Systematic Meta-Analysis of Social Robotics and AI: Comparative Interactions Among Humanoid Robots, Humans, and Digital Entities Across Embodiment and Interaction Contexts

Caroline Elisabeth Hallmann, Master Thesis, August 27, 2024[a]

[a]*Vrije Universiteit Amsterdam , Msc. Artificial Intelligence, Amsterdam*

Abstract

This systematic literature review examines the comparative effectiveness of social robots in various domains such as healthcare, education, and service. From an initial pool of 273 articles, 80 studies were included in the final analysis, following the PRISMA guidelines to ensure methodological rigor and transparency. The review focuses on comparing social robots to humans, other robots, other devices, and digital avatars across multiple interaction contexts. The findings indicate that while social robots often outperform avatars and other robotic systems in specific domains, they generally fall short of human capabilities in terms of social presence, attachment, and interaction quality. The study underscores the significance of embodiment and anthropomorphism in shaping user interactions with social robots. This review provides critical insights for the future development of social robots, emphasizing the need for continued advancements to bridge the gap between human and robotic social interactions.

Keywords: Embodiment, Anthropomorphism, Social Robotics, Devices, Human-Robot Interaction (HRI), Human-Computer Interaction (HCI)

1. Introduction

Human-Computer Interaction (HCI) is a multidisciplinary field encompassing various disciplines, dedicated to the study and development of interactive systems that facilitate effective communication and interaction between humans and computers [95]. HCI aims to enhance user experience and usability by incorporating design principles, cognitive processes, user behavior, and social dynamics into its framework. With the evolution of HCI, the scope has expanded beyond conventional computing systems to incorporate novel technologies such as conversational agents and social robots [38].

While HCI focuses on human-machine interactions across various technologies, it is important to differentiate it from Human-Robot Interaction (HRI) research. HRI specifically emphasizes interactions between humans and robots [45]. HRI investigates the design, behavior, and impact of robots in different contexts, including social interactions and collaborative tasks.

Within the realm of HRI, social robotics plays a prominent role, involving the study and development of robots designed to interact with humans in social settings [33]. Social robots are designed to exhibit social behaviors, engage in meaningful interactions, and elicit emotional responses from humans. Social robots are now found in various domains, including healthcare, education, entertainment, and customer service [100].

1.1. Anthropomorphism

A common theme that is both studied in HCI and HRI research is Anthropomorphism, defined by [16] as " the attribution of human-like characteristics and behaviors to non-human entities". In recent years, vast developments in social robotics and interactive systems have increased the need to understand the role of anthropomorphism in human-machine interactions. Anthropomorphism is of particular significance in the study of social robots, as these systems are designed to engage and interact with humans socially. By attributing human-like qualities to robots, such as facial expressions, gestures, and conversational abilities, anthropomorphism can enhance the perceived social presence and engagement of these machines [29].

Furthermore, in the HCI field, understanding the effects of anthropomorphism on user experiences and attitudes is crucial for designing more effective and user-centric interactive systems. Anthropomorphic features can influence users' perceptions of the system's trustworthiness, likability, and effectiveness, thereby impacting their acceptance and adoption of the technology [18]. Therefore, investigating the role of anthropomorphism in social robots and HCI is essential to advance our understanding of human-machine interactions and inform the design of future interactive systems.

1.2. Embodiment

Complementing the concept of anthropomorphism, embodiment is another fundamental aspect investigated in HCI research that significantly influences human-machine interactions. Embodiment, as defined by [72] refers to " the physical representation of digital entities, encompassing diverse configurations such as audio-based interactions, virtual avatars, and humanoid robots".

Humanoid robots are physical representations of digital entities that closely resemble human form and movement [35].

These robots offer unique opportunities for integrating non-verbal cues, gestures, and anthropomorphic features, thereby enhancing the immersive and engaging nature of user interactions (ibid.). Moreover, audio-based interactions utilize sound and voice to augment user experiences, fostering a sense of presence and immersion. Virtual avatars, on the other hand, provide users with the ability to interact with digital representations that simulate realistic human behavior and characteristics (ibid.).

These types of embodiment offer opportunities to integrate non-verbal cues, gestures, and anthropomorphic features, thereby enhancing the overall user experience. By creating immersive and engaging environments, embodiment contributes to a sense of presence and realism, allowing users to interact with technology more naturally and intuitively [97].

1.3. Evaluation Criteria

Understanding the impact of different types of embodiment, such as audio-based interactions, virtual avatars, and humanoid robots, on key evaluative criteria is of great importance. Criteria such as loneliness, social presence, and trustworthiness are crucial indicators of the effectiveness and acceptability of interactive systems [63]. While audio-based devices can facilitate communication, the embodiment in HCI and HRI research plays a significant role in influencing these criteria.

Research has shown that embodiment enhances social presence, allowing users to feel a stronger connection and sense of co-presence with the interactive system [2]. Moreover, the physical appearance and behaviors of embodied entities, such as virtual avatars or humanoid robots, can evoke emotional responses and shape perceptions of trustworthiness [119]. Embodiment enables the integration of non-verbal cues, gestures, and anthropomorphic features that contribute to a more natural and engaging user experience.

1.4. Previous Research

To gain a comprehensive understanding of the effects of different types of social AI systems, it is crucial to evaluate them in terms of their impact on human-machine interactions. This evaluation allows for the assessment of user experiences, social perceptions, and acceptance of these technologies [34]. The study conducted by [107] specifically investigates the social acceptance of humanoid robots, comparing two models, NAO and Pepper, to shed light on the factors influencing people's acceptance and attitudes towards different humanoid robots.

Furthermore, the study by [5] proposed an extended framework for characterizing social robots. The framework is used for categorizing social robots based on their physical appearance, functionality, interaction modalities, and user experience. This comprehensive approach facilitates a deeper understanding and analysis of the role of embodiment in human-machine interactions (ibid.)

In conclusion, the study of different types of social AI systems in the context of Human-Computer Interaction (HCI) is a rich and evolving field that explores the impact of AI embodiment on human-machine interactions. The previous literature highlights the significance of embodiment in social robots, such as humanoid robots, as well as audio-based systems and virtual avatars, in facilitating immersive and engaging interactions with users [34]. Moreover, the extended framework proposed by [5] provides a comprehensive approach to characterize social robots based on their physical appearance, functionality, interaction modalities, and user experience.

1.5. Purpose of the study

The overall aim of this study is to gain a better understanding and more insight into how social robots compare to humans and other devices in different domains and on various dimensions. To explore this topic, three research questions are formulated. The first research question **RQ1** investigates *which types of social robots and devices have been used in comparative studies*. To investigates this, existing work from [5] to classify social robots and introduce our taxonomy to categorize the other entities (including humans) to which those social robots are compared. The approach includes extraction of all specific information about the social robots that have been compared in the literature, and, to introduce general categories to classify with which types of entities they have been compared. The second research question **RQ2** investigates *how social robots have been compared with humans and other devices*. This question is addressed by extracting the primary measurement scales and data collection tools of the included articles that compare social robots with humans and other devices.

The extracted scales and tools are further classified into different categories to get a broader picture of what tools have been used to compare social robots.

The third and last research question **RQ3** investigates *how social robots compare to humans and other devices*. To this end, we zoom in on the set of papers that compare physically present social robots with humans and other devices. The chosen approach includes the extraction of the main results about how social robots compare from all papers that are part of the main scope of this review. We use our classification of entities to analyze how social robots compare to these entities.

Conducting a systematic literature review (SLR) of HCI and HRI domain-specific articles that compare humans, and social robots to other devices is the chosen study approach for this research paper. To provide relevant insights, only select articles that were published from year 2013 to 2024 are used. Furthermore, our method includes extracting social robotic-specific articles from 3 databases and refining them through the AI-powered SLR tool 'Rayyan' [87].

2. Systematic Literature Review

In this section, we elaborate on the method used to conduct a systematic literature review (SLR). This is an essential step in systematically reviewing and synthesizing existing academic literature that compares various types of robots, including humanoid, service, and social robots, within the context of Human-Robot Interaction (HRI).

The systematic literature review (SLR) includes the extraction of data from several studies that can be aggregated to create an overarching picture of what exists in the literature on a given topic [81]. It helps overcome the downfall of significance testing by accumulating related results so that even small or non-significant effects are included. Collecting and combining data also allows for the investigation of potential mediator and moderator variables across a larger data set (ibid.).

2.1. Methodology

To address the three research questions, the systematic literature review (SLR) was conducted as a foundational step. It involved the establishment of a systematic search methodology. The systematic search was established with the inclusion criteria which acted as the basis for generating literature. The inclusion criteria used on the 3 databases were defined, as shown in the forthcoming table (Table 1).

Criteria	Description
1.Language	Published in English
2.Source Type	Peer-reviewed journal articles or conference papers
3.Source Credibility	Retrieved from IEEE, ACM, Elsevier Scopus
4.Relevance to Search Key-terms	Must pertain to the specified search keywords
5.Publication Timeline	Published within the last decade (2014-2023)

Table 1: Inclusion Criteria for database article retrieval

2.1.1. Data Source

The data for the systematic literature review (SLR) was collected from three prominent academic databases: ACM, Elsevier Scopus, and IEEE. These databases were chosen due to their extensive coverage of AI and robotics research.

2.1.2. Data Collection and Selection

The research process followed a top-down methodology, starting with the initial investigation of the primary research question. To source pertinent literature for the first research question, **RQ1**: *"Which types of social robots and devices have been used in comparative studies?"* broad-spectrum keywords were employed to explore multiple academic databases. These keywords were consistently refined and augmented as the search process unfolded, with the resultant literature being efficiently managed using the 'Rayyan' literature search tool. Via Rayyan, the de-duplication function was used to eliminate duplicates.

2.1.3. Initial Article Retrieval and de-duplication

After an extensive search across the mentioned databases, the total number of articles per database, and the total number of articles after duplicate removal were retrieved.

The outcomes of the systematic search are summarized in Table 2, which presents the search results obtained from each of the specified databases from the year 2014 up to the year 2023.

Keywords	Elsevier	ACM	IEEE	Total
"Social robot*" AND compar*	1662	979	257	2509
"Humanoid robot*" AND compar*	1468	1004	1492	3309
"Service robot*" AND compar*	759	317	1750	2450
"Social robot*" OR "Humanoid robot*" OR "Service robot*" AND compar*	3889	2300	3499	7502

Table 2: Database search results and total after duplicate removal from year 2014 up to 2023

2.1.4. Round 1 - Title and Abstract Screening

The final set of articles retrieved from keywords "Social robot*" OR "Humanoid robot*" OR "Service robot*" AND compar* (Table 2) after de-duplication, eliminating 2186 articles, was downloaded and uploaded to a new 'Rayyan' systematically review worksheet. Two authors carried out title and abstract screening; Caroline Hallmann and Prof. Dr. Hindricks. The first step was to reduce the number of articles by applying the exclusion criteria defined below. During this stage, titles and abstracts were screened according to a comprehensive set of criteria to ensure the selection of only high-quality and relevant articles.

2.1.5. Round 1 - Article Exclusion Criteria

Below are the exclusion criteria along with the respective number of articles excluded based on each criterion:

1. *Not a Comparison of Two Entities* (6,355); Articles were excluded due to the absence of a comparative analysis involving two distinct entities, a fundamental prerequisite for our study.
2. *Short Paper (Less Than 6 Pages)* (475); Articles containing less than six pages of content were excluded.
3. *Wrong Publication Type* (205); Articles that did not align with the appropriate publication type for this meta-analysis were excluded.
4. *Survey* (165); Articles identified as survey papers rather than comparative studies were excluded.
5. *No Study Reported* (7); Articles that failed to report any substantive study were excluded.
6. *Additional duplicate* (17); Two or more articles containing identical information that should be unique for each entry were excluded.
7. *Foreign Language* (4); Articles not composed in English were excluded.

Following the exclusion criteria, 7,228 articles were excluded, resulting in a refined dataset of 273 articles.

2.1.6. Round 1 - Article Inclusion Criteria

The remaining 273 articles were screened and categorized based on specific inclusion criteria, each associated with its respective count of articles. The inclusion criteria are defined below:

1. Robot-Human (141): Articles that compare robots and humans. These articles were focused on interactions between robots and human subjects.

2. Robot-Robot (66): Articles that center on the interaction, comparison, or analysis of multiple robots.

3. Relevant (36): This category encompasses articles that, although not used for the primary analysis, held relevance for theoretical explanations and contextualization within the thesis.

4. Robot-Avatar (31): Articles where robots were compared or integrated with avatars, often within virtual or simulated environments.

5. Robot-Computer (10): Articles concerning interactions between robots and computer systems.

6. Robot-Speaker (16): Articles that feature robots in a role involving speech or vocal interactions.

7. Robot-Tablet (12): Articles involving the utilization of tablets in conjunction with robots, typically for human interaction.

8. Robot-Tele (9): Articles focusing on remote or teleoperated robots and their interactions.

9. Validity (9): Articles identified as having particular relevance for establishing the validity of the meta-analysis.

10. Robot-Smartphone (5): Articles involving the integration or comparison of robots with smartphone technology.

11. Robot-Animal (5): Articles that explore interactions between robots and animals.

Several articles met more than one inclusion criterion, resulting in a total count that exceeded the initial set of 281 articles, indicating a richer diversity within the dataset than initially identified

It is noteworthy that articles classified under "Relevant" and "Validity" were earmarked for theoretical and contextual utilization in the thesis but were not integrated into the primary analysis.

2.1.7. Data Refinement with SQL Organization

From an initial dataset of 7,502 articles, a refined selection process utilized exclusion criteria for title and abstract screening. This process resulted in the reduction of the dataset to 273 articles, eliminating 7,228 articles. Subsequently, through a rigorous screening procedure aligning with predefined inclusion criteria, 273 articles were identified as meeting the specified criteria, thereby affirming their relevance to the research objectives. SQL was used to efficiently organize and refine the dataset during this process. Structured queries allowed for the removal of duplicates, sorting articles based on key metadata such as publication year and research domain, and filtering entries that met inclusion criteria. Following this SQL-driven refinement, an Excel file was created to further analyze the 273 articles by extracting and evaluating key information across 8 categories. This structured approach provided a thorough understanding of the research landscape, facilitating insightful conclusions and contributions to the field.

2.2. Round 2 - Screening Round for Article Selection

All 273 remaining articles were downloaded as a PDF file for full article screening and the 8 categories were annotated via Excel for each article.

2.2.1. Round 2 - Description of 8 evaluation categories

1. *Year of publication*; The year the article was published.
2. *Participation task*; The experiment or task the participants had to complete during the study.
3. *Application domain (if any)*; The area of research the study is placed into.
4. *Study sample size*; The number of participants who took part in the study.
5. *Participant characteristics*; Age range and/or mean age of participants
6. *Number of (accepted) hypotheses*; The number of hypotheses that were identified and accepted from the study. Hypotheses were scanned by searching for "hypo" in the article. If that was not found, or locations that matched with "hypo" did not present a statement presented as a hypothesis, the number of (accepted) hypotheses was set to 0. Articles, for example, that only formulate a research question that is later in the article treated as a hypothesis were thus not counted as hypotheses.
7. *Outcome measurement*; Any variable recorded during a study to assess the effects of a treatment or experimental intervention.
8. *Social robot type*; The type of social robots that were used in the study.

To further reduce the number of articles, we added a refined set of exclusion criteria.

2.2.2. Round 2 - Article Exclusion Criteria

New exclusion criteria were added during the second screening and the exclusion criteria of the first round of title and abstract screening were also applied whenever applicable.

