Design Metaphors for Understanding User Expectations of Socially Interactive Robot Embodiments

NATHANIEL DENNLER, CHANGXIAO RUAN, JESSICA HADIWIJOYO, BRENNA CHEN, STEFANOS NIKOLAIDIS, and MAJA MATARIĆ, University of Southern California, USA

The physical design of a robot suggests expectations of that robot's functionality for human users and collaborators. When those expectations align with the robot's true capabilities, users are more likely to adopt the technologies for their intended use. However, the relationship between expectations and socially interactive robot design is not well understood. This paper applies the concept of *design metaphors* to robot design and contributes the *Metaphors for Understanding Functional and Social Anticipated Affordances* (MUFaSAA) dataset of 165 extant robots and the expectations users place on them. We used Mechanical Turk to crowd-source user expectations over three user studies. The first study (N=382) associated crowd-sourced design metaphors to different robot embodiments. The second study (N=803) assessed initial social expectations of robot embodiments. The final study (N=805) addressed the degree of abstraction of the design metaphors and the functional expectations projected on robot embodiments. We performed analyses to gain insights into how design metaphors can be used to understand social and functional expectations of robots and how these data can be visualized to be useful for study designers and robot designers. Together, these results can serve to guide robot designers toward aligning user expectations with true robot capabilities, facilitating positive human-robot interaction.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**; • **Computer systems organization** → Robotics.

Additional Key Words and Phrases: Socially Interactive Robots, Robot Morphology, Social Perceptions

ACM Reference Format:
Nathaniel Dennler, Changxiao Ruan, Jessica Hadiwijoyo, Brenna Chen, Stefanos Nikolaidis, and Maja Matarić. 2021. Design Metaphors for Understanding User Expectations of Socially Interactive Robot Embodiments. *ACM Trans. Hum.-Robot Interact.* 37, 4, Article 111 (August 2021), 42 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Human-robot interaction (HRI) research aims to develop robotic systems that can aid humans in a variety of different contexts. While advances in HRI have enabled robots to be more functionally performant and socially competent than ever, few robots are present in everyday life. The lack of adoption is in part due to concerns about user acceptance, which is linked to user expectations [19, 22, 59]. In socially interactive robots [33], these expectations are formed around two high-level concepts: the robot's functional capabilities (i.e., how well it can perform the task it is designed to do) and the robot's social capabilities (i.e., how natural interactions with the agent are) [19, 23, 33, 43]. Setting expectations too low results in robots that are expected to be useless while

Authors' address: Nathaniel Dennler, dennler@usc.edu; Changxiao Ruan, changxir@usc.edu; Jessica Hadiwijoyo, hadiwijo@usc.edu; Brenna Chen, brennajc@usc.edu; Stefanos Nikolaidis, nikolaid@usc.edu; Maja Matarić, mataric@usc.edu, University of Southern California, 941 Bloom Walk, Los Angeles, CA 90089, Los Angeles, California, USA, 90089.

2573-9522/2021/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

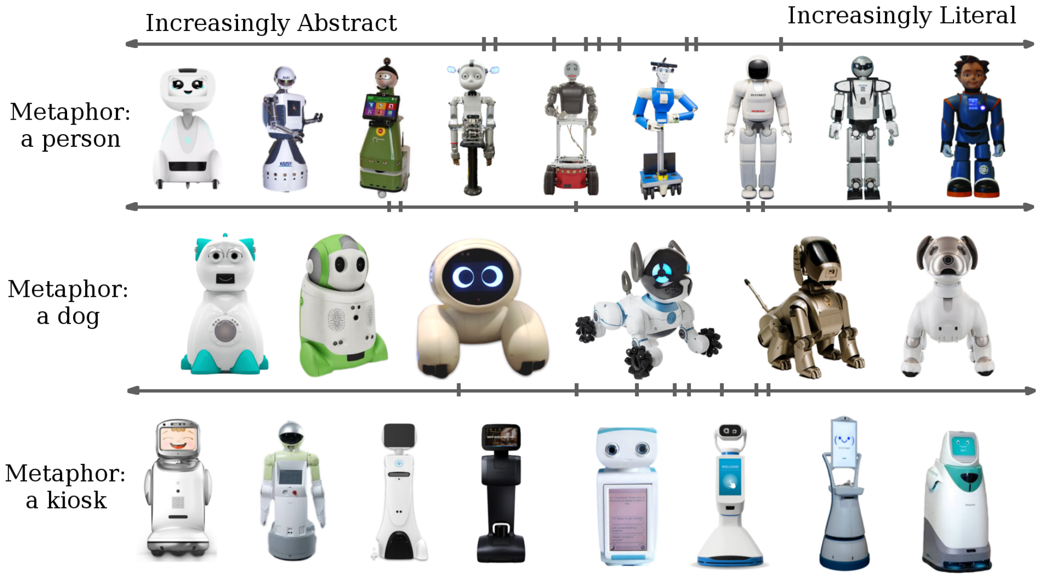


Fig. 1. Examples of robots' physical designs measured by abstraction level along three different design metaphors.

setting expectations too high leads to disappointment when robots fail to meet those expectations. Both scenarios inhibit acceptance and adoption. While software changes can rapidly change functional and social capabilities to align with user expectations, physical designs of robots have much longer development cycles. In order to inform robot embodiment design, it is crucial to understand the effect of robots' physical appearances on their social and functional expectations.

The problem of understanding the implications of design on system use has long been a topic of interest in human-computer interaction (HCI) [73, 112]. A powerful tool for addressing this problem is the concept of design metaphors [17, 21], which link novel designs with extant and familiar concepts or interactions. For example, a computer desktop shares many similarities with a physical desktop: a user can sort, organize, and label files by placing them in physically/visually co-located folders. By using these design metaphors, HCI practitioners are able to accurately set user expectation about actual system functionality. While previous work has focused on functional system behavior, design metaphors have recently been applied to social chatbots, indicating that selecting metaphors to align social expectations with true social capabilities drives user acceptance of these systems [51].

However, robots are significantly different from computers because robots are physically embodied [23]. While computers can also be treated as social actors [74], physical embodiment increases the social presence of robots by affording several signaling modalities that are not available to computers (e.g., gesture [85], gaze [4], and proxemics [71]). Physical expressiveness introduces complexities in user expectations, as specific aspects of a morphology designed for functional tasks may cause users to also expect competence in social tasks–a robot with an arm that grasps objects may reasonably be expected to use that arm to gesture. This duality of use for both functional and social expectations has lead to the exploration of robot embodiments as a sum of their low-level design features to understand how expectations are formed in anthropomorphic robots [79], zoomorphic robots [64], and rendered robot faces [49]. We argue that design metaphors, as illustrated in Figure

1, offer a broader and more holistic view of robot embodiment that can inform socially interactive robot design in complementary ways to component-based approaches. Particularly, robot designers and HRI practitioners can use design metaphors to quickly find similar robots to understand how people expect to interact with novel robots. Comparing robots based on design metaphors is much easier than comparing hundreds of low-level features.

In this paper, we introduce design metaphors as the conceptual tool for addressing the problem of understanding user expectations of robots based on their physical designs. We evaluate this tool by collecting a dataset of 165 extant robot designs and exploring four core research questions related to embodiment:

(RQ1.) How can we crowd-source design metaphors to describe how potential end-users conceptualize socially interactive robots?

(RQ2.) To what extent does a robot's embodiment establish *social* expectations in relation to its identity and social characteristics (e.g., role, likeability, and social perceptions) and how are these expectations moderated by design metaphors?

(RQ3.) To what extent does a robot's embodiment establish *functional* expectations in relation to its capabilities and expected use cases and how are these expectations moderated by design metaphors?

Because of the interplay of social and functional expectations identified in prior work [19, 43, 105], we also aim to understand:

(RQ4.) How are social and functional expectations related in socially interactive robots and what does this imply for the design of socially interactive robots?

In addressing these questions, this work contributes the *Metaphors for Understanding Functional and Social Anticipated Affordances* (MUFaSAA) dataset, an open-source collection of 165 robot embodiments and results of three crowd-sourced studies that provide insights toward the effect of robot design on user expectations of robot capabilities. This paper is organized as follows. Section 2 describes the related work in design metaphors, social expectations, and functional expectations. Section 3 describes the data collection process for the contributed dataset. The first study (N=382) is detailed in Section 4 and associates crowd-sourced design metaphors with robot embodiments. The second study (N=803) is detailed in Section 5 and assesses initial social expectations of robot embodiments. The third study (N=805) is detailed in Section 6 and addresses the degree of abstraction of the design metaphors and the functional expectations projected on robot embodiments. Section 7 examines trends in the designs of socially interactive robots across social and functional expectations. To evaluate the usefulness of the dataset, we show how viewing robots through design metaphors can contextualize and extend prior work in HRI, focusing on social and functional expectations. We discuss the implications of using metaphors as a tool for both interaction designers and robot designers in Section 8. To support replicable research, the collected dataset and interactive data visualizations to explore the dataset are made publicly available at `interaction-lab.github.io/robot-metaphors/`.

2 BACKGROUND AND RELATED WORK

This section provides background about design metaphors and understanding user expectations. Our work extends those methods toward studying the formation of user expectation about robot embodiment. We present a review of past work that aimed to achieve similar goals regrading robot design and identify how our contributions contextualize findings from that past work.

2.1 Understanding Design Through Metaphors

Design metaphors concisely describe complex ideas by associating unfamiliar objects with familiar objects that have similar characteristics. Design metaphors are extensively studied in human-computer interaction (HCI) as a way to help users develop mental models of the system they are interacting with in order to facilitate interaction [48, 51, 52, 102]. For example, HCI research shows that describing a chatbot with different design metaphors shaped user perceptions of the chatbot's warmth and competence, thereby affecting both the users' pre-interaction intention to use the system and their subsequent intention to adopt the system post-interaction [51].

The notion of design metaphors has also been recently applied to formalizing general design processes for socially interactive robots [24]. Deng et al. [23] provide a comprehensive review of HRI studies through the lens of design metaphors of the robot embodiments used and provide a design-metaphor based analysis of the relationships between different user studies and their outcomes. We apply this framework to explore how design metaphors shape the formulation of social and functional expectations of robots, aiming to enable HRI practitioners to contextualize their study findings relative to user expectations resulting from a robot's design.

2.2 Social Perception of Robot Embodiment

Past work in psychology, HRI, and HCI has shown that people form social expectations from initial impressions [15, 32, 59, 70]. Specifically in HRI, the embodiment hypothesis [103] states that a robot offers more channels (e.g., gesture and proxemics) to more strongly establish social expectations compared to computer-based agents. Thus, research in HRI must examine how those impressions are formed to understand how robots can socially interact with people while aligning with those expectations. Quantifying expectations, however, poses a significant challenge. Previous works have proposed specific measures that affect certain aspects of interaction, such as cuteness [18], credibility [110], animacy [7], and predictability [31]. However, these measures tend to focus on a single dimension of interaction in a particular context. To unify measures across contexts, Carpinella et al. [16] proposed RoSAS, a high-level scale that measures general attitudes toward robots through three constructs: Warmth, Competence, and Discomfort. In this work, we use the RoSAS scale to assess general social expectations of robots without tying the robots to a specific context.

In addition to social expectations that result from the robot's design, there are also components of social perception that arise from the interaction of the user's unique identity and the robot's design. Social Identity Theory includes a large body of research indicating that people react more favorably to other agents that share similar identity traits [98]. Higher degrees of identity closeness facilitate positive interaction metrics within people: cooperation [38], group cohesion [13], and moral evaluation [84]. Similar effects have been found when a robot is a partner in an interactive setting [34, 90]. While social identity is a highly complex phenomenon, recent research in HRI and HCI has increasingly focused on gender as a salient form of identity established through embodiment that may affect interaction in a variety of contexts [10, 57, 97]. Throughout this work, we examine how design metaphors may shape the identity of robots and how that impacts human-robot interaction.

Together, these general measures of social perception and identity enable HRI researchers to more easily compare perception of robots across disparate contexts. However, current comparisons of robot embodiment typically involve only a few robots performing similar tasks [54, 62, 101, 104]. To address this limitation, there is a need for a large-scale aggregation and comparison of embodiments in a centralized dataset in order to examine how design trends may affect social perception across contexts [65].

Efforts aimed at constructing such datasets to date focus on a specific construct of a robot's design (e.g., anthropomorphism [79] or animal-likeness [64]), and on quantifying that specific construct. Those data-driven approaches are crucial for understanding the studied constructs, but, in the context of design, several criteria may be optimized concurrently [24]. Kalegina et al. [49] addressed the multi-dimensionality of design by providing a dataset that relates multiple constructs from the Godspeed questionnaire [8] to the design of rendered robot faces, but did not focus on how the robot's embodiment may additionally modulate these constructs. In this work, we aim to address the gap in understanding multiple facets of embodiment design through the validated RoSAS scale for robot embodiments.

2.3 Functional Perception of Robot Embodiment

In addition to social expectations, functional performance of robots is the key evaluation metric of robotic systems. In HRI, robots are frequently meant to perform tasks that aid people, such as fetching objects [46], cooking [11, 111], folding clothes [106], and other physically assistive tasks [25, 50]. A robot's embodiment is naturally a key determinant of its functional performance on any such task. For example, a robot with no means of manipulation will not be able to retrieve an object, no matter how effective and robust its perceptual and planning algorithms may be. This work focuses on quantifying the expectations of functional performance that users place on robots based on their embodiment.

In robotics, an *affordance* is a fundamental type of interaction with the environment that is possible due to the robot's sensors and actuators [37]; a robot with wheels is *afforded* the ability to move through its environment, an action that an end-user may want the robot to perform. Affordances have been used in various methods of classifying robots relative to people's expectations of those robots' behaviors. One such method grouped robots by their sensors and actuators [47, 108]. Using that framing, a robot can be represented as a graph of sensors attached to the constituent robot components [47]. While this framing is useful, it may be incongruent with a user's perception of the robot's functionality. For example, a robot with a screen-rendered face with eyes may be expected to visually perceive its environment, despite not having a camera.

