being of college students. Fostering positive relationships is particularly important for companionship-based robots, such as general home robots or those addressing loneliness in elder care [15, 19, 57, 59, 106, 110].

*2.2.3 Robot Adoption and **Long-Term Engagement (LTE)**.* Long-term studies also offer valuable insights into the adoption of robots in various settings. Factors such as usage patterns, drop-off rates, and strategies for sustaining engagement [8, 160, 165] are crucial to understanding the effective integration of robots into daily life and society. One specific area of interest is the examination of the novelty effect and its impact on user interactions with robots [142, 146]. Consequently, a key challenge for long-term HRI is designing systems that continue to engage users even after the novelty effect has subsided [106, 141]. Only through studying long-term robot deployments can researchers explore strategies and design principles that promote both adoption and sustained user engagement, ensuring that robots maintain their effectiveness.

2.3 Challenges of Long-Term HRI Research

Despite the many benefits, there are several non-trivial challenges when considering long-term studies. We present a brief overview of many of the challenges faced by the studies in our corpus of long-term HRI research. Such hurdles reinforce the need for an understanding of the current state of the domain.

2.3.1 Autonomous, In-the-Wild Deployments. During longer-term deployments, there is often a greater need for robot autonomy, in order to avoid having a human operator always present. Not all long-term studies face this challenge, but many require building autonomous systems with more complex development such as supporting a fuller set of robotic features and behavioral planning, wireless connectivity, remote data logging, easy remote troubleshooting, and so on.

Many long-term studies are also conducted in in-the-wild environments such as homes, schools, or hospitals. These settings allow for the study of robots in real-world contexts and can also provide access to relevant participants at scale. However, in-the-wild deployments require significantly more adaptive systems, as the research team has no control over the environmental variables. Each home or school is very different from each other, and the robot needs to operate in these varied locations. For example, perception systems must accommodate variable noise, safety measures must predict a wider range of potential scenarios, and behavioral planning must work under variable conditions.

2.3.2 Recruitment and Adherence. Recruiting participants for long-term research studies can be significantly more complex compared to recruiting for single-session studies. Participants must be available for multiple sessions instead of a single day, and the longer a study, the more likely participant routines or circumstances may change. Depending on the setting, some participants will also need to be willing to accept the robot into relatively more private spaces, such as their homes and schools. All these factors can make it more difficult to find willing participants. When interacting with a robot for one or a limited number of sessions, most users can stay engaged and adhere to the protocols, especially with researcher supervision. With long-term systems, engagement must be sustained past initial novelty. This is a bigger challenge for in-the-wild settings, where there is limited researcher oversight and users must choose for themselves whether to interact with the system. In these manners, long-term systems have a deeper requirement to be engaging, or have visible benefits to the user, so that users continue to incorporate the robotic system as a part of their life.

2.3.3 Dynamic Content and Personalization. Creating sufficient content and robot behaviors can be cumbersome, especially if providing novel content for every interaction over multiple months or years. It is significantly more difficult to create sufficient and engaging content for long-term sessions than for short-term interactions as the user will not want to interact with a robot performing the same or too-similar utterances and actions each session.

Relatedly, many robotic studies have demonstrated the benefits of personalizing content for each user [44, 113]. However, content creation becomes especially difficult when considering personalization. Not only is sufficient long-term content needed, but it also requires consideration of diverse user characteristics such as age, preferences, and skill levels.[1] Additional challenges specific to personalization include creating robust user modeling systems and effective action selection mechanisms based on the user's state. On the whole, the development of novel robotic interactions is often lengthier and more complex for long-term interactions than single-session ones.

2.3.4 Cost. An additional factor when creating a long-term study is the cost of the systems. Social robots are expensive, and many are often needed at once when conducting long-term deployments with multiple users in parallel. If only one robot is used at a time, many studies would take years to complete if it only interacts with one user at a time. In addition to robots, multiple copies of computers, cameras, and other hardware must be acquired. The increased cost can limit the type, length, and number of long-term studies that can be conducted.

2.4 Prior Long-Term HRI Reviews

There have been two prior reviews in the domain on long-term HRI [71, 78]. The first, published by Leite et al. in 2012, focused on long-term social robots to the date of publishing [78] and included 24 papers. The authors include papers in which the robot interacts socially with users and engages them for extended periods of time (without defining in detail what is considered to be an extended period of time). The primary goal of the reviews was to highlight the main findings of the studies in their corpus and discuss open areas for future work in long-term HRI. Particular attention was paid to the results of studies that could show interaction with the novelty effect.

In their results, Leite et al. [78] categorize each of the 24 papers into 1 of 4 categories: health care and therapy, education, work environments and public spaces, and home robots. Regarding health care and therapy robots, overall there were positive results for therapeutic and health-related scenarios, but most studies had limited sample sizes in the domain. Their discussion regarding educational robots focused on children and stated that the robot's form factor and behavior greatly impacted the interactions. When discussing work environments and public spaces robots, the authors state that it is difficult to make general conclusions due to the great variety of studies and the relatively small sample sizes. Lastly, they discuss home robots and mention the importance of overcoming the novelty effect so people continue using them in the long run. The authors define the novelty effect as having worn off when the user becomes familiar with the robot and starts preferring new robot behaviors.

There are three main differences between the study of Leite et al. and ours. The first is the addition of new papers published between 2013 and mid-2023. Many important works have been done in the field over the past decade that we capture in our review corpus. The second difference lies in our qualifications for our corpus. While Leite et al. include studies of robots interacting with new users each day (e.g., a receptionist or mall guide), we limit our review to studies that track the same user over time with a robot. *In this way, we distinguish between long-term **interaction** deployments, where a user is experiencing the long-term nature of the robot, and long-term **research** deployments, where an individual user may only see the robot once. Our focus is on the former, and we mean this type of longitudinality when using the term "long-term" in this work.* Finally, we investigate characteristics around a study's location and domain separately, such as a health robot used in a

[1]It is important to note that the recent developments and applications of large-language models in HRI make these challenges easier to overcome now than they have been in the past. However, virtually all of the studies in our corpus did not have access to this technology at the time of study.

home versus a care facility. Given the smaller number of studies at the time, studies in the Leite et al. corpus were more correlated between their locations and domains.

The second prior review on long-term HRI is a book chapter on long-term SIAs published in 2022 [71]. Although their scope is broader than ours by including digital agents, Kory-Westlund et al. include a section on long-term embodied social robots. The authors use the criteria of "at least five sessions over any length of time per session and for any elapsed time from the first encounter to the fifth or more" and find 67 studies fitting these criteria. Like the 2012 review paper by Leite et al., they focus their discussion on the different applications of long-term robots and discuss the following three areas "social robots and children," "social robots and health and wellness," and "living with consumer robots." The authors discuss key takeaways from the results of studies in these domains as well as highlight the importance of building relationships over time with long-term SIAs. Our review is complementary to this work, with several key differences. Our review utilizes different inclusion criteria, specifically a minimum of 3 sessions with the same user over 3 days. We additionally analyze these studies for trends over the past two decades across seven key HRI domains and a multitude of study design characteristics as further described in Section 3.

3 Review Method

In this section, we describe the method for compiling our corpus. In order to systematically review social robots deployed in long-term situations over the past two decades, we developed the following criteria for paper inclusion:

(1) The study must involve social interaction between at least one robot and at least one human. We did not include studies in which the robot only had functional use with little to no social elements (e.g., [98, 103]).
(2) The participants and the robot must have interacted for at least 3 sessions (either as required by the study or from the user's free use) over a minimum of 3 separate days.
(3) The robot interacted with the same user for at least three sessions, and data were tracked for the user throughout the sessions. This excludes long-term studies in public spaces where the robot was interacting with a different user every time (e.g., [43, 63]), or where data were not clearly tracked for the same user (e.g., [48, 152]).
(4) The robot must be or appear to be an autonomous agent to the user. Wizard-of-Oz studies are included in the meta-analysis as long as participants believed the robot to be autonomous. We did not include telepresence studies (e.g., [17]) or studies where the participants mainly controlled or programmed the robot (e.g., [5]). Additionally, we did not include studies that were solely design-focused, in which the user did not interact with the robot in a deployed manner (e.g., [76]).
(5) The robot must be physically present (embodied) during the interaction. We included studies with multiple experimental conditions if at least one had an embodied robot. We did not include studies where the robot was only represented on a screen or as an avatar (e.g., [153]).

We first included all papers from Leite et al. [78], that fit our criteria. For papers after 2012, we proceeded with a keyword search on Google Scholar and relevant conference and journal proceedings. We used the following keywords for our search: *long-term; long term; in-home; in home; home deployment; repeat use; repeated use; multi-session; multisession; in the wild; in-the-wild; weeks months; and longitudinal*, alongside *robot* or *robotics*. Journal and conference proceedings included *Human-Robot Interaction (HRI), International Symposium on Robot and Human Interactive Communication (RO-MAN), International Journal of Social Robotics (IJSR), Journal of Human-Robot Interaction (JHRI), Transactions in Human-Robot Interaction (THRI)*, and *Science Robotics*. Papers were added to a queue based on their abstract and then manually assessed by the authors for each

element of our criteria. The earliest published paper included in the corpus was published in 2003 [135], and the survey cutoff date was April 2023.

In total, we found 118 papers that fit our criteria, which represent 120 studies. The number of studies and papers differ as some papers (e.g., [72]) presented several separate studies that fit our criteria, and others had multiple papers derived from the same data which were not included in our final corpus. Our research team reviewed the papers and annotated key information from each, as outlined below. When unsure or in disagreement, the annotators discussed a paper together to reach a consensus on which category or number best represented the situation. The following information was extracted from each paper.

Temporal Qualities.

- *Year*: The year the paper was published.
- *Study Period*: The average number of days the study was performed for each user. This was most often reported directly by the paper authors themselves. When the period was not listed in days, we assumed that 1 week is 7 days and 1 month is 30 days, in order to calculate an estimated number of days for the study. For example, "4 sessions within 2 weeks" [150] has a study period of 14 days, while 1 session every week for a total of 10 sessions [37] totaled a 70-day study period estimate (the sessions need not be on the same day per week). In the rare cases where the study period was unclear, such as "for a semester," we listed the study period as unknown.
- *Number of User Sessions*: The average number of sessions between a robot and participant, as relevant. Some studies were not session-based and were instead labeled as "free use" if the use was entirely up to the user or "daily use" if participants were instructed to generally use the robot on a daily basis.
- *Session Length*: The average number of minutes per session (as relevant).
- *Total Interaction Time*: The average total time, in minutes, that each participant spent with the robot during the study. When this information was not given explicitly in a paper, we manually calculated an estimated length of time based on other listed characteristics in the paper.

Application Domains, Participants, and Study Location.

- *Domain*: The main domain of the study. We classified studies into the following domains: physical health, mental and cognitive health, education, entertainment, service and workplace, general purpose (i.e., home robots that provide several general uses), and ASD. Although studies in the ASD domain can overlap with physical and mental health, we chose to identify it as its own distinct category due to the prevalence of studies focused on ASD.
- *Participants*: The number of data-generating participants in the study. We only consider the number of participants that interacted with a physical robot. For example, in [26], only 20 out of 61 participants interacted with an embodied robot; the other participants interacted with systems that did not include a robot. Therefore, we considered the number of participants to be 20. We also do not consider participants who were excluded from the analysis due to varying reasons such as technical difficulties or leaving the study prematurely.
- *Age Group*: The primary age group of the participants. We grouped participants into different age groups: infants (0–3 years), children (3–12), teens (13–17), adults (18–65), elderly (65+), and mixed. When over 80% of participants fell into one particular category, we classified it as the majority category instead of mixed.
- *Location*: The primary location where the study took place. We categorized each study into the following locations: *home*, *school*, *care home*, *hospital*, *rehabilitation facility*, *daycare*, *laboratory*,

workplace, *museum*, and *other*. We considered a paper as taking part in the *workplace* category when the study's location was an office, business location, or similar. For example, a corporate office for employees or a car rental agency. While schools and hospitals are workplaces for teachers and doctors, we categorized these papers as their own locations, given the uniqueness of these environments.

— *Country*: The primary country where the study occurred or participants were recruited.

Study and Robot Qualities.

— *Robot Platform*: The primary robot(s) or robotic platform(s) used in the study.
— *Autonomy*: The level of autonomy of the robot. We classified whether the robot was designed to be autonomous, semi-autonomous (mix of autonomous and operated by a person), or non-autonomous.
— *Interaction Dynamic*: How many people participated in the interaction. We classified interactions into six primary types: *dyadic* (one robot and one person), *triadic* (one robot and two people), *group* (one robot and a group or family), *observer* (one person observing others interacting with the robot; e.g., [145, 154]), *mixed* (varied among the previous categories; e.g., [80, 135]).
— *Personalization and Adaptation (Y/N)*: The level of personalization and adaption of the robot. We labeled *Yes* if the robot adapts its actions to that of the user or personalizes them longitudinally, and *No* if otherwise.

Results and Measures.

— *Qualitative Results (Y/N)*: Whether the study had any qualitative results. We labeled *Yes* if the paper presents a discussion or analysis of any qualitative findings, and *No* if the paper does not. The definition of qualitative for this review includes non-numerical subjective measures, including but not limited to: open survey responses, interviews, ethnographic methods, and behavioral observations. If some numerical data was collected but the predominant analysis, results, and discussion presented were qualitative, the paper was labeled as qualitative only.
— *Quantitative Results (Y/N)*: Whether the study had results from a quantitative analysis presented. We labeled *Yes* if the paper presents a statistical analysis of quantitative findings, or *No* if the paper does not report any quantitative findings. At times, studies presented quantitative analyses on qualitative results. These were judged as quantitative in addition to qualitative if the results presented multiple numerical analyses toward the authors' claims. Studies presenting descriptive metrics (e.g., interaction counts or survey averages) alongside predominantly qualitative results were not considered quantitative. A small number of studies that collected longitudinal data for training machine learning models were also classified as quantitative.
— *Conditions (Y/N)*: Whether the study had conditions in pursuit of statistically significant results. We labeled *Yes* if the paper presents a statistical analysis of separable and controlled conditions, or *No* if the paper does not have conditions or does not perform a statistical analysis of study conditions.
— *LTE (Y/N)*: Whether the study reported a user's engagement over time during the study. We labeled *Yes* if the paper reports LTE, or *No* if not.
— *Engagement Measure*: Details of how engagement was measured. We labeled the different types of engagement measures each paper used (e.g., user-reported survey, study drop-out rate, continued interaction with the robot) if user engagement is reported in the paper.

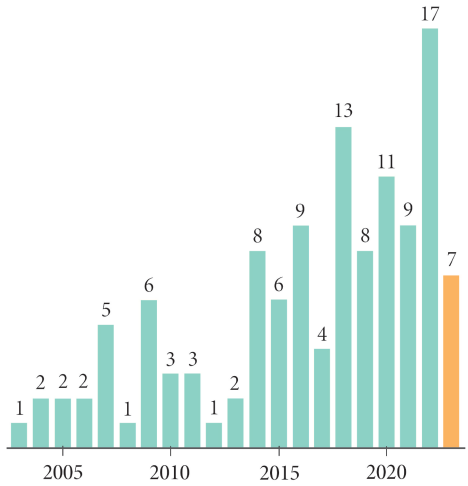


Fig. 2. Number of studies meeting our review criteria by year. The bar for 2023 highlights the partial data collected between January and April 2023, in contrast to the more complete data from previous years.

4 Findings

This section contains an analysis of the 120 studies included on the basis of our aforementioned criteria (Section 3). All studies, along with a subset of the key characteristics, can be found in the Appendix. As seen in Figure 2, the number of such studies has been increasing steadily over the past two decades. The types of long-term interactions investigated have, in turn, also expanded widely. Here, we report on the defining characteristics of long-term studies since 2012, as outlined in Section 3, including study lengths, participant types, interaction locations, interaction modalities, and application domains. We additionally highlight methods for measuring LTE and for improving long-term outcomes, such as personalization and adaptation. To analyze trends over time, we often report shifts in different study characteristics across two decades: 2003–2012 and 2013–2023.

4.1 Temporal Qualities

In this section, we first analyze the myriad ways researchers have reported on the temporal characteristics of long-term HRI studies.[2] These include the overall study period, the number of interactive sessions with a robot, the lengths of these sessions, and the total interaction time per user over the course of the study. As seen in Figure 3, these characteristics vary widely. For instance, some studies report on three sessions given across several weeks (e.g., [22, 31, 97]), while others (e.g., [8, 138]) investigate tens of sessions over multiple months.