1. *Participants do not interact with a physically present robot* (93); This means that all articles where the *participants* are interacting themselves, however minimal, with a social robot are included; using this criterion, we excluded articles that only asked participants to complete an online questionnaire or asked participants to watch videos of robots, for example.
2. *Participants do not interact with a social robot* (10); we excluded articles that reported on studies in which participants did not interact with a social robot. Social robots are physically embodied, life-like robots that interact in a human-like way [46].
3. *Participants are* 12 *years old or younger* (40); we excluded articles that reported on studies in which participants that were 12 years or younger took part, as it has been reported in the literature that they perceive robots differently from older people and do not have the capabilities to complete surveys, experimental tasks, and questionnaires accurately [25].
4. *Article does not report the results of a user study* (12); we excluded articles that only report on a protocol design for a user study or when it was not clear from the abstract that no user study was performed.
5. *Sample size is smaller than* 20 ***and*** *the article does not provide support that shows that a small sample size is sufficient* (33); we excluded articles with small sample sizes

using 20 as a lower bound as we could not compute this for articles that do not report on sample size estimation. Computing a sample size depends on the study design and most articles do not report how they estimated sample size themselves. Although a sample size of less than 20 can be sufficient for some studies, it more often than not is too small a sample size to reliably obtain a large enough effect size.

6. *Other* (5); Articles do not contain information of one or more of the 8 evaluation categories.

2.2.3. Round 2 - Article Inclusion Criteria

During the second screening of the 273 articles, the inclusion labels were redefined. The updated numbers of articles per inclusion criteria before and after applying the exclusion criteria are given below.

1. Robot-Human: Out of the 141 articles that report on a robot-human comparison, 105 were excluded, leaving 36 articles in the scope of this review.
2. Robot-Robot: Out of the 66 articles that report on a robot-robot comparison, 52 were excluded, leaving 14 in scope.
3. Robot-Avatar: Out of the 31 articles that report on a robot-avatar comparison, 13 were excluded, leaving 18 in scope.
4. Robot-Computer: Out of the 10 articles that report on a robot-computer comparison, 6 were excluded, leaving only 4 in scope.
5. Robot-Speaker: Out of the 16 articles that report on a robot-speaker comparison, 8 were excluded, leaving 8 in scope.
6. Robot-Tablet: Out of the 12 articles that report on a robot-tablet comparison, 6 were excluded, leaving 6 in scope.
7. Robot-Tele: Out of the 9 articles that report on a robot-tele-operated comparison, 2 were excluded, leaving 7 in scope.
8. Robot-Smartphone: Out of the 5 articles that report on a robot-smartphone comparison, 1 was excluded, leaving 4 in scope.
9. Robot-Animal: Out of the 5 articles that report on a robot-animal comparison, 5 were excluded, leaving only 0 in scope.

3. Overview of Articles and Review Findings

This section presents the findings of the systematic literature review which evaluates the comparative impact of social robots and other devices on human-robot interactions. The results are organized following the PRISMA guidelines to maintain clarity and transparency in the study selection process. The subsequent subsections provide a detailed overview of the search strategy, screening process, and final selection of studies, offering a comprehensive analysis of the collected evidence.

3.1. PRISMA Overview

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) is a standardized guideline that enhances the transparency and completeness of systematic reviews. It was first introduced by [68] to standardize reporting and to ensure clarity and replicability. The PRISMA guidelines provide a checklist to ensure that essential information is reported, along with a flow diagram to illustrate the stages of the review process [76]. By following the PRISMA guidelines, this study ensures methodological precision and transparency, making the selection process and results comprehensible and well-documented [77].

3.2. PRISMA

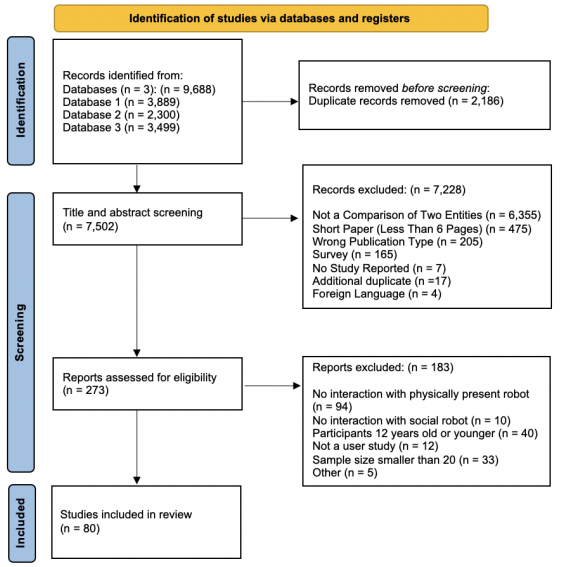

Figure 1: PRISMA

The PRISMA flow diagram in Figure 1 illustrates the systematic article selection process. The initial search across three academic databases yielded 9,688 articles. After removing 2,186 duplicates, 7,502 unique articles remained for screening.

During the title and abstract screening phase, articles were excluded due to a lack of comparative analysis (6,355 articles), short paper length (475 articles), wrong publication type (205 articles), survey-based methodologies (165 articles), absence of primary study data (7 articles), additional duplicates (17 articles), and non-English language (4 articles). This rigorous screening process left 273 articles for further evaluation.

In the eligibility assessment, 183 additional articles were excluded due to lack of interaction with physically present robots (94 articles), absence of interaction with social robots (10 articles), participants six years old or younger (30 articles), lack of a reported user study (12 articles), small sample size (33 articles), and other methodological limitations (5 articles).

Ultimately, 80 studies met the final inclusion criteria, forming the basis of the systematic literature review (SLR). These studies offer significant insights into the comparative analysis

of social robots and other devices within human-robot interactions. The subsequent sections provide detailed insights based on this curated dataset.

3.2.1. Descriptive Results

The articles assessed for eligibility (273 articles) were used for the analysis. Findings from all articles assessed for eligibility (273 articles) were compared to the final included articles (80 articles).

3.2.2. Number of Articles per Publication Year

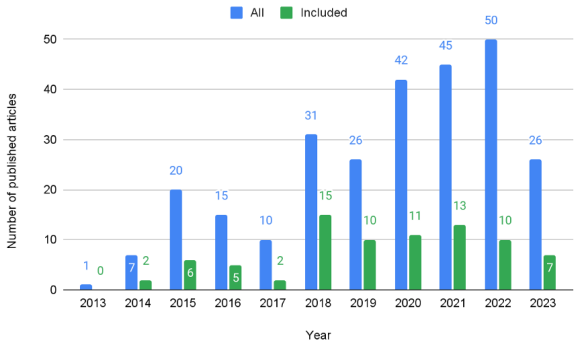

Figure 2: Distribution of the number of articles published between 2013 and 2023

Figure 2 depicts the number of published articles per year from 2013 to 2023, categorized into all reviewed articles (blue bars) and those included (green bars). This analysis provides a clear view of the trends and growth in research on social robotics over the past decade.

The data reveals a significant upward trend in the number of published articles, reflecting the growing interest and research efforts in the field of social robotics. This trend underscores the increasing recognition of the importance and potential of these technologies in various domains.

The inclusion rate of articles varied each year, with a noticeable peak in 2018 where almost half of the published articles were included (15 out of 31). In contrast, 2022 saw the highest number of publications but a lower inclusion rate (10 out of 50), suggesting more rigorous selection criteria or varied relevance of studies to the Systematic Literature Review (SLR).

The analysis of publication trends (Figure 2) demonstrates a clear and increasing interest in social robots and smart devices. The significant rise in publications over the years highlights the expanding scope and depth of research in this area. The variations in inclusion rates reflect the evolving focus and quality of research, indicating a maturing field with increasing scholarly contributions.

This trend analysis is critical for understanding the development and current state of research in social robotics. It underscores the importance of continued investment in research and development further to explore the potential and applications of these technologies. The growing body of literature provides a robust foundation for future studies and innovations in social robotics, paving the way for advancements that can significantly impact various sectors such as healthcare, education, and customer service.

3.2.3. Participant Country Distribution

Figure 3 and Figure 4 illustrate the global distribution of the countries of origin of the study participants involved in the reviewed articles. Figure 3 represents the participant countries for all 273 reviewed articles, while Figure 4 shows the distribution for the 80 included articles. The color intensity on the maps reflects the number of participants from each country, providing a visual representation of the geographical diversity of the study populations. These maps highlight the global nature of research on social robots and human-robot interactions, offering insights into the regional focus of participant recruitment in this field.

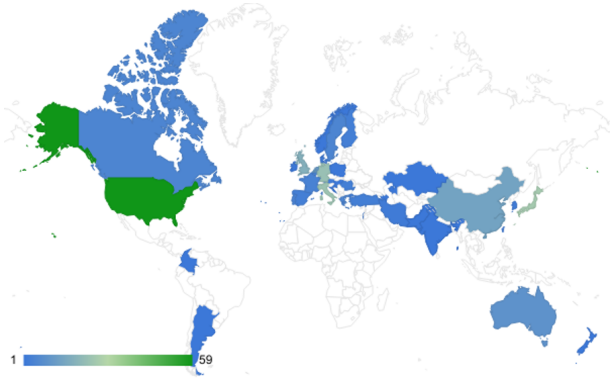

Figure 3: World Map of Participant Country of All Articles (273 Articles)

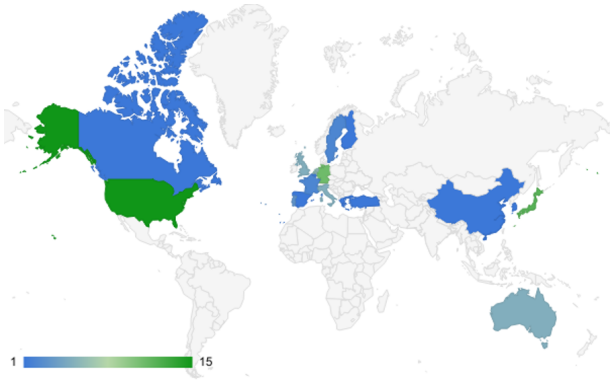

Figure 4: World Map of Participant Country of Included Articles (80 Articles)

The world maps reveal that research covering social robots to humans and other devices is globally distributed, with significant concentrations in certain regions.

Figure 3 shows that the highest number of participants (up to 59) originated from the United States, followed by significant representation from European countries, China, Japan, and South Korea. This broad participant distribution reflects the widespread global interest in the field, particularly in technologically advanced regions.

In Figure 4, which focuses on the 80 included articles, the distribution patterns remain similar but more specific. The

United States continues to have the highest number of participants (up to 15), underscoring its prominent role in studies involving social robotics. European countries such as the United Kingdom, Germany, and the Netherlands also show high participant representation, indicating strong involvement in research from these regions. Additionally, significant participant representation is observed from China, Japan, and South Korea, reflecting their active engagement in studies comparing social robots, humans, and other devices. The maps also show participants from other regions, such as Australia, Canada, and a few countries in South America and Africa, though with lower intensity. This indicates growing but less concentrated research participation in social robotics from these regions.

The maps also highlight the presence of research in other regions such as Australia, Canada, and a few countries in South America and Africa, albeit with lower intensity. This indicates a growing but less concentrated interest in social robotics in these regions.

4. Comparative Studies of Social Robots: Article and Appearance Classification

This section introduces the results addressing the first research question, **RQ1** *"Which types of social robots and devices have been used in comparative studies?"*. Figures 5 and 6 provide a comprehensive overview of the comparative contexts and physical designs of social robots examined in the literature. Together, these figures offer an understanding of the research landscape, underscoring the diversity and scope of comparative studies in the field of social robotics.

4.1. Study Focus Classification

Figure 5 presents the number of published articles categorized by their study focus. The categories include comparisons between robots and humans, robots and other robots, and various other comparative contexts such as avatars, animals, speakers, tablets, computers, and more. The figure distinguishes between all reviewed articles (blue bars) and those included in the SLR (green bars).

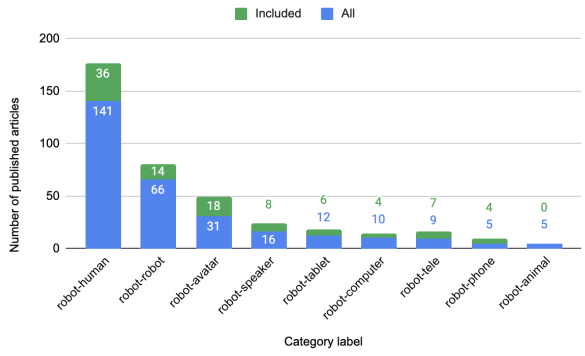

Figure 5: Number of Articles per Entity Category

The bar chart in Figure 5 shows that the entity category with the highest number of articles was 'robot-human', with a total of 141 articles reviewed and 36 articles included. This indicates a significant focus on understanding how robots interact with humans, which is crucial for applications in social settings.

4.1.1. Robot-Human Entity Category

The application domains under the 'robot-human' category were Education, Healthcare, Service, Human Resources, and Gaming. The Education domain accounted for the highest reviewed articles but not for the included articles. The second most dominant domain amongst reviewed articles was Healthcare which entailed the highest number of articles amongst the included articles. Amongst the 'robot-human' category, the humanoid robot NAO was used the most. NAO was compared to humans in 41 reviewed articles, of which 11 were included in the final scope.

The humans used for comparison varied depending on the application domain. For example, in the education sector, social robots such as Pepper and NAO, acting as teachers, were compared to human teachers in classroom settings. Various types of educational robots were evaluated by their interactions with children, where these robots actively taught the kids. In control studies, children's learning experiences were assessed using traditional methods with a human teacher for comparison.

In the healthcare sector, elderly care emerged as a significant area where therapeutic robots, such as the animal-inspired robot Paro or Pleo, were used. Specifically, research articles focusing on dementia analyzed the use of these therapeutic robots among patients. Traditional care methods involved direct interactions solely with a human caretaker. However, experimental approaches in elderly care incorporate the use of these robots, either by the patients independently or under the supervision of a caretaker.

4.1.2. Robot-Robot Entity Category

The 'robot-robot' comparison was the second most common category, with 66 articles reviewed and 14 included. This highlights the interest in evaluating interactions and performances between different robotic systems. The application domains included Education, Healthcare, Home, Human Resources, and Service. The Education and Healthcare domain scored highest.

Out of the 66 articles, 12 did not specify the brand or name of the social robot. The remaining 54 reviewed articles compared social robots to robots that differed in the degree of anthropomorphism. Out of the 11 included articles, 9 articles used either the humanoid robot NAO, Pepper, or both to compare to other humanoid robots, androids, geminoids, animal-inspired robots, mechanical robots, or functional robots. One included article did a comparison of 8 different animal-inspired robots in elderly care homes.

4.1.3. Robot-Avatar Entity Category

The 'robot-avatar' entity category had a notable presence, with 28 articles reviewed and 18 included. This indicates a growing interest in virtual embodiment and their interactions with physical robots in the HCI research field. Articles under the 'robot-avatar' category that used the physical NAO robot compared it to a virtual NAO avatar. The majority of articles

compared physical robots to their virtual avatar version. A total of 6 articles used various virtual avatars that varied in degrees of anthropomorphism, from simple geometric shapes to more human-like figures.

4.1.4. Robot-Speaker Entity Category

Comparisons between robots and speakers, tablets, and computers were represented to a lesser extent. The 'robot-speaker' category had 16 reviewed articles, with 8 included. The smart speaker that was compared the most was the Amazon Echo speaker (6 articles) followed by the Google Home speaker (4 articles).

4.1.5. Robot-Tablet Entity Category

The 'robot-tablet' category had 12 reviewed articles, with 6 included. The term 'tablet' was used in all reviewed articles, with the visual representation of Android tablets. However, no articles listed the brand name of the tablet.