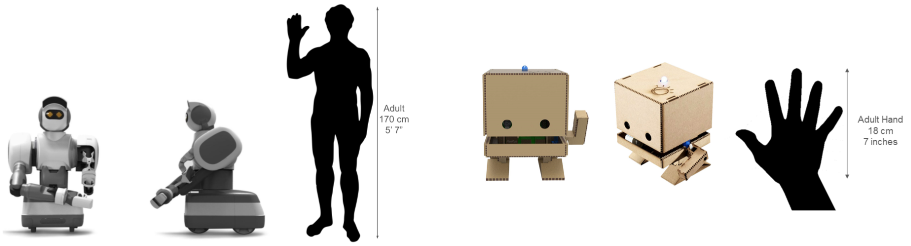
To address this issue, other researchers have grouped robots by the tasks people expect them to perform [12, 19, 49, 82], capturing user expectations within a robot's ability to perform those specific tasks. However, comparing robot designs across tasks remains challenging, in turn making it difficult to understand how general purpose robots may work across tasks. Hoffmann et al. [42] proposed the EmCorp scale–a high-level psychometric scale of expected robot capabilities from factor analysis that expresses perceived capability from three areas: perception, movement, and manipulation. By using this scale, disparate embodiments can be compared across tasks.

Accurately understanding human expectations of robot functional performance is important because functional performance is directly related to the trust users place in a robot [40]. Trust is a key component of adoption of a system, and is intensively studied in human factors and HRI [20, 40, 53, 76, 89]. Trust is usually evaluated through interactions with a robot [40]. Intuitively, when a robot performs successfully, trust is increased, while when a robot fails, trust decreases. However, the strength of this effect on trust depends greatly on how users perceive the robot's embodiment [40, 54]. While we were not able to have users interact with the system in the presented study (because of pandemic restrictions), we were able to provide initial expectations of the robot's functional affordances, which provide designers with insights about the aspects of the robot that need to align with user expectations and understand the effects robot failures.

3 COLLECTING A DATASET OF ROBOT EMBODIMENTS

To study user expectations of robots based on robot embodiments, we assembled the MUFaSAA dataset of 165 unique existing robots and codified those robots based on low-level design features. We then performed three studies on Amazon Mechanical Turk (MTurk) to evaluate our research questions (from Section 1) that relate design metaphors to social and functional expectations. By assembling the dataset to include robots from a variety of sources and outlining a process to create reproducible composite images as used in our dataset, we aim to enable replication and extension of this dataset to include novel robot designs as they continue to emerge. We describe the methodology and justifications for creating the composite images used in our dataset in the following section.

3.1 Data Collection Methodology



(a) Composite image for Aeolus.

(b) Composite image for TJBot.

Fig. 2. Example composite images from the MUFaSAA dataset.

Due to the immense cardinality of the design space of robots in widely varied contexts (i.e., drones, autonomous vehicles, industrial robots, etc.), we limited the scope of our dataset to those that fit the definition of *socially interactive robots* as proposed by Fong et al. [33]. Unlike Fong et al. [33], we do not require high-level dialogue so we can include non-humanoid embodiments. Thus, our dataset inclusion criteria were robots that had or could be perceived as having all of the following capabilities:

- The ability to perceive and express emotion.
- The ability to learn or recognize other agents.
- The ability to establish and maintain social relationships.
- The ability to use natural cues for social interaction (e.g., gaze or gesture).
- The ability to exhibit a distinctive personality or character.
- The ability to learn social competencies.

Using those guidelines, we assembled a collection of 165 robots from the IEEE "ROBOTS: Your guide to the world of robots" site [91] and Google searches of "Social Robot", "Socially Interactive Robot", "Socially Assistive Robot", "Robot Pet", and "Social Robot Animal". Google searches were performed under several user profiles and in incognito modes to mitigate the effects of prior search histories and stored user information. The data collection took place in June of 2020.

Each robot was represented with a composite image consisting of two high-resolution images, one of a front view and one of a side view of the robot, to convey the 3D structure of the robots' design. The sense of scale was provided by including a common reference image: a 170 centimeter tall gender-neutral silhouette for robots at/over 80 centimeters in height or a silhouette of a 18 centimeter tall human hand for robots under 80 centimeters in height. The image backgrounds were solid white, to control for contextual factors, cues, and influence. In addition, any objects that a robot was holding in the original image were edited out. We prioritized the use of images of robots in neutral poses with neutral facial expressions (for robots that had actuated faces). All composite images were created with identical aspect ratios and each view of the robot robot took up 30-40% of the composite image by width. Two example composite images from our dataset are shown in Figure 2.

3.2 Describing Embodiments as a Collection of Features

Similar to previous work in developing robot datasets [49, 64, 79], we codified robot embodiments with a series of manually labeled features derived from observed design patterns of the robots in the dataset and applicable features from previous studies [49, 64, 79, 97]. In total, we labeled 43 binary or categorical variables related to present/absent features, 4 ordinal variables related to feature counts, and 5 continuous variables. The continuous features were directly reported by robot data sheets, design documentation, or through manufacturer websites (e.g., height, weight, etc.). The categorical and binary features were evaluated through images of the robot. To address potential differences between observers of these manually defined features, two researchers independently coded all of the robots in the dataset. We calculated the interrater reliability of the attributed low-level design features that were not directly reported. The full set of coded features, descriptions, and interrater reliability are provided in Appendix B.

Our coding process identified some interesting trends in the design space. For example, the heights of the robots in this dataset appear strongly bimodal, with one peak at robots near 25cm in height and the other peak at robots 150cm in height. The most common color of robot in our dataset was white (101 robots), followed by blue (23 robots), and black (13 robots).

By manually describing embodiments in terms of low-level design features, designers can evaluate how specific design choices may affect user expectations. Using low-level features allows designers to adapt descriptions of embodiments to design patterns changes that come with evolving societal tastes and trends. Within this dataset, features were developed specifically for socially interactive robots. Designers from other areas of robotics can expand the design space with relevant features from other robot contexts.

In addition to the low-level coding of design features that allows this dataset to be compared with other datasets, we also placed an emphasis on using design metaphors as a conceptual tool to understand the expectations placed on a robot's embodiment. We collect and evaluate the utility of design metaphors in a series of three online studies.

3.3 Online Survey Design Preliminaries

Due to the nature and limitations of online surveys [86, 96, 109], we took several precautions in the three studies we administered through Amazon Mechanical Turk (MTurk). We used the following study participant inclusion criteria: user approval ratings of $\geq$ 99% for previously completed tasks, $\geq$ 1000 completed tasks, and normal or corrected to normal vision. Furthermore, we limited the participants to residents of the United States in order to control for cultural factors of design metaphors. The full set of survey questions is presented in Appendices A.2-A.4. In each study, participants answered questions for up to five separate robots. If survey participants appeared to stop providing meaningful answers as described below, their survey ended before five robots were viewed, and participants were paid for the robots they viewed, as approved by the IRB protocol (UP-18-00510).

In addition to the inclusion criteria, several measures were taken throughout the survey to protect against non-human MTurk participants. We employed a "honeypot" question that was invisible on the survey, but visible through the HTML files, thereby identifying non-human participants when answered. Participants who responds to this question were excluded from the study. Additionally, we employed random attention checks that instructed participants to answer a question by selecting a specific option to continue the survey, requiring a careful reading of the question. For quantitative questions, if the user responded with all neutral responses, their survey ended early and the data were discarded. For qualitative questions, if the user responded with identical text strings to ones that they had used previously (indicating copy-pasting), their survey ended early and their data were discarded.

As part of the study design process for all three studies, we first pilot-tested with 10 naïve users. The data from those pilot tests were used to confirm the reliability of any modified scales and to understand the expected time to complete the surveys in order to inform pricing, following recommendations from [87]. The pilot study participants were excluded from the analysis of the full studies. To analyze the studies, all statistics presented in this manuscript were calculated using the Pinguoin library for Python [100].

4 STUDY 1: ATTRIBUTING DESIGN METAPHORS TO EMBODIMENTS

The first study we conducted addresses RQ1: *How can we crowd-source design metaphors to describe how potential end-users conceptualize socially interactive robots?* We developed an MTurk study design to associate user-reported design metaphors with robots. We then compiled the findings for each robot in the dataset to create three user-reported design metaphors for each robot in the dataset.

4.1 Metaphor Survey Design

Participants were paid US$1.00 per robot for which they provided 2-5 design metaphors. Participants viewed up to five robots that were presented in a randomized and counter-balanced manner. Each response took about 3 minutes, and the whole survey took around 15 minutes.

4.1.1 Qualitative Measures.

To attribute design metaphors to each robot in the dataset, we developed three qualitative questions to allow participants to freely associate familiar concepts with the designs of robots in our dataset. In addition to specific metaphors, we asked users to explain their thought process by indicating what aspects of the robot represented the metaphors they provided. We asked the following three questions:

(1) *Description of Robot:* We provided an open-form response box with the prompt to describe the robot to a friend using two to three sentences.
(2) *Related Design Metaphors:* We provided an open-form response box to input at least two and up to five specific persons, animals, plants, characters, or objects that the robot looks like.
(3) *Reasoning for Related Design Metaphors:* We provided an open-form response box to describe why the aforementioned design metaphors were chosen. This box was immediately to the right of the previous response box.

4.2 Study 1: Design Metaphor Results

Due to the unconstrained nature of the design metaphor survey, our analysis focused on the qualitative analysis of the user-reported design metaphors. We present an overview of the specific metaphors used to describe the embodiments in our dataset and develop a coding scheme of the metaphors based on prior work. We evaluate if our coding system is reflective of findings from other works to demonstrate the ecological validity of our coding system.

4.2.1 Overview of Collected Data.

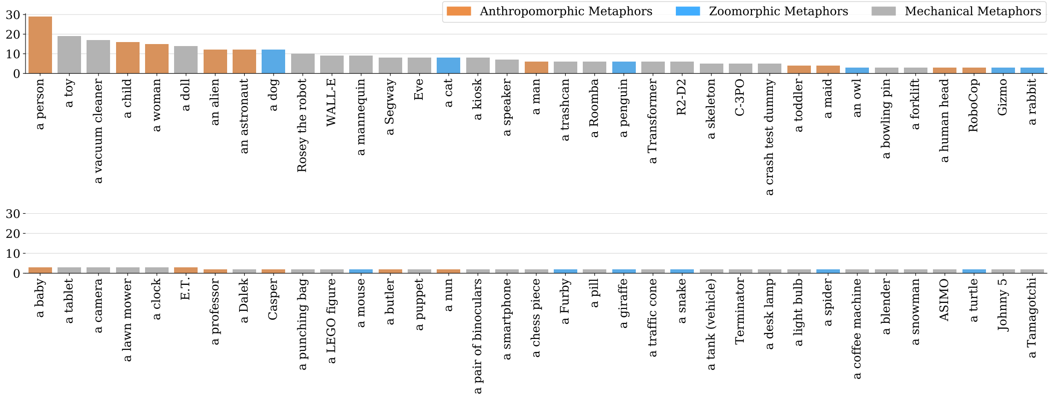


Fig. 3. Histogram of metaphor counts for metaphors that were associated with more than one robot in the dataset. The additional 129 singleton design metaphors are not shown. The specific sources of the referential metaphors (e.g., movies, shows) are removed for visualization, but were provided to participants during data collection.

A total of 382 participants took part in the study. Full demographic information is shown in Appendix A.1. We collected 1716 responses for the 165 robots in the dataset. Answers that did not provide specific persons, animals, plants, characters, or objects (i.e., "good", "nice", or paragraphs of copy-pasted text) were excluded from analysis. The removal of these responses did not result in a difference from the original distribution of responses, as evidenced by a chi-square test χ^2 (1, N = 165) = 12.62, $p > .999$. This indicates that the excluded answers followed a uniform distribution and that the assignment of robots did not cause the participants to provide non-specific answers.

4.2.2 *Metaphor Summary.*

The participant-sourced metaphors were unconstrained in the data collection with several metaphors appearing repeatedly throughout the dataset. For each robot, we calculated the top three metaphors by number of responses, grouping together similar responses for each robot; i.e., synonyms or hyponyms. This resulted in the final set of design metaphors that we consider in our dataset. The frequency of metaphors approximated an exponential distribution, as shown in Figure 3, with a few metaphors being highly used for several robots across the dataset.

In total, 199 metaphors were used by the participants to describe all the robots in the dataset. As in previous studies, we assigned each design metaphor to one of the following groups: anthropomorphic, zoomorphic, or mechanical [49, 62, 108]. These broad categories form the groups that will be compared in the subsequent user studies.

Metaphors were sorted into these categories by their literal interpretations; all nonliving metaphors were considered mechanical, living metaphors that represented animals were considered as zoomorphic, and living metaphors representing humanoids were considered anthropomorphic. Of the 199 metaphors, 46 were classified as anthropomorphic, 46 were zoomorphic, and 107 were mechanical. In addition to the broad types of metaphors, we observed that 38 of the metaphors were references to specific robots from popular media, such as Disney's WALL-E and The Jetsons' Rosey the Robot.

4.2.3 *Robot Metaphor Category Ascription and Manipulation Verification.*

To quantitatively compare between the different embodiments in the dataset, we categorized each embodiment into one of the three broad groups identified above. Embodiments were assigned to be either anthropomorphic, zoomorphic, or mechanical based on the majority of the top three design metaphors. For robots with one of each kind of metaphor, we chose the metaphor with the highest number of responses. Based on these criteria, the dataset consisted of 46 anthropomorphic, 28 zoomorphic, and 91 mechanical robots.

To verify that these groups are meaningful, we performed a manipulation check with the open-source ABOT database [79], a collection of similar robots that are rated on their human-likeness. We selected the robots that occurred in both datasets and compared their human-likeness to verify that the assignment of the three categories were meaningful. The intersection of the two datasets contained 44 mechanical, 36 anthropomorphic, and 6 zoomorphic robots. Using a Welch's ANOVA test for unequal group sizes, we found that robots with both mechanical metaphors ($M_{human-likeness} = 23.58$) and zoomorphic metaphors ($M_{human-likeness} = 30.61$) were significantly less human-like than robots with anthropomorphic metaphors ($M_{human-likeness} = 47.87$), with Welch's F(2,14.60) = 17.92, $p < .001$, $\eta^2 = .33$. This affirms that our assignment of robots to metaphor types is reflective of findings from other relevant datasets.