4.1.1 Study Period. The *study period* is defined as the number of days from the first study session a participant has with the robot system to the final session with the system. Given the intentional breadth of our inclusion criteria for long-term interaction, we find a large variety of study periods (M = 48.4 days, SD = 70.2 days). These ranged from the minimum 3 days [151] to 570 days [73], as illustrated in Figure 3(a). For seven papers, which total eight studies, the study period was not reported [27, 29, 31, 72, 85, 99, 129].

[2]It is important to note that not all studies in our corpus make claims of being "long-term" despite meeting our qualifications for inclusion. Some authors utilize terms such as "multi-session" (e.g., [38, 97]) or "longitudinal" (e.g., [13, 70]) to describe the nature of their study rather than "long-term." Our corpus includes all studies that met our criteria, separate from any authors' claims.



Fig. 3. Comparison of study length and frequency between 2003–2012 and 2013–2023. The distribution of the studies based on study length is shown in (a). For studies that were sessions-based rather than free use of the robot system, distributions of the studies based on the number of sessions (b) and session length in minutes (c) are shown. Lastly, the distribution of studies based on total study length in hours as reported or estimated by the reported number of sessions and session length is shown in (d). We compare the distributions across two decades to examine emerging trends in light of the rapid growth in long-term HRI research.

Investigating the primary motivations of studies that are shorter versus longer, we observe several patterns. For studies shorter than 2 weeks ($N = 17$), we find that the goal of the study was often exploratory or in the evaluation of a new technique. For instance, nine of these studies aimed to investigate user perceptions and usability of a robot in a new environment (e.g., [2, 39, 58, 107, 158]) or the relationship built between participants and the robot (e.g., [1, 68]). Four of the studies that were shorter than 2 weeks sought to validate new technical methods (e.g., [3, 32, 38, 151]), and four studies show participant improvements on a new skill or task (e.g., [2, 117]). In contrast, many of the studies that occurred over the span of 3 months or longer ($N = 14$, [8, 26, 36, 53, 72, 73, 102, 106, 120, 138, 139, 148, 160, 165]) explored the adoption, engagement, and usage patterns of robots in long-term settings. Such a divide makes sense, as studies aimed at longer-term engagement and usage necessitate that the study period lasts beyond any novelty effects.

Comparing study periods by decade, we find that more studies with shorter periods have occurred in the recent decade (2013–2023). For example, of the 23 studies in the first decade (2003–2012) that report a study period, only 5 (21.7%) of them have a study period of less than 1 month. In contrast, 48 (53.9%) of the 89 studies reporting a study period in 2013–2023 report one that is less than 1 month. The increase in long-term studies over shorter periods of time can be considered in one of two manners: either the lengths of long-term studies have decreased over time, or there has been a natural progression from single-session studies into multi-session studies. We find there is an emergence of studies that do not intend to examine the long-term effects of the robot interaction, but rather to investigate certain elements of usability and instruction that require a handful of sessions instead of just one (e.g., [2, 158]).

4.1.2 User Sessions. In addition to the number of days of a study, we report on the duration and number of user sessions with a robot. These measures capture the amount of direct interaction a participant has with a robot. A *session* is characterized as a specific duration when a participant is expected to interact with the robot system. Eighty-six studies (71.7%) in our corpus reported a particular number of sessions per participant and are thus considered to be *session-based*. For sessions-based studies, we find that the number of sessions varies widely from 3 sessions [10, 22, 31, 56, 80, 81, 85, 107] to 1,559 sessions [106], as illustrated in Figure 3(b). The average number of sessions for a study in our corpus, excluding the 2017 paper by Ostrowski et al. [106] as a clear outlier, is 15.5 sessions (SD = 23.9). For 21 studies, the number of sessions per participant was not reported.

Of the session-based studies, 77 (64.2% of the overall corpus) report a specific session length. For these studies, we find that the duration of each session varies widely from 1 minute [68, 106] to 3 hours [72] (M = 26.9 minutes, SD = 28.2 minutes). The remaining 32 studies did not report a specific time duration of the sessions. Similar to findings over time of study periods (Section 4.1.1), more studies occurred with fewer sessions and shorter session lengths during 2013–2023 than during 2003–2012 (Figure 3(c)).

In our corpus, 10 studies are neither session-specified nor session-based because the number and the length of sessions were not reported. An additional 11 studies reported that the user sessions were self-directed *daily use* (N = 9 [23, 30, 83, 87, 93, 102, 111, 119, 124]) or *free use*[3] across the entire study period (N = 2 [7, 39]).

4.1.3 Total Interaction Time. Three studies directly reported the total interaction time per participant [1, 55, 135]. Many of the remaining studies reported session lengths alongside the number of sessions. In order to compare sessions-based and free-use studies on the amount of participant-to-robot time, we present an estimate of the total interaction time per participant for each study. We estimated the total interaction time by multiplying the number of sessions by the session length for each study that reports both values (N = 76). Together, the 79 studies demonstrate an average interaction time per participant of 553.6 minutes (SD = 1,173.5), ranging from 10 minutes [122] to 120 hours [8]. As seen in Figure 3(d), we observe a slight decrease in the number of studies with less than 2 hours of interaction length published during 2013–2023 (N = 7, 41.2%) compared to 2003–2012 (N = 31, 50.0%).

4.2 Application Domains, Participants, and Study Locations

In this section, we classify our corpus of long-term studies based on their domain, participant information, and location (Figure 4). While the first few studies in long-term HRI focused primarily on education [61, 79], behavioral interventions for children with ASD [120, 129], or the general perception of robots [135, 145], more recent studies cover a wide range of applications. For instance, providing service to support employees in an office environment [155] or improving cognitive health by encouraging exercise [128]. Section 4.2.1 outlines seven categories of primary application domains: education, mental and cognitive health, general purpose, physical health, entertainment and game-play, and service and workplace.

In Sections 4.2.2 and 4.2.3, we additionally classify studies based on their participant age ranges (ranging from infants to the elderly) and the number of participants with reported data. Occasionally, the domain dictates the participant type of the study, such as children using an education robot, though this is not always the case. Often, the location of a study is related to its domain

[3] We define "free use" as users having complete control over when and how to use the robot. We acknowledge some researchers may not have reported casual guidance given to participants and that some participants may have felt implicit pressure to use the robot regularly.

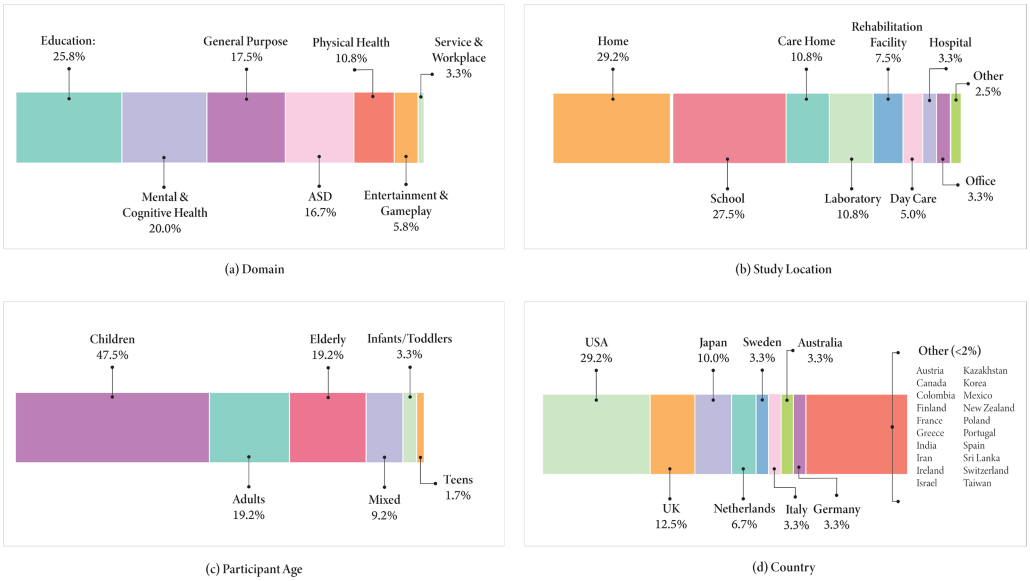


Fig. 4. Distribution of study qualities, an overview of key attributes in long-term HRI studies conducted over the past two decades. The distribution is presented across four main dimensions: study domains (a), study locations (b), participant age groups (c), and countries where the studies were conducted (d). It is important to note that the largest category, "Other," in (d) encompasses 20 countries, each contributing less than 2% to our current dataset.

and participant types, such as a rehabilitation robot for the elderly being deployed into a care home. Given that most long-term HRI research is motivated by preparing robots for real-world applications, considering a study's environment in relation to its participants and domain can be valuable.

4.2.1 HRI Domains. We find that long-term studies in our corpus fall into the following primary domain classifications: *Education* ($N = 31$, 25.8%), *Mental and Cognitive Health* ($N = 24$, 20.0%), *General Purpose* ($N = 21$, 17.5%), *ASD* ($N = 20$, 16.7%), *Physical Health* ($N = 13$, 10.8%), *Entertainment and Game-Play* ($N = 7$, 5.8%), and *Service and Workplace* ($N = 4$, 3.3%). A summary of areas of study per domain is below:

— *Education ($N = 31$, 25.8%)*: Within studies in the *Education* domain, robot interactions tend to be focused on tutoring or teaching skills in reading (e.g., [15, 19, 165]), math (e.g., [10, 56, 114]), language (e.g., [150, 158, 162]), handwriting (e.g., [18, 54]), and other common academic subjects (e.g., [60, 77]). For the majority of the studies in this domain ($N = 26$, 83.9%), the population of interest is children between the ages of 3 and 12, with the robot most often in a school ($N = 17$) or the home ($N = 6$).
— *Mental and Cognitive Health ($N = 24$, 20.0%)*: Studies within this domain can be organized into three categories: *elder care*, *condition-specific rehabilitation*, or *general wellness*. Ten studies target cognitive stimulation for elderly participants, ages 65 or older, either in the individual's home ($N = 2$; [2, 87]) or care homes ($N = 8$; [24, 33, 39, 40, 42, 156, 157]). Condition-specific studies ($N = 8$; [8, 30, 73, 102, 111, 126, 158, 159]) examine the robot interactions with users with a diagnosed cognitive disorder such as Dementia, Down Syndrome, or Alzheimer's. The

remaining seven studies [13, 58, 59, 68, 124, 128, 144] primarily examine improving mental and cognitive health in typically developing adults.

— *General Purpose (N = 21, 17.5%)*: Studies in the *General Purpose* domain are generally focused on investigating user perceptions of, and user engagement with, the robot. For instance, studies in this category have shown that users enjoyed the robotic interactions (e.g., [27, 57]), were motivated to interact with the robot (e.g., [29, 86, 96]), and demonstrated changes in their perception of (e.g., [72, 107, 146]) or engagement with the robot throughout the study (e.g., [70, 106, 119, 135, 160]). Further studies examined methods for behavioral personalization and adaptation of the robot due to user preferences (e.g., [3]), used the robot as a tool for understanding the target population better (e.g., [72, 125]), or evaluated the feasibility of the robot to provide living assistance (e.g., [7]). *General Purpose* studies target a wide range of participant groups, from infants to seniors, in a variety of settings such as the home, a laboratory, daycare, or school.

— *ASD (N = 20, 16.7%)*: Studies in this category explore the use of robots with individuals with ASD. Several studies in the *ASD* category would traditionally fall within the Mental and Cognitive Health or Education categories; however, we treat these studies as a separate category due to their large number. Within studies in the *ASD* domain, most studies either investigate skills development or validation of new measurement or prediction techniques. Those in the skills group often investigate interactions such as verbal communication or social skills (e.g., [94, 131, 138, 163]). Studies about new measurement techniques focus on validating methods that predict engagement, build user models, or develop metrics for specifically the ASD population (e.g., [12, 45, 55, 99, 112, 129]). The majority of *ASD* studies focus on interactions with children ($N = 16$, 80.0%) in a school ($N = 6$) or home ($N = 5$) setting. Within the past half-decade specifically (2018–2023), we observe new interest in long-term, in-home interactions [21, 55, 108, 117, 131, 136], as well as developing long-term social skills interventions for teenagers [20, 138] and adults with ASD [117, 139].

— *Physical Health (N = 13, 10.8%)*: Studies within the *Physical Health* domain primarily fall into two categories: *general wellness* and *condition-specific rehabilitation*. For general wellness, studies focused on supporting exercising and healthy eating ($N = 8$, e.g., [28, 66, 137, 147]) across a wide range of participant age groups, but primarily centered on adults and older populations ($N = 5$). These robots are often deployed in a home or laboratory setting. Condition-specific studies ($N = 5$) instead support individuals recovering from a medical event such as surgery, stroke, or a diagnosis of a life-impeding condition [35, 49, 53, 89, 151]. Robots for condition-specific support are more often found in hospitals or rehabilitation centers, almost entirely for elderly users ($N = 4$).

— *Entertainment and Game-Play (N = 7, 5.8%)*: The studies in this domain [36, 38, 62, 91, 110, 148, 149] primarily investigate the potential for engaging and maintaining user interest with a robotic system via games or other interactive media. A majority of studies in this domain ($N = 5$) focus on children younger than 12 years old in a range of settings, such as a laboratory, school, daycare, or home.

— *Service and Workplace (N = 4, 3.3%)*: In our corpus, there are currently only four studies [122, 135, 144, 154] that investigate robots in the *Service and Workplace* domain. These robots are designed to perform tasks related to workplace offices or other locations that offer services to customers. For example, Vishwanath et al. [154] explored how a humanoid robot receptionist could improve staff productivity in an office setting. While many studies included in the *Education*, *Physical Health*, and *Cognitive Health* domains are conducted in settings that

arguably offer services (e.g., schools and health clinics), these studies focus on a specific educational or health outcomes rather than on the workplace itself.[4]

Assessing study domains over the past 20 years, we find several trends over time. While the first 5 years of long-term studies (2003–2007) consisted of a small number of studies ($N = 12$), these studies were evenly distributed across six domains: *Mental and Cognitive Health*, *ASD*, *General Purpose*, *Entertainment and Game-Play*, *Education*, and *Service and Workplace*. In contrast, 50% of the studies published in the following 5 years (2008–2013) were in the *General Purpose* domain. A newfound interest in the *Physical Health* domain emerged in 2009 [66] and has steadily grown to the present year. *Physical Health*, *Mental and Cognitive Health*, *Education*, and *ASD* all have exhibited steady growth over the past 15 years, whereas the *Service and Workplace* domain has only recently emerged for long-term study.

Many of these domains lend themselves inherently to long-term study, for instance, with *Education*, *ASD*, and to a certain extent *Mental and Cognitive Health* and *Physical Health* studies, often the goal of the robotic intervention is to support skill acquisition. Developing a new skill, by its nature, requires time to develop and must be studied as such to show true value. Similar to skill acquisition, therapies such as rehabilitation and cognitive exercises also require the repeat practice of a certain activity. In contrast, *General Purpose* studies are often more related to the exploration of how social robots may integrate into our homes and daily lives. This type of study requires longitudinal study, as real-world robot applications will not happen in a single session. For researchers interested in understanding adoption of robots in society, long-term study is an important step. In the less-represented domains of *Entertainment and Game-Play* and *Service and Workplace*, such real-world requirements still apply, but there may be other challenges. As the *Service and Workplace* domain is still relatively new in long-term exploration, researchers may opt to first explore the problem space, initial usability, and initial user perception with a single-session study before transitioning to long-term study. The *Entertainment and Game-Play* domain offers an easy way to test user perceptions of a robot and algorithmic contributions around engagement; however, researchers may have less interest in this domain now as more specific and beneficial use cases for social robots have emerged (e.g., supporting social skills development, rehabilitation, home assistive tasks).

4.2.2 Participant Types. For the purposes of this review, we classify participants based on their age group corresponding to the general stages of psychological development ranging from infants to seniors. Across our corpus, 57 studies (47.5%) feature children (ages 3–12) as the primary participant type, 23 (19.2%) feature adults (ages 18–65), 23 (19.2%) feature seniors (above age 65), 4 (3.3%, [95, 96, 126, 148]) feature infants or toddlers (under age 3), and 2 studies (1.7%, [20, 138]) feature teenagers (ages 13–18). The remaining 11 studies (9.2%, [8, 13, 19, 27, 29, 32, 35, 66, 122, 150, 160]) report outcomes of the robot interaction to describe mixed demographics across participant age groups.