4.1.6. Robot-Computer Entity Category

The 'robot-computer' category had 9 articles reviewed with 4 included. All articles used a standard desktop computer. This category indicates a moderate interest in multi-modal interaction studies.

4.1.7. Other Entity Categories

Other categories such as 'robot-tele', 'robot-smartphone' and 'robot-animal' had fewer articles. The 'robot-tele' category included 9 reviewed articles with 7 included.

Entity category 'robot-smartphone' had 5 reviewed, with one included. Out of the 5 articles, 4 compared humanoid robots to smartphone applications.

The 'robot-animal' category had 5 articles reviewed but with 0 included. 2 articles compared real-life therapy dogs to the MiRo-E therapeutic robot [57].

4.2. Entity Category Summary

The distribution of articles across categories in Figure 5 highlight several key trends. The predominant focus on 'robot-human' comparisons indicates a central interest in understanding how robots can effectively interact with humans. This focus is crucial for applications in areas such as healthcare, education, and customer service, where robots need to engage meaningfully with human users [35].

The significant number of 'robot-robot' and 'robot-avatar' comparisons reflects ongoing efforts to refine robotic technologies and explore hybrid interactions involving virtual agents. These studies are vital for advancing multi-robot systems and understanding the potential of integrating physical and virtual interaction modalities [5].

The moderate representation of 'robot-speaker', 'robot-tablet', and 'robot-computer' comparisons suggests an interest in exploring multi-modal interaction contexts, which are essential for developing versatile and adaptable robotic systems.

The lower representation and inclusion of studies in other categories, such as tele-operated robots, smartphones, and animal comparisons, indicate niche areas that, while less explored, still contribute valuable insights into specific aspects of human-robot interaction.

4.3. Robot appearance classification

Figure 6 presents a detailed classification of robot appearances. Baraka's [5] adapted classification scheme was used to provide a comprehensive understanding of the diverse physical designs of social robots and their impact on human-robot interactions [5]. The data underpinning this figure are detailed in Appendix A.

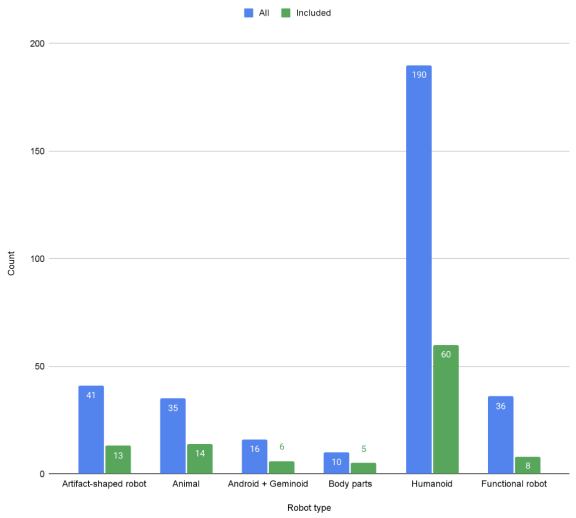

Figure 6: Robot appearance classification of identified robot types

The classification scheme organizes robots into three primary categories: Bio-inspired, Artifact-shaped, and Functional robots, each with specific subcategories. Figure 6 shows the distribution of robot types, with the number of robots classified in each category.

4.3.1. Bio-inspired robots: Human-inspired

Bio-inspired robots are the most prominent category, particularly human-inspired robots, which include androids, body parts, geminoids, and humanoids. Notably, 190 robots were classified as humanoid, with 60 included in the meta-analysis. This significant number reflects the high research interest in humanoid robots, as they closely resemble humans in structure, skin, and facial features. According to Appendix A, the most frequently used robots in this category are NAO and Pepper, indicating their popularity and versatility in research applications. Specifically, 80 NAO robots were used across all reviewed articles, with 29 included in the meta-analysis. Similarly, 37 Pepper robots were utilized in all articles, with 13 of these included in the review.

4.3.2. Bio-inspired robots: Animal-inspired

Animal-inspired robots within the bio-inspired category are also well-represented. 20 reviewed articles included 'Imaginary familiar animal-inspired' robots of which 8 were included. The 'Real familiar animal-inspired' robots were identified in

15 articles of which 7 were included (Appendix A). This category highlights the diversity of creative approaches in robot design aimed at eliciting emotional responses from users. The animal-inspired robot Paro, a therapeutic robot baby harp seal, was used the most (total 5 articles, 2 included), followed by the Imaginary: Unfamiliar robot, MiRo-E [57]. These types of robots are noted for their ability to engage users through familiar animal-like behaviors [5].

4.3.3. Artifact-shaped robots

Artifact-shaped robots derive their appearance from human-made objects and inventions. In Figure 6, 41 artifact-shaped robots are classified, with 13 included. These robots can resemble everyday items or complex apparatuses, such as the Amazon Echo and Google Home smart speakers. The practical utility and user acceptance of these designs are detailed in Appendix A, demonstrating their ability to blend into familiar environments effectively [5].

4.3.4. Functional robots

Functional robots, whose appearance is dictated by their functional components and technological necessities, constitute 36 robots in Figure 6, with 8 included. These robots typically have a more mechanical look, emphasizing practicality over anthropomorphic design. The functional robots that accounted for the highest count are 1) Industrial robot Baxter (total 4, 1 included), 2) Mobile manipulator robot TIAGo (total 3, 1 included), and 3) Lego Mindstorms (total 3, 2 included). Appendix A highlights the functional advantages of these designs, particularly in tasks requiring high precision and efficiency [5].

4.4. Robot appearance classification Summary

This systematic literature review examined the types of social robots and devices used in comparative studies, categorizing them into three main groups: Bio-inspired, Artifact-shaped, and Functional robots.

Bio-inspired robots were the most common, particularly human-inspired models like humanoids. Notable examples include NAO and Pepper, which were frequently used in the reviewed studies. A total of 190 humanoid robots were identified, underscoring a strong research interest in robots that closely resemble humans.

Animal-inspired robots, such as the therapeutic seal Paro and the robot MiRo-E, were also significant. These robots are designed to evoke emotional responses through familiar animal-like behaviors, demonstrating a creative approach to robot design.

Artifact-shaped robots, resembling everyday objects like Amazon Echo and Google Home, were noted for their practical integration into daily life, blending seamlessly into familiar environments. Functional robots, such as Baxter focuses on efficiency and precision, prioritizing functional design over appearance.

In summary, the review highlighted a diverse range of robot designs used in research, emphasizing the role of appearance in shaping human-robot interactions. The prevalence of humanoid and animal-inspired robots reflects a focus on creating engaging, relatable experiences, while artifact-shaped and functional robots underscore practical applications in daily and industrial settings.

5. Primary measurement scales

This section addresses the second research question, **RQ2**: *"How have social robots been compared with humans and other devices?"*. To answer this question, we categorized the measurement and data collection tools used in the 80 included articles. Table 3 summarizes these categories and their respective counts, offering a clear overview of the primary measures used in human-robot interaction research.

The table includes self-report measures, further divided into robot-specific, psychological-specific, and domain-specific categories, as well as measures of brain activity, behavioral responses, test and task performance, interviews, coding, and sensor data. This classification highlights the diverse methodologies employed in the field and provides insight into the prevalent tools used to compare social robots, humans, and other devices.

Category Measure	Measure Count	Use Count
Self-report: Robot-specific	26	50
Self-report: Psychological	55	71
Self-report: Domain-specific	21	24
Self-report: Technology-specific	20	27
Brain Activity	2	3
Behavioral	-	100
Test + Task	8	10
Interview	3	6
Video	1	12
Sensors	15	22

Table 3: Primary Measurement Scales in the 80 Included Articles

5.1. Description of Category Measures

Table 3 provides a comprehensive overview of the measurement tools used in the 80 articles included in the SLR. Each category in the table encapsulates a distinct dimension of measurement, reflecting the diverse methodologies employed in human-robot interaction (HRI) research.

5.1.1. Self-report: Robot-specific

The category *Self-report: Robot-specific* includes measures specifically designed to assess interactions and attitudes towards robots. This category included 26 different measures that were used 50 times, assessing interactions and attitudes specifically towards robots. This category represents a substantial portion of the measures used. The tools in this category were tailored to capture the unique aspects of human-robot interaction, such as trust, usability, and social attributes. Notable examples include the Negative Attitude toward Robots Scale (NARS) by [83] (7 counts), Multi-Dimensional Measure of Trust (MDMT) by [71] (3 counts) and, the Godspeed questionnaire by [7] (11 counts). A detailed list of these measures, along with their sources and counts, is provided in Appendix B.

5.1.2. Self-report: Psychological-specific

The *Self-report: Psychological-specific* measures focus on evaluating general psychological constructs that influence or are influenced by interactions with robots. This category had 55 measures used 71 times, focusing on general psychological constructs like motivation, anxiety, and personality traits, and was used most frequently. Prominent examples included the Positive And Negative Affect Scale (PANAS) by [115] (4 counts) the Intrinsic Motivation Inventory (IMI) by [27] (4 counts), and the Big 5 Inventory (BFI) by [51] (4 counts). The comprehensive list of psychological-specific measures can be found in Appendix C.

5.1.3. Self-report: Domain-specific

The *Self-report: Domain-specific* measures evaluate domain-specific aspects. Comprising 21 measures used 24 times, these tools assessed specific domains like user experience and physical health status. Examples included the NASA Task Load Index by [43] (6 counts), the User Experience Questionnaire (UEQ) by [64], and the Foreign Language Classroom Anxiety Scale (FLCAS) by [49]. A total of 20 measures accounted for one count, and only one was adapted. For a detailed outline of these domain-specific measures, please refer to Appendix D.

5.1.4. Self-report: Technology-specific

The *Self-report: Technology-specific* measures focus on the usability and acceptance of technology. This category included 20 measures, used 27 times. The System Usability Scale (SUS) by [17] with a count of 8 and adapted count of 2 is the most frequently *Self-report: Technology-specific* measure used.

5.1.5. Brain Activity

Moving beyond self-report measures, the *Brain Activity* category (3 use counts) included measures that involve monitoring brain activity to understand cognitive and emotional responses during interactions with robots. This category comprises functional near-infrared spectroscopy (fNIRS)(2 counts) and electroencephalography (EEG) (1 count). These measures provide insights into the neural correlates of human-robot interaction, showing the brain's response to robotic stimuli.

5.1.6. Behavioral

The *Behavioral* category, with 100 measures, stands out as the most frequently used category. Behavioral measures are employed to observe and quantify the actions and reactions of participants when interacting with robots. This high count reflects the emphasis on observable behavior in HCI research, allowing researchers to capture a range of interactions and responses. Examples include eye gaze, task efficiency, and facial expressions for positive and negative affect.

5.1.7. Test + Task

The *Test + Task* category includes 8 types of measures. Standardized tests and specific tasks are designed to measure cognitive or physical abilities in the context of human-robot interaction. These measures provide structured and objective assessments of participants' capabilities and performance during robot interactions. Notable examples include the Peabody Picture Vocabulary Test (PPVT-4) by [30] with 2 counts, and the variant of the Speech Perception In Noise (SPIN) test by [32] with a count of 3. These tasks are critical for evaluating specific competencies and the effectiveness of robot-assisted interventions.

5.1.8. Interview

The *Interview* category, with 3 different types of measures, involves qualitative data gathered through structured or semi-structured interviews. These interviews aim to gain deeper insights into participants' experiences and perceptions of robots, offering rich, descriptive data that complements quantitative findings. Examples include interviews (2 counts), semi-structured interviews (3 counts), and focus group interviews (1 count). 4 out of 6 interviews followed guidelines and applied coding schemes. For the focus group interview, a thematic analysis was used. One interview applied the coding scheme method. For one semi-structured interview, the data was mapped onto mental/emotional and a social other scale. This qualitative approach provides a nuanced understanding of user attitudes and experiences. In summary, out of the 6 articles that applied interviews,4 followed a guide, and 5 applied coding schemes.

5.1.9. Video

The *Video* category is used 12 times amongst the 80 included articles. The video measurement scale followed a coding scheme amongst 10 articles. Out of the 12 articles, 10 used video measures that followed a coding scheme. One article followed standardized criteria, 7 articles applied coding schemes, and 2 used coding rules. These coding tools involve the systematic coding of behaviors or verbalization to quantify specific aspects of human-robot interaction. This method allows for detailed analysis of interaction patterns and communication behaviors, contributing to a nuanced understanding of the interaction dynamics.

Examples include coding schemes (8 counts), and thematic analysis (2 counts). These coding practices are essential for rigorous and reproducible behavioral analysis.

5.1.10. Sensors

Lastly, the *Sensors* measures used various sensors to objectively measure physiological responses such as heart rate, skin conductance, or motion tracking during interactions with robots. This category included 15 different measures used 22 times. These measures provide real-time, objective data on the physiological states of participants, offering valuable insights into the physical and emotional impact of human-robot interactions. Examples include motion tracking systems (3 counts) and audio sensors (4 counts). These sensor-based measures are crucial for capturing detailed and continuous data during interactions.

5.2. Summary of Category measures

In addressing **RQ2**: "How have social robots been compared with humans and other devices?", this systematic literature re-

view identified and categorized the primary measurement scales used across 80 studies. These measures provide a comprehensive overview of the methodologies employed in human-robot interaction (HRI) research, highlighting the tools used to evaluate and compare social robots with humans and other devices.

The measures were grouped into several categories. Self-report measures were the most diverse, including robot-specific, psychological-specific, domain-specific, and technology-specific tools. These measures assessed various dimensions such as user attitudes, emotional responses, usability, and specific experiences with robots. Behavioral measures emerged as the most frequently used category, offering objective data on user actions and interactions, such as eye gaze and task performance, to evaluate how users engage with robots compared to humans and other devices.

Additional categories included brain activity measures, such as EEG and fNIRS, which explored the cognitive and emotional responses elicited by robot interactions. Test and task performance measures assessed specific cognitive and physical abilities, providing insights into the practical capabilities and impacts of robot-assisted interventions. Interviews and video analysis were utilized to gather qualitative data, offering deep insights into user experiences and perceptions. Lastly, sensor data, including physiological metrics like heart rate and skin conductance, provided real-time, objective measures of users' physical and emotional states during interactions with robots.

Overall, the review underscored the use of a diverse range of methodologies in HRI research, combining subjective self-reports with objective behavioral, physiological, and cognitive data. This multi-faceted approach allows for a comprehensive comparison of social robots with humans and other devices, offering a nuanced understanding of human-robot interactions.

Each of the 10 categories provides a different lens through which researchers can assess the complexities of human-robot interaction, contributing to a holistic understanding of this multifaceted field. For detailed information on the self-report measures, please refer to the appendices as specified.

6. How do robots compare to humans and other devices?

To address the final research question, **RQ3** "How do social robots compare to humans and other devices?", the findings and the robots used in each article are listed in entity tables (as shown below).

The entity category 'robot-animal' dropped out from the main scope of the review as no paper compared social robots with animals in the refined scope (see Fig. 5). This comparison category was initially small, but no paper was found that compared animals to physically present social robots. Thus, the three biggest categories of comparison include *humans* (36 papers), other robots (14 papers), and, avatars (18 papers).

6.1. How do social robots compare to humans?

A total of 36 articles compared robots to humans. Only two of those papers compared two [40, 105] and one compared three different robots [52] to a human.