4.3 Discussion of Metaphor Findings

In the first study, we developed a method for users to freely associate robots with known concepts. We found that collecting free-response data reliably probed users' expectations and similar metaphors coalesced across several users for the same robot. By cross-referencing the metaphors used to describe the robots in our dataset with the ABOT database [79], we found that the user-reported design metaphors were meaningful representations of the robots. in particular, we find that measurements of robot anthropomorphism align with the design metaphors we collected.

Anthropomorphism is a salient aspect of a robot's embodiment that impacts the robot's likeability, expected intelligence, and expected empathy [113], and has been studied in HRI [27]. However, highly anthropomorphic robots may elicit unreasonable expectations that are not achievable in current systems, inhibiting adoption [27]. Considering the socially interactive robots that comprised our dataset, we found that the majority were not perceived as representing anthropomorphic design metaphors. Design metaphors therefore offer an extension to measurements of anthropomorphism and allow the comparison of a more diverse range of robots.

Interestingly, we observed that the robots in the dataset were described with metaphors that ranged from generic concepts to specific examples of characters from popular media. The relatively high and repeated occurrences of design metaphors that represented fictional characters point to popular media as a source of understanding robots. This mirrors the body of work in Cultivation Theory [36] that has identified that media shapes people's perception of the world. A previous study by Banks [6] applied this theory to robots and found that the amount of fictional robots that participants could recall from media affected intentions to interact with a single real socially interactive robot. Other studies have found similar links between the types of robots that participants can recall from media and the participant's social evaluation of robots [45, 94]. Design metaphors may serve as a helpful conceptual tool to understand how exposure to robots from the media may shape users' expectations across robots, especially those that resemble fictional robots from the media. In the second study, we examined in detail how social expectations may be formed based on these design metaphors.

5 STUDY 2: SOCIAL EXPECTATIONS

The goal of the second study was to address *RQ2: To what extent does a robot's embodiment establish social expectations of robots in relation to the robot's identity and social characteristics and how are these moderated by design metaphors?* We measured the social attributes of robots that formed the expectation of how a robot should socially behave. The interface for the study is described in Appendix A.3.

5.1 Study Design

The study followed a mixed design wherein each participant provided ratings for up to five robots in the dataset. Participants that did not pass the attention checks ended the study early. The assignment of robots was randomized and counter-balanced. Participants were paid US$0.20 per robot they rated, and took a median of 1.5 minutes per robot for an expected maximum length of 7.5 minutes to complete the survey.

5.1.1 Quantitative Measures.

To evaluate the social expectations of robot embodiment, we assembled a collection of questionnaires from relevant areas of HRI to measure general social constructs that can to be applied to a wide variety of robots. Our goal is to enable other researchers to contextualize robot embodiments. Guided by that goal, we collected quantitative evaluations of the following constructs:

(1) *RoSAS Scale:* We used a modified version of the validated RoSAS scale [16] to assess the constructs originally defined in RoSAS that were confirmed to be reliable in the pilot study (Section 3.3). All items followed the prompt "Indicate how closely the following words are associated with the robot" and were rated on a 7-point Likert scale of "strongly disagree" to "strongly agree". The scale measured the following constructs:
   - *Warmth* is related to the perception that another agent may want to help or harm us.
   - *Competence* is related to the perception that another agent has the ability to help or harm us.
   - *Discomfort* is related to the awkwardness of a robot.

(2) *Robot Gender Expression:* While gender is a complex social phenomenon, we measured *perceived gender expression* as proposed by the Bem Sex-Role Inventory Scale [9], using two axes– masculinity and femininity–as 7-point Likert scales. This approach allowed for perceptions of androgyny and non-gendered robots within the two axes.

(3) *Social Role:* The social role is a measure of the interaction dynamics between the person and robot in an interaction [23, 81]. We used a 9-point differential scale from Deng et al. [23], where 1 labeled the robot as "a subordinate", 5 labeled the robot as "a peer", and 9 labeled the robot as "a superior".

(4) *Identity Closeness:* Identity closeness measures the degree of in-group identification of the person with the robot [95]. We used a 9-point differential scale where 1 corresponded to the rater viewing the robot as "not at all like me", and 9 corresponded to the rater identifying the robot as "exactly like me". This scale has been shown to achieve high validity and reliability in related contexts [83].

(5) *Likeability:* Likeability measures the general attitude toward a robot, and has been used in other robot assessment studies [49, 68]. It was assessed using a 9-point differential scale, where 1 indicated the rater "strongly dislikes" the robot, and 9 indicated that the rater "strongly likes" the robot, adapted from the Godspeed Scale [8].

5.1.2 Qualitative Measures.

In addition to quantitative measures, we also employed qualitative evaluations of social perception. In particular, we were interested in open-ended responses to what participants liked about the robot, in order to glean participants' thought processes behind their quantitative ratings. We asked for qualitative evaluation of the following:

(1) *Reasoning for Likeability Rating:* In addition to the likeability rating, we collected an optional open-ended response about the reasons for liking or disliking a robot.

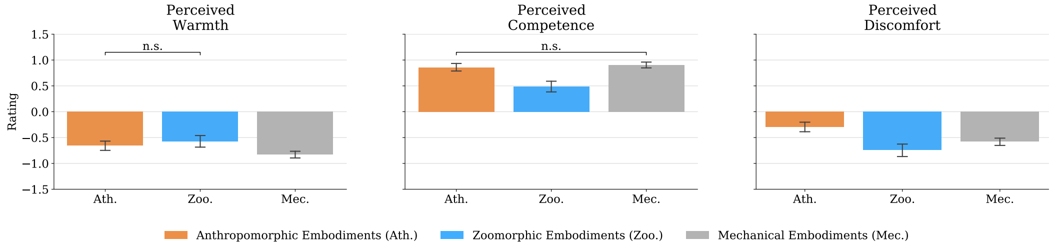


Fig. 4. Differences in means for anthropomorphic, zoomorphic, and mechanical embodiments for social constructs of embodiment. All differences are significant with $p < 0.001$ unless marked otherwise. Error bars represent 95% CI of means.

5.2 Study 2: Social Perception Results

Our analysis of the second study addresses social characteristics of robot embodiments motivated by prior work and our research question, namely social expectations and social identity. First, we explored how design metaphors shape high-level social expectations of robots. Second, we explored how design metaphors may shape a robot's identity through two perspectives: the robot's ascribed gender (robot-centric) and the participants' self-reported identity closeness (participant-centric).

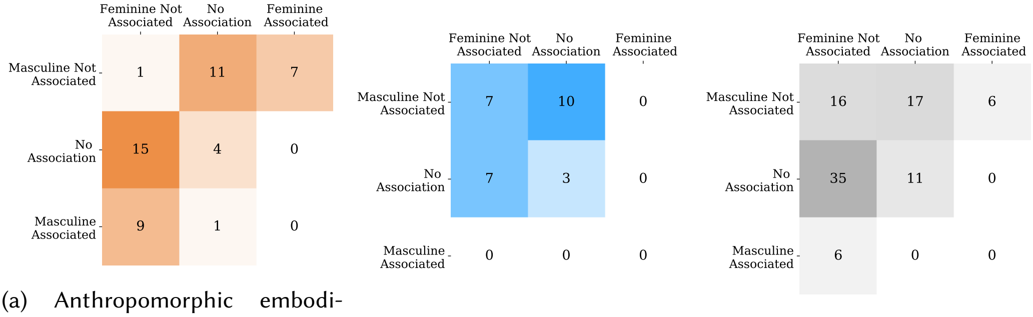
5.2.1 *Overview of Collected Data.*

A total of 803 participants took part in the study. Full demographic information is shown in Appendix A.1. We collected 3481 ratings from the participants for the 165 robots in the dataset. Participants who failed random attention checks ended the survey early. Entries that provided nonsensical answers to the qualitative questions or behaved randomly on the questionnaire were excluded. A total of 3155 responses were ultimately included in the analysis. A chi-square test showed that the exclusion of these 326 responses did not significantly affect the uniform distribution of assignment, χ^2 (1, N = 165) = 72.83, $p > .999$, indicating that it is unlikely that certain robots are more associated with excluded answers than other robots. The modified version of the RoSAS scale showed high reliability with Cronbach's alphas of $\alpha = 0.84$ for Warmth, $\alpha = 0.87$ for Competence, and $\alpha = 0.81$ for Discomfort. The values of the four questions that measured each construct are averaged for analysis.

5.2.2 *Metaphor Type and Social Expectation.*

For the robots in our dataset, the different categorizations of robot types had significant effects on participants' broad social expectations with respect to the RoSAS Scale, as shown in Figure 4. For warmth, the main effect of group type was significant, Welch's F(2, 1382.94) = 9.28, $p < .001$, $\eta_p^2 = .005$. Post hoc analysis revealed that the mean Warmth of anthropomorphic embodiments (M=-0.58) was significantly higher than the mean warmth of mechanical embodiments (M=-.83), $p = .001$, $\eta^2 = .007$, and the mean Warmth of zoomorphic embodiments (M=-.65) was significantly higher than that of mechanical embodiments, $p = .003$, $\eta^2 = .004$.

The main effect of group type for competence was also significant with Welch's F(2, 1353.95) = 24.09, $p < .001$, $\eta_p^2 = .016$. The difference between perceived competence in anthropomorphic embodiments (M=.86) was significantly higher than the mean competence of zoomorphic embodiments (M=.50), $p = .001$, $\eta^2 = .022$, and the mean competence of mechanical embodiments (M=.90) was significantly higher than the mean competence of zoomorphic embodiments, $p = .001$, $\eta^2 = .026$.



	Feminine Not Associated	No Association	Feminine Associated
Masculine Not Associated	1	11	7
No Association	15	4	0
Masculine Associated	9	1	0

(a) Anthropomorphic embodiments.

	Feminine Not Associated	No Association	Feminine Associated
Masculine Not Associated	7	10	0
No Association	7	3	0
Masculine Associated	0	0	0

(b) Zoomorphic embodiments.

	Feminine Not Associated	No Association	Feminine Associated
Masculine Not Associated	16	17	6
No Association	35	11	0
Masculine Associated	6	0	0

(c) Mechanical embodiments.

Fig. 5. A visualization of the space of gender expression by robot metaphor type.

Significant differences in discomfort were also observed across robot types, with Welch's F(2, 1399.77) = 16.13, $p < .001$, $\eta_p^2 = .010$. Zoomorphic embodiments (M=-.73) were rated significantly lower in discomfort than mechanical embodiments (M=-.57), $p = .04$, $\eta^2 = .003$, followed by anthropomorphic embodiments (M=-.31), $p = .001$, $\eta^2 = .007$, and zoomorphic embodiments were rated significantly lower in discomfort than anthropomorphic embodiments, $p = .001$, $\eta^2 = .042$.

5.2.3 *The Space of Robot Gender Expression.*

In addition to broad social characteristics, we were interested in identifying how robots may form a social identity through their embodiment. We examined user-reported perceptions of gender as a form of a robot's identity. To examine the space of gender expression in robots (i.e., how masculinity and femininity are embodied [3]; see Section 5.1.1 regarding the selection of these axes), we constructed the space according to the results of a two-tailed Wilcoxon signed-rank test. For each robot, we independently determined if the robot's average ratings for femininity and masculinity were significantly above zero, significantly below zero, or the null hypothesis that the value is zero could not be rejected. A value of zero corresponded to masculinity or femininity being neither associated with the robot nor not associated with the robot. The cutoffs for the robots were around average values of ±1, corresponding to "slightly agree" and "slightly disagree". The results of this analysis are shown in Figure 5.

By separating across design metaphor classifications, we observed patterns in how robot gender expression was perceived. Anthropomorphic embodiments were more likely to be perceived as having a significant association with either femininity or masculinity. Zoomorphic robots were unlikely to be associated with either masculinity or femininity. Mechanical embodiments were more likely to have no gender expression association, but in some cases they were associated with either masculinity or femininity.

5.2.4 *Formation of Ingroups and Outgroups.*

Because gender is one of many axes of identity, we additionally explored how the participants related to the robot through our measure of identity closeness. We observed that many of the robots in the dataset exhibited bimodal distributions, indicating that the formation of ingroups and outgroups may occur based on the robots' embodiments. Responses to Likert scale questions have been shown to follow binomial distributions in past work [2]. To evaluate this possibility in our dataset, we modeled the responses as coming from two possible models: a unimodal binomial model and a bimodal binomial model, with priors as described below. All models were developed and fit to the observed data using the pymc3 framework [88]. The following describes the unimodal model:

$$p \sim Beta(1, 1)$$
$$Y \sim Binomial(p, N)$$

where p represents the probability of success of a binomial trial with N repetitions. We used a 9-point Likert scale, thus N was set to 8. The prior for p was characterized as an uninformative Beta distribution. Y represented the values that users responded with. The following describes the bimodal model:

$$p_{ingroup} \sim Beta(2, 1)$$
$$p_{outgroup} \sim Beta(1, 2)$$
$$w \sim Dirichlet(1, 1)$$
$$Y \sim \sum_{i \in \{ingroup, outgroup\}} w_i \cdot Binomial(p_i, N)$$

The bimodal model was a weighted sum of two binomial distributions characterized with two different probabilities of binomial trial success, $p_{ingroup}$ and $p_{outgroup}$, which have weights corresponding to w. The ingroup and outgroup distributions had equal but opposite uninformative priors to ensure stability across different threads of MCMC sampling, since prior research has shown that ingroup membership is related to being closer than outgroup membership [83, 95]. All models were evaluated with two independent sampling chains with 20,000 iterations to guarantee convergence to the observed posterior distribution.

The binomial model was selected over the unimodal model if it was more than 10 times more likely based on the Watanabe Aikiake Information Criterion (WAIC) for the observed data. We found that of the 166 total robots, 60 were best described with the unimodel model of group membership, and 106 were best described with the bimodal model of group membership. A Chi Square test revealed that the distributions of metaphor types within these groups were significantly different from each other, $\chi^2(2, N = 166) = 59.33, p < .001$, with anthropomorphic and mechanical metaphors being more represented by unimodal model and zoomorphic robots being more often represented by the bimodal model. This suggests that zoomorphic robots were more likely to form ingroups and outgroups than anthropomorphic or mechanical designs. However, the majority of robots across all categories formed ingroups and outgroups based on their design.