It is expected that the domain of a study is related to the population of interest of the study, or vice versa (Figure 5(b)). For example, the majority of *Mental and Cognitive Health* studies target the senior population ($N = 16$, 64.0%) because many studies in this domain target cognitive stimulation for users with a diagnosed cognitive disorder such as Dementia or Alzheimer's. Similarly, the vast majority of studies within the *Education* and *ASD* domains focus on children ($N = 42$, 82.4%). Studies in the *General Purpose* and *Physical Health* domains contain a diverse range of participant age groups.

[4]For example, we classify work such as Rueben et al. [122] within the *Service and Workplace* domain because the authors investigated the impact of a mobile shoe rack at a yoga studio on client satisfaction. Had the authors investigated the impact of the robot on the individual performance of yoga students at the studio, we would have classified this work under *Physical Health*.

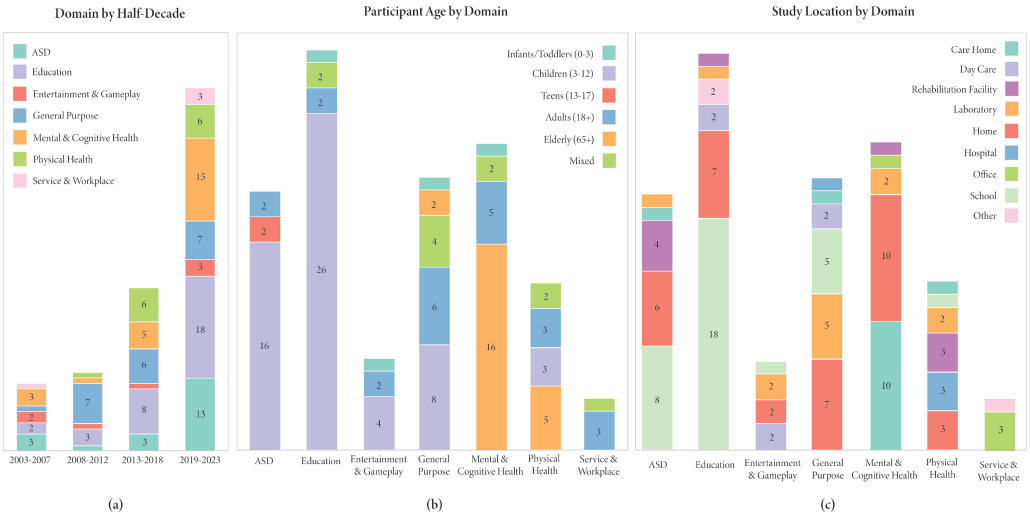


Fig. 5. Distribution of study qualities by domain. The change in the distribution of study domains every 5 years (a), the distribution of participant age groups by domain (b), and the distribution of study location by domain (c) are illustrated.

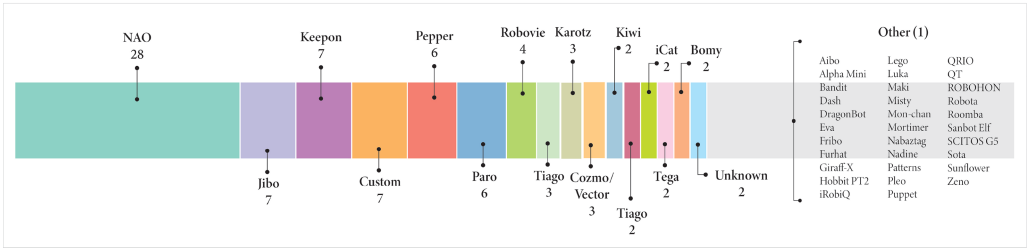


Fig. 6. Robot platforms. Illustrated is the distribution of robots employed in long-term HRI studies. The largest category, "Other," encompasses 34 platforms that were each represented only once in our dataset.

4.2.3 Participant Counts. In our corpus, the number of participants per study varies widely (M = 24.8, SD = 30.4), from a single participant [2, 53, 126, 135] to as many as 228 participants [61]. The distributions of participant counts differ by domain: *ASD* (M = 14.8, SD = 13.0), *Education* (M = 36.8, SD = 44.4), *Entertainment and Game-Play* (M = 15.3, SD = 10.5), *General Purpose* (M = 35.8, SD = 32.8), *Mental and Cognitive Health* (M = 19.5, SD = 21.9), *Physical Health* (M = 12.8, SD = 9.4), and *Service and Workplace* (M = 11.3, SD = 10.8). The majority of the 14 studies with less than five participants [2, 12, 22, 31, 35, 53, 68, 72, 94, 120, 126, 135, 136, 138] were focused on providing therapies to protected or sensitive populations, such as those with diagnosed physical or mental disabilities (N = 10). On the other hand, for the six studies with more than 100 participants [27, 29, 51, 61, 85, 155], research was performed either through educational systems [61, 85, 155] or with an off-the-shelf commercial home robot [27, 29]. The four largest participant studies [51, 61, 85, 155] were conducted in both an education setting and for children. While these studies are outliers in their respective domains due to their sample size, performing studies through established school structures likely provides easier access to large numbers of student participants. Similarly, using a pre-existing commercial home robot with minimal customization likely reduces

the burden of designing and deploying systems in the unstructured and diverse environment that is the users' homes.

Participant count influences the types of statistical tests that can be run based on having an adequate sample size. Among studies that report statistical results ($N = 70$; e.g., either between experimental conditions or between different population segments), the distribution of participant sample sizes tends toward larger counts ($M = 31.7$, SD = 35.6), with 65.7% with 15 participants or more. Among the remaining studies that do not conduct statistical analyses ($N = 50$), the distribution of participant counts shifts toward smaller counts ($M = 17.0$, SD = 19.3), with 72.0% below 15 participants. The motivation, expected outcomes, and structure of a specific study can inform the analyses (e.g., qualitative, quantitative, or statistical) researchers choose to conduct. In Section 4.4.1, we examine the potential factors influencing the results reported in our corpus of studies.

4.2.4 Study Locations. As long-term studies seek to emulate more real-world scenarios for HRI, the setting where robot interactions occur is an important study design consideration. Studies performed in a research facility or laboratory have the advantage of controlling for environmental variables in order to isolate specific components of interaction. In contrast, "in-the-wild" environments such as homes or classrooms are more likely to produce findings that are generalizable to real-world interactions and contexts. In-the-wild environments are dynamic, unstructured settings, however, and thus often have greater technical demands and challenges for robot use, design, and deployment.

In our corpus, the majority ($N = 103, 85.8\%$) of studies were performed in-the-wild, often either conducted in homes ($N = 35, 34.0\%$) or in schools ($N = 33, 32.0\%$). The remaining 17 studies (14.2%) were conducted in a laboratory research setting. Of these 17 studies, 10 studies [13, 32, 69, 72, 91, 99, 107, 126, 147, 150] were performed in a laboratory environment that intended to simulate a natural, in-the-wild environment, such as the home or workplace. Figure 4(b) presents a full breakdown of location types.

As with participant types, the location of a study typically aligns with the study's domain, as shown in Figure 5(c). For instance, many educational robots are deployed in schools (e.g., [6, 18, 60, 84, 115, 116, 140]), with a subset deployed in the home for tutoring outside of the classroom (e.g., [92, 93, 164, 165]). Similarly, robot interactions that target physical health rehabilitation are often conducted in hospitals or care facilities (e.g., [1, 35, 53]).

Beyond the immediate environment of the robot, the geographic and cultural environment may influence the results of long-term studies. According to the distribution of countries in our corpus, we find that 41.7% ($N = 50$) of studies were conducted in Europe, primarily in the UK ($N = 15$); 32.5% ($N = 39$) of the studies were conducted in North America, primarily in the United States ($N = 35$); 18.3% ($N = 22$) of the studies were conducted in Asian countries, primarily in Japan ($N = 12$); and 5.0% of the studies were conducted in Australia ($N = 4$) or New Zealand ($N = 2$). The remaining studies do not report the location [34, 99, 154].

4.3 Study and Robot Qualities

Researchers have explored many different modalities of long-term robot interactions across domains, participant types, and study locations. We report here on patterns in our corpus based on the robot used, the robot's level of autonomy, the interaction dynamic (e.g., dyadic, triadic, group), and the use of behavioral adaptation and personalization of the robot. By classifying these types of HRI qualities, we can observe the landscape of the many varying ways long-term HRI can manifest.

4.3.1 Robot Platform. A robot's form and affordances highly impact the types of interactions researchers can study. There are many benefits to using a commercially available (i.e., "off-the-shelf") robot, from physical robustness to the comparability of results across studies using the same robot. At the same time, off-the-shelf robots may have limitations in their design that can prohibit achieving certain research goals (e.g., Ramnauth et al. [117] utilized supplementary sensors since certain hardware of the popular Jibo robot [52] was inaccessible).

As shown in Figure 6, we find that the NAO robot [47] is most popular for long-term study (e.g., [6, 12, 45, 54, 60, 97, 147]), with 30 (25.0%) long-term studies utilizing this robot platform. The next most popular robots include: Keepon [11] (N = 7, 5.8%; [72, 80, 81, 84, 128]), Jibo (N = 7, 5.8%; [19, 58, 59, 106, 117, 131, 164]), Paro [112] (N = 7, 5.8%; [42, 73, 139, 156–159], and Pepper [146] (N = 6, 5.0%; [1, 13, 31, 35, 83, 119]. Beyond these 5 off-the-shelf robots, 7 studies (5.8%; [66, 70, 95, 122, 124, 135, 149]) feature novel, custom robots, or prototypes. We provide a brief description of these most popular robots below:

- *NAO* (25.0%; N = 30): NAO [47] is a small, programmable, table-top humanoid robot with a rich sensor suite and built-in interactions for complex natural language and facial and gesture recognition. Its popularity in HRI research is likely due to its diversity of affordances and capabilities, the ability to easily purchase the robot off-the-shelf, as well as its existing popularity lending itself to easier research reproducibility.
- *Keepon* (5.8%; N = 7): Keepon [11] is a small, minimal table-top robot with a soft bulbous body that can move side-to-side, up-and-down, and rotationally. It features audio and visual inputs and can be custom programmed, while being relatively affordable in comparison to more complex robots.
- *Jibo* (5.8%; N = 7): Jibo [52] is a table-top robot with a more cylindrical form factor and a circular display that can rotate expressively in multiple directions. Similar to NAO, it is equipped with audio and video inputs for complex natural language and computer vision.
- *Paro* (5.8%; N = 7): Paro [109] is a soft therapeutic robot designed to look like a seal and responds to touch, light, and sound in order to express certain emotional states.
- *Pepper* (5%; N = 6): Pepper [143], a humanoid on wheels, is the largest form factor of the top four robots, reaching almost human height with the ability to navigate an environment. It also features expressive gestures, cameras, and a conversational interface, however the robot's cost is significantly higher.

4.3.2 Robot Autonomy. Future real-world robots deployed for long-term use in homes, workplaces, and other areas of daily life will need to be fully autonomous. However, achieving autonomy in natural, in-the-wild environments presents many technical and perceptual demands. As a result, making the choice of whether to automate a robot for research, and how much, is a key study design consideration. In our corpus, researchers have used non-autonomous, semi-autonomous, and fully autonomous deployments. While a *fully autonomous* robot is closer to real-world scenarios where the robot's architecture alone solely directs its behaviors, researchers may choose to implement a *non-autonomous* system via Wizard of Oz[5] or teleoperated techniques. Such techniques enable researchers to imply complex levels of autonomy without the technical requirements of building an autonomous system. In between non-autonomous and fully autonomous systems lies *semi-autonomous* systems, in which researchers have access to a human-in-the-loop method of updating robot behaviors while deployed. Such a system can enable researchers to correct robot errors or introduce deeper levels of personalized interactions.

[5]A method in which participants interact with a robot system that users believe to be autonomous, but is actually operated or partially operated by another human.

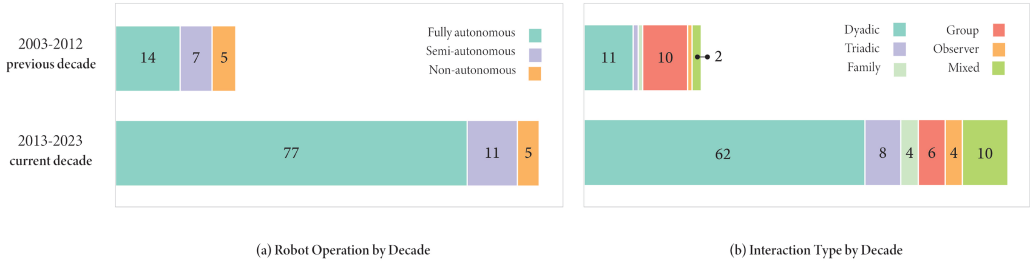


Fig. 7. Distributions of robot operation (a) and study interaction dynamic (b) by decade.

We find that 91 studies (75.8%) in our corpus utilize *fully-autonomous* systems, 18 (15.0%) use *non-autonomous* systems, and 10 (8.3%) use *semi-autonomous* systems. We further observe a remarkable growth of 450% in autonomous design between 2003–2012 ($N = 14$) and 2013–2023 ($N = 77$). We do not observe this significant growth in semi-autonomous or non-autonomous design between 2003–2012 ($N = 12$) and 2013–2023 ($N = 16$).

Such growth may be attributed to a desire by the field to explore the questions and challenges surrounding interactive robot behaviors in real-world settings. For instance, maintaining LTE without an assisting researcher at hand is a known challenge. As discussed in Section 4.3.1, designing robot behaviors toward this end depends on the nature of the robotic system, as the affordances and capabilities of a robot impact the type and diversity of interactions possible. A simpler robot may be easier to implement, but may have limitations in the amount of behaviors possible to engage a user longitudinally. Conversely, a more complex robot may have so many options that a user may be overwhelmed by choice or not know the best ways to interact. Another core element of behavior design for autonomy is how pro-active the robot should be in responding to its environment, or in how much to try and entice a user to engage with it. A robot that is entirely passive risks being forgotten about. However, a robot that is too pro-active can become quickly annoying or in the way, if it is unable to determine the best contexts for engaging. In these ways, and many more, autonomous robot behavior design is crucial and ripe for long-term exploration.

4.3.3 Interaction Dynamics. HRI can take on many forms with respect to the social context of the interaction. We investigate the differences in studies that are *dyadic* (one-on-one between the human and robot), *triadic* (two humans interacting with one robot), *family* (one robot in a household interacting with multiple family members), *group* (one robot interacting with multiple humans in a group setting), *observer* (the human is observing a different human or group interacting with a robot), and *mixed* (a combination of any of the prior four categories) interactions. Overall, the majority of long-term studies are dyadic ($N = 73$, 60.8%), with 16 studies (13.3%) of group interactions, 12 studies (10.0%) of mixed interaction dynamics, 9 studies (7.5%) of triadic interactions, 5 studies (4.2%) of family interactions, and 5 studies (4.2%) of observer interactions. In comparing the distribution of interaction dynamics between the prior decade and 2013–2023 (Figure 7), we see a shift in the percentage of dyadic studies and group studies. Dyadic studies have grown in representation, from 42.3% to 66.0%, while group interaction studies have decreased in representation, from 38.5% to 6.4%.

One may have considered that as the field of long-term HRI matures, we would see a rise in more diverse interaction dynamics beyond dyadic—particularly since group interactions are rising in popularity in HRI [134]. There are many possible reasons for this trend. For one, longitudinal studies are challenging to run and deploy, and it is usually easier to run a dyadic one. Researchers need only design interactions for one user, rather than many, and need not consider the more

complex group social factors of an interaction. Recruiting and consent also become simpler for a single user rather than many in a group. Another potential explanation corresponds to the rise in health- and education-focused studies over the last decade. These domains may lend themselves more readily to isolating one-on-one robotic interactions with a single student or patient. Even though parents, teachers, and therapists are often key stakeholders, a key end-measure is often from a single individual's performance on a new skill, activity, or task. A general purpose robot will more often need to, in contrast, appeal to the entire family unit. A workplace robot may similarly need to appeal to an entire office. Finally, we also see an increase in quantitative studies (as discussed in Section 4.4.1) over the years. Measuring quantitative metrics may be easier to do in a one-on-one user-to-robot interaction, as it requires less individual identification and tracking. In a similar vein, we see a rise in studies with personalization and adaptation methods (Section 4.3.4), which is easier to design for and track over a single user than many.