Social robot	Total	Robot	Human	Neither
NAO	11	[111]	[55], [50], [110], [60], [73], [40], [15]	[36], [118], [6]
Pepper	6	-	[105], [104], [88], [94], [12]	[13]
Nadine	2	[74]	[4]	-
Telenoid	2	-	-	[58], [59]
AIDA	1	-	-	[117]
Arduino UNO	1	-	-	[102]
Baxter	1	-	[40]	-
Cozmo	1	-	-	[23]
Elenoide	1	-	[105]	-
EMYS	1	-	[86]	-
EVA	1	-	[65]	-
FACE	1	-	[24]	-
Glin+	1	-	[86]	-
HRP-2Kai	1	[109]	-	-
iCub	1	-	[90]	-
Jibo	1	-	[91]	-
mObi	1	-	-	[48]
Nomad Scout	1	-	[52]	-
Poppy	1	-	[11]	-
Robovie	1	-	[53]	-
Ryan	1	-	[78]	-
Vizzy	1	-	[19]	-
Total	*39*	*3*	*26*	*10*

Table 4: Social robots compared to humans

Table 4 presents a categorization of findings into whether the social robot, the human, or neither was preferred in robot-human comparison studies. The robots most frequently involved in comparisons include NAO, Pepper, Nadine, and Telenoid. NAO was the most frequently studied, with 11 comparisons, followed by Pepper with 6, and Nadine with 2. Overall, humans were preferred over social robots in the majority of studies, with 26 out of 39 comparisons favoring human interaction. Social robots were found to perform better than humans in 3 out of the 36 articles, while in 10 articles, neither the robot nor the human was distinctly preferred.

The social robot NAO was the most frequently studied, with 11 comparisons made against humans. Out of the 11 studies, 7 showed higher ratings for humans, particularly in terms of expressiveness and interaction quality. This may suggest that despite NAO's popularity and advancements, it still struggles to match human performance in these areas [55, 50, 110, 60, 73, 40, 15]. However, for one article within the education domain, NAO was found to be more effective than humans, especially in tasks where consistency and lack of bias were crucial [111].

In the 6 studies comparing Pepper to humans, various measures were used, including the Almere model, a 3-item customer satisfaction scale, the Conversational Skills Rating Scale (CSRS), and the 18-question TOPICS-SF [12, 104, 105, 13]. Across these measures, humans consistently outperformed Pepper, particularly in areas of social presence, conversational skills, and overall user satisfaction.

For the two articles that compared humans to telenoids, both used *Brain Activity measure* fNIRS [59, 58] where the results were non-significant.

The remaining articles used a variety of robots that differed in degree of anthropomorphism and compared them to humans. The android Nadine was compared in two different studies that yielded different results. In article [74], elderly people engaged with Nadine and scored highest in terms of actions and states in the robot condition compared to the human condition. For the

other study, where Nadine was used as a job interviewer and engaged with humans, the human condition was preferred.

Based on Table 4, it can be seen that overall, humans outperformed social robots, especially niche robots that are still in the development phase and had a total count of 1 in the table.

Measures used. Amongst the 11 articles that compared NAO to humans, 5 used *Self-report: Psychological* measures such as PANAS, Intrinsic Motivation Inventory (IMI) and the Big 5 Inventory (BFI) [50, 40, 118]. The results either showed a preference for the human over the robot condition or neither. Interestingly, the paper where NAO outperformed the human relied on *Domain measures* such as the STIMEY questionnaire (STQ) [111].

The two other robots that were preferred over humans were HRP-2Kai and android Nadine. Motion detection was used during tasks with HRP-2Kai where data on movement frequency, press accuracy, and form was recorded [109]. HRP-2Kai also scored higher on *Self-report measures* that encompass perception and robot exposure.

Video coding methods were used for interactions between elderly people and the android Nadine. Various states and actions (happiness, movement, activity) of participants were recorded [74]. Results showed that participants scored higher in the robot condition compared to the human condition.

6.2. How do social robots compare to other robots?

Social Robot	Compared with	Main findings	Ref
Pepper	Elenoide	Elenoide succeeded in inducing a significantly higher expressiveness in employees than Pepper	[105]
Pepper	Custom-made AGV	Overall perceived comfort was slightly higher for custom-made AGV compared to Pepper	[80]
Pepper	Smart wheelchair	Pedestrians behave more conservatively around the wheelchair compared to Pepper	[121]
Pepper	Sawyer	Perceived anthropomorphism was highest for Pepper and least for Sawyer	[79]
Pepper	NAO	Not significant	[66]
Pepper	NAO	Not significant	[56]
Pepper	Fake dog	Pepper scored highest on Perceived Safety compared to fake dog	[56]
NAO	Fake dog	NAO and Pepper scored highest on GODSPEED analysis compared to fake dog	[56]
NAO	Lego Mindstorm	NAO was more effective in creating intrinsic motivation across genders	[37]
NAO	Lego Mindstorm	NAO was preferred over Lego Mindstorm for computational thinking	[8]
NAO	Paro	NAO but not Paro was able to improve irritability and neuropsychiatric symptoms	[108]
Giraff	Magabot	Magabot was perceived as safer while Giraff was not perceived as more intelligent	[52]
Misty	QTrobot	Misty was perceived more positively and participants felt more connected with it	[103]
JfA cat	JfA dog;Paro	JfA cat was most preferred followed by JfA dog and Paro	[14]
Geminoid-F	Robi;MyKeepon	Geminoid-F was perceived as more anthropomorphic and intelligent than the other two, but less likable and safe	[41]
CR700	TIAGo	CR700 was perceived as the most assertive and least polite strategy	[3]
Lio	TIAGo	Lio was rated more humanlike than TIAGo	[3]
Total	*29*		*14*

Table 5: Social robots compared to other robots (*yellow* means outperforming)

A total of 14 articles compared social robots to other types of robots. Within some studies, multiple social robots were compared to different robots, accumulating to a total robot count of 29 (see Table 5). Table 5 categorizes these comparisons by listing the social robot, the robot it was compared with, the main findings of each comparison, and highlights the cases where the social robot outperformed the other robot (indicated by yellow highlights).

The findings reveal that in 10 out of the 29 comparisons, the social robots were rated more favorably than their counterparts. The social robots that were favored the most were humanoid NAO (4 counts) and Pepper (2 counts).

Interestingly, from all robot types, the android robot Elenoide outperformed the humanoid robot Pepper. In the study, the results showed that android robots evoke a higher level of expressiveness in employees than humanoid robots [105] which was measured using the Conversational Skills Rating Scale (CSRS).

The humanoid NAO and Pepper scored highest compared to functional robots (Custom-made AGV, Sawyer, Lego Mindstorm). As humanoid robots are more embodied than functional robots, perceived anthropomorphism scores were highest for the humanoid condition [79]. Other *Self-report: Robot specific measures* such as the GODSPEED analysis yielded higher ratings in the humanoid condition compared to less embodied robots [79].

Some articles compared humanoid to other humanoid robots. When Pepper was compared to NAO, no significant differences were found (2 counts). However, when participants interacted with the humanoid Misty and humanoid QTrobot, Misty was preferred. Misty was perceived more positively than the QTrobot and these results could be explained by the form function attribution bias as Misty is more toy-like, smaller, and less humanoid than QTrobot [103].

A total of 2 articles compared Paro, the animal-inspired robot which was outperformed by JfA Cat and NAO [14, 108].

One article compared the geminoid Geminoid-F to humanoid Robi and functional MyKeepon. The geminoid robot with the highest degree of anthropomorphism was perceived as more anthropomorphic and intelligent than the other two types. However, likeability and safety scores were lower for Geminoid-F compared to Robi and Mykeepon.

Measures used. The studies comparing social robots to other types of robots employed a wide variety of measures, categorized primarily into *robot-specific measures, psychological and behavioral scales*. These measures were critical in evaluating various attributes such as anthropomorphism, social presence, trust, and task performance.

The *Robot-specific measure* GODSPEED was used in 5 robot-robot comparison articles. It was used to assess how human-like the robots were perceived to be and was employed in comparisons between Pepper, NAO, and other robots like the Fake Dog and Sawyer. Pepper and NAO scored highest compared to Sawyer and Fake Dog [79, 56]. When compared to TIAGo, the zoomorphic robot Lio was rated more humanlike [3]. Furthermore, TIAGo was rated as the most uncanny robot. Potentially, the uncanniness of TIAGo overshadowed its human likeness, while for example, the human 'eyes' of Lio might have made it look more humanlike than TIAGo [3].

Only one article [103] applied the Robotic Social Attributes Scale questionnaire (RoSAS) to compare two types of humanoid robots that varied in level of anthropomorphism. This scale was used to evaluate social attributes of robots such as warmth, competence, and discomfort.

As seen in Table 5, 2 articles compared Paro, the animal-inspired robot. Paro was outperformed by NAO when engaging with elderly people in nursing homes and was evaluated by the Global Deterioration Scale (GDS), the Severe Mini Mental State Examination (sMMSE), the Mini Mental State Examination (MMSE), the Neuropsychiatric Inventory (NPI), the Apathy Scale for Institutionalized Patients with Dementia Nursing Home version (APADEM-NH), the Apathy Inventory (AI) and the Quality of Life Scale (QUALID) [108].

The other article [14] that compared Paro did a comparison study between multiple animal-inspired robots. A qualitative user-centered design study was applied and transcripts of group recordings were analyzed [14]. The results showed that the most preferred device was the Joy for All cat, followed by the Joy for All dog.

6.3. How do robots compare to avatars?

A total of 18 papers compared social robots with virtual avatars. These avatars were either displayed on a computer or were integrated with virtual reality (VR).

Table 6 categorizes these comparisons by listing the social robot, the avatar it was compared with, and the main findings, with yellow highlights indicating where the social robot outperformed the avatar.

The findings revealed that out of the 18 articles, 14 preferred physical present robots over digital avatars. NAO was the most frequently compared robot, with 7 comparisons to avatars. Of these, 6 studies directly compared the physical NAO to its digital NAO avatar [1, 114, 99, 20, 69, 96].

The results show that physical robots, particularly NAO, were often preferred over their digital counterparts. For example, participants felt more attachment to the physical NAO and rated its perceived quality higher compared to the NAO avatar [1, 114]. Additionally, physical EMAR outperformed its VR and avatar versions in reducing stress and was rated higher on embodiment and interaction quality [9, 10].

However, there were also cases where avatars performed similarly or better. For instance, in the comparison between the Nicole avatar and the robot Nadine [4], participants showed more careful speech patterns with Nadine, reflected in higher Harmonics-to-Noise Ratio (HNR) values, which suggests that the avatar's characteristics influenced speech interaction [4]. Similarly, VR avatars like Pepper VR induced more discomfort and a lower sense of presence compared to the physical Pepper, highlighting that VR does not always enhance user experience [67].

Some comparisons, such as those involving Temi and its avatars, yielded non-significant results, indicating that both the physical and virtual versions were perceived similarly in certain contexts [89].

These findings underscore the strengths of physically present robots in creating attachment, reducing stress, and maintaining perceived quality, while also recognizing that avatars, particularly in virtual environments, can sometimes match or even outperform robots in specific scenarios.

Social Robot	Avatar	Main findings	Ref
NAO	NAO avatar	No difference between physical NAO and virtual NAO on trust, intelligence, and ethics. Perceived quality was higher for physical NAO over virtual NAO	[1]
NAO	Virtual human	NAO partly outperformed the virtual agents in assisting humans in real-world home-settings	[75]
NAO	NAO avatar	Physical NAO was preferred over virtual NAO and also elicited more attachment	[114]
NAO	NAO avatar	People emphasize more with physical NAO and fail to emphasize with simulations	[99]
NAO	NAO face avatar	Where mistreatment occurred, the witnesses sympathized with physical NAO but not the avatar	[20]
NAO	NAO avatar	Not significant	[69]
NAO	NAO avatar	Virtual NAO and real agent were evaluated equally, but users preferred a real or virtual embodied agent over a non-embodied system	[96]
Emys;Glen	Disembodied agents	Higher levels of group identification in physical robot condition compared to virtual. The embodied selfish agent was rated more competent than the disembodied selfish agent	[26]
EMAR	EMAR avatar	Physical EMAR scored highest on Robot Embodiment Attribute Ratings compared to virtual avatar	[10]
EMAR	EMAR VR	Highest preference for VR condition and VR interactions resulted in similar therapeutic disclosure when compared to a face-to-face	[10]
EMAR	EMAR avatar	All robot interactions were stress-reducing but the physical EMAR reduced stress most significantly for the group regardless of interaction order	[9]
EMAR	EMAR VR	Highest reduction in momentary stress for physical EMAR, followed by VR condition	[9]
Reeti	Karen avatar	Not significant	[101]
CommU	CommU avatar	Elderly engaged in the conversation with physical CommU more than virtual CommU	[82]
Temi	Temi avatar	Not significant	[89]
Temi	Temi VR	Not significant	[89]
Nadine	Nicole avatar	Participants spoke slower and softer to Nadine and had the highest value of Harmonics-to-Noise Ratio (HNR)	[4]
Ryan	Avatar	Highest emotion recognition rates for physical Ryan compared to virtual agent	[78]
EVA	EVA avatar	Strong confusion between the emotions disgust and anger in both static and dynamic expressions of the robot and virtual avatar	[65]
ASIMO	ASIMO VR	Expectations for the VR robots' ability to perform instrumental functions that are useful for humans were lower compared to physical robot	[54]
Pepper	Pepper VR	Participants felt more discomfort and lower sense of presence in VR compared to real life. Visual familiarity in VR did not affect proxemic preferences	[67]
Total	41		18

Table 6: Social robots compared to avatars (*yellow* means outperforming)

Measures used. The most widely used measure categories under *robot-avatar* were *Psychological-specific* [54, 96, 69, 4, 99, 9, 82, 89], followed by *Robot-specific* measures [1, 114, 67, 10, 26].

Psychological-specific measures were instrumental in capturing the nuanced emotional and cognitive responses of participants. The Empathy Questionnaire [99] provided critical in-

sights into the empathetic connections formed between users and the robot or avatar, while the Perceived Stress Scale (PSS-10) [9] quantified the stress experienced during these interactions. To explore the intrinsic motivations driving user engagement, the Intrinsic Motivation Inventory (IMI) [89] was employed, helping understand the underlying factors influencing participants' interactions. Additionally, the widely recognized Godspeed Questionnaire [1] was used to assess key attributes such as anthropomorphism, likability, perceived intelligence, and safety—factors pivotal in evaluating the social acceptability of both robots and avatars.

In contrast, *Robot-specific* measures provided a more targeted perspective on the unique characteristics and perceived efficacy of robots compared to avatars. The Negative Attitudes toward Robots Scale (NARS) [114, 10] was particularly significant in measuring pre-existing biases and attitudes towards robots, which could influence the outcomes of human-robot interactions. The Inclusion of Other in the Self scale (IOS) [82], a measure traditionally used in social psychology, was adapted to assess the degree of social closeness and identification participants felt towards robots or avatars, offering valuable insights into the emotional bond formation. The Robot Social Attributes Scale (RoSAS) [67, 26], was another crucial measure, capturing dimensions such as warmth, competence, and discomfort, which are essential in understanding the broader social implications of robot integration in human environments.

6.4. How do social robots compare to other devices?

A total of 19 papers compared social robots with other devices. Eights papers compared social robots with *smart speakers* (10 comparisons), 6 with *tablets* (7 comparisons), 4 with *computers/laptops*, and 3 with *smartphones* (Figure 5).

Table 7 shows that, except for the Bono bot, humanoid robots were rated more favorably than smart speakers. These robots were perceived as more likable, intelligent, and trustworthy, and, unsurprisingly, they scored higher in anthropomorphism. Similar results were observed when comparing humanoid robots with smartphones [117, 93]. Brain imaging studies indicated that the humanoid Telenoid yielded better predictions for perceived story difficulty compared to a smart speaker [58, 59]. Likewise, Furhat was generally preferred over a smart speaker, although task performance results were more ambiguous in [61, 92]. Notably, in the 10 robot-speaker comparisons, a smart speaker was rated higher only once, specifically when compared with the non-humanoid Fake Dog [56], particularly on several sub-scales of the Godspeed questionnaire.