5.3 Discussion of Social Expectation Findings

The social perception study revealed three key insights in the space of social expectations and identity characteristics of robots. First, the robots in our dataset showed clear differences in social expectations based on the categories of their design metaphors. Second, design metaphor categories were related to differences in how gender was attributed to the robots. Third, design metaphor type affected the formation of ingroups and outgroups in user-reported identification with the robot.

The finding that social expectations vary by metaphor type extends work that found differences in social attributes of different robot morphologies [62]. Design metaphors offer a framework to attribute robots to different conceptual categories, which may not be clear from simply looking at the robot design. Additionally, the social expectations data we collected can help to understand how expectations may be unmet or exceeded based on the robot's actual social skills. Furthermore, the results can inform discrepancies that may arise in studies using different robots for similar social tasks (e.g., those highlighted in sign language learning [55], cooking instruction [54], and conversational tasks [77]), and can support reproducibility in HRI research[99].

In addition to general social expectations, we found that robots from different metaphor categories have differences in perceived gender. Previous work on understanding gender in robots has focused on anthropomorphic embodiments [29, 57, 78, 80, 97]. Our work shows how perception of gender may be different for other forms of embodiments. We show that the design space of gender in socially interactive robots merits further exploration, similar to Perugia et al. [78], and replicate the findings that anthropomorphism is an important factor in participants perceiving a robot as gendered. There are relatively few robots in the dataset that portrayed both highly masculine and highly feminine characteristics of gender expression. One anthropomorphic robot did exhibit masculine characteristics and some degree of feminine characteristics. Interestingly, we also observed that anthropomorphic and mechanical embodiments tended toward not being associated with femininity, while zoomorphic embodiments tended slightly toward not being associated with masculinity. This suggests future directions of design research that could explore how to balance these trends in the current design space, in particular toward improving interaction metrics with marginalized genders [107].

Beyond gender, our results also extend work on sharing identities with robot partners. Previous work found that group membership can be established through personality modulation [30]. We found that the differences in embodiment may also establish different forms of group membership. This supports past findings that establish groups based on robot color [56], but also applies to robot embodument more broadly. Our dataset provides insights into how these groups may form for different embodiments, enabling research into group membership questions, since embodiment is difficult to change compared to behavior, personality, and cosmetic details. In the next study, we examined the functional affordances of what a robot can do in the world, without the context of its social capabilities.

6 STUDY 3: FUNCTIONAL EXPECTATIONS

The goal of the third study was to address *RQ3: To what extent does a robot's embodiment establish functional expectations in relation to its capabilities and expected use cases and how are these moderated by design metaphors?* The third study measured the robots' expected functional affordances and the attribution of expected tasks to the different robot embodiments. The interface for the study is provided in Appendix A.4.

6.1 Study Design

The study followed a mixed design where each participant provided ratings for up to five robots. Each rating was paid US$0.50 and took approximately 2 minutes to provide, and the whole survey had an expected length of 10 minutes. The robots each participant saw were randomized and counter-balanced to mitigate ordering effects.

6.1.1 Quantitative Measures.

(1) *EmCorp Measures:* We used a modified version of the 7-point Likert EmCorp-Scale [42] that has been validated in online survey contexts. We focused on the constructs of Shared Perception and Interpretation, Tactile Interaction and Mobility, and Nonverbal Expressiveness. The Corporeality construct was not studied because it represents how co-present a robot is in the room with the observer, and the robots in the dataset are 2D images. All items were rated on a scale from "strongly disagree" to "strongly agree". The scale measured the following constructs.

- *Shared Perception and Interpretation* is a measure of a robot's perceived perceptual capabilities, such as vision and hearing.
- *Tactile Interaction and Mobility* is a measure of a robot's perceived ability to move around and manipulate objects in space.
- *Non-verbal Expressiveness* is a measure of a robot's ability to use natural cues such as gestures and facial expressions.

(2) *Design Ambiguity and Design Atypicality Measures:* Design ambiguity and atypicality have been linked to aversion toward different robot designs in prior work [92]. In this work, we defined *ambiguity* as the difficulty of placing a robot in a single category, and *atypicality* as a robot having embodiment features not usually associated with the category it represents. We quantify these measures with differential scales valued from 1 to 9.

(3) *Metaphor Abstraction Measures:* The abstraction level of a metaphor provides a way to quantify how abstractly or literally the robot embodiment follows the metaphor. We quantified these values as a 9-point differential scale where 1 represented "highly abstract" interpretations of the design metaphor, and 9 represented "highly literal" interpretations of the design metaphor.

6.1.2 Qualitative Measures.

(1) *Task Descriptions:* We required participants to report two to five kinds of tasks each robot would be appropriate for, using open-ended responses.

6.2 Study 3: Functional Perception Results

Our analysis of the third study addresses both how embodiment shapes perceptions of functional affordances and of expected tasks. We examined how design metaphors shaped high-level perceptions of a robot's functional affordances through the EmCorp Scale. We additionally examined how the abstraction level of those metaphors may affect user perception, as previous work has proposed that abstraction is a key metric that affects functional outcomes of studies [23]. We performed qualitative analysis to understand how embodiment type affects tasks robots are expected to perform, and for whom robots are expected to perform those tasks.

6.2.1 Overview of Collected Data.

A total of 805 participants took part in the study. Full demographic information shown in Appendix A.1. We collected 3,435 ratings for the 165 robots in the dataset. Participants who failed random attention checks ended the survey early. Some responses were excluded for exhibiting disengaged or automated behaviors, as outlined in Section 3.3.

A total of 3,092 responses were ultimately included in the analysis. A chi-square test confirmed that the exclusion of the 343 responses did not significantly affect the uniform distribution of assignment, χ^2 (1, N = 165) = 25.39, $p > .999$, indicating that it is unlikely that certain robots are more associated with excluded answers than others. The modified version of the EmCorp-Scale showed high reliability with Cronbach's alphas of $\alpha = 0.91$ for Shared Perception and Interpretation, $\alpha = 0.84$ for Tactile Interaction and Mobility, and $\alpha = 0.87$ for Nonverbal Expressiveness. The values of the four questions that measured each construct were averaged.

6.2.2 Metaphor Type and Functional Expectation.

The different categorizations of metaphor type had clear effects on how participants expected a robot to perceive and interpret the world, with Welch's F(2, 1297.52) = 94.36, $p < .001$, $\eta_p^2 = .053$. Post hoc analysis revealed significance between all pairwise comparisons with $p < .001$. Zoomorphic embodiments were perceived as having the lowest perceptual capabilities (M=-.14), followed by mechanical embodiments (M=.26), $\eta^2 = .018$, and then followed by anthropomorphic embodiments

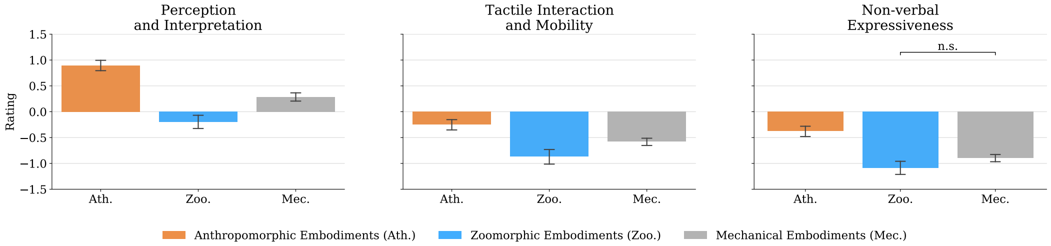


Fig. 6. Differences in means for anthropomorphic, zoomorphic, and mechanical embodiments for functional constructs of embodiment. All differences are significant with $p < 0.001$, unless marked otherwise. Error bars represent 95% CI of means.

(M=.95), $\eta^2 = .046$. Therefore, anthropomorphic embodiments had a much larger difference in perceived perceptual abilities than zoomorphic embodiments, $\eta^2 = .118$.

Tactile interaction and mobility also showed differences in metaphor types, with Welch's F(2, 1209.25) = 27.81, $p < .001$, $\eta_p^2 = .019$. All pairwise comparisons were significant in the post hoc analysis, with $p < .001$. Zoomorphic embodiments were perceived as having the lowest ability to manipulate objects in the world (M=.84), followed by mechanical embodiments (M=-.58), $\eta^2 = .007$, then followed by anthropomorphic embodiments (M=-.22), $\eta^2 = .013$. Zoomorphic embodiments therefore had much lower perceived tactile abilities than anthropomorphic embodiments, $\eta^2 = .037$.

The different forms of embodiment showed different expectations for non-verbal communication, with Welch's F(2, 1239.08) = 49.65, $p < .001$, $\eta_p^2 = .032$. Zoomorphic embodiments (M=-1.06) were perceived as less capable of communicating non-verbally than anthropomorphic embodiments (M=-.33), $p = .001$, $\eta^2 = .055$. Mechanical embodiments (M=-.91) were also viewed as having lower non-verbal communicative abilities than zoomorphic embodiments, $p = .001$, $\eta^2 = .033$.

6.2.3 *Abstraction and Functionality.*

To investigate the level of abstraction of a metaphor, we selected the top two most frequent metaphors from each category. For anthropomorphic metaphors, the two were "a person" and "a child"; for zoomorphic metaphors they were "a dog" and "a cat"; and for mechanical metaphors, they were "a toy" and "a vacuum". For all metaphors, the rating of functional expectations were regressed onto the level of abstraction of the given metaphor. Significant regressions are shown in Figure 7. The equation of the regression line is given along with the corresponding r^2 value.

For all anthropomorphic metaphors, we observed a significant increase in all perceived functional constructs as the robots were seen as more human-like. This supports findings from Section 6.2.2, where anthropomorphic metaphors were consistently rated as having the highest functional expectation. This trend of increasing functional expectation as embodiments were perceived more literally across all constructs was shown in most of the anthropomorphic metaphors that we measured. This finding aligns with the idea of anthropomorphization of robots as assigning more human-like abilities to these embodiments, not only socially but functionally as well [27].

For zoomorphic metaphors, we observed different trends across metaphors. For dog-like robots, the perception and interpretation and the non-verbal expressiveness constructs significantly increased as the robots appeared more like real dogs. For cat-like robots, however, only tactile interaction and mobility construct increased with increasingly literally perceived implementations. This difference may be a result of commonly held views about those animals in the United States (where the study participants were from); dogs are typically seen as more attentive to their owners

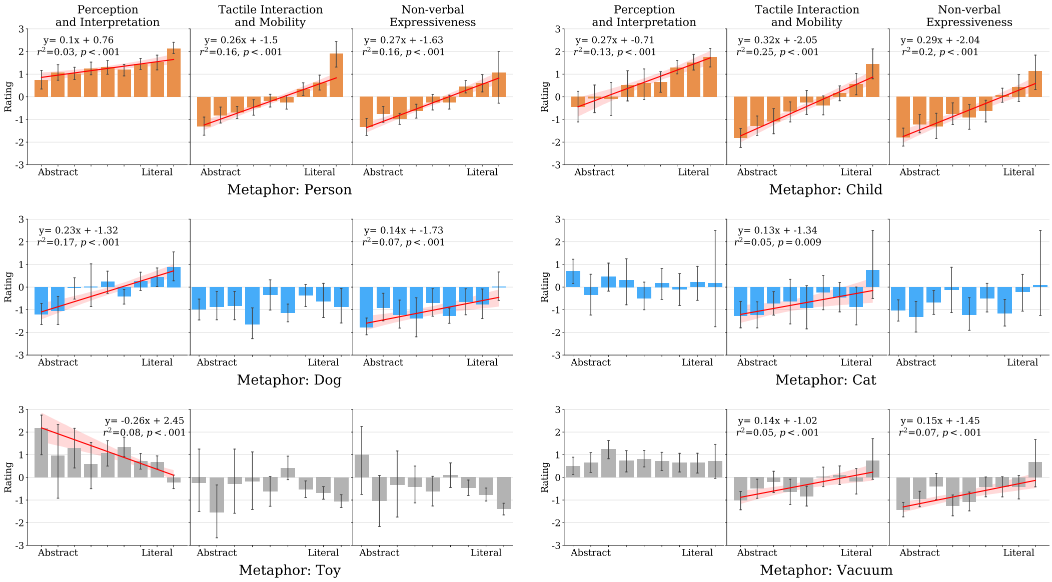


Fig. 7. Plots showing the effect of metaphor abstraction on a robot's perceived functional characteristics. Regression lines are shown for significant effects, after Bonferroni correction for 18 hypotheses.

Assigned Task		Specific Population
Companion	Home Assistant	Children
Customer Service	Informant	Elderly
Educator	Manufacturer	Persons with Disabilities
Entertainer	Surveillant	

Table 1. A table of codes developed through qualitative analysis of the user-reported tasks for socially interactive robot embodiments.

and are non-verbally expressive through embodied modalities such as tail-wagging [69]. Cats, on the other hand, are seen as more passive in interaction but also as more dexterous [69].

Mechanical metaphors also exhibited different trends across metaphors in terms of their level of abstraction. Robots described as "a toy" were seem as significantly *less* perceptive and interpretive when they looked more like literal toys. A similar phenomenon was observed by Hegel et al. [41], where users reported a Lego robot as looking like "a toy" and perceived it as simply pushing buttons, whereas they described computers as performing calculations. Robots resembling vacuum cleaners instead showed an increase in perceived tactile interaction and mobility as well as non-verbal communication capabilities as they appeared more like vacuum cleaners. This is likely because vacuums typically move around as part of the cleaning process, and robot vacuums prevalent in popular culture (e.g., the iRobot Roomba) exhibit non-verbal communicative abilities through modalities such as colored lights and beeping sounds.