4.3.4 Personalization and Adaptation. Personalization [50] and adaptation [4], two methods to tailor the robot's behaviors and interaction with a particular user, have been shown to improve outcomes for long-term HRI [21, 84, 86]. Such methods foster a sense of rapport, trust, and enjoyment in users, leading to more meaningful and engaging interactions. In addition, personalization and adaptation can allow the robot to maintain its relevance and usefulness over time, adapting to varying contexts, user preferences, and situational demands. We find that 47 studies (39.2%) in our corpus utilize personalization or adaptation in their robot design. They are distributed relatively evening across domains.

Some common methods for personalization and adaptation include adjusting task difficulty based on prior performance (e.g., [114, 116, 131]) or utilizing affect recognition or physiological data to improve user engagement (e.g., [33, 38]). Many studies specifically investigate the effects of personalization and adaptation methods as their core research aims (e.g., [3, 10, 15, 21, 31, 56, 84–86, 91, 116, 136]. These studies have demonstrated that adaptive or personalized system results in increased user satisfaction, improved performance and effectiveness, enhanced user engagement, and improved flexibility in changing environments.

Overall, we find a rise in studies with personalization and adaptation over time, with 43% ($N = 40$) in the current decade versus 23% ($N = 6$) versus the previous decade. This may be caused by a few factors. One, as long-term studies become more common in a domain, they may transition from more exploratory to more targeted in their measures and techniques. Given that engagement is a known challenge, techniques to specifically counter them are warranted. Additionally, technology advances over the past 20 years have provided far more tools to support individual tailoring of HRIs, from computer vision to biometric feedback.

4.4 Result Types and Measures

Studies in long-term HRI are varied in their type of research findings, with a mixture of quantitative and qualitative insights. In our corpus, case studies and ethnographies are present alongside quantitative studies and those that seek statistically significant differences between conditions. In this section, we assess the ways long-term HRI researchers have pursued different types of measured results. We additionally report patterns on two specific measures that are particularly relevant to long-term studies: pre/post analyses and LTE.

4.4.1 Qualitative Versus Quantitative Approaches. A fundamental decision that long-term HRI researchers must make is whether to pursue a qualitative or quantitative methodology, or a blend of both approaches. This choice is multi-faceted and hinges on several motivations that should be carefully weighed. Qualitative research often involves exploratory investigations of research questions rather than setting experimental conditions to test a hypothesis. Reporting qualitative data

can provide insights into user preferences, challenges, and needs as well as serve as a foundation for new hypotheses or theories [117, 122, 149]. Examples of qualitative measures in our corpus include interviews, open-ended questionnaires, experimenter observation, video labeling, and diaries. In contrast, quantitative research typically facilitates objective measurements and evaluations of specific outcomes or performance metrics. Examples of quantitative measures in our corpus include: test performance, game scores, robot usage rates, interaction type counts, standardized surveys, custom surveys often containing Likert or similar scales, and so on. Such metrics can be crucial in assessing the effectiveness or efficiency of robot interventions, measuring user perceptions, or evaluating task completion rates.

In our corpus, 49 (40.8%) studies report both qualitative and quantitative results, 33 (27.5%) report only qualitative results, and 25 (20.8%) studies report only quantitative results. The measures and metrics employed by researchers are often specific to the domain, participant type, and setting, and we encourage readers to consult the Appendix to find long-term studies with similar deployments as examples.

We find that the percentage of studies with quantitative results (either quantitative only or both quantitative and qualitative) has risen over the years, with 58.3% ($N = 70$) in the current decade versus 13.3% ($N = 16$) in the prior. With regard to domains with at least 10 total studies: we find that 77.4% of *Education* studies, 76.9% of *Physical Health* studies, 75.0% of *ASD* studies, 72% of *Mental and Cognitive Health* studies, and 68.2% of *General Purpose* studies have quantitative results. In contrast, the percentage of studies with qualitative results (either qualitative only or both quantitative and qualitative) has decreased from the previous (76.9%, $N = 20$) to the current decade (67.0%, $N = 63$). We find the distribution of these result types varying more widely across domains in comparison to quantitative results. The large majority of *Mental and Cognitive Health* (84.0%, $N = 21$), *Physical Health* (76.9%, $N = 10$), and *ASD* (75.0% $N = 15$) studies contain qualitative studies, whereas a smaller majority or minority of *General Purpose* (63.6%, $N = 14$) and *Education* studies (48.4%, $N = 15$) contain qualitative results.

These differences may be explained by the inherent nature of the domains. For instance, *Educational* contexts often feature student testing and standardized measures of academic skills, which are easy to quantify. This domain has also been established for longer, and may be moving away from initial exploratory studies to more targeted proofs of efficacy. In contrast, *Mental and Cognitive Health* studies are a newer domain that may seek to gain early qualitative insight into new population needs. For *ASD*, there is a combination of standardized benchmarks as well as high individual variance within the population and the need for qualitative assessment from caregivers. Both qualitative and quantitative measures may be particularly helpful for this domain.

4.4.2 Study Conditions. For quantitative studies, conducting experimental conditions can help establish causal relationships and make comparisons between different conditions. By including a control condition, researchers can determine whether the experimental manipulation or intervention likely causes the observed effects. In our corpus, 47 studies (39.2%) used quantitative methods with preset experimental conditions in pursuit of statistical significance. Often, these conditions were on (a) the use of a robot versus no robot (e.g., using a robot with a set of smart sensors for elder care versus just the sensors themselves [87]), (b) inclusion of a specific interaction or approach versus without (e.g., including personalization or not with a general purpose robot for children [86]), or (c) between different types of populations (e.g., special needs children and typically developing children [161]). We find that the percentage of studies with conditions-based experimentation has increased in the current decade (43.6%; $N = 41$) versus the previous (23.1%; $N = 6$), likely following the trend of increasing quantitative studies.

4.4.3 Pre/Post Experiment Analyses. To evaluate the long-term effects of robotic interventions, a common approach is to analyze specific metrics collected in the same way before and after the experiment. Among the studies in our corpus that employed quantitative analysis, 20.9% ($N = 18$) utilized such pre/post experiment comparisons.

There are multiple common methods used for this approach. For instance, single-case experimental designs (e.g. ABA or ABAB) are frequently used in medical and psychological research to assess behavioral changes due to experimental conditions. Two studies from our corpus employed ABA analysis with one showing that a companion-like robot resulted in significant improvements in participants' psychological well-being [59], and another showing that a robot improved clinical measures in autism-related scored for children with autism [131]. In educational contexts, the most common metric involves comparing educational test responses and scores before and after the robotic deployment (also known as "pretests" and "posttests"). For instance, several studies assessed the effectiveness of fixed versus personalized tutoring assistance [10, 84, 115, 116] and examined the impact of different scaffolding behaviors exhibited by robots [6]. Other methods of assessing before-and-after effects included analyzing behavioral changes in school children following robot interaction sessions [83], comparing pre- and post-experiment questionnaires [107, 144], and conducting pre- and post-experiment interviews [93].

4.4.4 Measuring LTE. While there is a wide variety of qualitative and quantitative measures in long-term HRI studies, we highlight one type in particular in this review: LTE. In our corpus, 45 studies (37.5%) measured LTE in some way. This measure is of notable importance to long-term HRI papers for several reasons. First, LTE can be a critical tool for quantifying when robotic interactions may have passed the novelty effect (Section 2.1), as determined by changes in the measure of user engagement over time. If researchers do not account for such an effect, they may mistakenly attribute positive outcomes to the robot's inherent effectiveness rather than the initial novelty. Of note, six studies in our corpus attempted to characterize the possible effect of novelty on the results of their study [2, 7, 8, 13, 111, 160].

LTE is also key for designing adaptative and personalized systems (Section 4.3.4). Such data allows for a deeper understanding of what longitudinal changes in an interaction maintain user interest and satisfaction, ultimately improving the overall quality and impact of HRI experiences. In our corpus, 24 studies discuss the effect of personalization alongside reported engagement measures. These can be found in the Appendix. Finally, LTE can also be in itself a research aim for a study, such as understanding when users drop off from using a robot regularly in the home.

Measurements of LTE can take many forms, and no single standard has emerged. Overall, the choice of measure depends on the interaction being studied. We identify seven categories of LTE measurement methods based on our corpus:

— *Self-Reported*: This method uses surveys to directly ask users about their levels of engagement over time. Likert scales were often used with this approach, with participants rating questions such as: "*How often have you used the robot in the last period?*" [27], "*I think I would like to use this system frequently*" [87], and "*I think I could spend a good time with [the robot]*" [34]. Some studies instead reported insights from a series of user interviews that asked qualitatively about engagement with a robot. Overall, 28.9% ($N = 13$) [15, 22, 27, 34, 36, 53, 83, 86, 87, 89, 124, 140, 149] of studies measuring LTE used self-reported methods.
— *Interaction Times*: This method uses temporal measures of HRIs to gauge engagement in the moment and compare these metrics over time. A common method was measuring the duration of robot usage per usage instance, with longer durations as an indication of higher engagement. For instance, Scassellati et al. [131] showed that children played with a robot for a similar average amount of time during the first 5 sessions of use in comparison to the last

five of 23 sessions. Another method was to measure the amount of additional time participants chose to spend with the robot (e.g., [56]). Out of the studies measuring LTE, eight (17.8%) measured it using an interaction time approach [23, 56, 91, 93, 97, 119, 131, 137].

— *Annotations*: This method was used on recordings of study sessions to hand annotate user engagement labels during robotic interactions. Trends in the annotations were then compared over time. For instance, Clabaugh et al. [21] annotated a video of children with ASD with a social robot, basing engagement levels on whether the child was "paying full attention to the interaction, immediately responding to the robot's prompts, or seeking further guidance or feedback from others in the room." In total, 13.3% ($N = 6$; [21, 102, 112, 116, 148, 163]) of studies measured LTE using annotations.

— *Count-Based*: This method includes counting the number or rate of certain types of interactions, such as games or activities, that the user performed with the robot over time. For example, Kanda et al. [61] utilized wireless tags to identify individual children who used the robot, and how often, in order to find patterns of drop-off. The *dropout rate* of users between two periods of time can also be calculated from counts of robot versus control use. For example, Barco et al. [8] used this method to show that a robot-supported rehabilitation program had less user dropout than without a robot. A counting technique often used in consumer electronics is reporting the *daily active users* of a device longitudinally. Zhao and McEwen [165] used this method of reporting to find that the daily active users of a Luka robot for reading with children dropped from 20 to 6 over the course of 180 days. Overall, 11.1% ($N = 5$; [8, 26, 61, 62, 160]) of studies measuring LTE used some count-based approach.

— *Sensor-Based Modeling*: A number of studies used vision or audio as inputs to machine learning methods for predicting levels of engagement in real-time or *post hoc*. The most common features used to predict engagement were affect or mood, body pose, vocal tone, and gaze. These measures were then taken and compared over time. In total, 11.1% ($N = 5$; [38, 55, 116, 136, 164]) of studies measuring LTE primarily used this type of sensor-based approach.

— *Behavioral Observation*: In one study [94], researchers used their own live observations of users to note a subjective interpretation of engagement. The authors noted patterns in how children would proactively assist the robot when the robot did not appear to react correctly to certain stimuli.

— *Mixed*: Seven studies (15.5%) [3, 10, 30, 54, 77, 130, 161] utilized a combination of the above measures for LTE.

5 Discussion

The amount of research dedicated to conducting long-term studies has witnessed remarkable growth, experiencing a significant increase from 26 published papers spanning the period between 2003 and 2013 to 94 papers over the past decade. This surge illustrates a growing interest in and commitment to comprehending the ways in which robots interact with and impact users over extended periods of time.

Our analysis revealed numerous promising trends over the past two decades. For instance, our findings show a diverse representation of age groups in long-term studies, spanning from toddlers to the elderly (however, we did notice a lack of representation of studies conducted with teenagers, which is discussed in Section 5.1.1). Another positive trend is that the majority of the studies we analyzed involved the robot operating entirely *autonomously* (75.8%) and *in situ* (85.8%), mirroring the real-world environments and contexts in which they will need to function. Even in lab-based studies (14.2%), the researchers often used simulated home or school environments closely resembling the real deployment locations. This approach fosters more authentic interactions, more likely eliciting responses from participants that closely resemble their behavior in real-world

scenarios. Lastly, there was often a good match between the domain and the location. For example, educational robots were frequently placed in schools, therapy robots were frequently based in rehabilitation facilities, and eldercare robots were frequently placed in eldercare homes. While this point may seem trivial, the amount of effort to place robots in such locations is not.

Moving forward, our discussion will delve into the gaps and opportunities for long-term HRI study based on how it has evolved over the past two decades. We additionally present several design recommendations for researchers planning to conduct their own long-term social robotic studies.

5.1 Gaps and Opportunities in Long-Term HRI Study

From this review, we have identified several gaps in the literature, whether pertaining to overlooked populations or understudied domains. We detail several opportunities for future HRI research below.

5.1.1 Teenaged Participants. Our analysis only identified two long-term HRI papers (1.7% of our corpus) that focused their study on teenagers (those aged between 13 and 17). The first study focused on teenagers with autism and severe developmental disabilities. It analyzed the impact of a robot on their communication skills in secondary school and showed the potential for robots to improve communication [138]. The second paper also focused on teenagers with autism, and the authors found a link between the teenagers' sensory profile and their capabilities to imitate a robot [20].

Considering the very low number of studies focusing on long-term robots for teenagers, we highlight this as an open area of research. Furthermore, we have not identified any papers that have analyzed how neurotypical teenagers engage with a robot in the long run. There are many open questions to explore with this population, such as how teenagers engage with robots in their homes with family around, or how teenagers may share their experiences interacting with robots with their peers and friends. We especially highlight the areas of how a robot can have an impact on a teenager's education, given the many studies that show benefits for younger children in the education domain. Education is generally a crucial topic in a teenager's life, as they are deciding whether and what future educational opportunities to pursue. Additionally, we point out mental health for teenagers as a critical area of exploration. The World Health Organization has identified one in seven teenagers (between 10 and 19) as having a mental health disorder [105]. Given the positive effects robots have had on mental health with other ages, we believe it is important also to investigate how robots can help this population longitudinally.

Another question is whether teenagers will adopt robotic technologies in a similar manner to other populations. Teenagers have been shown to have different relationships with technology compared to other users [64], and their technology usage is often strongly influenced by their peers [46, 132]. Furthermore, they are usually amongst the first adopters of new technology [74]. Therefore, it's crucial to study how teenagers interact with robots in the long term, especially to understand the novelty effect in this population and how it may fade in comparison to other age groups.

5.1.2 Exploration of the Workplace Location and Domain. A second area we have identified as needing further exploration is the long-term deployment of robots in workplace settings, particularly focusing on office or business environments (rather than schools or hospitals, as explained in Section 3). With this categorization, we identified only four papers where robots were tested for extended periods in workplaces. The first was an experiment in which a robot acted as an assistant in a collaborative workspace, helping workers with routine day-to-day tasks [154]. The second paper compared different types of robots as they coached employees on mental health, specifically in the workplace [144]. The third paper investigated robots that provided break management at

desks [124]. Lastly, the fourth paper explored the social aspects of a fetch-and-carry robot designed to assist motion-impaired users in an office environment [135].

Despite the limited amount of prior research in workplace environments, most individuals spend a significant portion of their lives at work, dedicating approximately 40 hours or more per week to it. Therefore, we believe it is important to study how robots can assist and interact with us in the workplace. Many open questions remain about how robots might impact work environments—Will they be accepted and incorporated organically by employers and employees? Will robots increase or decrease productivity? How will they affect the well-being of employees?

In addition to office-like environments, another common workplace is the industrial and factory setting. While numerous studies have explored collaborative and manufacturing robots in short-term, single-session investigations, we have not uncovered any comprehensive, long-term examinations within this specific sub-domain. Manufacturing accounts for a significant portion of jobs; for example, around 8% of the UK population works in this sector [104]. Moreover, many robots are already deployed in manufacturing settings, and their numbers continue to grow [118]. Important HRI questions in manufacturing include whether robots will be accepted by workers in the long run, especially given concerns about job displacement [25], and how these robots can develop effective working relationships with their users.

Lastly, future investigations should account for the wide variety of workplaces not yet studied, including restaurant kitchens, service industry environments, building sites, and more. How can robots benefit each of these places in the long run? While students and patients have been well-studied as participants, considering the perspectives of teachers and staff in schools, as well as doctors, nurses, and staff in hospitals and care homes, can provide a more holistic view of social robots in the workplace. Given the potential for robots to assist workers regarding workplace mental health [144], productivity [154], and healthy work practices [124], we as a field must continue to investigate how robots can support users in a variety of real-world workplace settings.