Vyo, an artifact-shaped robot, was the only robot assessed against multiple other devices in a smart-home scenario [70]. It was rated as offering higher levels of flow, enhancing users' focus, involvement, and enjoyment during interactions. However, users felt that the other devices (a speaker, smartphone, and tablet) provided them with a greater degree of control over the interaction.

The findings from comparison studies within the *robot-tablet* and *robot-computer* categories did not consistently indicate that humanoid social robots were rated higher overall. While there were instances where humanoid robots received more favorable ratings, this was not universally the case. For example, REEM was associated with fewer privacy concerns [112], and NAO demonstrated greater social presence when facilitating interactions between two people [120]. However, in other scenarios, the tablet or computer outperformed the robots. Specifically, a tablet was rated higher for usability in providing information in a clinical radiology setting [106], and a computer collected survey data more effectively than Opie, a humanoid-like torso robot with a tablet displaying eyes as its head [44]. Furthermore, in 7 out of 11 comparisons (> 60%) involving humanoid robots, the results were either inconclusive or did not show any significant differences.

Social robot	Compared with	Main findings	Ref
NAO	smart speaker	During perimetry, NAO was preferred to a speaker providing the same feedback	[73]
NAO	smart speaker (Google Home)	NAO was more likeable and intelligent, higher anthropomorphism and animacy	[56]
Pepper	smart speaker (Google Home)	Pepper was more likeable and intelligent, higher anthropomorphism and animacy	[56]
Fake Dog	smart speaker (Google Home)	Speaker was more likeable and intelligent, higher anthropomorphism and animacy	[56]
Telenoid	smart speaker	Communicating through Telenoid induced a higher pattern of frontal brain activation	[58]
Telenoid R4	smart speaker	Telenoid enabled better prediction of perceived difficulty of story	[59]
Furhat	smart speaker (Google Home)	Furhat was trusted more and helped increase task engagement and performance	[92]
Furhat	smart speaker (Amazon Echo)	Furhat was preferred and more socially present; using speaker was more efficient	[61]
Bono bot	smart speaker	Bono bot had higher overall UX but talkative participants preferred the speaker	[98]
Vyo	smart speaker	Vyo was more enjoyable; speaker enabled control from any location but felt uneasy and provided low situation awareness	[70]
Vyo	smartphone	Vyo had higher flow; phone gave more sense of control	[70]
Vyo	tablet	Vyo had higher flow; tablet gave more sense of control and was less distracting	[70]
Reeti	tablet	Reeti was more social and preferred; tablet had higher usability and lower workload	[28]
FURo-i Home	tablet	Sender presence is lower for the telepresence FURo-i Home robot but the robot itself is perceived as more present	[22]
REEM	tablet	REEM elicits fewer privacy concerns when privacy policies are communicated to users	[112]
Pepper	tablet	Usability of tablet for information provision was rated higher	[106]
Pepper	tablet	No evidence for social desirability effect; Pepper was more socially present	[66]
NAO	tablet	No evidence for social desirability effect	[66]
NAO	computer	No influence on preference learning; NAO rated higher on animacy, anthropomorphism, intelligence, likeability, and safety	[96]
NAO	computer	NAO was a more socially present facilitator and personal topics were talked about more	[120]
NAO	computer	A Q&A session revealed no effects on interaction, social or usability measures	[47]
Opie	computer	Quality and completeness of survey data collected with a computer form was better	[44]
AIDA	smartphone	AIDA was able to promote safe driving behaviors and reduce cognitive load better	[117]
NAO	smartphone	NAO was preferred but did not increase movie recommendation acceptance	[93]
Count	*24*		*19*

Table 7: Social robots compared to other devices (*yellow* means outperforming)

Measures used. A considerable number of studies [28, 44, 47, 70, 92, 96, 106, 112] comparing social robots with other devices employed usability or user experience (UX) self-reporting scales. These included adapted versions of well-established measures such as the Interactive Experiences Ques-

tionnaire [122], PARADISE [113], System Usability Scale [17], and User Experience Questionnaire [64], as well as custom-designed UX questionnaires [73]. In total, 9 out of 19 papers (47%) employed these methodologies.

Overall, other devices were rated higher in usability or user experience (UX) than robots [28, 70, 106], or no significant differences were found, except for [112], which reported a significant main effect of robot embodiment. Additionally, several studies [28, 47, 56, 96, 98, 120] (6 out of 19 papers; 32%) employed adapted versions of more robot-specific measures, such as the Almere model [45], the Godspeed questionnaire [7], the Robotic Social Attributes Scale (RoSAS) [21], or USUS (Usability, Social acceptance, User experience, and Societal impact) scale [116].

Based on the findings from [28, 56, 96, 98], social robots were generally rated higher on these robot-specific scales compared to other devices. However, [47] did not observe significant differences on any of the Godspeed or Almere model scales. In total, 12 out of 19 papers (63%) employed such technology-oriented self-reporting scales.

Several of the remaining studies [22, 61, 66, 117, 120] (26%) focused on co-presence and social presence, using scales from sources like [42, 84]. Notably, [120] employed both the RoSAS scales and a social presence scale. Overall, social robots were reported as significantly more present than other devices, with one exception: [66] found that while Pepper was experienced as more socially present than a tablet, NAO was not. Additionally, [58, 59] used brain imaging (fNIRS) to demonstrate that Telenoid improved the prediction of perceived story difficulty compared to a speaker. Lastly, [93] primarily relied on behavioral measures to show that NAO was preferred over a smartphone.

Overall, social robots were reported as significantly more socially present than other devices, with one exception: [66] found that while Pepper was perceived as more socially present than a tablet, NAO was not. Additionally, [58, 59] used brain imaging (fNIRS) to demonstrate that Telenoid improved the prediction of perceived story difficulty compared to a speaker. Finally, [93] primarily utilized behavioral measures to show that NAO was preferred over a smartphone.

In summary, there is a clear trend indicating that social robots are perceived as more likable, more anthropomorphic, more animated, and more socially present than other devices. However, from a usability and user experience (UX) perspective, there is no evidence to suggest that social robots enhance UX compared to other devices.

To conclude, slightly more than half of the studies, 10 out of 19 (53%) employed behavioral measures. These ranged from facial landmark analysis to metrics such as the number of accepted recommendations. Only one study [70] utilized a semi-structured interview to collect and analyze qualitative data from participants.

6.5. How do robots compare to entities that are tele-present?

A total of 7 papers compared social robots to telepresence robots. The term 'telepresence' is defined as "a sense of transportation to a space created by technology" that "occurs when a user perceives that he or she is physically present in a remote environment" [31]. Out of the 7 papers, only one paper provided supporting evidence for the telepresent robot condition [79]. Results of the remaining articles showed that a physically present entity is favored over telepresent entities.

Measures used. A variety of measures were employed in the *robot-tele* category, encompassing both subjective assessments, such as trust, anthropomorphism, and believability, as well as objective metrics like fNIRS data and behavioral compliance [59]. For example, [79] utilized a 7-point Likert scale and the Godspeed questionnaire to evaluate trust and anthropomorphism, while [59] employed fNIRS to monitor brain activity during interactions, providing insights into cognitive engagement. Other studies, such as [62], focused on language learning fluency and emotional responses, using rubrics and Likert scales to capture the nuances of participant experiences.

Custom questionnaires were also widely used; for instance, [118, 85] explored well-being and social skills and adapted existing believability metrics to fit their experimental designs. Additionally, [31, 39] integrated measures of source credibility, affective learning, and presence to further analyze the complexities of human interaction with both physical and telepresence robots. Collectively, these measures provided a comprehensive understanding of how the physical presence of robots influences user perception, engagement, and overall interaction quality compared to telepresence systems.

6.6. Novelty Effect

Given that social robots are a relatively new and unfamiliar technology compared to humans and other devices, we examined whether any of the studies mentioned novelty as a factor influencing their results. Slightly more than half of the papers (54%; 43 out of all 80) referenced either *novelty* or *familiarity*. Interestingly, the frequency of these references varies across different comparison categories.

The highest percentage of papers mentioning novelty or familiarity (79%) were found in studies comparing social robots to other devices (15 out of 19 papers). This was followed by studies comparing social robots with avatars and telepresence systems, where 67% of the papers in both categories referenced novelty or familiarity (12 out of 18 for avatar comparisons and 4 out of 6 for telepresence system comparisons). In contrast, only 50% of the papers comparing social robots with other robots (7 out of 14 papers) and 41% of those comparing social robots to humans (15 out of 37 papers) addressed these factors.

Many papers recognize the potential impact of novelty, but most only mention it briefly without providing detailed discussion.

7. Conclusion and Discussion

This paper addresses the **3 primary research questions** guiding the systematic literature review of social robots in comparative studies.

RQ1: *"Which types of social robots and devices have been used in comparative studies?"*. The analysis reveals that most

studies compare social robots with humans, reflecting a significant interest in enhancing human-robot interactions, particularly in domains like healthcare, education, and customer service. Comparisons involving robots and avatars, as well as other devices such as tablets and smartphones, demonstrate the expanding exploration of multi-modal interaction contexts, which are essential for advancing robotic technologies and hybrid interactions. Less frequently explored categories, including teleoperated robots and artifact-shaped robots, although niche, offer valuable insights into specific aspects of human-robot interaction, thereby contributing to a broader understanding of social robotics. The classification of robot appearances emphasizes the critical role that physical form plays in shaping human perceptions and interactions, with humanoid robots like NAO and Pepper being the most frequently utilized due to their ability to closely replicate human features and behaviors.

RQ2: *"How have social robots been compared with humans and other devices?"*. In addressing this question, the review identified and categorized the primary measurement scales used across 80 studies, providing a comprehensive overview of the methodologies employed in human-robot interaction research. These measures were grouped into several categories, including self-report measures (encompassing robot-specific, psychological-specific, domain-specific, and technology-specific tools), behavioral measures, brain activity measures, test and task performance, interviews, video analysis, and sensor data. The diversity of these methodologies allows for a nuanced comparison of social robots with humans and other devices, offering a multi-faceted understanding of human-robot interactions. The integration of subjective self-reports with objective behavioral, physiological, and cognitive data is crucial in capturing the complexities of these interactions.

RQ3: *"How do social robots compare to humans and other devices?"*. The outcomes of these comparative studies indicate that social robots, particularly humanoid ones, often excel in domains requiring social presence and emotional engagement, such as education and healthcare. However, when compared to other devices like tablets and smartphones, social robots do not consistently outperform them in terms of usability or user experience. The findings also reveal that novelty and familiarity significantly influence user perceptions, with social robots often benefiting from the novelty effect, although this advantage tends to diminish as users become more familiar with the technology. In direct comparisons with humans, social robots are generally perceived as less capable in areas requiring complex emotional or cognitive tasks. Based on Table 4, interactions with humans consistently scored highest, highlighting a clear preference for human interaction over social robots in these contexts.

This systematic literature review provides a comprehensive understanding of the current landscape of social robot research. The findings underscore the growing interest in developing robots that can effectively interact with humans, particularly in socially and emotionally engaging ways. The diversity of measures used across studies highlights the complexity of human-robot interaction research, where subjective experiences are complemented by objective data to provide a holistic view. The review also identifies areas for further exploration, particularly in underrepresented robot categories and in the inclusion of diverse demographic groups in research. As social robotics continues to evolve, these insights will be instrumental in guiding the development of robots that are not only functional but also capable of enhancing human experiences across various domains.

7.1. Limitations

While this systematic literature review provides valuable insights into social robot research, several limitations must be acknowledged.

First, the methodology may be subject to bias in the coding and classification of studies. The absence of an interrater reliability check raises concerns about the consistency and objectivity of the data analysis process. Future reviews should incorporate interrater reliability measures to ensure the validity and reproducibility of coding decisions, minimizing the potential for subjective interpretation by individual researchers.

Second, there are limitations in the scope of the review. It is unclear whether the studies compared social robots solely based on their embodiment (e.g., physical form and presence) or their general capabilities (e.g., cognitive and functional performance). This lack of distinction may affect the interpretation of the results. Future research should explicitly define and control for these factors, ensuring that comparisons are contextually appropriate and that differences in embodiment or functionality are accounted for in the analysis.

Finally, the reviewed studies may have limited generalizability due to the homogeneity of demographic groups included in the research. Many studies do not account for cultural, age-related, or experiential diversity among participants, which could influence perceptions and interactions with social robots. Broader representation in future studies would enhance the applicability of findings across diverse populations and contexts.

Addressing these limitations in future work will strengthen the reliability and applicability of human-robot interaction research, enabling a more nuanced understanding of the potential and limitations of social robots in various domains.

Acknowledgements

I would like to express my sincere gratitude to my thesis supervisor, Professor Koen Hindriks, for his invaluable guidance, support, and encouragement throughout this research. His expertise and insights have been instrumental in shaping this work.

Appendices

Appendix A. Robot appearance classification of robot type

Extended Framework for Characterizing Social Robots (Adapted from K.Baraka., 2020)