6.2.4 *Embodiment and Task.*

To evaluate the task expectations of the different robot embodiments in the dataset, we developed a coding scheme based on an iterative axial coding approach [93] and applied it to the participants' free-response answers to the question regarding what task the robot appeared to be useful for. We

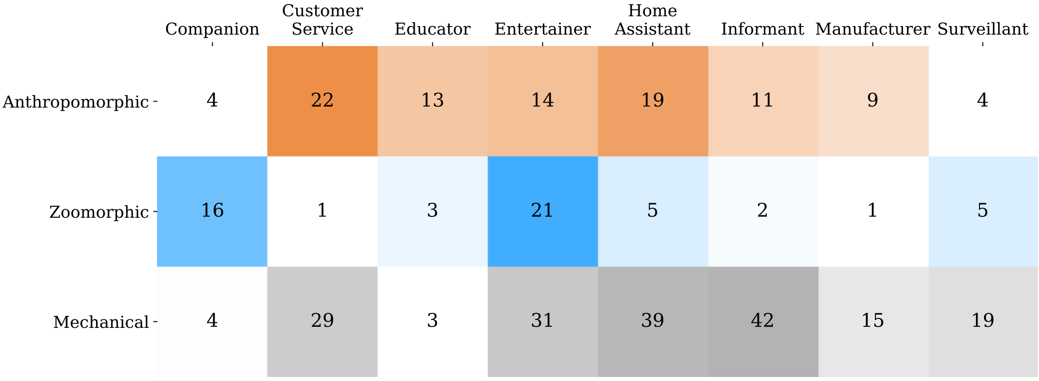


Fig. 8. A heat map of the distribution of the top two tasks for robots in our dataset separated by their metaphor type.

observed both task-related and intended population remarks from the participants. Eight main task-related codes were developed and three specific population labels were identified as trends in the design space of socially interactive robots, as summarized in Table 1. Interestingly, these codes have considerable alignment with the task categorization used by Kalegina et al. [49], despite being collected in an open format. The key differences we found were that we did not observe high numbers of responses for performing research, nor for health-related tasks. Furthermore, we observed two additional categories: being used as a companion and being used to collect or provide information.

The *companion* context was characterized by tasks involving the robot acting socially to improve mood or mental health over long periods. Examples of common tasks participants provided for this context were robots that "provide warmth and comfort", "are an interactive friend for my child", and "being a conversation partner". Most commonly, zoomorphic robots were described as being appropriate for this task, with task descriptions of 16 robots aligning with this category. This finding aligns with the zoomorphic robots' tendency to be perceived as comforting and warm (as found in Section 5.2.2), a key component of these tasks where functional expectation are not as important.

Robots ascribed to *customer service* contexts were defined as directly interacting with people in public places such as stores, restaurants, or hotels. Example tasks were robots that function as a "greeter or a receptionist", "a waiter" and "a museum guide". Both anthropomorphic embodiments (22 robots) and mechanical embodiments (29 robots) were described as being useful for customer service-type tasks. This aligns with the high expected functionalities of these embodiments to perform the services those tasks require.

Educator tasks were defined to involve knowledge transfer from or through the robot to a person interacting with the robot. Tasks fitting this category involved robots that could be used "in language education", "to interact with students in class", and to provide "light educational lessons like spelling or math". Interestingly, a robot's embodiment was often related to the topic that the robot was meant to teach. For example, the baby-like robot Babyloid was described as a "a training baby for expecting mothers", and the cat-like robot MarsCat saw suggested "to help educate about cats". Most commonly anthropomorphic embodiments (13 robots) were assigned to tasks relating to the educator category. This is consistent with the high perceived competence and functionality of anthropomorphic embodiments (as identified in Section 5.2.2).

For robots that played the role of *entertainers*, expected tasks aligned with short-term entertainment purposes. For example, robots in this category were expected to "play music", "be used like a toy", and "tell jokes". This category was common across all types of metaphors, however each metaphor was described as entertaining in a specific way. Anthropomorphic metaphors were described as being used as "a game-playing partner", zoomorphic metaphors were most often seen as functioning like "a pet that doesn't require attention when not in use", and mechanical metaphors fulfilled roles that are common in other forms of technology such as "playing music".

Home assistant robots were described as being able to work within the household, performing chores and other daily tasks, including "cleaning up after kids", "making coffee", and "carrying groceries". These tasks are similar to the customer service task, but are distinct in that they occur in the home and consist of repeated interaction with a few people. Similar to customer service tasks, both mechanical embodiments (39 robots) and anthropomorphic embodiments (19 robots) were found to be well-suited for the home assistant task.

Robots that act as *informants* were described with tasks that answer questions or otherwise provide information. Common tasks in this category were robots that "verbally answer questions", "tell time", or "report daily events like news or weather". Mechanical embodiments (42 robots) were most frequently described as being useful for these impersonal and intellectual tasks, consistent with their perceived high competence (as identified in Section 5.2.2).

Manufacturer robots were described in contexts where they build or move objects, typically without constant direct human interaction. These robots were expected to "carry heavy objects", "be a factory worker", and "pack in a warehouse". 15 Mechanical embodiments and 9 anthropomorphic embodiments were selected for tasks like these, primarily for their functional capabilities, as these tasks were perceived to not require social interaction.

Robots that fell in the *surveillant* category were those that monitor behavior, and were typically expected to provide security in some way. These robots were expected to be similar to "security alarms", "spy cameras", or "a sentry". Mechanical embodiments (19 robots) were most frequently attributed to this task. Similar to informants, these types of tasks are impersonal but require high levels of competence and perceptual capabilities, qualities that were attributed to mechanical embodiments.

6.3 Discussion of Functional Expectation Findings

In the study of functional expectations of robots, we found three main insights that relate design metaphors and expectations. First, we identified differences in high-level expectations of functional affordances from EmCorp-Scale measures. Second, we showed that the level of abstraction of different metaphors is an important consideration for establishing user expectations. Third, we found how design metaphors and expected robot tasks were related.

Our first key finding shows that overall anthropomorphic embodiments are perceived as having the most functional affordances, followed by mechanical embodiments, and finally followed by zoomorphic embodiments, with the fewest functional affordances. This result is consistent with the extensive body of work in anthropomorphism in robots [27]. The differences in functional expectation of robots also has implications for understanding how user trust evolves. Previous work has found that different robot error mitigation strategies were effective under different expectations of robots [61]. Using this dataset, interaction designers will be able to understand what user expectations are to more accurately select recovery strategies that work for the robot's specific embodiment.

The second key finding is related to the level of abstraction of the metaphors we collected for each robot. We found that increasingly literal metaphors (i.e., how closely the robot resembles a stereotypical version of the metaphor) significantly effect user expectations. Specifically, more literal metaphor implementations are more closely associated with the literal interpretation of that metaphor; literal dog-like robots shared similar expectations to actual dogs. This finding highlights how design metaphors may be used to holistically evaluate functional perceptions of a robot. Designers may use the data we collected to understand how novel designs may be perceived in terms of their level of abstraction by comparing to similar existing robots in our dataset. This opens interesting new directions in understanding how mental models of robots may be formed through these design metaphors. Understanding mental models is crucial for creating robots that are perceived as useful, a key factor in adoption.

Our qualitative findings of expected tasks further expand understanding of system use characteristics by providing the contexts in which people expect different design metaphors to be the most appropriate. This can help align robots to their expected use case which can help reduce the atypicality of a robot operating in an unexpected context, which has been shown to lead to negative evaluations of the robot [39]. It may also be used to help explain why different robots performing the same task can often be perceived differently [54]. In our final section we address how social and functional expectations may be jointly explored and offer interactions with this dataset for two communities: researchers in HRI and practitioners of robot design.

7 THE INTERACTION OF SOCIAL AND FUNCTIONAL EXPECTATIONS

The final investigation of our dataset addressed *RQ4: How are social and functional expectations related in socially interactive robots and what does this imply for the design of socially interactive robots?* In this section, we provide an overview of the space of socially interactive robots. We consider how low-level design features and high-level design metaphors may affect user expectations. We then show that both perspectives can generate insights into the design of socially interactive robots.

7.1 Predictive Features for Each Construct

To relate the low-level physical aspects of a robot's design to its expected social and functional affordances, we performed statistical feature selection on the manually coded features for each of the measured quantitative perceptual constructs. We used the Boruta algorithm [58] because it aims to find *all-relevant* features (i.e., all features that carry information on the modeled construct) as opposed to *minimal-optimal* features (i.e., the minimum set of features that maximize predictive accuracy for some specific model). The Boruta algorithm selects important attributes and is stable and unbiased when feature importance is measured with random forests of unbiased weak classifiers [58].

We performed feature selection by creating "shadow features" of the true features by randomly permuting the true values. Both the true and shadow features were used to predict the value of a construct. If a true feature was given importance that was higher than its shadow feature, it was considered useful in classification. This process was performed 500 times, for statistical validity. Due to the exploratory nature of this work, we selected features that were more relevant than their shadow features with a probability of 0.5. By selecting features that were relevant to specific constructs, we presented possible directions for investigating the relationship between robots' embodiments and their perceived expectations. Table 2 shows the selected features as they relate to the measured constructs.

These selected features can be classified into two categories: unobservable and observed. From the unobservable features we can analyze trends in the design space that are reflected in the overall design of robot embodiments as opposed to specific aspects of embodiments. The main unobservable features of importance were: *Year*, the year of release of the embodiment and *Industry?* whether or not the robot was at one point commercially available. The year descriptor allowed us to capture the non-stationary nature of design practices over time. Most notably, newer robots in our dataset were

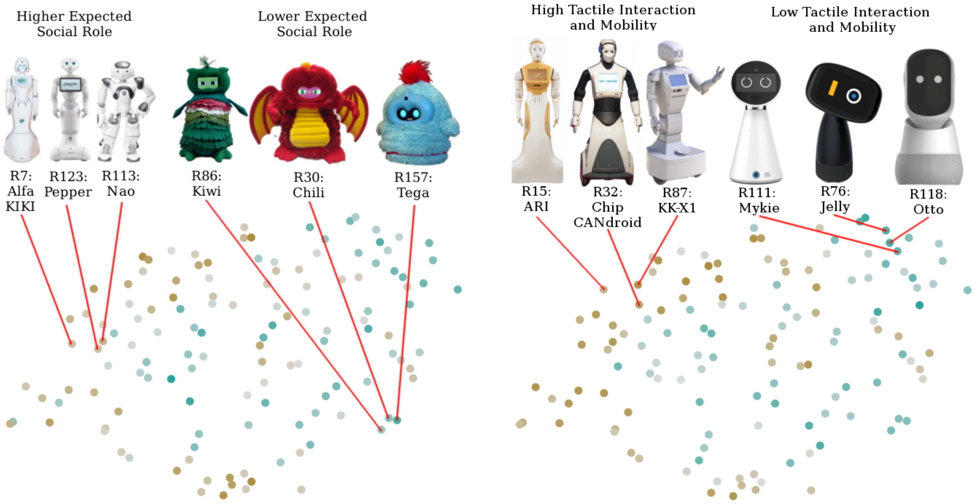
Construct	**Relevant Selected Features (and Relationship)**
Warmth	*Mouth?* (+)
Competence	*Height* (+), *Humanoid Embodiment?* (+)
Discomfort	*Height* (+), *Year* (-), *Mechanical Face?* (+), *Industry?* (-)
Femininity	*Height* (-), *Weight* (-), *Most Prominent Color = Beige* (+), *Blush?* (+), *High Waist-Hip Ratio?* (+), *Curved Embodiment?* (+)
Identity Closeness	*Height* (+), *Humanoid Embodiment?* (+)
Likeability	*Height* (-), *Industry?* (+)
Masculinity	*Height* (+), *Weight* (+), *Year* (-), *Curved Embodiment* (-), *Jointed Limbs?* (+)
Social Role	*Height* (+), *Year* (-), *Humanoid Embodiment?* (+), *Number of Arms* (+)
Perception and Interpretation	*Height* (+), *Humanoid Embodiment?* (+), *Dominant Classification*=Anthropomorphic (+)
Tactile Interaction and Mobility	*Height* (+), *Mobile?* (+), *Number of Wheels* (+), *Number of Arms* (+), *Jointed Limbs?* (+)
Nonverbal Communication	*Height* (+), *Year* (-), *Humanoid Embodiment?* (+), *Number of Wheels* (-), *Number of Legs* (+), *Number of Arms* (+), *Dominant Classification = Anthropomorphic* (+), *Jointed Limbs?* (+)
Design Ambiguity	*Height* (-), *Weight* (-), *Number of Legs* (-), *Dominant Classification = Anthropomorphic* (-)
Design Atypicality	*Weight* (-), *Number of Legs* (-), *Dominant Classification = Anthropomorphic* (-)

Table 2. The important features as selected by the Boruta algorithm from our manually coded feature set that corresponded to the constructs measured in the survey.

in general seen as less discomforting, less stereotypically masculine, had a lower expected social role, and had fewer non-verbal communicative capabilities. The commercially available attribute is related to the effect that larger teams of designers were likely involved with the development of the robot. Robots that were commercially available (had the *Industry?* attribute) were found to be more likeable and less discomforting. This suggests that robots used in settings that require the robot to be a comforting partner in interaction may benefit from using robots that are have had the benefit of well-resourced and thorough design processes.

Of the observed features of importance, height was the most frequently selected feature across all of the measured constructs. Height has previously been related to increased expected social role [81] in controlled settings. Consequently, identifying height as an important characteristic in our non-lab setting allows for generalizing its importance. Furthermore, the relationship between height and other constructs has not been studied closely, suggesting interesting research questions for future work.

Another trend we noted is the importance of anthropomorphism in functional constructs. In general, robots that are seen as human-like are expected to have higher degrees of functional capability and readiness to take on superior social roles. Such elevated expectations, however, must be met. Thus, when using anthropomorphic embodiments, care should be taken to ensure that robots operate to their expectation.



(a) t-SNE plot of expected social role.

(b) t-SNE plot of tactile interaction and mobility.

Fig. 9. A t-SNE visualization of the design space of robot embodiments. Each point represents one robot in the dataset. Brown represents high values and teal represents lower values of the measured ratings. Here we show only the front view of robots; study participants viewed composite images that included scaling information, as described in Section 3.1. The fully interactive version of this plot is located at `interaction-lab.github.io/robot-metaphors/`, where researchers, designers, and others interested in these findings may hover over points to view robots and click on a specific robot to view its social and functional expectations.

Our results replicated findings that related body shape to the expression of femininity in robots. Previous work has similarly linked the relationship between robots' waist-to-hip ratio to their perceived gender expression [10, 97]. Additionally, Kalegina et al. [49] found a relationship between perceived gender and the presence of blush, which we also found in our study's ratings of femininity. However, blush did not make robots appear less masculine in our dataset. This suggests that the axes of femininity and masculinity in robots are not diametrically opposed.