5.1.3 Standardization of Measures and Metrics Related to Long-Term Study. During our analysis of the papers in this review, we frequently noted the use of common HRI surveys, such as the Godspeed [9] (to measure participant perceptions of robots), the RoSAS [16] (to measure participant judgments of the social attributes of robots), and the NARS questionnaires [100] (to measure participant negative attitudes towards robots). These standardized surveys have been extraordinarily useful in providing researchers with common ground to compare results across different robot, participant, and interaction types. These same surveys are also frequently used repeatedly during a longitudinal study; however, they have not been validated specifically as repeated measures. The Godspeed questionnaire, for instance, was initially conceived to help developers choose different robot features or configurations during the design phase. The authors themselves note that human perception is "not stable" [9], and results are likely to change as humans are more accustomed to a robot. The recent proliferation of long-term HRI studies now begs the question: Is there a need for a new, updated measurement system specifically for more longitudinal HRI?

Thus, our final suggested future opportunity is in the exploration of standardized measurement tools for long-term HRI. We posit that current tools are limited in their ability to capture the uniqueness of how a human–robot relationship evolves and endures over time, as well as the pervasiveness of the novelty effect and maintaining LTE. During our analysis, we did not find a questionnaire or measurement tool that covered such topics in a standardized way. Updating existing standard surveys like Godspeed, RoSAS, and NARS to cover these critical long-term HRI inquiries is one avenue for future research. Alternatively, future work can validate these commonly used questionnaires for repeated use and show that they remain applicable over time.

Another avenue is diving deeper into a specific type of long-term quality, such as standardizing the way researchers consider the novelty effect to have "worn off" or how to measure LTE. For the studies in our corpus that measured the latter (Section 4.4.4), the methods were widely varied across observation, survey, and sensor-based techniques. While it is likely that unique research goals, participant types, locations, and domains influence the type of LTE measurement, attempts to standardize or create common ground across studies of similar types are well worthwhile. As more HRI studies continue to be long-term in their nature, having standardized bases for comparison will better enable future researchers to develop new, longitudinal robotic interactions.

5.2 Study and Design Recommendations

In this section, we provide several design recommendations for those planning on conducting a long-term HRI study. These recommendations come from both the analyzed data in our corpus and the experience of the authors' reading of the corpus.

5.2.1 Choice of Study Length, Participant Count, and Data Type. For researchers embarking on a long-term HRI study, the choice of just how long to run a study is an important one, with implications on timing, funding, recruiting, and technical complexity. In a meta-analysis, we found that this choice is frequently tied to the number of participants in a study as well as the type of data being collected (quantitative or qualitative).

Figure 8 shows four primary categories, as determined by observation. Group I encompasses session-based studies with fewer than 10 sessions and fewer than 20 participants ($N = 35$ studies; 40.7%). These studies may be motivated by a blend of practical considerations. Here, resource limitations or the inherent complexity of sustained HRI interactions could steer researchers toward concise study durations and small participant pools. Group II features studies with 10 or more sessions but fewer than 20 participants ($N = 14$ studies; 16.3%). Such studies may aim to balance depth with sample size, with researchers prioritizing statistical power or striving for diverse participant representation. In the third category, characterized by studies with fewer than 10 sessions and 20 or more participants ($N = 29$ studies; 33.7%), researchers may seek to gather insights from diverse participants despite the study's comparative brevity. Lastly, the smallest cluster comprises studies with both 10 or more sessions and 20 or more participants ($N = 8$ studies; [6, 51, 59, 72, 96, 106, 125, 165]). A high number of participants for a long amount of time is ideal from a research perspective, but a practical and logistical challenge. In this group, researchers may be driven by a desire for comprehensive exploration within a more specialized context.

We present these categories not to provide a formal classification system but instead to encourage researchers to reflect on which relative category they strive to be in with their work. In pursuit of the occasionally elusive question, *how long should my study be?* researchers can consider their participants and measure types in tandem. A qualitative study with a handful of users may require longer study times in order to provide actionable insights. A quantitative study pursuing statistical significance may require a large sample size that logistically necessitates a shorter study length. Researchers are encouraged to use our corpus as a guide for how these choices have been made for their domain, location, and participant type in the past.

5.2.2 Including Novel Behaviors and Personalization. As discussed in Section 4.3.4, robots that exhibit diverse and varied behaviors tend to engage their users more effectively compared to those displaying repetitive patterns. One particularly effective approach to achieving such adaptation is by personalizing the robot's behavior to suit the individual user it interacts with [10, 116]. By creating a model of the user's preferences and needs, the robot can adapt and select the most appropriate behavior for that specific user. Prior work has shown that robots that personalize their behavior increase learning gains [84], engagement, and rapport [75].

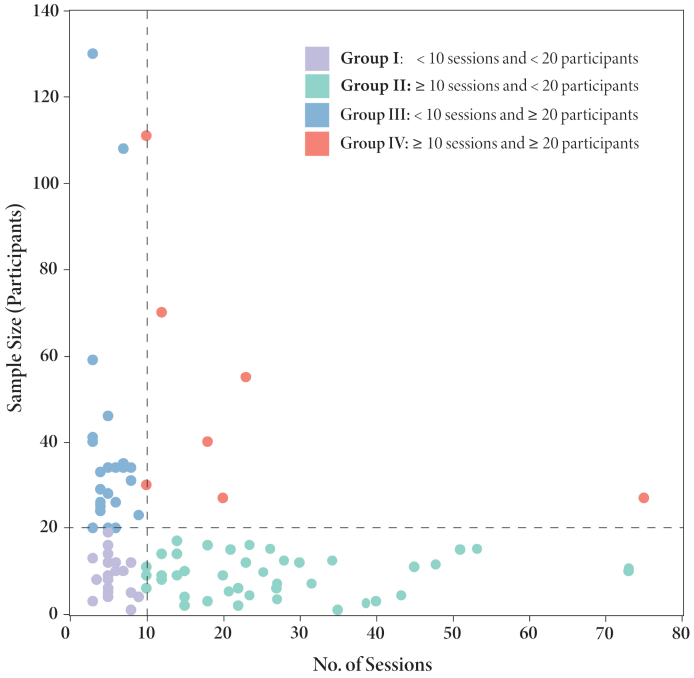


Fig. 8. Distribution of session-based studies by sessions and sample size. Four primary categories emerge: studies with < 10 sessions and < 20 participants (Group I, lower left, purple), studies with ≥ 10 sessions and < 20 participants (Group II, lower right, green), studies with < 10 sessions and ≥ 20 participants (Group III, upper left, blue), and studies with ≥ 10 sessions and ≥ 20 participants (Group IV, upper right red). Four studies [52, 60, 107, 139] were considered outliers and are excluded from this plot for clarity.

Our analysis reveals that a minority of our analyzed studies (39.2%) either have the robot adapt or personalize throughout the interaction, though this trend is increasing. We strongly encourage future researchers to continue incorporating adaptation and personalization mechanisms into their robotic systems where appropriate, as these aspects can be fundamental to ensuring long-term user acceptance and usage. Numerous approaches exist to introduce adaptation in long-term systems. For instance, some studies have successfully implemented a continuous backstory for the robot, carrying over from one session to another [131]. Others have enabled robots to advance and acquire more complex skills over time [18], while some have integrated multiple activities to switch between [75]. There are also different ways to achieve personalization: recording participant details such as names for use in subsequent sessions [86], modeling the user's skills to personalize tutoring [31], taking into account what activity the user is currently engaged with [30], learning the user's preferences [32], and so on.

We encourage researchers to use this review to find prior long-term studies for their domain, population, and interaction type as inspiration for methods of adaptation and personalization. Additional reviews specifically on adaptive and personalized interactions and technical methods in HRI include Gasteiger et al. [41], Ahmad et al. [4], and Hellou et al. [50]. At the same time, there is an open opportunity to define new methods, particularly given new technological advances in artificial intelligence. What is most successful is likely dependent on the interaction context at hand. For instance, personalizing and adapting a robot used with many different participants in the same setting, such as a school, may provide a challenge based on identifying the correct user. Group interactions may provide challenges where a certain personalization technique is acceptable with a

user within a certain sub-group, but not when there are other group members are around. Working with the *Physical Health* domain may be a challenge if users have widely differing physical abilities to interact with a robot. Personalizing a non-verbal Paro robot with limited features is likely more challenging to personalize than a robot like NAO, which has many different outputs to customize. In these ways, there is no one-size-fits-all approach to personalize and adapt robotic behaviors, and the considerations are many. We encourage researchers to identify in their work what parts of their personalization and adaptation techniques may be transferrable to other settings, participants, and domains, and which are unique to their own set of study.

5.2.3 Report Full Data. Many of the papers that were part of this analysis had missing information or statistics describing the long-term nature of robotic interactions. For session-based studies, the most common missing details that we identified in our review were the number of sessions, the session length, and the total amount of time the user interacted on average with the robot. These details are essential for the HRI community to understand the effects of long-term interaction and how engagement and novelty interact with the number of sessions and time spent with the robot. We were able to estimate some of the missing information, such as the average number of minutes spent with the robot, for the purpose of this review. Nonetheless, it is important for future work to provide accurate numbers for each of their studies in their papers.

Of note, several of the studies we examined were categorized as *free-use*, allowing users the autonomy to determine when and how they engage with the robot. These insightful studies shed light on the natural interactions people have with robots in their homes and schools without the many constraints usually present in more traditional HRI studies. They are particularly instrumental in helping us comprehend the demographics of users who engage with the robot and in discerning the specific contexts in which these interactions occur. Furthermore, these studies enable us to discern emerging patterns in the decline of engagement with these systems over time. While most free-use papers provided data on the robot's duration in users' homes, we frequently noticed a lack of details regarding how often or long the robot was used per day, which users or family members it interacted with, and which times of the day and situations the robot was most used. We encourage future researchers in this area to capture this valuable information in their studies.

One particular aspect that was very rarely reported was whether participants were offered compensation for participating in the study. This is especially relevant for longer-term studies investigating the long-term use of robots and user engagement or compliance. While some participants could be motivated to complete all sessions or interactions because they are enjoying the robot and are engaged, others could be completing the sessions only as a prerequisite for study compensation or due to not wanting to disappoint the research team. We suggest all long-term papers clearly define the incentivization methods during their study so readers can fully evaluate the reasons why participants engaged with the robot and whether there were any confounding factors.

5.3 Limitations

There are several limitations to our review. Despite a rigorous search methodology, there will ultimately be several papers that we have missed. We did not extensively search through the proceedings of more traditional robotic conferences, or general human-computer interaction and technological methods venues, and instead focused on HRI conferences and journals. A second limitation is that we excluded papers that did not interact with the same user throughout the duration of the study, such as robots deployed in a museum. However, there is also value in analyzing how robots are designed to interact with many unique users over multiple weeks or months. A third limitation is that we focus on the primary characteristics of each paper, leading to potentially abstracting away the nuances of papers investigating multiple populations, multiple domains, or

with multiple phases in their study investigating different temporal aspects. A final limitation is that we had to estimate values for several papers, such as the number of days for the interaction and the average lengths of the interaction. These estimations could be a misrepresentation of the original study.

6 Conclusions

There has been a significant increase in the number of studies analyzing long-term social robotic interactions in the past two decades. In this review, we report on studies in which a robot socially interacted with a user for at least 3 sessions spread across 3 days. Our search led to a corpus of 120 studies that align with our inclusion criteria since 2003. For each of these studies, we extracted and categorized several factors, including how long users interacted with the robot, the domain of the study, the location where the study took place, the number and types of participants, the level of autonomy of the robot, interaction modality, and the type of measures and results in the study.

Our analysis shows that there has been a steady increase in papers published in the area of long-term HRI with social robots. More specifically, we see a significant increase in the number of papers published in the past decade compared to the prior decade. We highlight that there has been good variability in age groups, domains, and locations. Most studies were also conducted *in situ*, and most used a fully autonomous robot.

We identified several gaps in the corpus of data, which we present as opportunities for future work. The first opportunity we identify is regarding teenagers. Very few studies have conducted long-term social robotic interactions with teenagers, despite this population interacting and adopting new technology in unique ways. The second identified opportunity regards the workplace. Very few studies have been conducted in workplace environments, despite most people spending a great proportion of their time at work. Our final opportunity centers around the development of standardized questionnaires or metrics specifically tailored for long-term HRI scenarios. Most current questionnaires focus on single-session studies and may be inappropriate for use over repeated interactions. We close with a number of recommendations for researchers designing long-term HRI studies based on patterns seen in our corpus, including surrounding study duration, interaction adaptation, and data reporting.

In the ever-evolving landscape of social robotics, the examination of long-term HRI has taken center stage. By sharing the insights and design recommendations in this review, we aspire to propel long-term HRI research forward more efficiently and facilitate more profound investigations into the dynamics of extended HRI. We hope the knowledge from these studies will serve as a compass for more meaningful, effective, and sustainable human–robot relationships.

References

[1] A. Afyouni, I. Ocnarescu, I. Cossin, E. Kamoun, A. Mazel, and C. Fattal. 2022. Living one week with an autonomous Pepper in a rehabilitation center: Lessons from the field. In *Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 554–559.

[2] Roxana Agrigoroaie and Adriana Tapus. 2018. The outcome of a week of intensive cognitive stimulation in an elderly care setup: A pilot test. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 814–819.

[3] Muneeb Imtiaz Ahmad, Omar Mubin, and Joanne Orlando. 2017. Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *International Journal of Human–Computer Interaction* 33, 12 (2017), 943–962.

[4] Muneeb Imtiaz Ahmad, Omar Mubin, and Joanne Orlando. 2017. A systematic review of adaptivity in human-robot interaction. *Multimodal Technologies and Interaction* 1, 3 (2017), 14.

[5] Aino Ahtinen, Nasim Beheshtian, and Kaisa Väänänen. 2023. Robocamp at home: Exploring families' co-learning with a social robot: Findings from a one-month study in the wild. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 331–340.

[6] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. 2014. Employing humanoid robots for teaching English language in Iranian junior high-schools. *International Journal of Humanoid Robotics* 11, 03 (2014), 1450022.
[7] Markus Bajones, David Fischinger, Astrid Weiss, Paloma De La Puente, Daniel Wolf, Markus Vincze, Tobias Körtner, Markus Weninger, Konstantinos Papoutsakis, Damien Michel, et al. 2019. Results of field trials with a mobile service robot for older adults in 16 private households. *ACM Transactions on Human-Robot Interaction* 9, 2 (2019), 1–27.
[8] Alex Barco, Jordi Albo-Canals, Carles Garriga-Berga, Xavier Vilasís-Cardona, Laura Callejón, Marc Turón, Claudia Gómez, and Anna López-Sala. 2014. A drop-out rate in a long-term cognitive rehabilitation program through robotics aimed at children with TBI. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 186–192.
[9] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1 (2009), 71–81.
[10] Paul Baxter, Emily Ashurst, Robin Read, James Kennedy, and Tony Belpaeme. 2017. Robot education peers in a situated primary school study: Personalisation promotes child learning. *PloS One* 12, 5 (2017), e0178126.
[11] BeatBots. 2024. My Keepon. Retrieved November 5, 2024 from https://www.beatbots.net/my-keepon
[12] Momotaz Begum, Richard W. Serna, David Kontak, Jordan Allspaw, James Kuczynski, Holly A. Yanco, and Jacob Suarez. 2015. Measuring the efficacy of robots in autism therapy: How informative are standard HRI metrics. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 335–342.
[13] Indu P. Bodala, Nikhil Churamani, and Hatice Gunes. 2021. Teleoperated robot coaching for mindfulness training: A longitudinal study. In *Proceedings of the 2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 939–944.
[14] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. In *Springer Handbook of Robotics*. B. Siciliano and O. Khatib (Eds.), Springer Handbook, Springer, Cham, 1935–1972. DOI: https://doi.org/10.1007/978-3-319-32552-1
[15] Bengisu Cagiltay, Nathan Thomas White, Rabia Ibtasar, Bilge Mutlu, and Joseph Michaelis. 2022. Understanding factors that shape children's long term engagement with an in-home learning companion robot. In *Proceedings of the 21st Annual ACM Interaction Design and Children Conference (IDC '22)*. Association for Computing Machinery, New York, NY, USA, 362–373. DOI: https://doi.org/10.1145/3501712.3529747
[16] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 254–262.
[17] Amedeo Cesta, Gabriella Cortellessa, Andrea Orlandini, and Lorenza Tiberio. 2016. Long-term evaluation of a telepresence robot for the elderly: Methodology and ecological case study. *International Journal of Social Robotics* 8 (2016), 421–441.
[18] Shruti Chandra, Raul Paradeda, Hang Yin, Pierre Dillenbourg, Rui Prada, and Ana Paiva. 2018. Do children perceive whether a robotic peer is learning or not? In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 41–49.
[19] Huili Chen, Anastasia K. Ostrowski, Soo Jung Jang, Cynthia Breazeal, and Hae Won Park. 2022. Designing long-term parent-child-robot triadic interaction at home through lived technology experiences and interviews. In *Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 401–408.
[20] Pauline Chevalier, Gennaro Raiola, Jean-Claude Martin, Brice Isableu, Christophe Bazile, and Adriana Tapus. 2017. Do sensory preferences of children with autism impact an imitation task with a robot? In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 177–186.
[21] Caitlyn Clabaugh, 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. *Frontiers in Robotics and AI* 6 (2019), 110.
[22] Alexandre Coninx, Paul Baxter, Elettra Oleari, Sara Bellini, Bert Bierman, O. Henkemans, Lola Cañamero, Piero Cosi, Valentin Enescu, R. Espinoza, et al. 2016. Towards long-term social child-robot interaction: Using multi-activity switching to engage young users. *Journal of Human-Robot Interaction* 5, 1 (2016), 32–67.
[23] Serhan Coşar, Manuel Fernandez-Carmona, Roxana Agrigoroaie, Jordi Pages, François Ferland, Feng Zhao, Shigang Yue, Nicola Bellotto, and Adriana Tapus. 2020. ENRICHME: Perception and interaction of an assistive robot for the elderly at home. *International Journal of Social Robotics* 12 (2020), 779–805.
[24] Dagoberto Cruz-Sandoval, Arturo Morales-Tellez, Eduardo Benitez Sandoval, and Jesus Favela. 2020. A social robot as therapy facilitator in interventions to deal with dementia-related behavioral symptoms. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 161–169.
[25] Eric Dahlin. 2019. Are robots stealing our jobs? *Socius* 5 (2019), 2378023119846249.