Robot Type	Robot appearance classification	ALL	IN
No social robot	-	18	
Not specified	-	16	
AV1	Artifact-shaped robot: Imaginary	1	
EMAR	Artifact-shaped robot: Imaginary	2	2
EMYS	Artifact-shaped robot: Imaginary	3	2
Jibo	Artifact-shaped robot: Imaginary	4	1
KIBO	Artifact-shaped robot: Imaginary	1	
Kuri	Artifact-shaped robot: Imaginary	1	1
Magabot	Artifact-shaped robot: Imaginary	1	1
Maslo	Artifact-shaped robot: Imaginary	1	
Misty	Artifact-shaped robot: Imaginary	2	1
Sphero	Artifact-shaped robot: Imaginary	1	
Vyo	Artifact-shaped robot: Imaginary	1	
Zenbo	Artifact-shaped robot: Imaginary	2	
Astro	Artifact-shaped robot: Object-inspired	1	
Clocky	Artifact-shaped robot: Object-inspired	1	
Cozmo	Artifact-shaped robot: Object-inspired	3	1
Ed robot	Artifact-shaped robot: Object-inspired	1	
ElliQ	Artifact-shaped robot: Object-inspired	1	
Google Home	Artifact-shaped robot: Object-inspired	4	2
Amazon Echo	Artifact-shaped robot: Object-inspired	6	2
IdeaBot	Artifact-shaped robot: Object-inspired	1	
Robot via Arduino UNO board	Artifact-shaped robot: Object-inspired	1	
Vector	Artifact-shaped robot: Object-inspired	1	
DragonBot	Bio-inspired: Animal-inspired: Imaginary: Familiar	2	
Furby	Bio-inspired: Animal-inspired: Imaginary: Familiar	1	1
Patricc	Bio-inspired: Animal-inspired: Imaginary: Familiar	2	
MiRo-E	Bio-inspired: Animal-inspired: Imaginary: Unfamiliar	4	1
My Keepon	Bio-inspired: Animal-inspired: Imaginary: Unfamiliar	1	1
Probo	Bio-inspired: Animal-inspired: Imaginary: Unfamiliar	1	
Reeti	Bio-inspired: Animal-inspired: Imaginary: Unfamiliar	2	2
Biscuit	Bio-inspired: Animal-inspired: Real: Familiar	1	
Blue-Bot	Bio-inspired: Animal-inspired: Real: Familiar	2	
Dog smart speaker	Bio-inspired: Animal-inspired: Real: Familiar	1	1
ELE	Bio-inspired: Animal-inspired: Real: Familiar	1	
Hedgehog	Bio-inspired: Animal-inspired: Real: Familiar	1	1
iCat	Bio-inspired: Animal-inspired: Real: Familiar	1	
JFA cat	Bio-inspired: Animal-inspired: Real: Familiar	3	1
JFA dog	Bio-inspired: Animal-inspired: Real: Familiar	2	1
Perfect Petzzz dog	Bio-inspired: Animal-inspired: Real: Familiar	1	1
Pleo	Bio-inspired: Animal-inspired: Real: Familiar	2	2
Huggable	Bio-inspired: Animal-inspired: Real: Unfamiliar	2	
Paro	Bio-inspired: Animal-inspired: Real: Unfamiliar	5	2
ACTROID-F	Bio-inspired: Human-inspired: Android	2	
AIDA	Bio-inspired: Human-inspired: Android	2	1
Elenoide	Bio-inspired: Human-inspired: Android	1	1
Female android	Bio-inspired: Human-inspired: Android	1	
Philip K. Dick (PKD)	Bio-inspired: Human-inspired: Android	1	
Repliee Q2	Bio-inspired: Human-inspired: Android	1	
Telenoid R4	Bio-inspired: Human-inspired: Android	5	3
Custom robot face	Bio-inspired: Human-inspired: Body parts	1	
Floka	Bio-inspired: Human-inspired: Body parts	1	
Furhat	Bio-inspired: Human-inspired: Body parts	3	2
Lio	Bio-inspired: Human-inspired: Body parts	1	1
Maki	Bio-inspired: Human-inspired: Body parts	1	
MH-2	Bio-inspired: Human-inspired: Body parts	1	1
NDX-A hand	Bio-inspired: Human-inspired: Body parts	1	
Opie	Bio-inspired: Human-inspired: Body parts	1	1
Geminoid DK	Bio-inspired: Human-inspired: Geminoids	1	
Geminoid HI- 1	Bio-inspired: Human-inspired: Geminoids	1	
Geminoid-F	Bio-inspired: Human-inspired: Geminoids	1	1
Vizzy	Bio-inspired: Human-inspired: Humanoid	1	
Bono bot	Bio-inspired: Human-inspired: Humanoid	1	1
Nadine	Bio-inspired: Human-inspired: Humanoid	3	2
NAO	Bio-inspired: Human-inspired: Humanoid	80	29
AELOS	Bio-inspired: Human-inspired: Humanoid	1	
ALPHA 1P	Bio-inspired: Human-inspired: Humanoid	1	
Alpha 2	Bio-inspired: Human-inspired: Humanoid	2	
Arash	Bio-inspired: Human-inspired: Humanoid	1	
Asimo	Bio-inspired: Human-inspired: Humanoid	3	1
Atlas	Bio-inspired: Human-inspired: Humanoid	5	
Bandit	Bio-inspired: Human-inspired: Humanoid	1	
Casper	Bio-inspired: Human-inspired: Humanoid	1	
CHARLI	Bio-inspired: Human-inspired: Humanoid	1	
Cody	Bio-inspired: Human-inspired: Humanoid	1	
CommU	Bio-inspired: Human-inspired: Humanoid	3	1
Diego-san	Bio-inspired: Human-inspired: Humanoid	1	
EVA	Bio-inspired: Human-inspired: Humanoid	1	1
FACE	Bio-inspired: Human-inspired: Humanoid	1	1
Gina	Bio-inspired: Human-inspired: Humanoid	1	1
Glin+	Bio-inspired: Human-inspired: Humanoid	1	1
Hanson Robokind	Bio-inspired: Human-inspired: Humanoid	1	
HRP-2Kai	Bio-inspired: Human-inspired: Humanoid	1	1
iClooney	Bio-inspired: Human-inspired: Humanoid	1	
iCub	Bio-inspired: Human-inspired: Humanoid	6	1
InMoov robot	Bio-inspired: Human-inspired: Humanoid	1	
JD Humanoid	Bio-inspired: Human-inspired: Humanoid	1	
KASPAR	Bio-inspired: Human-inspired: Humanoid	1	
Kojiro	Bio-inspired: Human-inspired: Humanoid	1	
Mechanical humanoid robot	Bio-inspired: Human-inspired: Humanoid	1	
MecWilly	Bio-inspired: Human-inspired: Humanoid	3	
Meka	Bio-inspired: Human-inspired: Humanoid	2	
mObi	Bio-inspired: Human-inspired: Humanoid	2	1
Pepper	Bio-inspired: Human-inspired: Humanoid	37	13
Poppy	Bio-inspired: Human-inspired: Humanoid	1	1
QTrobot	Bio-inspired: Human-inspired: Humanoid	1	1
REEM	Bio-inspired: Human-inspired: Humanoid	1	1
Robi	Bio-inspired: Human-inspired: Humanoid	1	1
RoBoHoN	Bio-inspired: Human-inspired: Humanoid	2	
Robonova	Bio-inspired: Human-inspired: Humanoid	1	
Robovie	Bio-inspired: Human-inspired: Humanoid	6	1
Roboy	Bio-inspired: Human-inspired: Humanoid	2	
RUBI-6	Bio-inspired: Human-inspired: Humanoid	1	
Ryan	Bio-inspired: Human-inspired: Humanoid	1	
Sonny	Bio-inspired: Human-inspired: Humanoid	1	
Sophia	Bio-inspired: Human-inspired: Humanoid	2	
Teo	Bio-inspired: Human-inspired: Humanoid	1	
Zeno R25	Bio-inspired: Human-inspired: Humanoid	1	
Tessa	Functional robot	1	
FCR	Functional robot	1	
Kiropi	Functional robot	1	
AGV	Functional robot	1	1
Anti-Sedentary Robot	Functional robot	1	
ARC robot	Functional robot	1	
Baxter	Functional robot	4	1
Beam	Functional robot	1	
FURo-i Home	Functional robot	1	1
Genie	Functional robot	1	
Giraff	Functional robot	1	
Hobbit	Functional robot	1	
Julia	Functional robot	1	
Lego Mindstorms	Functional robot	3	2
MantaroBot	Functional robot	1	1
Nomad Scout	Functional robot	1	1
PeopleBot	Functional robot	1	
R2-D2	Functional robot	1	
Relay	Functional robot	1	
Roomba	Functional robot	1	

Appendix B. Self-report: Robot-specific Measures

Robot-specific measures are self-reporting scales that make explicit reference to robots, at least in their original, non-adapted form.

Appendix C. Self-report: Psychological-specific Measures

SAM	Functional robot	2	
Sawyer	Functional robot	1	
Smart wheelchair	Functional robot	1	
Sympartner	Functional robot	1	
Temi	Functional robot	1	
TIAGo	Functional robot	3	1
Turtlebot 2	Functional robot	1	
Wall-E	Functional robot	1	
PR2	Functional robot	1	

Appendix A. Robot appearance classification of robot type

Name	Ref	Use Count	Times adapted
Godspeed questionnaire	[7]	11	5
Robotic Social Attributes Scale (RoSAS)	[21]	6	2
Negative Attitudes toward Robots Scale (NARS)	[83]	6	1
Almere model	[45]	3	2

Table B.8: Robot-specific self-reporting measures

Self-report: Robot-specific	Source of measure	Count	Adapted
Multi-Dimensional Measure of Trust (MDMT)	Malle, B.F. and Ullman, D., 2021	3	1
Godspeed	Bartneck, C., et al., 2009	11	5
NASA Task Load Index	Hart, S.G., and Staveland, L.E., 1988	3	0
Robotic Social Attributes Scale (RoSAS)	Carpinella, C.M., et al., 2017	6	2
System Usability Scale (SUS)	Brooke, J., 1996	8	2
Almere model	Heerink, M., et al., 2010	3	2
Negative Attitudes toward Robots Scale (NARS)	Nomura, T., et al., 2006	6	1
Social presence scale	Biocca, F., et al., 2003	5	2
4-item social presence scale	Lee, K.M., et al., 2006	2	1
4-item automated social presence	Bailenson, J.N., et al., 2001 and Heerink, M., et al., 2010	1	1
Co-presence questionnaire	Lombard, M. et. al., 2000	1	0
Enjoyability questionnaire	Takayama, L., et al., 2009	1	0
Sociability questionnaire	Takayama, L., et al., 2009	1	0
Automation bias scale	Singh, I.L., et al., 1993	1	1
Perceived Information Quality and Openness	Nass, C., et al., 1996	1	0
5-item relationship with robots and technology	Velentza, A.M., et al., 2019	1	0

Appendix D. Self-report: Domain-specific Measures

12-item determine user's trust in the robot (credibility)	Rau, P.L.P., et al., 2013	1	0
5-item for attachment	Madsen, M. and Gregor S., 2000	1	0
Faith (confidence in the skills of the robots)	Madsen, M. and Gregor S., 2000	1	0
Believability metric	Poel, M., et al., 2009	1	1
15-item IDAQ scale (Individual Differences in Anthropomorphism Questionnaire)	Waytz, A., et al., 2010	1	0
9-item quality of interaction questionnaire	Wullenkord, R., 2017	1	0
Technology affinity: 19-item TA-EG questionnaire	Karrer, K., et al., 2009	1	0
Attitudes toward robots, physical and social attractiveness, human likeness, and trust questionnaire	Joosse, M.P.,et al., 2013	1	0
Human-Robot Interaction Comfort and Safety Scale	Brscic D, et al., 2015	1	1
Social and Moral Interaction with Robots Scale	Kahn PH, et al., 2012	1	1
1-item for human-likeness	Nitsch, V. and Glassen, T., 2015	1	0
Affinity for Technological Interaction (ATI)	Franke, T., et al., 2019	1	0
Perceived intentionality of the robot	Stenzel, et al., 2012	1	0

Ability to make its own decisions	van der Woerdt and Haselager, 2017	1	0
measure of awareness of the robot's morphology	Bornstein, B.H. and Zickafoose, D.J., 1999	1	1
Trust Perception Scale-HRI questionnaire (TPS-HRI)	Schaefer, 2016	1	0
16 items inspired by the Usability and UX - USUS	Weiss, A., et al., 2009	1	1
6-item assessment of acceptance	Van Der Laan, J.D., et al., 1997	1	0
6-item for uncanniness	Ho, C.C. and MacDorman, K.F., 2017	1	0
Computational thinking questionnaire	Atmatzidou, S. and Demetriadis, S., 2015	1	0
Interactive Experiences Questionnaire	Castro-González, A., et al., 2016	1	0
11 adjectives rated for enjoyableness and usefulness of the interaction 2 statements and 3 questions for suitability as central user interface	Fasola, J., and Mataric, M.J., 2012	1	1
Unified theory of acceptance and use of technology (UTAUT) model	Venkatesh, V., et al., 2003	1	0

Appendix B. Self-report: Robot-specific Measures

References

[1] Muneeb I. Ahmad and Reem Refik. "no chit chat!" a warning from a physical versus virtual robot invigilator: Which matters most? *Frontiers in Robotics and AI*, 9, 2022.

[2] Luis Almeida, Paulo Menezes, and Jorge Dias. Telepresence social robotics towards co-presence: A review. *Appl. Sci. (Basel)*, 12(11):5557, May 2022.

[3] Franziska Babel, Johannes Kraus, Philipp Hock, and Martin Baumann. Verbal and Non-Verbal conflict resolution strategies for service robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, August 2022.

Self-report: Psychological	Source of measure	Count	Adapted
Positive And Negative Affect Scale (PANAS)	Watson, D., et al., 1988	4	0
Intrinsic Motivation Inventory (IMI)	Deci, E.L. and Ryan, R.M., 2003	4	1
Social Interaction Anxiety Scale	Heimberg, R.G., et al., 1992	1	1
Conversational Skills Rating Scale (CSRS)	Spitzberg, B.H., 1995	1	0
Flow scale	Zuckerman, O. and Gal-Oz, A., 2013	1	1
Positive and Negative Syndrome Scale (PANSS)	Kay et al., 1987	1	0
Inclusion of Other in the Self (IOS)	Aron, A., et al., 1992	3	1
Mind Perception Questionnaire (MPQ)	Gray et al., 2007	1	0
UCLA Loneliness Questionnaire	Russell, D. and Peplau, L., 1980	1	0
Three-Item Loneliness Scale	Derrick et al., 2009	1	1
Big 5 Inventory (BFI)	John, O.P. et al., 1991	4	2
Agency rating	Sidarus, N., et al., 2017	1	0
Evaluation apprehension	Bolin, A. U., and Neuman, G. A., 2006	1	1
PHQ-2 (mental health status)	Löwe, B., et al., 2005	1	0
Relationship Closeness Inventory	Berscheid, E., et al., 1989	1	0
Social Support Questionnaire	Norbeck, J.S., et al., 1983	1	0
Subjective Happiness	Lyubomirsky, S. and Lepper, H., 1999	1	0
PERNOD (Perceived Norms of Deservingness)	Komorita, 1963; Matell and Jacoby, 1972	1	0
7-point Likert scale questionnaire for trust	Jian, J.Y., et al., 2000	1	1
Neo-FFI-30 personality scale	Korner, A., et al., 2008	1	0
Conversation experience items	Kardas, M., et al., 2021	1	0
Ten Item Personality Measure (TIPI)	Gosling, S.D., et al., 2003	1	0
Geriatric Depression Scale (GDS)	Sheikh, J.I. and Yesavage, J.A., 1986	1	0
Kikuchi's scale of social skills (Kiss-18)	Kikuchi, A., 1988	1	0
Subjective Well-Being Scale (SWBS)	Ito, J., et al., 2003	1	0
16-item of SES-17 Social Desirability Scale	Stöber, J., 1999	1	0
Reysen likability scale	Reysen, S., 2005	1	1
Level of enjoyment	Velentza et al. 2019	1	0
Self-Assessment Manikin (SAM) scale	Bradley & Lang, 1994	1	0
Empathy questionnaire	Batson, C.D., et al., 1997	1	0
International Test on Risk Attitudes (INTRA tests)	Rieger, M. O., et al., 2015	1	0
Personal Wellbeing Index	Van Beuningen, J. and de Jonge, T., 2011	1	0
Resilience Index	Wagnild, G.M. and Young, H.M., 1993	1	0
Mood and readiness to change scale	Carey, K.B., et al., 2001	1	0
Ryff's Psychological Well-being Scale (RPWS)	Van Dierendonck, D., 2004	1	0
Satisfaction with Life Scale	Pavot, W. and Diener, E., 2008	1	0
Working Alliance Inventory Short Revised covering task, goal, and bond (WAI-SR)	Munder, T., et al., 2010	1	0
6-item perceived warmth perceived competence	Cuddy, A.J., et al., 2008	1	0
Affective Learning Measure	McCroskey, 1994	1	0
Measure of Source Credibility	McCroskey & Teven, 1999	1	0
32-item horizontal and vertical individualism, and collectivism	Singelis, et al., 1995	1	0
12-item fulfillment of belonging, self-esteem, control, and meaningful existence	Zadro, et al., 2004	1	0
Perceived Stress Scale (PSS-10)	Cohen, S., et al., 1994	1	0
Group identification scale	Leach, C.W., et al., 2008	1	0
Group trust scale	Allen, K. and Bergin, R., 2004	1	0
1-item on participant's feeling of closeness	Lee and Choi, 2017	1	1
2-item Modified Interactant Satisfaction Survey	Riek & Robinson, 2008	1	1
Perceived sociability scale (4 item subscale)	Heerink, M., et at., 2009	1	1
Modified Interpersonal Attraction Scale	McCroskey, J.C. and McCain, T.A., 1974	1	0
5-item holistic evaluation of the interaction	Hoffmann, L., et al., 2013	1	1
14-item Schaeffer trust scale	Schaefer, 2013	1	-
4-item scale for user satisfaction (adapted from PARADISE)	Walker et al., 1997	1	1

Appendix C. Self-report: Psychological-specific Measures

[4] Evangelia Baka, Nidhi Mishra, Emmanouil Sylligardos, and Nadia Magnenat-Thalmann. Social robots and digital humans as job interviewers: A study of human reactions towards a more naturalistic interaction. In *International Conference on Human-Computer Interaction*, pages 455–474. Springer, 2022.