7.2 Visualization of the Design Space

To enable other researchers and robot designers to readily benefit from our findings, we developed an intuitive open-source visualization of our findings. Specifically, we used the hand-crafted features developed in Section 3.2 as descriptions of the physical attributes of the robots in our dataset. To learn a mapping without supervision from that high-dimensional feature space to 2D, we used t-Stochastic Neighbors Embedding [66] that preserves distances between points from high-D to 2D space. Figure 9 demonstrates that robots mapped near each other share similar characteristics. We show evaluations of different robots with different color values in 2D space. Higher values are concentrated in different parts of the space, indicating differences in social and functional expectations of the robot embodiments. This visualization technique can be used as a design tool to rapidly explore different robot embodiments for a desired set of expectations related to specific tasks.

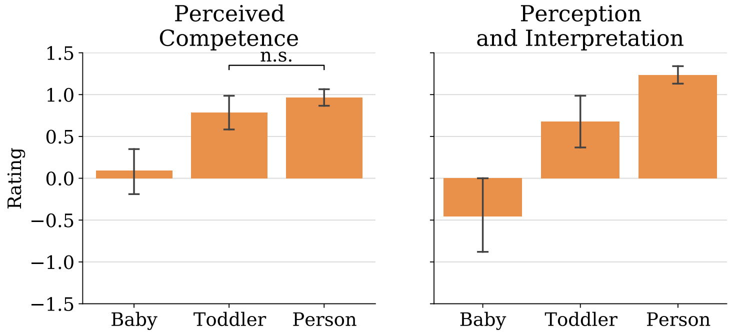


Fig. 10. Perceived competence and perceived perceptual ability by metaphors for different maturity levels.

7.3 Design Metaphor Semantics

We were interested in exploring whether design metaphors were semantically meaningful in terms of user perceptions. While the semantic space of metaphors is difficult to describe, there are some locally ordered areas. To examine the effects of social and functional perceptions, we selected three metaphors: "a baby", "a toddler", and "a person". Because age is associated with competence [51] and interpretation of the world, we expected that robots described with more mature metaphors would have higher competence and perceptual capabilities.

As expected, we found that a main effect was present on metaphor name and competence with Welch's F(2, 127.81) = 16.55, $p < .001$, $\eta_p^2 = .057$. The perceived competence is lower for robots labeled with the metaphor "a baby" (M=.08), followed by robots described with the metaphor "a toddler" (M=.78), $p = .001$, $\eta^2 = .09$, and then followed by robots described with the metaphor "a person" (M=.96), $p = .001$, $\eta_p^2 = .110$.

There was an additional effect on the perceived perceptual abilities of robots with Welch's F(2, 96.60) = 30.81, $p = .001$, $\eta_p^2 = .120$. Robots described as babies were assumed to have lower expected perceptual capabilities (M=-.45) than robots described as toddlers (M=.67), $p = .001$, $\eta_p^2 = .118$. Robots associated with the toddler metaphor were, in turn, perceived as having lower perceptual abilities than robots described as persons (M=1.23), $p = .001$, $\eta_p^2 = .110$. Additionally, robots associated with the baby metaphor had significantly lower perceived perceptual abilities than robots associated with a person metaphor, $p = .001$, $\eta_p^2 = .212$.

7.4 Correlations of Measures

We show the correlations between all measures in Figure 11. While many correlations in the large collected dataset are significant, the coefficients of correlation are relatively small. We consider values of Pearson's $r > 0.5$ to be of practical importance for reporting and discussion, representing a moderate correlation in similar psychological contexts [1]. The key findings are described in detail, with all reported correlations being significant with $p < .001$ after correcting for 110 pairwise comparisons of the 11 measures. The following are the main trends in the results.

7.4.1 Identity closeness is correlated with positive social perceptions. Identity closeness had a moderately strong correlation with warmth (r(163)=.55) and competence (r(163)=.52). Interestingly, the correlation with the discomfort construct was small. This is possibly related to the idea that discomfort is a construct that uniquely applies to robots, while warmth and competence apply to people as well as robots [16]. Furthermore, the identity closeness of the user with the robot was strongly correlated with the robot's likeability (r(163)=.63) and its expected social role (r(163)=.53).

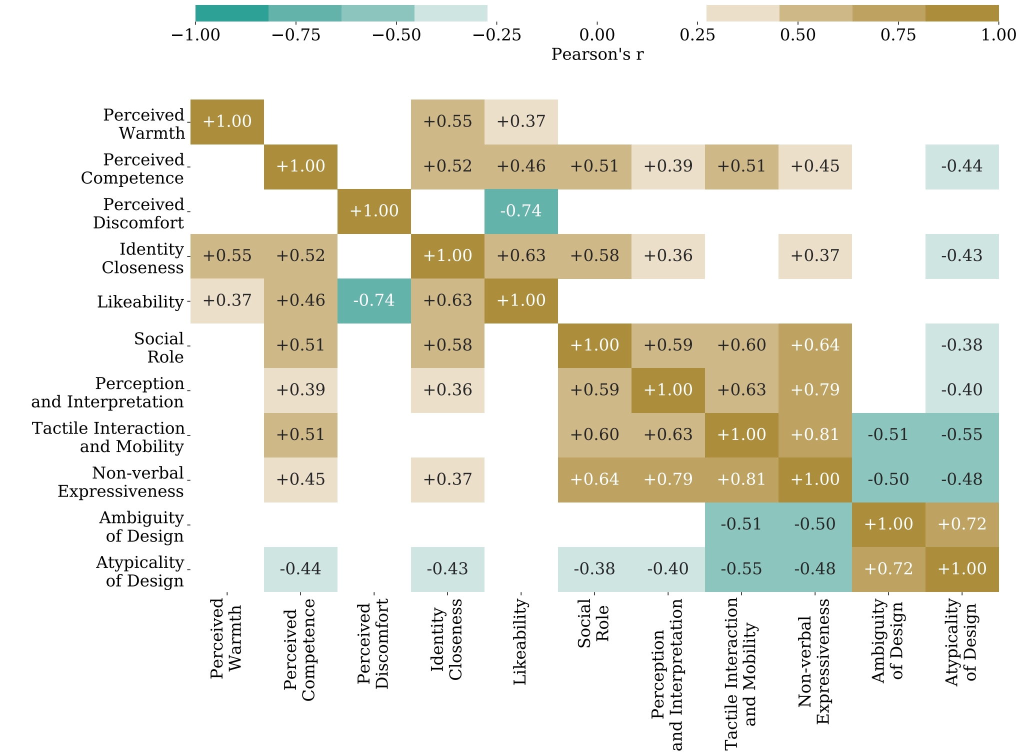


Fig. 11. Correlations between attributes collected through surveys. All shown correlations are significant, with $p < 0.001$. Pearson's r value is shown in each square. The first six items are from the survey of social perceptions, and the next five are from the survey of functional perceptions.

For the socially interactive robots we tested, we found that the closer the raters identified with the robot, the more positively they viewed the robot and the more likely the were to view the robot as a peer or superior.

7.4.2 Likeability is not significantly correlated with perceived functionality. Likeability responses were moderately correlated with warmth (r(163)=.37) and competence (r(163)=.46), strongly anticorrelated with discomfort (r(163)=-.74), and strongly correlated with identity closeness (r(163)=.63). However, the reported likeability was not strongly correlated with any measures of perceived functional ability. In general, as raters felt more socially close to the robot and the design of the robot was less discomforting, users reported liking the robot more. Changes in perceived physical capabilities of the robot, however, did not correspond with a discernible change in how much users liked the robot.

7.4.3 A robot's role is correlated with its functionality. The expected social role of the robot was strongly correlated with perception and interpretation (r(163)=.59), non-verbal communication (r(163)=.64), and tactile interaction and mobility (r(163)=.60). The social role was also correlated with the competence (r(163)=.51) and social identity (r(163)=.63). As functional abilities increased for robots in our dataset, raters were more likely to view them as peers or superiors.

7.4.4 The perceived functionalities of robots are entangled. For all pairwise comparisons of the functional constructs from the modified EmCorp-Scale, we observed correlation values larger than r(163)=.63. Thus, increases of one construct from this scale were associated with increases in the other two constructs of the scale for the set of socially interactive robots we tested. This implies that robots that appear more capable of moving through space *additionally* elicit higher expectations of perceptual and interpretive abilities, as well as higher expectations of non-verbal expressivity than robots that do not appear as capable of moving through space. These three functional constructs can be interpreted together as a generic measure of a robot's holistic capability to interact with other agents and the world.

8 GENERAL DISCUSSION

The analysis of the large dataset resulting from the three studies conducted in this work demonstrates many nuances of the design space of socially interactive robot embodiments. The results include multiple insights that inform design processes for various robot use-cases. Specifically, this work suggests how measuring social and functional attributes of embodiment via design metaphors can be used to estimate and evaluate designs of robot embodiments. Toward advancing HRI research, we identified several use-cases for this dataset specifically aimed at informing study designers and robot designers. These tools inform the development of future socially interactive robots and effective human-centered HRI.

8.1 Implications for Study Designers

For study designers in HRI, this dataset offers a data-driven methodology for understanding how the embodiment of a robot may affect study outcomes. This dataset is useful for study designers by enabling three main interactions: robot selection, task selection, and contextualizing results.

A common stage of the study design process in research labs is to choose a robot to investigate how its interaction with humans affect outcome measures of importance. Due to the high cost of most robots, research labs tend to have only a few different embodiments that can be potential candidates for study. By using this dataset, researchers can make informed decisions in selecting a robot that aligns most closely with the specific requirements of the task and research topic of interest. Because our dataset provides multiple measures, researchers can choose those that are most relevant to their needs and interests. For example, researchers exploring trust in delivery robots may investigate functional expectations of the robots in the dataset, whereas research exploring conversational dynamics may be more interested in the social measures of our dataset. By selecting robots that align with their use-case of interest, researchers and designers will not require as much time to learn about a robot. This is especially critical for short-term studies, where first impressions of a robot are important for the success of the intervention.

Interaction design is a key step in the HRI study design process. In that process, it is critical to consider user expectations of robot performance in the specific interaction. Our dataset and study results inform researchers about baseline social and functional expectations of robots. Using these expectations, researchers can modify their study designs to more closely align with the specific expectations about their robots. For example, if the specific robot is not expected to have a high degree of mobility, the interaction can be modified and simplified to not require the robot to move. Additionally, our dataset and study results can be used to provide directions for slight but principled modifications to robots in cases where changes to study design are not possible but where a robot's design can be modified. For example, a robot's mouth can be removed if the robot is not meant to be perceived as friendly within the study's context, since we found that mouths increased perceived warmth (see Table 2).

Our dataset and results can also provide insights that inform interpretation of HRI study results. Findings from one specific embodiment may be more easily replicated in similar embodiments, which can be explored through the t-SNE visualization we developed. More generally, our work can validate if study findings may be less tied to the specific low-level design choices, and more influenced by the design metaphors of the embodiment, or even more broadly by overarching social and functional expectations of the robot. Thus, follow up studies could explore using this dataset as a tool for comparing different robot embodiments. This work provides a way forward to more effectively evaluate theories in HRI in the context of robot embodiment.

8.2 Implications for Robot Designers

For robot designers focusing on HRI, this dataset and study results offer a data-driven methodology for determining robot designs given a set of intended task/use-case constraints. The dataset is particularly useful to research through design [112], a paradigm that highlights the importance of the exploratory implementation of systems to solve real-world scenarios. Such specific implementations are called *design artifacts* and their production creates knowledge of different design patterns, design processes, and other forms of design knowledge. Our dataset and findings can be interpreted as a form of intermediate-level design knowledge [44, 65], which can be used to guide the creation of individual design artifacts akin to other forms of intermediate level design works, such as annotated design portfolios [35], design guidelines [63], and design patterns [60]. To effectively use this type of intermediate-level knowledge, we describe the following three potential interactions with this dataset inspired by data-driven visualization design [72]: browse, discover, and compare.

The *browse* interaction with our dataset describes looking through the extant designs that we have collected. Designers can explore entries to discover what robots exist, and what expectations they elicit. By looking through the data without specific aims, designers can identify holes in the design space as well as over-saturated regions of the space. Novel designs can then be added to the dataset to expand the space as desired. This is especially useful in iterative design processes where a robot's physical structure becomes increasingly stable as iterations progress [24]. Searching can be done at different levels of granularity; initially, searching can happen across social and functional expectations, then more specifically over design metaphors, and finally over low-level design features. As the design becomes more concrete, the search becomes more localized in the design space.

The *discover* interaction with this dataset describes the process of finding similar robots for comparison, and modifying a design to more closely match or move away from those alternate designs. By locating robots that are nearby in design space, designers can explore potential design metaphors to see how closely they align with the intended use-case. For example, a designer may examine the trend in functional expectation of making a more literal human-like robot, finding that this increases the expected functionality of the robot. Designers can then use this information to see how the design may need to change to reinforce desirable traits or minimize unwanted traits for a given design context. Our dataset also allows this interaction to happen across differing levels of granularity: at the expectation level by selecting robots of similar expectations, at the metaphor level by examining robots with similar design metaphors, and at the feature level by exploring robots nearby in feature space.

The *compare* interaction can be used by designers to evaluate the effect of specific design decisions. If a designer needs to make a decision on whether or not to include a mouth on the robot, they can examine how the presence or absence of a mouth may change the social perception of the robot while also considering other constraints of the robot, such as cost and space. Because our dataset includes an image of each whole robot, designers can use those images to define new binary features over the set of extant robots to determine how the inclusion of a previously unexplored feature may affect social and functional expectations. Similar interactions can decide between broad metaphor categories (e.g., anthropomorphic, zoomorphic, or mechanical) or individual design metaphors to numerically evaluate differences in potential embodiment design choices.

Interactions with the dataset can be composed into a full design process to facilitate informed design decisions. In early iterations, designers may engage with the existing robots at a high level, by *browsing* through the design space by exploring the t-SNE visualization to see different clusters of visually similar robots and their social and functional expectations. Initial designs for a specific context may start to take specific functional and social constraints for that context, which may be informed by exploratory prototyping. These social and functional constraints can then be evaluated over the robots within the dataset, and specific mechanisms can be explored to address these constraints to *explore* potential modifications the design. In further iterations, as the robot takes a more concrete form, design metaphors may be applied to *compare* how cosmetic changes can be made to more effectively communicate robot capabilities.