[26] Daniel P. Davison, Frances M. Wijnen, Vicky Charisi, Jan van der Meij, Vanessa Evers, and Dennis Reidsma. 2020. Working with a social robot in school: A long-term real-world unsupervised deployment. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 63–72.

[27] Maartje De Graaf, Somaya Ben Allouch, and Jan Van Dijk. 2017. Why do they refuse to use my robot? Reasons for non-use derived from a long-term home study. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 224–233.

[28] Maartje Ma De Graaf, Somaya Ben Allouch, and Tineke Klamer. 2015. Sharing a life with Harvey: Exploring the acceptance of and relationship-building with a social robot. *Computers in Human Behavior* 43 (2015), 1–14.

[29] Maartje M. A. De Graaf, Somaya Ben Allouch, and Jan A. G. M. van Dijk. 2016. Long-term evaluation of a social robot in real homes. *Interaction Studies* 17, 3 (2016), 462–491.

[30] Claudia Di Napoli, Giovanni Ercolano, and Silvia Rossi. 2023. Personalized home-care support for the elderly: A field experience with a social robot at home. *User Modeling and User-Adapted Interaction* 33, 2 (2023), 405–440.

[31] Melissa Donnermann, Philipp Schaper, and Birgit Lugrin. 2022. Investigating adaptive robot tutoring in a long-term interaction in higher education. In *Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 171–178.

[32] M. M. S. N. Edirisinghe and A. G. B. P. Jayasekara. 2018. Use of autobiographical memory for enhancing adaptive HRI in multi-user domestic environment. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1118–1123.

[33] Jing Fan, Akshith Ullal, Linda Beuscher, Lorraine C. Mion, Paul Newhouse, and Nilanjan Sarkar. 2021. Field testing of Ro-Tri, a robot-mediated triadic interaction for older adults. *International Journal of Social Robotics* 13 (2021), 1711–1727.

[34] Juan Fasola and Maja J. Matarić. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction* 2, 2 (2013), 3–32.

[35] Ronit Feingold Polak and Shelly Levy Tzedek. 2020. Social robot for rehabilitation: Expert clinicians and post-stroke patients' evaluation following a long-term intervention. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 151–160.

[36] Ylva Fernaeus, Maria Håkansson, Mattias Jacobsson, and Sara Ljungblad. 2010. How do you play with a robotic toy animal? A long-term study of Pleo. In *Proceedings of the 9th International Conference on Interaction Design and Children*, 39–48.

[37] Dorothée François, Stuart Powell, and Kerstin Dautenhahn. 2009. A long-term study of children with autism playing with a robotic pet: Taking inspirations from non-directive play therapy to encourage children's proactivity and initiative-taking. *Interaction Studies* 10, 3 (2009), 324–373.

[38] Edwinn Gamborino, Hsiu-Ping Yueh, Weijane Lin, Su-Ling Yeh, and Li-Chen Fu. 2019. Mood estimation as a social profile predictor in an autonomous, multi-session, emotional support robot for children. In *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–6.

[39] Norina Gasteiger, Ho Seok Ahn, Christine Fok, JongYoon Lim, Christopher Lee, Bruce A. MacDonald, Geon Ha Kim, and Elizabeth Broadbent. 2022. Older adults' experiences and perceptions of living with Bomy, an assistive dailycare robot: A qualitative study. *Assistive Technology* 34, 4 (2022), 487–497.

[40] Norina Gasteiger, Ho Seok Ahn, Chiara Gasteiger, Christopher Lee, Jongyoon Lim, Christine Fok, Bruce A. Macdonald, Geon Ha Kim, and Elizabeth Broadbent. 2021. Robot-delivered cognitive stimulation games for older adults: Usability and acceptability evaluation. *ACM Transactions on Human-Robot Interaction* 10, 4 (2021), 1–18.

[41] Norina Gasteiger, Mehdi Hellou, and Ho Seok Ahn. 2023. Factors for personalization and localization to optimize human–robot interaction: A literature review. *International Journal of Social Robotics* 15, 4 (2023), 689–701.

[42] Leonardo Giusti and Patrizia Marti. 2006. Interpretative dynamics in human robot interaction. In *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (ROMAN '06)*. IEEE, 111–116.

[43] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C. Schultz, et al. 2005. Designing robots for long-term social interaction. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1338–1343.

[44] Goren Gordon, 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[45] Jillian Greczek, Edward Kaszubski, Amin Atrash, and Maja Matarić. 2014. Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 561–566.

[46] Rebecca E. Grinter and Leysia Palen. 2002. Instant messaging in teen life. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, 21–30.

[47] Aldebaran United Robotics Group. 2022. *NAO*. Retrieved June 30, 2023 from https://www.aldebaran.com/en/nao

[48] Marc Hanheide, Denise Hebesberger, and Tomáš Krajník. 2017. The when, where, and how: An adaptive robotic info-terminal for care home residents. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 341–349.

[49] Denise Hebesberger, Tobias Koertner, Christoph Gisinger, Juergen Pripfl, and Christian Dondrup. 2016. Lessons learned from the deployment of a long-term autonomous robot as companion in physical therapy for older adults with dementia a mixed methods study. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 27–34.

[50] Mehdi Hellou, Norina Gasteiger, Jong Yoon Lim, Minsu Jang, and Ho Seok Ahn. 2021. Personalization and localization in human-robot interaction: A review of technical methods. *Robotics* 10, 4 (2021), 120.

[51] Eunja Hyun, Hyunmin Yoon, and Sooryun Son. 2010. Relationships between user experiences and children's perceptions of the education robot. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 199–200.

[52] Indiegogo. 2014. Jibo: The World's First Social Robot for the Home. Retrieved November 5, 2024 from https://www.indiegogo.com/projects/jibo-the-world-s-first-social-robot-for-the-home#/

[53] Bahar Irfan, Nathalia Céspedes Gomez, Jonathan Casas, Emmanuel Senft, Luisa F. Gutiérrez, Monica Rincon-Roncancio, Marcela Munera, Tony Belpaeme, and Carlos A. Cifuentes. 2020. Using a personalised socially assistive robot for cardiac rehabilitation: A long-term case study. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 124–130.

[54] Alexis Jacq, Séverin Lemaignan, Fernando Garcia, Pierre Dillenbourg, and Ana Paiva. 2016. Building successful long child-robot interactions in a learning context. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 239–246.

[55] Shomik Jain, Balasubramanian Thiagarajan, Zhonghao Shi, Caitlyn Clabaugh, and Maja J. Matarić. 2020. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Science Robotics* 5, 39 (2020), eaaz3791.

[56] Joris B. Janssen, Chrissy C van der Wal, and Mark Neerincx. 2011. Motivate to learn: Effects of performance adaptation on child motivation of robot interaction (2011).

[57] Kwangmin Jeong, Jihyun Sung, Hae-Sung Lee, Aram Kim, Hyemi Kim, Chanmi Park, Yuin Jeong, JeeHang Lee, and Jinwoo Kim. 2018. Fribo: A social networking robot for increasing social connectedness through sharing daily home activities from living noise data. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 114–122. DOI: https://doi.org/10.1145/3171221.3171254

[58] Sooyeon Jeong, Sharifa Alghowinem, Laura Aymerich-Franch, Kika Arias, Agata Lapedriza, Rosalind Picard, Hae Won Park, and Cynthia Breazeal. 2020. A robotic positive psychology coach to improve college students' wellbeing. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 187–194.

[59] Sooyeon Jeong, Laura Aymerich-Franch, Sharifa Alghowinem, Rosalind W. Picard, Cynthia L. Breazeal, and Hae Won Park. 2023. A robotic companion for psychological well-being: A long-term investigation of companionship and therapeutic alliance. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 485–494.

[60] Aidan Jones and Ginevra Castellano. 2018. Adaptive robotic tutors that support self-regulated learning: A longer-term investigation with primary school children. *International Journal of Social Robotics* 10 (2018), 357–370.

[61] Takayuki Kanda, Takayuki Hirano, Daniel Eaton, and Hiroshi Ishiguro. 2004. Interactive robots as social partners and peer tutors for children: A field trial. *Human–Computer Interaction* 19, 1–2 (2004), 61–84.

[62] Takayuki Kanda, Rumi Sato, Naoki Saiwaki, and Hiroshi Ishiguro. 2007. A two-month field trial in an elementary school for long-term human–robot interaction. *IEEE Transactions on Robotics* 23, 5 (2007), 962–971.

[63] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro, and Norihiro Hagita. 2010. A communication robot in a shopping mall. *IEEE Transactions on Robotics* 26, 5 (2010), 897–913.

[64] Raul L. Katz, Max Felix, and Madlen Gubernick. 2014. Technology and adolescents: Perspectives on the things to come. *Education and Information Technologies* 19 (2014), 863–886.

[65] James Kennedy, Séverin Lemaignan, Caroline Montassier, Pauline Lavalade, Bahar Irfan, Fotios Papadopoulos, Emmanuel Senft, and Tony Belpaeme. 2017. Child speech recognition in human-robot interaction: Evaluations and recommendations. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 82–90.

[66] Cory D. Kidd and Cynthia Breazeal. 2008. Robots at home: Understanding long-term human-robot interaction. In *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3230–3235.

[67] Elizabeth Kim, Rhea Paul, Frederick Shic, and Brian Scassellati. 2012. Bridging the research gap: Making HRI useful to individuals with autism. *Journal of Human-Robot Interaction* 1, 1 (July 2012), 26–54. DOI: https://doi.org/10.5898/JHRI.1.1.Kim

[68] Tineke Klamer, Somaya Ben Allouch, and Dirk Heylen. 2011. "Adventures of Harvey"–Use, acceptance of and relationship building with a social robot in a domestic environment. In *Proceedings of the 3rd International Conference on Human-Robot Personal Relationships (HRPR '10), Revised Selected Papers*. Springer, 74–82.

[69] Kheng Lee Koay, Dag Sverre Syrdal, Wan Ching Ho, and Kerstin Dautenhahn. 2016. Prototyping realistic long-term human-robot interaction for the study of agent migration. In *Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 809–816.

[70] Kheng Lee Koay, Dag Sverre Syrdal, Michael L. Walters, and Kerstin Dautenhahn. 2007. Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '07)*. IEEE, 564–569.

[71] Jacqueline M. Kory-Westlund, Hae Won Park, Ishaan Grover, and Cynthia Breazeal. 2022. Long-term interaction with relational SIAs. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application (1st ed.)*. Association for Computing Machinery, New York, NY, USA, 195–260. DOI: https://doi.org/10.1145/3563659.3563667

[72] Hideki Kozima, Marek P. Michalowski, and Cocoro Nakagawa. 2009. Keepon: A playful robot for research, therapy, and entertainment. *International Journal of Social Robotics* 1 (2009), 3–18.

[73] Geoffrey W. Lane, Delilah Noronha, Alexandra Rivera, Kathy Craig, Christina Yee, Brent Mills, and Eimee Villanueva. 2016. Effectiveness of a social robot, "Paro," in a VA long-term care setting. *Psychological Services* 13, 3 (2016), 292.

[74] Alexis R. Lauricella, Drew P. Cingel, Courtney Blackwell, Ellen Wartella, and Annie Conway. 2014. The mobile generation: Youth and adolescent ownership and use of new media. *Communication Research Reports* 31, 4 (2014), 357–364.

[75] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis, and Sarun Savetsila. 2012. Personalization in HRI: A longitudinal field experiment. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 319–326.

[76] Min Kyung Lee, Jodi Forlizzi, Paul E. Rybski, Frederick Crabbe, Wayne Chung, Josh Finkle, Eric Glaser, and Sara Kiesler. 2009. The Snackbot: Documenting the design of a robot for long-term human-robot interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 7–14.

[77] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2014. Empathic robots for long-term interaction: Evaluating social presence, engagement and perceived support in children. *International Journal of Social Robotics* 6 (2014), 329–341.

[78] Iolanda Leite, Carlos Martinho, and Ana Paiva. 2013. Social robots for long-term interaction: A survey. *International Journal of Social Robotics* 5 (2013), 291–308.

[79] Iolanda Leite, Carlos Martinho, Andre Pereira, and Ana Paiva. 2009. As time goes by: Long-term evaluation of social presence in robotic companions. In *Proceedings of the 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '09)*. IEEE, 669–674.

[80] Iolanda Leite, Marissa McCoy, Monika Lohani, Daniel Ullman, Nicole Salomons, Charlene Stokes, Susan Rivers, and Brian Scassellati. 2015. Emotional storytelling in the classroom: Individual versus group interaction between children and robots. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 75–82.

[81] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. 2015. Comparing models of disengagement in individual and group interactions. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 99–105.

[82] Iolanda Leite, André Pereira, Samuel Mascarenhas, Carlos Martinho, Rui Prada, and Ana Paiva. 2013. The influence of empathy in human–robot relations. *International Journal of Human-Computer Studies* 71, 3 (2013), 250–260.

[83] Séverin Lemaignan, Nigel Newbutt, Louis Rice, and Joe Daly. 2022. "It's important to think of Pepper as a teaching aid or resource external to the classroom": A social robot in a school for autistic children. *International Journal of Social Robotics* (2022), 1–22.

[84] Daniel Leyzberg, Aditi Ramachandran, and Brian Scassellati. 2018. The effect of personalization in longer-term robot tutoring. *ACM Transactions on Human-Robot Interaction* 7, 3 (2018), 1–19.

[85] Mike E. U. Ligthart, Simone M. de Droog, Marianne Bossema, Lamia Elloumi, Kees Hoogland, Matthijs H. J. Smakman, Koen V. Hindriks, and Somaya Ben Allouch. 2023. Design specifications for a social robot math tutor. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 321–330.

[86] Mike E. U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. 2022. Memory-based personalization for fostering a long-term child-robot relationship. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 80–89.

[87] Matteo Luperto, Javier Monroy, Jennifer Renoux, Francesca Lunardini, Nicola Basilico, Maria Bulgheroni, Angelo Cangelosi, Matteo Cesari, Manuel Cid, Aladar Ianes, et al. 2022. Integrating social assistive robots, IoT, virtual communities and smart objects to assist at-home independently living elders: The MoveCare project. *International Journal of Social Robotics* 15, 3 (2023), 517–545.

[88] Hamza Mahdi, Sami Alperen Akgun, Shahed Saleh, and Kerstin Dautenhahn. 2022. A survey on the design and evolution of social robots—Past, present and future. *Robotics and Autonomous Systems* 156 (2022), 104193.

[89] Felip Martí Carrillo, Joanna Butchart, Sarah Knight, Adam Scheinberg, Lisa Wise, Leon Sterling, and Chris McCarthy. 2018. Adapting a general-purpose social robot for paediatric rehabilitation through in situ design. *ACM Transactions on Human-Robot Interaction* 7, 1 (2018), 1–30.