[5] Kim Baraka, Patrícia Alves-Oliveira, and Tiago Ribeiro. An extended framework for characterizing social robots. *Human-Robot Interaction: Evaluation Methods and Their Standardization*, pages 21–64, 2020.

[6] Zeynep Barlas. When robots tell you what to do: Sense of agency in human- and robot-guided actions. *Consciousness and Cognition*, 75:102819, 2019.

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Int. J. Soc. Robot.*, 1(1):71–81, January 2009.

[8] Bianca Bergande and Anne Gressmann. Towards computational thinking beliefs of computer science non-major students in introductory robotics - a comparative study. In *Advances in Intelligent Systems and Computing*, Advances in intelligent systems and computing, pages 69–80. Springer International Publishing, Cham, 2021.

[9] Elin A Björling, Honson Ling, Simran Bhatia, and Kimberly Dziubinski. The experience and effect of adolescent to robot stress disclosure: A mixed-methods exploration. In *Social Robotics: 12th International Conference, ICSR 2020, Golden, CO, USA, November 14–18, 2020, Proceedings 12*, pages 604–615. Springer, 2020.

[10] Elin A. Björling, Honson Ling, Simran Bhatia, and Jeff Matarrese. Sharing stressors with a social robot prototype: What embodiment do adolescents prefer? *International Journal of Child-Computer Interaction*,

Self-report: Domain specific	Source of measure	Count	Adapted
User Experience Questionnaire (UEQ)	Laugwitz, B., et al., 2008	1	0
PAR-Q (physical health status)	Thompson, P.D., et al., 2013	1	0
Foreign Language Classroom Anxiety Scale (FLCAS)	Horwitz, E.K., et al., 1986	1	0
ABC Questionnaire for Affective Benefits	Powell, L.E. and Myers, A.M., 1995	1	0
Apathy Scale for Institutionalized Patients with Dementia Nursing Home version (APADEM-NH)	Agüera-Ortizetal., 2015	1	0
Quality of Life in Late-stage Dementia(QUALID)	Weiner, et al.,2000; Garre-Olmoetal, 2010	1	0
3-item scale for customer satisfaction	Homburg, C., and Stock, R. M., 2004	1	1
4-item scale for innovative service behavior	Stock, R. M. , 2016	1	0
SF-12 Health Status	Bjorner, J.B. and Turner-Bowker, D.M., 2009	1	0
STIMEY questionnaire (STQ)	Pnevmatikos, D., et al., 2022	1	0
18-question TOPICS-SF (The Older Persons and Informal Caregivers Survey Short Form)	Searle, S.D., et al., 2008	1	0
8 items for hedonic value perception and utilitarian value perceptions	Childers, T.L., et al., 2002; Mullen, S.P., et al., 2011	1	0
Dallas Pain (DPQ) Questionnaire	Marty, M., et al., 1998	1	0
Fear-Avoidance Beliefs Questionnaire (FABQ)	Chaory, K., et al., 2004	1	0
Roland Morris (RMQ) Questionnaire	Zerkak, D., et al., 2013	1	0
Apathy Inventory(AI)	Robert, et al., 2002	1	0
The Global Deterioration Scale(GDS)	Reisbergetal.,1982	1	0
The Severe MiniMental State Examination (sMMSE)	Harrell et al.,2000; Buizaetal.,2011	1	1
Dallas Pain (DPQ) Questionnaire	Marty, M., et al., 1998	1	0
Mini Mental State Examination (MMSE)	Folstein et al.,1975; Loboetal.,1999	1	0
Neuropsychiatric Inventory(NPI)	Cummings, et al., 1994; Vilalta-Franchetal.,1999; Boadaetal.,2005	1	0

Appendix D. Self-report: Domain-specific Measures

28:100252, 2021.

[11] Agathe Blanchard, Sao Mai Nguyen, Maxime Devanne, Mathieu Simonnet, Myriam Le Goff-Pronost, and Olivier Rémy-Néris. Technical feasibility of supervision of stretching exercises by a humanoid robot coach for chronic low back pain: The R-COOL randomized trial. *Biomed Res. Int.*, 2022:5667223, March 2022.

[12] Roel Boumans, Fokke van Meulen, Koen Hindriks, Mark Neerincx, and Marcel Olde Rikkert. A feasibility study of a social robot collecting patient reported outcome measurements from older adults. *International Journal of Social Robotics*, 12:259–266, 2020.

[13] Roel Boumans, Fokke van Meulen, Koen Hindriks, Mark Neerincx, and Marcel G M Olde Rikkert. Robot for health data acquisition among older adults: a pilot randomised controlled cross-over trial. *BMJ Quality & Safety*, 28(10):793–799, 2019.

[14] Hannah Louise Bradwell, Katie Edwards, Deborah Shenton, Rhona Winnington, Serge Thill, and Ray B Jones. User-centered design of companion robot pets involving care home resident-robot interactions and focus groups with residents, staff, and family: Qualitative study. *JMIR Rehabil. Assist. Technol.*, 8(4):e30337, November 2021.

[15] Paul Bremner and Ute Leonards. Iconic gestures for robot avatars, recognition and integration with speech. *Front. Psychol.*, 7:183, February 2016.

[16] Elizabeth Broadbent, I Han Kuo, Yong In Lee, Joel Rabindran, Ngaire Kerse, Rebecca Stafford, and Bruce A MacDonald. Attitudes and reactions to a healthcare robot. *Telemed. J. E. Health.*, 16(5):608–613, June 2010.

[17] John Brooke. Sus: A quick and dirty usability scale. *Usability Eval. Ind.*, 189, 11 1995.

[18] J K Burgoon, J A Bonito, B Bengtsson, C Cederberg, M Lundeberg, and L Allspach. Interactivity in human–computer interaction: a study of credibility, understanding, and influence. *Comput. Human Behav.*, 16(6):553–574, November 2000.

[19] Martina Čaić, João Avelino, Dominik Mahr, Gaby Odekerken-Schröder, and Alexandre Bernardino. Robotic versus human coaches for active aging: An automated social presence perspective. *Int. J. Soc. Robot.*, 12(4):867–882, August 2020.

[20] Zachary Carlson, Louise Lemmon, MacCallister Higgins, David Frank, Roya Salek Shahrezaie, and David Feil-Seifer. Perceived mistreatment and emotional capability following aggressive treatment of robots and computers. *International journal of social robotics*, 11:727–739, 2019.

[21] Colleen M. Carpinella, Alisa B. Wyman, Michael A. Perez, and Steven J. Stroessner. The robotic social attributes scale (rosas): Development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 254–262, New York, NY, USA, 2017. Association for Computing Machinery.

[22] Jung Ju Choi and Sonya S. Kwak. Can you feel me?: How embodiment levels of telepresence systems affect presence. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 606–611, 2016.

[23] Francesca Ciardo, Frederike Beyer, Davide De Tommaso, and Agnieszka Wykowska. Attribution of intentional agency towards robots reduces one's own sense of agency. *Cognition*, 194:104109, 2020.

[24] Lorenzo Cominelli, Francesco Feri, Roberto Garofalo, Caterina Giannetti, Miguel A Meléndez-Jiménez, Alberto Greco, Mimma Nardelli, Enzo Pasquale Scilingo, and Oliver Kirchkamp. Promises and trust in human-robot interaction. *Sci. Rep.*, 11(1):9687, May 2021.

[25] L Coombes, K Bristowe, C Ellis-Smith, J Aworinde, L K Fraser, J Downing, M Bluebond-Langner, L Chambers, F E M Murtagh, and R Harding. Enhancing validity, reliability and participation in self-reported health outcome measurement for children and young people: a systematic review of recall period, response scale format, and administration modality. *Qual. Life Res.*, 30(7):1803–1832, July 2021.

[26] Filipa Correia, Samuel Gomes, Samuel Francisco Mascarenhas, Francisco S. Melo, and Ana Paiva. The dark side of embodiment - teaming up with robots vs disembodied agents. *Robotics: Science and Systems XVI*, 2020.

[27] E L Deci, H Eghrari, B C Patrick, and D R Leone. Facilitating internalization: the self-determination theory perspective. *J. Pers.*, 62(1):119–142, March 1994.

[28] Andrea Deublein and Birgit Lugrin. (expressive) social robot or tablet?–on the benefits of embodiment and non-verbal expressivity of the interface for a smart environment. In *Persuasive Technology. Designing for Future Change: 15th International Conference on Persuasive Technology, PERSUASIVE 2020, Aalborg, Denmark, April 20–23, 2020, Proceedings 15*, pages 85–97. Springer, 2020.

[29] Brian R Duffy. Anthropomorphism and the social robot. *Rob. Auton. Syst.*, 42(3-4):177–190, March 2003.

[30] L M Dunn and D M Dunn. *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4) [Database record*. APA PsycTests, 2007.

[31] Autumn Edwards, Chad Edwards, Patric R. Spence, Christina Harris, and Andrew Gambino. Robots in the classroom: Differences in students' perceptions of credibility and learning between "teacher as robot" and "robot as teacher". *Computers in Human Behavior*, 65:627–634, 2016.

[32] Lois L Elliott. Verbal auditory closure and the speech perception in noise

(SPIN) test. *J. Speech Lang. Hear. Res.*, 38(6):1363–1376, December 1995.

[33] David Feil-Seifer and Maja J Matarić. Human robot interaction. In *Encyclopedia of Complexity and Systems Science*, pages 4643–4659. Springer New York, New York, NY, 2009.

[34] Marcel Finkel and Nicole C Krämer. Humanoid robots – artificial. human-like. credible? empirical comparisons of source credibility attributions between humans, humanoid robots, and non-human-like devices. *Int. J. Soc. Robot.*, 14(6):1397–1411, August 2022.

[35] Jesse Fox and Andrew Gambino. Relationship development with humanoid social robots: Applying interpersonal theories to human-robot interaction. *Cyberpsychol. Behav. Soc. Netw.*, 24(5):294–299, May 2021.

[36] Julia Geerts, Jan de Wit, and Alwin de Rooij. Brainstorming with a social robot facilitator: Better than human facilitation due to reduced evaluation apprehension? *Frontiers in Robotics and AI*, 8, 2021.

[37] Anne Gressmann, Erica Weilemann, Dany Meyer, and Bianca Bergande. Nao robot vs. lego mindstorms. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research*, New York, NY, USA, November 2019. ACM.

[38] Andrea L Guzman and Seth C Lewis. Artificial intelligence and communication: A Human–Machine communication research agenda. *New Media Soc.*, 22(1):70–86, January 2020.

[39] Hyun-Tae Han, Yoshimune Nonomura, and Yuichi Tsumaki. Communication capability of telepresence system with the miniature humanoid mh-2. *Artificial Life and Robotics*, 23:328–337, 2018.

[40] Kerstin S Haring, Kelly M Satterfield, Chad C Tossell, Ewart J de Visser, Joseph R Lyons, Vincent F Mancuso, Victor S Finomore, and Gregory J Funke. Robot authority in human-robot teaming: Effects of human-likeness and physical embodiment on compliance. *Front. Psychol.*, 12:625713, May 2021.

[41] Kerstin S Haring, David Silvera-Tawil, Tomotaka Takahashi, Katsumi Watanabe, and Mari Velonaki. How people perceive different robot types: A direct comparison of an android, humanoid, and non-biomimetic robot. In *2016 8th International Conference on Knowledge and Smart Technology (KST)*. IEEE, February 2016.

[42] Chad Harms and Frank Biocca. Internal consistency and reliability of the networked minds measure of social presence. In *Seventh annual international workshop: Presence*, volume 2004. Universidad Politecnica de Valencia Valencia, 2004.

[43] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, Advances in psychology, pages 139–183. Elsevier, 1988.

[44] Scott Heath, Jacki Liddle, and Janet Wiles. The challenges of designing a robot for a satisfaction survey: Surveying humans using a social robot. *International Journal of Social Robotics*, 12(2):519–533, 2020.

[45] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. Measuring acceptance of an assistive social robot: a suggested toolkit. In *RO-MAN 2009 - The 18th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, September 2009.

[46] Anna Henschel, Guy Laban, and Emily S Cross. What makes a robot social? a review of social robots from science fiction to a home or hospital near you. *Curr. Robot. Rep.*, 2(1):9–19, February 2021.

[47] Damith C. Herath, Nicole Binks, and Janie Busby Grant. To embody or not: A cross human-robot and human-computer interaction (hri/hci) study on the efficacy of physical embodiment. In *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, pages 848–853, 2020.

[48] Guy Hoffman, Jodi Forlizzi, Shahar Ayal, Aaron Steinfeld, John Antanitis, Guy Hochman, Eric Hochendoner, and Justin Finkenaur. Robot presence and human honesty: Experimental evidence. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 181–188, New York, NY, USA, 2015. Association for Computing Machinery.

[49] Elaine K Horwitz, Michael B Horwitz, and Joann Cope. Foreign language classroom anxiety. *Mod. Lang. J.*, 70(2):125–132, June 1986.

[50] Sarah Jessup, Anthony Gibson, August Capiola, Gene Alarcon, and Morgan Borders. Investigating the effect of trust manipulations on affect over time in human-human versus human-robot interactions. In *Proceedings of the Annual Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2020.

[51] O P John, E M Donahue, and R L Kentle. *The Big-Five Inventory-Version 4a and 54*. Berkeley, CA, 1991.

[52] Michiel Joosse, Manja Lohse, Niels Van Berkel, Aziez Sardar, and Vanessa Evers. Making appearances. *ACM Trans. Hum. Robot Interact.*, 10(1):1–24, March 2021.

[53] Peter H Kahn, Jr, Takayuki Kanda, Hiroshi Ishiguro, Brian T Gill, Solace Shen, Heather E Gary, and Jolina H Ruckert. Will people keep the secret of a humanoid robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, March 2015. ACM.

[54] Hiroko Kamide, Yasushi Mae, Tomohito Takubo, Kenichi Ohara, and Tatsuo Arai. Direct comparison of psychological evaluation between virtual and real humanoids: Personal space and subjective impressions. *International Journal of Human-Computer Studies*, 72(5):451–459, 2014.

[55] Junko Kanero, Cansu Oranç, Sümeyye Koşkulu, G Tarcan Kumkale, Tilbe Göksun, and Aylin C Küntay. Are tutor robots for everyone? the influence of attitudes, anxiety, and personality on robot-led language learning. *Int. J. Soc. Robot.*, 14(2):297–312, March 2022.

[56] Haruka Kasuga and Yuichiro Ikeda. Gap between owner's perceptions and dog's behaviors toward the same physical agents: Using a dog-like speaker and a humanoid robot. HAI '20, page 96–104, New York, NY, USA, 2020. Association for Computing Machinery.

[57] Patricia A. Kelly, L. Annette Cox, Sandra F. Petersen, Richard E. Gilder, Amy Blann, Ashley E Autrey, and Kathryn MacDonell. The effect of paro robotic seals for hospitalized patients with dementia: A feasibility study. *Geriatric Nursing*, 42(1):37–45, Jan 2021.