8.3 Limitations and Future Work

A key limitation of this work is the use of images to convey representations of inherently 3D robot embodiments. Multiple views of the robots were shown in an attempt to mitigate this, but a 2D screen cannot fully reconstruct the impression of real-world embodiments. This study design can be readily applied to other representations of robot embodiments to expand the use of the findings. Future work may explore alternate forms of presentation, such as 3D renderings, videos, virtual reality, and augmented reality. Applying methods introduced in this paper allows designers and researchers to flexibly trade off realism and data collection costs.

Additionally, social and physical contexts are not considered in this work. Contextual information can have great impact on how a user expects a robot to behave (e.g., [5, 75]). We additionally restricted our participants to the United States; since many metaphors and perceptions may be culturally situated, it is not certain how the results generalize to other cultures. This issue could be addressed by following a similar design in future examinations of design metaphors across different cultural contexts.

There are several limitations associated with recruiting though Amazon Mechanical Turk. Although we took several precautions to ensure high data quality, it is difficult to ensure realistic responses to questions that require study participant effort to answer. Some participants may aim to answer questions as quickly as possible and consequently may only superficially address more introspective questions. Online surveys also do not allow users to interact with the robots in person. We cannot, therefore, infer how user expectation may be altered through real-world interactions and over time. These findings are best viewed as *priors* on robot expectation before interaction occurs.

In-the-wild robot deployments will best address concerns about the task, context, and population of the MUFaSAA dataset. They may also utilize different robot representation methods (especially augmented and virtual reality) to validate expectations of embodiments in the dataset. Deployments in locations such as museums, cafes, and school campuses solidify social and functional expectations that allow users to make better judgements of robot performance. Finally, interaction in physically co-located contexts facilitate the verification of personhood and elicit more genuine responses to robot behavior.

This work utilizes the ontology of anthropomorphic, zoomorphic, and mechanical metaphors as a means of analysis. While that was a useful classification for our analysis, as the space of robot design expands and other metaphors are used in design processes, that ontology can be restructured to overcome its limitations (for example to capture robots that look like plants). Importantly, the utility of design metaphors as a tool for understanding user expectations does not directly depend on this ontology, allowing for the evolution of classification systems.

9 CONCLUSION

This work provides a framework for understanding and informing robot design to set realistic expectations through the use of design metaphors. We contributed a methodology for determining the design metaphors and collecting social and functional expectations of a given robot embodiment. We set up the MUFaSAA dataset of 165 socially interactive robot embodiments, and collected a rich dataset of participant responses about social, functional, and other relevant expectations for of those embodiments. We also developed an open-source visualization of the dataset and study findings toward broadly facilitating HRI research and human-centered robot and interaction design. The analysis results offer general guidelines for designing socially interactive robots for different contexts and ways in which user expectation of functional and social capabilities are impacted by robot embodiments. They also point to new and fruitful research directions in the context of robot embodiment.

10 STATEMENTS

10.1 Ethical Impact Statement

The collection of the MUFaSAA dataset was approved by the IRB at the University of Southern California under the protocol UP-18-00510. We discussed ethical implications of administering this survey in Section 3.3. We note that the participants are recruited from all areas of the US though MTurk. This results in a dataset that is skewed toward young white college-educated Americans and the metaphors used to describe robots may not be reflective of metaphors used by other cultures. We provide full demographic information in Appendix A.1, which can be used to assess the similarity of this dataset to the target audience in design contexts. Additionally, because we collected free-response answers, there may responses that are not factually correct if users have incorrectly remembered certain references. Thus this dataset should not be viewed as an objective truth, but rather how perceptions of robots are established naturalistically.

By introducing the MUFaSAA dataset, we hope to assist in the design process of socially interactive robots by providing insights into how users expect these robots to act. While this is useful to reduce the number of iterations in design cycles, this also has the potential to design robots that reflect potentially negative stereotypes. Thus we urge designers to consider how the robot's design and the contexts that robots are deployed in may be affecting harmful stereotypes in society at large.

10.2 Citation Diversity Statement

Many fields have seen gender-based biases in citations of other works [14, 26, 28, 67]. In an effort to make the citation practices of our work transparent, we present the gender proportions of our works cited. We report the genders of the first and last authors as determined by names, images, and public research websites. While this practice may under-report identities that are not directly disclosed through these modalities, it offers some insight into the makeup of our works cited at the time of writing this manuscript. Excluding self-citations, our manuscript citations are comprised of 23.2% woman/woman papers, 22.2% woman/man papers, 21.3% man/woman papers, and 33.3% man/man papers. We look forward to future work that can better support equitable citation practices in science.

REFERENCES

[1] Haldun Akoglu. 2018. User's guide to correlation coefficients. *Turkish journal of emergency medicine* 18, 3 (2018), 91–93.

[2] Jüri Allik. 2014. A mixed-binomial model for Likert-type personality measures. *Frontiers in psychology* 5 (2014), 371.

[3] Steph M. Anderson. 2020. Gender Matters: The Perceived Role of Gender Expression in Discrimination Against Cisgender and Transgender LGBQ Individuals. *Psychology of Women Quarterly* 44, 3 (2020), 323–341. https://doi.org/10.1177/0361684320929354 arXiv:https://doi.org/10.1177/0361684320929354

[4] Sean Andrist, Bilge Mutlu, and Adriana Tapus. 2015. Look like me: matching robot personality via gaze to increase motivation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 3603–3612.

[5] Siddhartha Banerjee, Andrew Silva, Karen Feigh, and Sonia Chernova. 2018. Effects of interruptibility-aware robot behavior. *arXiv preprint arXiv:1804.06383* (2018).

[6] Jaime Banks. 2020. Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI* 7 (2020), 62.

[7] Christoph Bartneck, Takayuki Kanda, Omar Mubin, and Abdullah Al Mahmud. 2007. The perception of animacy and intelligence based on a robot's embodiment. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*. IEEE, 300–305.

[8] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81.

[9] Sandra L Bem. 1981. Bem sex role inventory. *Journal of Personality and Social Psychology* (1981).

[10] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. 2017. Shape it–the influence of robot body shape on gender perception in robots. In *International Conference on Social Robotics*. Springer, 75–84.

[11] Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. 2013. Interpreting and executing recipes with a cooking robot. In *Experimental Robotics*. Springer, 481–495.

[12] Guido Bugmann and Simon N Copleston. 2011. What can a personal robot do for you?. In *Conference Towards Autonomous Robotic Systems*. Springer, 360–371.

[13] Donald T Campbell. 1958. Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Behavioral science* 3, 1 (1958), 14.

[14] Neven Caplar, Sandro Tacchella, and Simon Birrer. 2017. Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy* 1, 6 (2017), 1–5.

[15] Julie Carpenter, Joan M Davis, Norah Erwin-Stewart, Tiffany R Lee, John D Bransford, and Nancy Vye. 2009. Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics* 1, 3 (2009), 261–265.

[16] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[17] John M Carroll, Robert L Mack, and Wendy A Kellogg. 1988. Interface metaphors and user interface design. In *Handbook of human-computer interaction*. Elsevier, 67–85.

[18] Catherine Caudwell, Cherie Lacey, and Eduardo B Sandoval. 2019. The (Ir) relevance of Robot Cuteness: An Exploratory Study of Emotionally Durable Robot Design. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 64–72.

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

[20] Meia Chita-Tegmark, Theresa Law, Nicholas Rabb, and Matthias Scheutz. 2021. Can you trust your trust measure?. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 92–100.

[21] Nazli Cila. 2013. Metaphors we design by: The use of metaphors in product design. (2013).

[22] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[23] Eric Deng, Bilge Mutlu, and Maja Mataric. 2019. Embodiment in socially interactive robots. *arXiv preprint arXiv:1912.00312* (2019).

[24] Eric C Deng, Bilge Mutlu, and Maja J Matarić. 2018. Formalizing the design space and product development cycle for socially interactive robots. In *Workshop on Social Robots in the Wild at the 2018 ACM Conference on Human-Robot Interaction (HRI)*.

[25] Nathaniel Dennler, Eura Shin, Maja Matarić, and Stefanos Nikolaidis. 2021. Design and Evaluation of a Hair Combing System Using a General-Purpose Robotic Arm. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 3739–3746.

[26] Michelle L Dion, Jane Lawrence Sumner, and Sara McLaughlin Mitchell. 2018. Gendered citation patterns across political science and social science methodology fields. *Political analysis* 26, 3 (2018), 312–327.

[27] Brian R Duffy. 2003. Anthropomorphism and the social robot. *Robotics and autonomous systems* 42, 3-4 (2003), 177–190.

[28] Jordan D Dworkin, Kristin A Linn, Erin G Teich, Perry Zurn, Russell T Shinohara, and Danielle S Bassett. 2020. The extent and drivers of gender imbalance in neuroscience reference lists. *Nature neuroscience* 23, 8 (2020), 918–926.

[29] Friederike Eyssel and Frank Hegel. 2012. (s) he's got the look: Gender stereotyping of robots 1. *Journal of Applied Social Psychology* 42, 9 (2012), 2213–2230.

[30] Friederike Eyssel and Dieta Kuchenbrandt. 2012. Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology* 51, 4 (2012), 724–731.

[31] Friederike Eyssel, Dieta Kuchenbrandt, and Simon Bobinger. 2011. Effects of anticipated human-robot interaction and predictability of robot behavior on perceptions of anthropomorphism. In *Proceedings of the 6th international conference on Human-robot interaction.* 61–68.

[32] Susan T Fiske, Amy JC Cuddy, and Peter Glick. 2007. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences* 11, 2 (2007), 77–83.

[33] Terrence Fong, Illah Nourbakhsh, and Kerstin Dautenhahn. 2003. A survey of socially interactive robots. *Robotics and autonomous systems* 42, 3-4 (2003), 143–166.

[34] Marlena R Fraune, Selma Šabanović, and Eliot R Smith. 2017. Teammates first: Favoring ingroup robots over outgroup humans. In *2017 26th IEEE international symposium on robot and human interactive communication (RO-MAN).* IEEE, 1432–1437.

[35] Bill Gaver and John Bowers. 2012. Annotated portfolios. *interactions* 19, 4 (2012), 40–49.

[36] G Gerbner. [n.d.]. (1976, Mar 20). Living with television: The violence profile. *Journal of Communication* ([n. d.]), 172–199.

[37] James J Gibson. 1977. The theory of affordances. *Hilldale, USA* 1, 2 (1977), 67–82.

[38] Lorenz Goette, David Huffman, and Stephan Meier. 2006. The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review* 96, 2 (2006), 212–216.

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

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

[41] Frank Hegel, Soren Krach, Tilo Kircher, Britta Wrede, and Gerhard Sagerer. 2008. Understanding social robots: A user study on anthropomorphism. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication.* 574–579. https://doi.org/10.1109/ROMAN.2008.4600728

[42] Laura Hoffmann, Nikolai Bock, and Astrid M Rosenthal vd Pütten. 2018. The Peculiarities of Robot Embodiment (EmCorp-Scale) Development, Validation and Initial Test of the Embodiment and Corporeality of Artificial Agents Scale. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 370–378.

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

[44] Kristina Höök and Jonas Löwgren. 2012. Strong concepts: Intermediate-level knowledge in interaction design research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 3 (2012), 1–18.

[45] Aike C Horstmann and Nicole C Krämer. 2019. Great expectations? Relation of previous experiences with social robots in real life or in the media and expectancies based on qualitative and quantitative assessment. *Frontiers in psychology* 10 (2019), 939.

[46] Advait Jain and Charles C Kemp. 2010. EL-E: an assistive mobile manipulator that autonomously fetches objects from flat surfaces. *Autonomous Robots* 28, 1 (2010), 45–64.

[47] Alex Juarez, Christoph Bartneck, and Loe Feijs. 2011. Using semantic technologies to describe robotic embodiments. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI).* 425–432. https://doi.org/10.1145/1957656.1957812

[48] Heekyoung Jung, Heather Wiltse, Mikael Wiberg, and Erik Stolterman. 2017. Metaphors, materialities, and affordances: Hybrid morphologies in the design of interactive artifacts. *Design Studies* 53 (2017), 24–46.

[49] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the design space of rendered robot faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction.* 96–104.

[50] Ariel Kapusta, Zackory Erickson, Henry M Clever, Wenhao Yu, C Karen Liu, Greg Turk, and Charles C Kemp. 2019. Personalized collaborative plans for robot-assisted dressing via optimization and simulation. *Autonomous Robots* 43,

8 (2019), 2183–2207.

[51] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey Hancock, and Michael Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *arXiv preprint arXiv:2008.02311* (2020).

[52] Jingoog Kim and Mary Lou Maher. 2020. Conceptual Metaphors for Designing Smart Environments: Device, Robot, and Friend. *Frontiers in Psychology* 11 (2020), 198.

[53] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.

[54] Dimosthenis Kontogiorgos, Sanne van Waveren, Olle Wallberg, Andre Pereira, Iolanda Leite, and Joakim Gustafson. 2020. Embodiment Effects in Interactions with Failing Robots. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[55] Hatice Köse, Pınar Uluer, Neziha Akalın, Rabia Yorgancı, Ahmet Özkul, and Gökhan Ince. 2015. The effect of embodiment in sign language tutoring with assistive humanoid robots. *International Journal of Social Robotics* 7, 4 (2015), 537–548.

[56] Dieta Kuchenbrandt, Friederike Eyssel, Simon Bobinger, and Maria Neufeld. 2013. When a robot's group membership matters. *International Journal of Social Robotics* 5, 3 (2013), 409–417.

[57] Dieta Kuchenbrandt, Markus Häring, Jessica Eichberg, Friederike Eyssel, and Elisabeth André. 2014. Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *International Journal of Social Robotics* 6, 3 (2014), 417–427.

[58] Miron B Kursa, Witold R Rudnicki, et al. 2010. Feature selection with the Boruta package. *J Stat Softw* 36, 11 (2010), 1–13.

[59] Minae Kwon, Malte F Jung, and Ross A Knepper. 2016. Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 463–464.

[60] James A Landay, Jason I Hong, et al. 2003. *The design of sites: patterns, principles, and processes for crafting a customer-centered Web experience*. Addison-Wesley Professional.

[61] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 203–210.

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

[63] William Lidwell, Kritina Holden, and Jill Butler. 2010. *Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design*. Rockport Pub.

[64] Diana Löffler, Judith Dörrenbächer, and Marc Hassenzahl. 2020. The Uncanny Valley Effect in Zoomorphic Robots: The U-Shaped Relation Between Animal Likeness and Likeability. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 261–270.