[90] Johnny L. Matson and Matthew J. Konst. 2013. What is the evidence for long term effects of early autism interventions? *Research in Autism Spectrum Disorders* 7, 3 (2013), 475–479.

[91] Louis McCallum and Peter W. McOwan. 2015. Face the music and glance: How nonverbal behaviour aids human robot relationships based in music. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 237–244.

[92] Joseph E. Michaelis, Bengisu Cagiltay, Rabia Ibtasar, and Bilge Mutlu. 2023. "Off Script:" Design opportunities emerging from long-term social robot interactions in-the-wild. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 378–387.

[93] Joseph E. Michaelis and Bilge Mutlu. 2018. Reading socially: Transforming the in-home reading experience with a learning-companion robot. *Science Robotics* 3, 21 (2018), eaat5999.

[94] François Michaud, Tamie Salter, Audrey Duquette, Henri Mercier, Michel Lauria, Helene Larouche, and Francois Larose. 2007. Assistive technologies and child-robot interaction. In *AAAI Spring Symposium on Multidisciplinary Collaboration for Socially Assistive Robotics*.

[95] Javier Movellan, Micah Eckhardt, Marjo Virnes, and Angelica Rodriguez. 2009. Sociable robot improves toddler vocabulary skills. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 307–308.

[96] Junya Nakanishi, Jun Baba, and Hiroshi Ishiguro. 2022. Robot-mediated interaction between children and older adults: A pilot study for greeting tasks in nursery schools. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 63–70.

[97] Marco Nalin, Ilaria Baroni, Ivana Kruijff-Korbayová, Lola Canamero, Matthew Lewis, Aryel Beck, Heriberto Cuayáhuitl, and Alberto Sanna. 2012. Children's adaptation in multi-session interaction with a humanoid robot. In *Proceedings of the 2012 21st IEEE International Symposium on Robot and Human Interactive Communication (IEEE RO-MAN)*. IEEE, 351–357.

[98] Amal Nanavati, Nick Walker, Lee Taber, Christoforos Mavrogiannis, Leila Takayama, Maya Cakmak, and Siddhartha Srinivasa. 2022. Not all who wander are lost: A localization-free system for in-the-wild mobile robot deployments. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 422–431.

[99] Guangtao Nie, Zhi Zheng, Jazette Johnson, Amy R. Swanson, Amy S. Weitlauf, Zachary E. Warren, and Nilanjan Sarkar. 2018. Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–4.

[100] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. 2006. Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 7, 3 (2006), 437–454.

[101] S.L. NTT Disruption Europe. 2020. Jibo–Together for You. Retrieved June 30, 2023 from https://jibo.com/

[102] Kazuko Obayashi, Naonori Kodate, and Shigeru Masuyama. 2022. Assessing the impact of an original soft communicative robot in a nursing home in Japan: Will softness or conversations bring more smiles to older people? *International Journal of Social Robotics* 14, 3 (2022), 645–656.

[103] Cagatay Odabasi, Florenz Graf, Jochen Lindermayr, Mayank Patel, Simon D. Baumgarten, and Birgit Graf. 2022. Refilling water bottles in elderly care homes with the help of a safe service robot. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 101–110.

[104] House of Commons Library. 2023. Manufacturing: Key Economic Indicators. Retrieved from https://commonslibrary.parliament.uk/research-briefings/sn05206/#:~:text=Manufacturing%20is%20one%20of%20the,employment%20in%20January%E2%80%93March%202023

[105] World Health Organization. 2021. Mental Health of Adolescents. Retrieved from https://www.who.int/news-room/fact-sheets/detail/adolescent-mental-health

[106] Anastasia K. Ostrowski, Cynthia Breazeal, and Hae Won Park. 2022. Mixed-method long-term robot usage: Older adults' lived experience of social robots. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 33–42.

[107] Maike Paetzel, Giulia Perugia, and Ginevra Castellano. 2020. The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 73–82.

[108] Roxanna Pakkar, 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 *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–7.

[109] PARO Robots U.S. Inc. 2014. PARO Therapeutic Robot. Retrieved November 5, 2024 from http://www.parorobots.com/
[110] Hannah R. M. Pelikan, Mathias Broth, and Leelo Keevallik. 2020. "Are you sad, Cozmo?" How humans make sense of a home robot's emotion displays. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 461–470.
[111] Joanna Piasek and Katarzyna Wieczorowska-Tobis. 2018. Acceptance and long-term use of a social robot by elderly users in a domestic environment. In *Proceedings of the 2018 11th International Conference on Human System Interaction (HSI). IEEE*, 478–482.
[112] Nazerke Rakhymbayeva, Aida Amirova, and Anara Sandygulova. 2021. A long-term engagement with a social robot for autism therapy. *Frontiers in Robotics and AI* 8 (2021), 669972.
[113] Aditi Ramachandran. 2018. *Building Effective Robot Tutoring Interactions for Children*. Ph.D. Dissertation. Yale University.
[114] Aditi Ramachandran, Chien-Ming Huang, and Brian Scassellati. 2019. Toward effective robot–child tutoring: Internal motivation, behavioral intervention, and learning outcomes. *ACM Transactions on Interactive Intelligent Systems* 9, 1 (2019), 1–23.
[115] Aditi Ramachandran, Alexandru Litoiu, and Brian Scassellati. 2016. Shaping productive help-seeking behavior during robot-child tutoring interactions. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 247–254.
[116] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. 2019. Personalized robot tutoring using the assistive tutor pOMDP (AT-POMDP). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 8050–8057.
[117] Rebecca Ramnauth, Emmanuel Adéníran, Timothy Adamson, Michal A. Lewkowicz, Rohit Giridharan, Caroline Reiner, and Brian Scassellati. 2022. A social robot for improving interruptions tolerance and employability in adults with ASD. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 4–13.
[118] Grand View Research. 2023. Collaborative Robots Market Size, Share & Trends Analysis Report By Payload Capacity, By Application (Assembly, Handling, Packaging, Quality Testing), By Vertical, By Region, And Segment Forecasts, 2023–2030. Retrieved from https://www.grandviewresearch.com/industry-analysis/collaborative-robots-market
[119] Claire Rivoire and Angelica Lim. 2016. Habit detection within a long-term interaction with a social robot: An exploratory study. In *Proceedings of the International Workshop on Social Learning and Multimodal Interaction for Designing Artificial Agents*, 1–6.
[120] Ben Robins, Kerstin Dautenhahn, R. Te Boekhorst, and Aude Billard. 2005. Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills? *Universal Access in the Information Society* 4 (2005), 105–120.
[121] Astrid Marieke Rosenthal-von der Pütten, Astrid Weiss, and Selma Šabanović. 2016. The challenge (not) to go wild! Challenges and best practices to study HRI in natural interaction settings. In *Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*. IEEE Press, 583–584.
[122] Matthew Rueben, Jeffrey Klow, Madelyn Duer, Eric Zimmerman, Jennifer Piacentini, Madison Browning, Frank J. Bernieri, Cindy M. Grimm, and William D. Smart. 2021. Mental models of a mobile shoe rack: Exploratory findings from a long-term in-the-wild study. *ACM Transactions on Human-Robot Interaction* 10, 2 (2021), 1–36.
[123] Selma Sabanovic, Marek P. Michalowski, and Reid Simmons. 2006. Robots in the wild: Observing human-robot social interaction outside the lab. In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control*. IEEE, 596–601.
[124] Selma Šabanović, Sarah M. Reeder, and Bobak Kechavarzi. 2014. Designing robots in the wild: In situ prototype evaluation for a break management robot. *Journal of Human-Robot Interaction* 3, 1 (2014), 70–88.
[125] Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. 2011. A conversational robot in an elderly care center: An ethnographic study. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, 37–44.
[126] Ipsita Sahin, Leena Bashir, Amanda Arnold, Herbert G. Tanner, and Elena Kokkoni. 2021. Network analysis of an infant's motor actions performed in a robot-assisted learning environment. In *Proceedings of the 2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 757–762.
[127] Jelle Saldien, Kristof Goris, Bram Vanderborght, Johan Vanderfaeillie, and Dirk Lefeber. 2010. Expressing emotions with the social robot Probo. *International Journal of Social Robotics* 2 (2010), 377–389.
[128] Nicole Salomons, Tom Wallenstein, Debasmita Ghose, and Brian Scassellati. 2022. The impact of an in-home co-located robotic coach in helping people make fewer exercise mistakes. In *Proceedings of the 2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 149–154.
[129] Tamie Salter, Kerstin Dautenhahn, and R. Bockhorst. 2004. Robots moving out of the laboratory-detecting interaction levels and human contact in noisy school environments. In *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN '04) (IEEE Catalog No. 04TH8759)*. IEEE, 563–568.

[130] Anara Sandygulova, Aida Amirova, Zhansaule Telisheva, Aida Zhanatkyzy, and Nazerke Rakhymbayeva. 2022. Individual differences of children with autism in robot-assisted autism therapy. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 43–52.

[131] Brian Scassellati, 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. *Science Robotics* 3, 21 (2018), eaat7544.

[132] Diane J. Schiano, Coreena P. Chen, Ellen Isaacs, Jeremy Ginsberg, Unnur Gretarsdottir, and Megan Huddleston. 2002. Teen use of messaging media. In *Proceedings of the CHI'02 Extended Abstracts on Human Factors in Computing Systems*, 594–595.

[133] Arielle A. J. Scoglio, Erin D. Reilly, Jay A. Gorman, and Charles E. Drebing. 2019. Use of social robots in mental health and well-being research: Systematic review. *Journal of Medical Internet Research* 21, 7 (2019), e13322.

[134] Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F. Jung. 2020. Robots in groups and teams: A literature review. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–36.

[135] Kerstin Severinson-Eklundh, Anders Green, and Helge Hüttenrauch. 2003. Social and collaborative aspects of interaction with a service robot. *Robotics and Autonomous Systems* 42, 3–4 (2003), 223–234.

[136] Zhonghao Shi, 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. *ACM Transactions on Human-Robot Interaction* 11, 4 (2022), 1–28.

[137] Elaine Short, 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 *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 924–929.

[138] David Silvera-Tawil, DanaKai Bradford, and Christine Roberts-Yates. 2018. Talk to me: The role of human-robot interaction in improving verbal communication skills in students with autism or intellectual disability. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–6.

[139] David Silvera-Tawil and Christine Roberts-Yates. 2018. Socially-assistive robots to enhance learning for secondary students with intellectual disabilities and autism. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 838–843.

[140] Divyanshu Kumar Singh, Manohar Kumar, Eduard Fosch-Villaronga, Deepa Singh, and Jainendra Shukla. 2023. Ethical considerations from child-robot interactions in under-resourced communities. *International Journal of Social Robotics* 15, 12 (2023), 2055–2071.

[141] Catharina Vesterager Smedegaard. 2019. Reframing the role of novelty within social HRI: From noise to information. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE*, 411–420.

[142] Catharina V. Smedegaard. 2022. Novelty knows no boundaries: Why a proper investigation of novelty effects within SHRI should begin by addressing the scientific plurality of the field. *Frontiers in Robotics and AI* 9 (2022), 741478.

[143] SoftBank Robotics. 2024. Pepper the Humanoid Robot. Retrieved November 5, 2024 from https://us.softbankrobotics.com/pepper

[144] Micol Spitale, Minja Axelsson, and Hatice Gunes. 2023. Robotic Mental Well-being Coaches for the Workplace: An In-the-Wild Study on Form. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*. Association for Computing Machinery, New York, NY, USA, 301–310. DOI: https://doi.org/10.1145/3568162.3577003

[145] Kristen Stubbs, Debra Bernstein, Kevin Crowley, and Illah R. Nourbakhsh. 2005. Long-term human-robot interaction: The personal exploration rover and museum docents. In *Proceedings of International Conference on Artificial Intelligence in Education*, 621–628.

[146] JaYoung Sung, Henrik I. Christensen, and Rebecca E. Grinter. 2009. Robots in the wild: Understanding long-term use. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 45–52.

[147] Luise Süssenbach, Nina Riether, Sebastian Schneider, Ingmar Berger, Franz Kummert, Ingo Lütkebohle, and Karola Pitsch. 2014. A robot as fitness companion: Towards an interactive action-based motivation model. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 286–293.

[148] Fumihide Tanaka, Aaron Cicourel, and Javier R. Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences* 104, 46 (2007), 17954–17958.

[149] Lloyd Taylor, Angela Downing, Gabriel Aguiar Noury, Giovanni Masala, Marco Palomino, Conor McGinn, and Ray Jones. 2021. Exploring the applicability of the socially assistive robot Stevie in a day center for people with dementia. In *Proceedings of the 2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 957–962.

[150] Nils F. Tolksdorf and Katharina J. Rohlfing. 2020. Parents' views on using social robots for language learning. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 634–640.

[151] Thanh Q. Trinh, Alexander Vorndran, Benjamin Schuetz, Bianca Jaeschke, Anke Mayfarth, Andrea Scheidig, and Horst-Michael Gross. 2020. Autonomous mobile gait training robot for orthopedic rehabilitation in a clinical environment. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 580–587.

[152] Nathan Tsoi, Joe Connolly, Emmanuel Adéníran, Amanda Hansen, Kaitlynn Taylor Pineda, Timothy Adamson, Sydney Thompson, Rebecca Ramnauth, Marynel Vázquez, and Brian Scassellati. 2021. Challenges deploying robots during a pandemic: An effort to fight social isolation among children. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 234–242.

[153] Esther J. G. Van Der Drift, Robbert-Jan Beun, Rosemarijn Looije, Olivier A. Blanson Henkemans, and Mark A. Neerincx. 2014. A remote social robot to motivate and support diabetic children in keeping a diary. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, 463–470.

[154] Ajay Vishwanath, Aalind Singh, Yi Han Victoria Chua, Justin Dauwels, and Nadia Magnenat-Thalmann. 2019. Humanoid co-workers: How is it like to work with a robot? In *Proceedings of the 2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–6.

[155] Paul Vogt, Rianne van den Berghe, Mirjam De Haas, Laura Hoffman, Junko Kanero, Ezgi Mamus, Jean-Marc Montanier, Cansu Oranç, Ora Oudgenoeg-Paz, Daniel Hernández García, et al. 2019. Second language tutoring using social robots: A large-scale study. In *Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 497–505.

[156] Kazuyoshi Wada and Takanori Shibata. 2006. Robot therapy in a care house-its sociopsychological and physiological effects on the residents. In *Proceedings 2006 IEEE International Conference on Robotics and Automation (ICRA '06)*. IEEE, 3966–3971.

[157] Kazuyoshi Wada and Takanori Shibata. 2007. Living with seal robots—Its sociopsychological and physiological influences on the elderly at a care house. *IEEE Transactions on Robotics* 23, 5 (2007), 972–980.

[158] Kazuyoshi Wada, Yuta Takasawa, and Takanori Shibata. 2013. Robot therapy at facilities for the elderly in Kanagawa prefecture-A report on the experimental result of the first week. In *Proceedings of the 2013 IEEE RO-MAN*. IEEE, 757–761.

[159] Kazuyoshi Wada, Yuta Takasawa, and Takanori Shibata. 2014. Robot therapy at facilities for the elderly in Kanagawa prefecture—A report on the experimental result of the first month. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 193–198.

[160] Astrid Weiss, Anna Pillinger, and Christiana Tsiourti. 2021. Merely a conventional 'diffusion' problem? On the adoption process of Anki Vector. In *Proceedings of the 2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 712–719.

[161] J. Kory Westlund, Goren Gordon, Samuel Spaulding, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2015. Learning a second language with a socially assistive robot. In *Conference Proceedings New Friends 2015. The 1st international conference on social robots in therapy and education*. Marcel Heerink and Michiel de Jong (Eds.), Windesheim Flevoland, Almere, The Netherlands, 32–33.

[162] Jacqueline Kory Westlund, Goren Gordon, Samuel Spaulding, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Lessons from teachers on performing HRI studies with young children in schools. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 383–390.

[163] Aida Zhanatkyzy, Zhansaule Telisheva, Aida Amirova, Nazerke Rakhymbayeva, and Anara Sandygulova. 2023. Multi-purposeful activities for robot-assisted autism therapy: What works best for children's social outcomes? In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 34–43.

[164] Xiajie Zhang, Cynthia Breazeal, and Hae Won Park. 2023. A social robot reading partner for explorative guidance. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 341–349.

[165] Zhao Zhao and Rhonda McEwen. 2022. "Let's read a book together": A long-term study on the usage of pre-school children with their home companion robot. In *Proceedings of the 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 24–32.