[58] Soheil Keshmiri, Hidenobu Sumioka, Ryuji Yamazaki, Masataka Okubo, and Hiroshi Ishiguro. Similarity of the impact of humanoid and in-person communications on frontal brain activity of older people. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2286–2291, 2018.

[59] Soheil Keshmiri, Hidenobu Sumioka, Ryuji Yamazaki, Masahiro Shiomi, and Hiroshi Ishiguro. information content of prefrontal cortex activity quantifies the difficulty of narrated stories. *Scientific Reports*, 9(1):17959, 2019.

[60] Helena Kiilavuori, Veikko Sariola, Mikko J Peltola, and Jari K Hietanen. Making eye contact with a robot: Psychophysiological responses to eye contact with a human and with a humanoid robot. *Biol. Psychol.*, 158(107989):107989, January 2021.

[61] Dimosthenis Kontogiorgos, Andre Pereira, Olle Andersson, Marco Koivisto, Elena Gonzalez Rabal, Ville Vartiainen, and Joakim Gustafson. The effects of anthropomorphism and non-verbal social behaviour in virtual assistants. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, page 133–140, New York, NY, USA, 2019. Association for Computing Machinery.

[62] Evan Krisdityawan, Sho Yokota, Akihiro Matsumoto, Daisuke Chugo, Satoshi Muramatsu, and Hiroshi Hashimoto. Effect of embodiment and improving japanese students' english pronunciation and prosody with humanoid robot. *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–6, 2022.

[63] Christos Kyrlitsias and Despina Michael-Grigoriou. Social interaction with agents and avatars in immersive virtual environments: A survey. *Front. Virtual Real.*, 2, January 2022.

[64] Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 63–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[65] Nicole Lazzeri, Daniele Mazzei, Maher Ben Moussa, Nadia Magnenat-Thalmann, and Danilo De Rossi. The influence of dynamics and speech on understanding humanoid facial expressions. *Int. J. Adv. Robot. Syst.*, 15(4):172988141878315, July 2018.

[66] Benedikt Leichtmann and Verena Nitsch. Is the social desirability effect in human–robot interaction overestimated? a conceptual replication study indicates less robust effects. *International Journal of Social Robotics*, 13(5):1013–1031, 2021.

[67] Rui Li, Marc van Almkerk, Sanne van Waveren, Elizabeth Carter, and Iolanda Leite. Comparing human-robot proxemics between virtual reality and the real world. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 431–439, 2019.

[68] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mul-

row, Peter C Gøtzsche, John P A Ioannidis, Mike Clarke, P J Devereaux, Jos Kleijnen, and David Moher. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *J. Clin. Epidemiol.*, 62(10):e1–34, October 2009.

[69] Mike Ligthart and Khiet P. Truong. Selecting the right robot: Influence of user attitude, robot sociability and embodiment on user preferences. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 682–687, 2015.

[70] Michal Luria, Guy Hoffman, and Oren Zuckerman. Comparing social robot, screen and voice interfaces for smart-home control. CHI '17, page 580–628, New York, NY, USA, 2017. Association for Computing Machinery.

[71] Bertram F. Malle and Daniel Ullman. Measuring human-robot trust with the mdmt (multi-dimensional measure of trust). 2023.

[72] Paul Marshall and Eva Hor Necker. Theories of embodiment in HCI. In *The SAGE Handbook of Digital Technology Research*, pages 144–158. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, 2013.

[73] Allison M. McKendrick, Astrid Zeman, Ping Liu, Dilek Aktepe, Illham Aden, Daisy Bhagat, Kieren Do, Huy D. Nguyen, and Andrew Turpin. Robot Assistants for Perimetry: A Study of Patient Experience and Performance. *Translational Vision Science Technology*, 8(3):59–59, 06 2019.

[74] Nidhi Mishra, Gauri Tulsulkar, and Nadia Magnenat Thalmann. Nadine robot in elderly care simulation recreational activity: Using computer vision and observations for analysis. In *Human Aspects of IT for the Aged Population. Technology in Everyday Living*, Lecture notes in computer science, pages 29–51. Springer International Publishing, Cham, 2022.

[75] SM Mizanoor Rahman. Collaboration between a physical robot and a virtual human through a unified platform for personal assistance to humans. *Personal Assistants: Emerging Computational Technologies*, pages 149–177, 2018.

[76] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, and PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.*, 6(7):e1000097, July 2009.

[77] David Moher, Larissa Shamseer, Mike Clarke, Davina Ghersi, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A Stewart, and PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst. Rev.*, 4(1):1, January 2015.

[78] Ali Mollahosseini, Hojjat Abdollahi, Timothy D Sweeny, Ron Cole, and Mohammad H Mahoor. Role of embodiment and presence in human perception of robots' facial cues. *Int. J. Hum. Comput. Stud.*, 116:25–39, August 2018.

[79] Manisha Natarajan and Matthew Gombolay. Effects of anthropomorphism and accountability on trust in human robot interaction. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, March 2020. ACM.

[80] Margot M E Neggers, Raymond H Cuijpers, Peter A M Ruijten, and Wijnand A IJsselsteijn. Determining shape and size of personal space of a human when passed by a robot. *Int. J. Soc. Robot.*, 14(2):561–572, March 2022.

[81] Alison Nightingale. A guide to systematic literature reviews. *Surgery*, 27(9):381–384, September 2009.

[82] Toshiaki Nishio, Yuichiro Yoshikawa, Kazuki Sakai, Takamasa Iio, Mariko Chiba, Taichi Asami, Yoshinori Isoda, and Hiroshi Ishiguro. The effects of physically embodied multiple conversation robots on the elderly. *Front. Robot. AI*, 8:633045, March 2021.

[83] Tatsuya Nomura, Tomohiro Suzuki, Takayuki Kanda, and Kensuke Kato. Negative attitudes toward robots scale. *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, 2006.

[84] Kristine L. Nowak and Frank Biocca. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 12(5):481–494, 10 2003.

[85] Masataka Okubo, Hidenobu Sumioka, Soheil Keshmiri, and Hiroshi Ishiguro. Intimate touch conversation through teleoperated android: Toward enhancement of interpersonal closeness in elderly people. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 23–28, 2018.

[86] Raquel Oliveira, Patrícia Arriaga, Patrícia Alves-Oliveira, Filipa Correia, Sofia Petisca, and Ana Paiva. Friends or foes? socioemotional support and gaze behaviors in mixed groups of humans and robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 279–288, New York, NY, USA, 2018. Association for Computing Machinery.

[87] Mourad Ouzzani, Hossam Hammady, Zbys Fedorowicz, and Ahmed Elmagarmid. Rayyan—a web and mobile app for systematic reviews. *Syst. Rev.*, 5(1), December 2016.

[88] Thomas Platz, Ann Louise Pedersen, Philipp Deutsch, Alexandru-Nicolae Umlauft, and Sebastian Bader. Analysis of the therapeutic interaction provided by a humanoid robot serving stroke survivors as a therapeutic assistant for arm rehabilitation. *Front. Robot. AI*, 10:1103017, March 2023.

[89] Jana Plomin, Paul Schweidler, and Astrid Oehme. Virtual reality check: a comparison of virtual reality, screen-based, and real world settings as research methods for hri. *Frontiers in Robotics and AI*, 10, 2023.

[90] Stéphane Raffard, Catherine Bortolon, Mahdi Khoramshahi, Robin N Salesse, Marianna Burca, Ludovic Marin, Benoit G Bardy, Aude Billard, Valérie Macioce, and Delphine Capdevielle. Humanoid robots versus humans: How is emotional valence of facial expressions recognized by individuals with schizophrenia? an exploratory study. *Schizophr. Res.*, 176(2-3):506–513, October 2016.

[91] Gaia Rancati and Isabella Maggioni. Neurophysiological responses to robot–human interactions in retail stores. *J. Serv. Mark.*, 37(3):261–275, February 2023.

[92] David A. Robb, José Lopes, Muneeb I. Ahmad, Peter E. McKenna, Xingkun Liu, Katrin Lohan, and Helen Hastie. Seeing eye to eye: trustworthy embodiment for task-based conversational agents. *Frontiers in Robotics and AI*, 10, 2023.

[93] Silvia Rossi, Mariacarla Staffa, and Anna Tamburro. Socially assistive robot for providing recommendations: Comparing a humanoid robot with a mobile application. *International Journal of Social Robotics*, 10:265–278, 2018.

[94] Aïsha Sahaï, Emilie Caspar, Albert De Beir, Ouriel Grynszpan, Elisabeth Pacherie, and Bruno Berberian. Modulations of one's sense of agency during human-machine interactions: A behavioural study using a full humanoid robot. *Q. J. Exp. Psychol. (Hove)*, 76(3):606–620, March 2023.

[95] Kuntal Saroha, Sheela Sharma, and Gurpreet Bhatia. Human computer interaction: An intellectual approach. *International Journal of Computer Science and Management Studies*, 11, 08 2011.

[96] Sebastian Schneider and Franz Kummert. Exploring embodiment and dueling bandit learning for preference adaptation in human-robot interaction. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1325–1331, 2017.

[97] Ulrike Schultze. Embodiment and presence in virtual worlds: a review. *J. Inf. Technol.*, 25(4):434–449, December 2010.

[98] Katie Seaborn, Takuya Sekiguchi, Seiki Tokunaga, Norihisa P Miyake, and Mihoko Otake-Matsuura. Voice over body? older adults' reactions to robot and voice assistant facilitators of group conversation. *International Journal of Social Robotics*, 15(2):143–163, 2023.

[99] Stela H. Seo, Denise Geiskkovitch, Masayuki Nakane, Corey King, and James E. Young. Poor thing! would you feel sorry for a simulated robot? a comparison of empathy toward a physical and a simulated robot. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, page 125–132, New York, NY, USA, 2015. Association for Computing Machinery.

[100] Elaheh Shahmir Shourmasti, Ricardo Colomo-Palacios, Harald Holone, and Selina Demi. User experience in social robots. *Sensors (Basel)*, 21(15):5052, July 2021.

[101] Candace L Sidner, Timothy Bickmore, Bahador Nooraie, Charles Rich, Lazlo Ring, Mahni Shayganfar, and Laura Vardoulakis. Creating new technologies for companionable agents to support isolated older adults. *ACM Trans. Interact. Intell. Syst.*, 8(3):1–27, September 2018.

[102] Nathan J. Smyk, Staci Meredith Weiss, and Peter J. Marshall. Sensorimotor oscillations during a reciprocal touch paradigm with a human or robot partner. *Frontiers in Psychology*, 9, 2018.

[103] Micol Spitale, Minja Axelsson, and Hatice Gunes. Robotic mental well-

being coaches for the workplace. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, March 2023. ACM.

[104] Ruth Maria Stock and Moritz Merkle. Can humanoid service robots perform better than service employees? a comparison of innovative behavior cues. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. Hawaii International Conference on System Sciences, 2018.

[105] Ruth Stock-Homburg, Martin Hannig, and Lucie Lilienthal. Conversational flow in human-robot interactions at the workplace: Comparing humanoid and android robots. In *Social Robotics*, Lecture notes in computer science, pages 578–589. Springer International Publishing, Cham, 2020.

[106] Dietrich Stoevesandt, Patrick Jahn, Stefan Watzke, Walter A Wohlgemuth, Dominik Behr, Christian Buhtz, Irina Faber, Stephanie Enger, Karsten Schwarz, and Richard Brill. Comparison of acceptance and knowledge transfer in patient information before an mri exam administered by humanoid robot versus a tablet computer: a randomized controlled study. In *RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, volume 193, pages 947–954. Georg Thieme Verlag KG, 2021.

[107] Sofia Thunberg, Sam Thellman, and Tom Ziemke. Don't judge a book by its cover. In *Proceedings of the 5th International Conference on Human Agent Interaction*, New York, NY, USA, October 2017. ACM.

[108] Meritxell Valentí Soler, Luis Agüera-Ortiz, Javier Olazarán Rodríguez, Carolina Mendoza Rebolledo, Almudena Pérez Muñoz, Irene Rodríguez Pérez, Emma Osa Ruiz, Ana Barrios Sánchez, Vanesa Herrero Cano, Laura Carrasco Chillón, Silvia Felipe Ruiz, Jorge López Alvarez, Beatriz León Salas, José M Cañas Plaza, Francisco Martín Rico, Gonzalo Abella Dago, and Pablo Martínez Martín. Social robots in advanced dementia. *Front. Aging Neurosci.*, 7:133, September 2015.

[109] Ashesh Vasalya, Gowrishankar Ganesh, and Abderrahmane Kheddar. More than just co-workers: Presence of humanoid robot co-worker influences human performance. *PLoS One*, 13(11):e0206698, November 2018.

[110] Anna-Maria Velentza, Nikolaos Fachantidis, and Ioannis Lefkos. Learn with surprize from a robot professor. *Comput. Educ.*, 173(104272):104272, November 2021.

[111] Anna-Maria Velentza, Sofia Pliasa, and Nikolaos Fachantidis. Future teachers choose ideal characteristics for robot peer-tutor in real class environment. In *Communications in Computer and Information Science*, Communications in computer and information science, pages 476–491. Springer International Publishing, Cham, 2021.

[112] Jonathan Vitale, Meg Tonkin, Sarita Herse, Suman Ojha, Jesse Clark, Mary-Anne Williams, Xun Wang, and William Judge. Be more transparent and users will like you: A robot privacy and user experience design experiment. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '18, page 379–387, New York, NY, USA, 2018. Association for Computing Machinery.

[113] Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. Paradise: A framework for evaluating spoken dialogue agents. *arXiv preprint cmp-lg/9704004*, 1997.

[114] Bingcheng Wang and Pei-Luen Patrick Rau. Influence of embodiment and substrate of social robots on users' decision-making and attitude. *International Journal of Social Robotics*, 11:411–421, 2019.

[115] D Watson, L A Clark, and A Tellegen. Development and validation of brief measures of positive and negative affect: the PANAS scales. *J. Pers. Soc. Psychol.*, 54(6):1063–1070, June 1988.

[116] Astrid Weiss, Regina Bernhaupt, Michael Lankes, and Manfred Tscheligi. The usus evaluation framework for human-robot interaction. In *AISB2009: proceedings of the symposium on new frontiers in human-robot interaction*, volume 4, pages 11–26, 2009.

[117] Kenton Williams, José Acevedo Flores, and Joshua Peters. Affective robot influence on driver adherence to safety, cognitive load reduction and sociability. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '14, page 1–8, New York, NY, USA, 2014. Association for Computing Machinery.

[118] Jin Xu, De'Aira G. Bryant, Yu-Ping Chen, and Ayanna Howard. Robot therapist versus human therapist: Evaluating the effect of corrective feedback on human motor performance. In *2018 International Symposium on Medical Robotics (ISMR)*, pages 1–6, 2018.

[119] Wenjing Yang and Yunhui Xie. Can robots elicit empathy? the effects of social robots' appearance on emotional contagion. *Computers in Human Behavior: Artificial Humans*, (100049):100049, February 2024.

[120] Alex Wuqi Zhang, Ting-Han Lin, Xuan Zhao, and Sarah Sebo. Ice-breaking technology: Robots and computers can foster meaningful connections between strangers through in-person conversations. CHI '23, New York, NY, USA, 2023. Association for Computing Machinery.

[121] Bingqing Zhang, Javad Amirian, Harry Eberle, Julien Pettré, Catherine Holloway, and Tom Carlson. From HRI to CRI: Crowd robot interaction—understanding the effect of robots on crowd motion. *Int. J. Soc. Robot.*, 14(3):631–643, April 2022.

[122] Álvaro Castro-González, Henny Admoni, and Brian Scassellati. Effects of form and motion on judgments of social robots animacy, likability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies*, 90:27–38, 2016.