Received 16 May 2024; revised 9 November 2024; accepted 8 December 2024

Appendix

A Dataset and Study Characteristics

Table A1. A Compilation of Data and Key Characteristics from the Studies Reviewed in This Article

		Per User									
Citation	**Year**	**Period (Days)**	**# Sessions**	**Participants**	**HRI Type**	**Domain**	**Robot Used**	**Location**	**Age Group**	**Results**	**Operation**
Klamer et al. [68]	2011	10	40	3	Diadic	Mental Health	Nabaztag	Home	Adults	Ql	A
Pelikan et al. [110]	2020	11	unk.	20	Group	Entertainment	Cozmo	Home	Children	Ql	A
Lemaignan et al. [83]	2022	21	Daily	30	Mixed	ASD	Pepper	School	Children	Ql, Qn	A
Zhao and McEwen [165]	2022	180	75	27	Diadic	Education	Luka	Home	Children	Ql, Qn	S
Michaelis et al. [92]	2023	28	27	6	Family	Education	Misty	Home	Children	Ql	A
Sabelli et al. [125]	2011	105	23	55	Group	General Purpose	Robovie2	Care Home	Elderly	Ql	N
Barco et al. [8]	2014	180	120	15	Diadic	Mental Health	Lego	Home	Mixed	Qn	A
Rakhymbayeva et al. [112]	2014	21	10	11	Triadic	ASD	NAO	Facility	Children	Ql, Qn	A
François et al. [37]	2009	70	10	6	Triadic	ASD	Aibo	School	Children	Ql	S
Süssenbach et al. [147]	2014	18	18	16	Diadic	Physical Health	NAO	Lab	Adults	Ql, Qn	A
Jeong et al. [59]	2023	28	12	70	Diadic	Mental Health	Jibo	Home	Adults	Ql, Qn	A
Jeong et al. [58]	2020	12	7	35	Diadic	Mental Health	Jibo	Home	Adults	Ql, Qn	A
Cruz-Sandoval et al. [24]	2020	45	12	8	Group	Mental Health	Eva	Care Home	Elderly	Ql, Qn	A
Ramnauth et al. [117]	2022	7	73	10	Diadic	ASD	Jibo	Home	Adults	Ql, Qn	A
Zhang et al. [164]	2023	21	8	31	Diadic	Education	Jibo	Home	Children	Qn	A
Fasola and Matarić [34]	2013	14	4	33	Diadic	Physical Health	Bandit	Lab	Elderly	Qn	A
Kanda et al. [62]	2007	32	unk.	37	Mixed	Entertainment	Robovie	School	Children	Ql, Qn	A
Piasek and Wieczorowska-Tobis [111]	2018	70	Daily	10	Diadic	Mental Health	Tiago	Home	Elderly	Ql	A
Carrillo et al. [89]	2018	14	14	9	Diadic	Physical Health	NAO	Hospital	Children	Ql, Qn	A
Jones and Castellano [60]	2018	30	4	24	Diadic	Education	NAO	School	Children	Qn	A
Ahmad et al. [3, 4]	2017	10	9	23	Diadic	General Purpose	NAO	School	Children	Ql, Qn	N
Leite et al. [79]	2009	35	5	5	Diadic	Education	iCat	Other	Children	Ql, Qn	A
Obayashi et al. [102]	2022	180	Daily	34	Group	Mental Health	Mon-chan	Care home	Elderly	Ql, Qn	A
Michaud et al. [94]	2007	49	22	2	Group	ASD	Puppet	Facility	Children	Ql	N
Trinh et al. [151]	2020	3	5	20	Diadic	Physical Health	Patterns	Hospital	Elderly	Ql, Qn	A
Jacq et al. [54]	2016	19	3.5	8	Diadic	Education	NAO	Facility	Children	Ql, Qn	A
Nalin et al. [97]	2012	35	3	13	Diadic	General Purpose	NAO	Hospital	Children	Ql, Qn	S
Leite et al. [81]	2015	21	3	40	Mixed	Education	Keepon	School	Children	Qn	A
Ligthart et al. [85]	2023	unk.	3	130	Diadic	Education	NAO	School	Children	Qn	A
Pakkar et al. [108]	2019	30	unk.	8	Mixed	ASD	Kiwi	Home	Children	Ql	A
Chen et al. [19]	2022	32	6	12	Triadic	Education	Jibo	Home	Mixed	Ql	S

(Continued)

Table A1. Continued

Citation	Year	Per User		Participants	HRI Type	Domain	Robot Used	Location	Age Group	Results	Operation
		Period (Days)	# Sessions								
Šabanović et al. [124]	2014	28	Daily	6	Diadic	Mental Health	Custom	Office	Adults	Ql, Qn	A
Chandra et al. [18]	2018	30	4	25	Diadic	Education	NAO	School	Children	Qn	A
Chevalier et al. [20]	2017	56	5	12	Diadic	ASD	NAO	Care home	Teens	Ql, Qn	A
Sung et al. [146]	2010	75	Daily	48	Family	General Purpose	Roomba	Home	Adults	Ql, Qn	A
Lane et al. [73]	2016	570	Daily	23	Diadic	Mental Health	Paro	Facility	Elderly	Ql, Qn	A
Leite et al. [80, 81]	2015	21	3	40	Mixed	Education	Keepon	School	Children	Qn	A
Leite et al. [77]	2014	35	5	16	Diadic	Education	iCat	School	Children	Ql, Qn	A
Alemi et al. [6]	2014	35	10	30	Group	Education	NAO	School	Children	Qn	A
Coşar et al. [23]	2020	70	Daily	15	Diadic	Mental Health	Tiago	Home	Elderly	Ql, Qn	A
Singh et al. [140]	2022	35	30	12	Triadic	Education	Cozmo	School	Children	Ql, Qn	N
Taylor et al. [149]	2021	10	20	9	Observer	Entertainment	Custom	Daycare	Adults	Ql	S
McCallum and McOwan [91]	2015	42	6	10	Diadic	Entertainment	Mortimer	Lab	Adults	Qn	A
Fan et al. [33]	2021	21	21	15	Diadic	Mental Health	NAO	Care home	Elderly	Ql, Qn	N
Jeong et al. [57]	2018	28	unk.	12	Mixed	General Purpose	Fribo	Home	Adults	Ql	A
Greczek et al. [45]	2014	18	5	12	Diadic	ASD	NAO	School	Children	Ql, Qn	A
Rivoire et al. [119]	2016	56	Daily	10	Family	General Purpose	Pepper	Home	Adults	Ql, Qn	A
Fernaeus et al. [36]	2010	180	unk.	6	Group	Entertainment	Pleo	Home	Children	Ql	A
Short et al. [137]	2014	21	6	26	Diadic	Physical Health	DragonBot	School	Children	Qn	N
Vishwanath et al. [154]	2019	30	unk.	12	Observer	Service and Workplace	Nadine	Office	Adults	Ql	A
Scassellati et al. [131]	2018	30	23	12	Triadic	ASD	Jibo	Home	Children	Ql, Qn	A
Sandygulova et al. [130]	2022	21	7	34	Triadic	ASD	NAO	Facility	Children	Ql, Qn	S
Luperto et al. [87]	2022	14	Daily	13	Diadic	Mental Health	Giraff-X	Home	Elderly	Ql, Qn	A
Kanda et al. [61]	2004	14	9	228	Group	Education	Robovie	School	Children	Ql	A
Giusti and Marti [42]	2006	30	8	5	Group	Mental Health	Paro	Care home	Elderly	Ql	A
Donnermann et al. [32]	2022	unk.	3	41	Diadic	Education	Pepper	School	Adults	Ql, Qn	A
Kozima et al. [72]	2009	unk.	unk.	25	Diadic	General Purpose	Keepon	Lab	Children	Ql, Qn	N
Kozima et al. [72]	2009	unk.	20	27	Group	General Purpose	Keepon	School	Children	Ql	N
Kozima et al. [72]	2009	150	15	2	Mixed	General Purpose	Keepon	School	Children	Ql	N
Kory-Westlund et al. [161]	2015	60	8	34	Diadic	General Purpose	Tega	School	Children	Qn	A
Kory-Westlund et al. [162]	2016	60	7	34	Diadic	Education	Tega	Daycare	Children	Ql	A
Hebesberger et al. [49]	2016	30	7	10	Group	Physical Health	SCITOS G5	Care home	Elderly	Ql, Qn	A
Afyouni et al. [1]	2022	7	unk.	8	Diadic	Physical Health	Pepper	Facility	Adults	Ql	A
Koay et al. [69]	2007	35	8	12	Diadic	General Purpose	Custom	Lab	Adults	Qn	S
de Graaf et al. [29]	2016	unk.	unk.	102	Mixed	General Purpose	Karotz	Home	Mixed	Qn	A
Stubbs et al. [145]	2005	105	unk.	11	Observer	Education	PER Rover	Other	Adults	Qn	A

(Continued)

Table A1. Continued

Citation	Year	Per User		Participants	HRI Type	Domain	Robot Used	Location	Age Group	Results	Operation
		Period (Days)	# Sessions								
Clabaugh et al. [21]	2019	41	14	17	Diadic	ASD	Kiwi	Home	Children	Qn	A
Begum et al. [12]	2015	8	18	3	Triadic	ASD	NAO	School	Children	Ql, Qn	N
Ligthart et al. [85]	2022	60	5	46	Diadic	General Purpose	NAO	School	Children	Qn	A
Rueben et al. [122]	2021	42	5	6	Mixed	Service and Workplace	Custom	Other	Mixed	Ql	N
Weiss et al. [160]	2021	210	unk.	8	Family	General Purpose	Vector	Home	Mixed	Ql	A
Ostrowski et al. [106]	2022	150	1,559	28	Diadic	General Purpose	Jibo	Home	Elderly	Ql	A
Jain et al. [55]	2020	30	unk.	7	Diadic	ASD	Kiwi	Home	Children	Qn	A
Gamborino et al. [38]	2019	5	5	14	Diadic	Entertainment	ROBOHON	Lab	Children	Ql, Qn	A
Janssen et al. [56]	2011	14	3	20	Diadic	Education	NAO	School	Children	Qn	N
Zhanatkyzy et al. [163]	2023	21	6	34	Triadic	ASD	NAO	Facility	Children	Qn	N
Sahin et al. [126]	2021	28	8	1	Diadic	Mental Health	NAO and Dash	Lab	Infants and Toddlers	Qn	N
Gasteiger et al. [39]	2021	6	Free	6	Diadic	Mental Health	Bomy	Care home	Elderly	Ql	A
Tolksdorf et al. [150]	2020	14	4	29	Observer	Education	NAO	Lab	Mixed	Ql	N
Napoli et al. [30]	2022	65	Daily	7	Diadic	Mental Health	Sanbot Elf	Home	Elderly	Ql, Qn	A
Ramachandran et al. [114, 116]	2019	21	5	28	Diadic	Education	NAO	School	Children	Ql, Qn	A
Nie et al. [99]	2018	unk.	5	34	Diadic	ASD	NAO	Lab	Children	Qn	A
Koay et al. [69]	2016	35	10	9	Diadic	General Purpose	Sunflower	Lab	Adults	Ql, Qn	A
Michaelis et al. [92]	2018	14	Daily	24	Diadic	Education	Maki	Home	Children	Ql, Qn	A
Hyun et al. [51]	2010	14	10	111	Group	General Purpose	iRobiQ	Daycare	Children	Qn	A
Bajones et al. [7]	2019	21	free	16	Diadic	Physical Health	Hobbit PT2	Home	Elderly	Ql, Qn	A
Ahtinen et al. [5]	2023	28	unk.	32	Family	Education	Alpha Mini	Home	Children	Ql	S
Baxter et al. [10]	2017	14	3	59	Diadic	Education	NAO	School	Children	Qn	A
Wada et al. [159]	2014	30	unk.	64	Diadic	Mental Health	Paro	Care home	Elderly	Ql	A
Wada et al. [158]	2013	7	unk.	80	Diadic	Mental Health	Paro	Care home	Elderly	Ql	A
Wada et al. [157]	2007	52	unk.	12	Group	Mental Health	Paro	Care Home	Elderly	Ql, Qn	A
Wada et al. [156]	2006	52	unk.	12	Group	Mental Health	Paro	Care home	Elderly	Ql, Qn	A
Gasteiger et al. [39]	2021	84	15	10	Diadic	Mental Health	Bomy	Care home	Elderly	Ql	A
Nakanishi et al. [96]	2022	9	18	40	Triadic	General Purpose	Sota	Daycare	Infants and Toddlers	Qt	N
Robins et al. [120]	2005	180	9	4	Diadic	ASD	Robota	School	Children	Ql, Qn	N
Spitale et al. [144]	2023	28	4	26	Diadic	Mental Health	Misty and QT	Office	Adults	Ql, Qn	A
Kidd and Breazeal [66]	2008	42	51	15	Diadic	Physical Health	Custom	Home	Mixed	Qn	A
Salter et al. [129]	2004	unk.	5	8	Diadic	ASD	Unknown	School	Children	Ql, Qn	A

(Continued)

Table A1. Continued

Citation	Year	Per User		Participants	HRI Type	Domain	Robot Used	Location	Age Group	Results	Operation
		Period (Days)	# Sessions								
Vogt et al. [155]	2019	21	7	108	Diadic	Education	NAO	School	Children	Qn	S
Ramachandran et al. [115]	2016	14	4	29	Diadic	Education	NAO	School	Children	Qn	A
de Graaf et al. [28]	2015	30	22	6	Diadic	Physical Health	Karotz	Home	Elderly	Ql	A
Movellan et al. [95]	2009	12	12	9	Diadic	Education	Custom	Daycare	Infants and Toddlers	Qn	A
Severinson -Eklundh et al. [135]	2003	90	unk.	1	Diadic	General Purpose	Custom	Office	Adults	Ql	S
Polak and Levy-Tzedesk [35]	2020	41	15	4	Diadic	Physical Health	Pepper	Facility	Mixed	Ql	A
Tanaka et al. [148]	2007	150	45	11	Group	Entertainment	QRIO	Daycare	Infants and Toddlers	Ql, Qn	S
Silvera-Tawil and Yates [139]	2018	150	unk.	45	Observer	ASD	NAO and Paro	School	Adults	Ql	A
Silvera-Tawil et al. [138]	2018	280	160	3	Mixed	ASD	NAO	School	Teens	Ql	A
Bodala et al. [13]	2021	35	5	9	Group	Mental Health	Pepper	Lab	Mixed	Ql, Qn	A
Leyzberg et al. [84]	2018	14	5	19	Diadic	Education	Keepon	School	Children	Qn	A
Salomons et al. [128]	2022	14	14	14	Diadic	Mental Health	Keepon	Home	Adults	Qn	A
Agrigoroaie and Tapus [2]	2018	4	8	1	Diadic	Mental Health	Tiago	Home	Elderly	Qn	A
Paetzel et al. [107]	2020	13	3	40	Diadic	General Purpose	Furhat	Lab	Adults	Qn	A
Ramachandran et al. [114, 116]	2019	14	4	29	Diadic	Education	NAO	School	Children	Qn	A
Shi et al. [136]	2022	30	5	4	Diadic	ASD	Kiwi	Home	Children	Qn	A
Coninx et al. [22]	2016	45	3	3	Diadic	Physical Health	NAO	Hospital	Children	Ql, Qn	N
Cagiltay et al. [15]	2022	28	12	14	Diadic	Education	Misty	Home	Children	Ql	A
Edirisinghe and Jayasekara [32]	2018	7	unk.	4	Mixed	General Purpose	Unknown	Lab	Mixed	Qn	A
Irfan et al. [53]	2020	126	35	1	Diadic	Physical Health	NAO	Facility	Adults	Ql, Qn	A
de Graaf et al. [27]	2017	unk.	unk.	102	Mixed	General Purpose	Karotz	Home	Mixed	Ql	A
Davison et al. [26]	2020	120	6	20	Diadic	Education	Zeno	School	Children	Qn	A

Each entry provides the data extracted from a study and used in our analysis. This includes the study period as measured in days, number of sessions, number of participants with full data, type of HRI interaction, study domain, robot platform used, location of the study, age group of the participants, type of results reported, and robot operation type. To improve the clarity of the table, abbreviations were used as follows: *unk.*, unknown or not reported by the study authors; *Ql*, qualitative data; *Qn*, quantitative data; *A*, fully-autonomous; *S*, semi-autonomous; *N*, non-autonomous. The study domains are also shortened in the table for space but correspond with the categories described in Section 4.2.