[65] Maria Luce Lupetti, Cristina Zaga, and Nazli Cila. 2021. Designerly ways of knowing in HRI: Broadening the scope of design-oriented HRI through the concept of intermediate-level knowledge. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 389–398.

[66] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[67] Daniel Maliniak, Ryan Powers, and Barbara F Walter. 2013. The gender citation gap in international relations. *International Organization* 67, 4 (2013), 889–922.

[68] Maya B Mathur and David B Reichling. 2016. Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition* 146 (2016), 22–32.

[69] Laura Menchetti, Silvia Calipari, Gabriella Guelfi, Alice Catanzaro, and Silvana Diverio. 2018. My dog is not my cat: Owner perception of the personalities of dogs and cats living in the same household. *Animals* 8, 6 (2018), 80.

[70] Youngme Moon and Clifford Nass. 1996. How "real" are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication research* 23, 6 (1996), 651–674.

[71] Jonathan Mumm and Bilge Mutlu. 2011. Human-robot proxemics: physical and psychological distancing in human-robot interaction. In *Proceedings of the 6th international conference on Human-robot interaction*. 331–338.

[72] Tamara Munzner. 2014. *Visualization analysis and design*. CRC press.

[73] Brad Myers. 1994. Challenges of HCI design and implementation. *interactions* 1, 1 (1994), 73–83.

[74] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[75] Aastha Nigam and Laurel D Riek. 2015. Social context perception for mobile robots. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 3621–3627.

[76] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. 2017. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research* 36, 5-7 (2017), 618–634.

[77] Anastasia K. Ostrowski, Vasiliki Zygouras, Hae Won Park, and Cynthia Breazeal. 2021. Small Group Interactions with Voice-User Interfaces: Exploring Social Embodiment, Rapport, and Engagement. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (Boulder, CO, USA) *(HRI '21)*. Association for Computing Machinery, New York, NY, USA, 322–331. https://doi.org/10.1145/3434073.3444655

[78] Giulia Perugia, Stefano Guidi, Margherita Bicchi, and Oronzo Parlangeli. 2022. The Shape of Our Bias: Perceived Age and Gender in the Humanoid Robots of the ABOT Database. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 110–119.

[79] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. 2018. What is human-like? Decomposing robots' human-like appearance using the anthropomorphic roBOT (ABOT) database. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 105–113.

[80] Aaron Powers, Adam DI Kramer, Shirlene Lim, Jean Kuo, Sau-lai Lee, and Sara Kiesler. 2005. Eliciting information from people with a gendered humanoid robot. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 158–163.

[81] Irene Rae, Leila Takayama, and Bilge Mutlu. 2013. The influence of height in robot-mediated communication. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[82] Céline Ray, Francesco Mondada, and Roland Siegwart. 2008. What do people expect from robots?. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3816–3821.

[83] Stephen Reysen, Iva Katzarska-Miller, Sundé M Nesbit, and Lindsey Pierce. 2013. Further validation of a single-item measure of social identification. *European Journal of Social Psychology* 43, 6 (2013), 463–470.

[84] Marjorie Rhodes and Lisa Chalik. 2013. Social categories as markers of intrinsic interpersonal obligations. *Psychological science* 24, 6 (2013), 999–1006.

[85] Danielle Rifinski, Hadas Erel, Adi Feiner, Guy Hoffman, and Oren Zuckerman. 2020. Human-human-robot interaction: robotic object's responsive gestures improve interpersonal evaluation in human interaction. *Human–Computer Interaction* (2020), 1–27.

[86] Steven V Rouse. 2015. A reliability analysis of Mechanical Turk data. *Computers in Human Behavior* 43 (2015), 304–307.

[87] Matthew Rueben, Shirley A Elprama, Dimitrios Chrysostomou, and An Jacobs. 2020. Introduction to (Re) Using Questionnaires in Human-Robot Interaction Research. In *Human-Robot Interaction*. Springer, 125–144.

[88] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science* 2 (apr 2016), e55. https://doi.org/10.7717/peerj-cs.55

[89] Lindsay Sanneman and Julie A Shah. 2020. Trust considerations for explainable robots: A human factors perspective. *arXiv preprint arXiv:2005.05940* (2020).

[90] Catherine E Sembroski, Marlena R Fraune, and Selma Šabanović. 2017. He said, she said, it said: Effects of robot group membership and human authority on people's willingness to follow their instructions. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 56–61.

[91] IEEE Spectrum. 2018. All Robots. https://robots.ieee.org/robots/

[92] Megan K Strait, Victoria A Floerke, Wendy Ju, Keith Maddox, Jessica D Remedios, Malte F Jung, and Heather L Urry. 2017. Understanding the uncanny: both atypical features and category ambiguity provoke aversion toward humanlike robots. *Frontiers in psychology* 8 (2017), 1366.

[93] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).

[94] S Shyam Sundar, T Franklin Waddell, and Eun Hwa Jung. 2016. The Hollywood robot syndrome media effects on older adults' attitudes toward robots and adoption intentions. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 343–350.

[95] Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social science information* 13, 2 (1974), 65–93.

[96] Kyle A Thomas and Scott Clifford. 2017. Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior* 77 (2017), 184–197.

[97] Gabriele Trovato, Cesar Lucho, and Renato Paredes. 2018. She's electric—the influence of body proportions on perceived gender of robots across cultures. *Robotics* 7, 3 (2018), 50.

[98] John C Turner, Michael A Hogg, Penelope J Oakes, Stephen D Reicher, and Margaret S Wetherell. 1987. *Rediscovering the social group: A self-categorization theory.* basil Blackwell.

[99] Daniel Ullman, Salomi Aladia, and Bertram F Malle. 2021. Challenges and opportunities for replication science in HRI: a case study in human-robot trust. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 110–118.

[100] Raphael Vallat. 2018. Pingouin: statistics in Python. *Journal of Open Source Software* 3, 31 (2018), 1026.

[101] J Ventre-Dominey, G Gibert, M Bosse-Platiere, A Farnè, PF Dominey, and F Pavani. 2019. Embodiment into a robot increases its acceptability. *Scientific reports* 9, 1 (2019), 1–10.

[102] Stephen Voida, Elizabeth D Mynatt, and W Keith Edwards. 2008. Re-framing the desktop interface around the activities of knowledge work. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. 211–220.

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

[104] Michael L Walters, Kerstin Dautenhahn, Rene te Boekhorst, Kheng Lee Koay, and Sarah N Woods. 2007. Exploring the Design Space of Robot Appearance and Behavior in an Attention-SeekingLiving Room'Scenario for a Robot Companion. In *2007 IEEE Symposium on Artificial Life*. IEEE, 341–347.

[105] Xijing Wang and Eva G Krumhuber. 2018. Mind perception of robots varies with their economic versus social function. *Frontiers in psychology* 9 (2018), 1230.

[106] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. 2022. FabricFlowNet: Bimanual Cloth Manipulation with a Flow-based Policy. In *Conference on Robot Learning*. PMLR, 192–202.

[107] Katie Winkle, Gaspar Isaac Melsión, Donald McMillan, and Iolanda Leite. 2021. Boosting Robot Credibility and Challenging Gender Norms in Responding to Abusive Behaviour: A Case for Feminist Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 29–37.

[108] Holly A Yanco and Jill Drury. 2004. Classifying human-robot interaction: an updated taxonomy. In *2004 IEEE international conference on systems, man and cybernetics (IEEE Cat. No. 04CH37583)*, Vol. 3. IEEE, 2841–2846.

[109] JD Zamfirescu-Pereira, David Sirkin, David Goedicke, Ray LC, Natalie Friedman, Ilan Mandel, Nikolas Martelaro, and Wendy Ju. 2021. Fake it to make it: Exploratory prototyping in HRI. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 19–28.

[110] Debora Zanatto, Massimiliano Patacchiola, Jeremy Goslin, and Angelo Cangelosi. 2016. Priming anthropomorphism: Can the credibility of humanlike robots be transferred to non-humanlike robots?. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 543–544.

[111] Hejia Zhang and Stefanos Nikolaidis. 2019. Robot learning and execution of collaborative manipulation plans from YouTube cooking videos. *arXiv preprint arXiv:1911.10686* (2019).

[112] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through Design as a Method for Interaction Design Research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. https://doi.org/10.1145/1240624.1240704

[113] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics* 7, 3 (2015), 347–360.

A SURVEY INFORMATION

A.1 Demographic Information

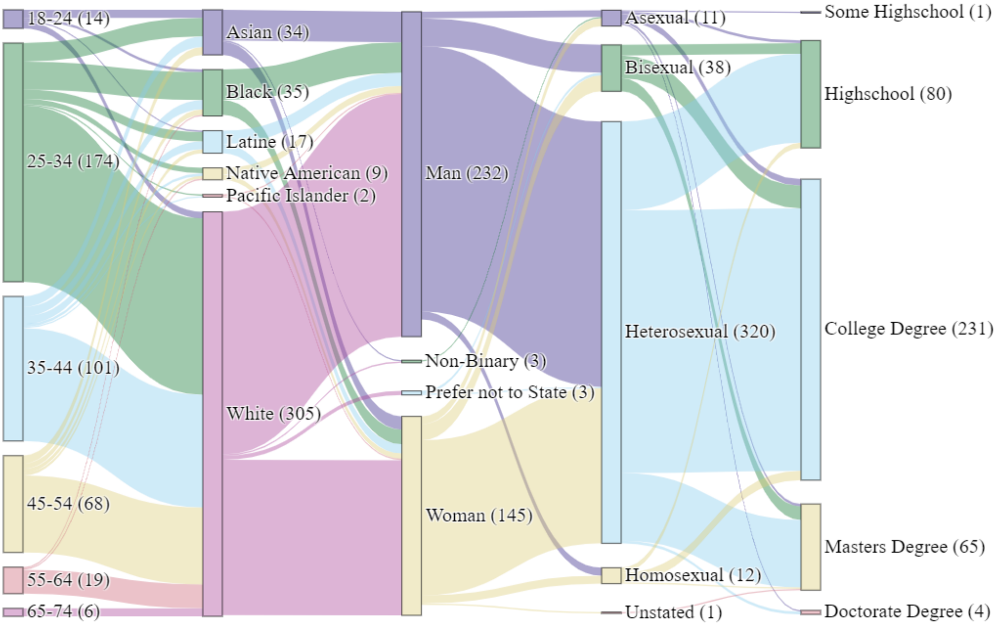


Fig. 12. The intersection of participant identities from the design metaphor survey (Study 1).

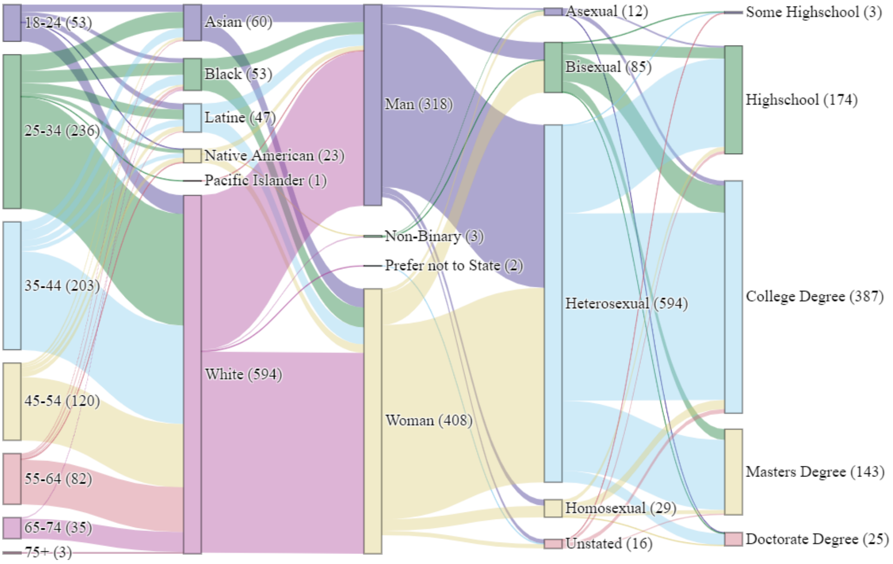


Fig. 13. The intersection of participant identities from the social expectation survey (Study 2).

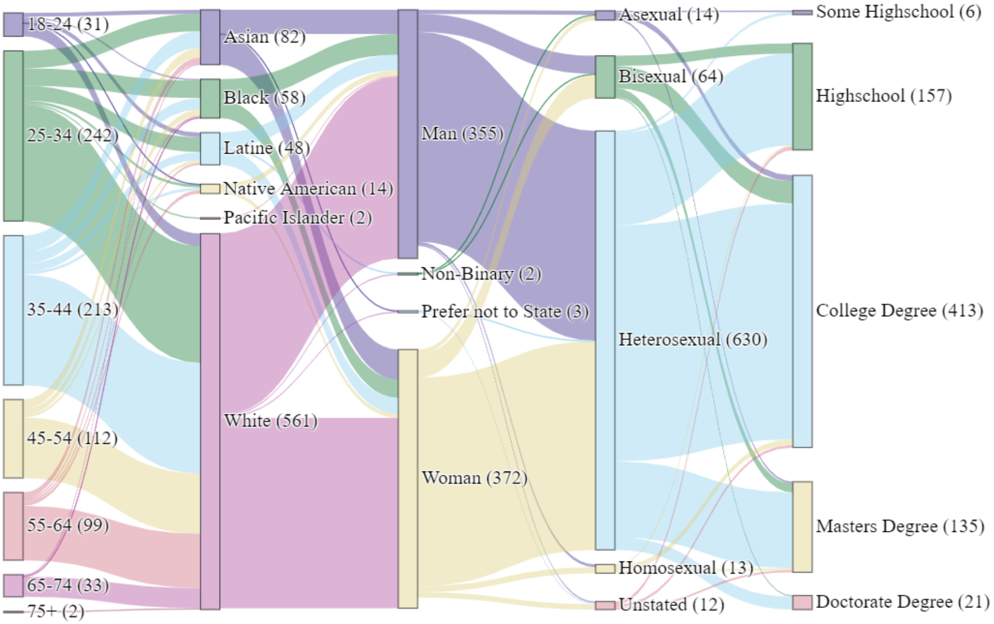
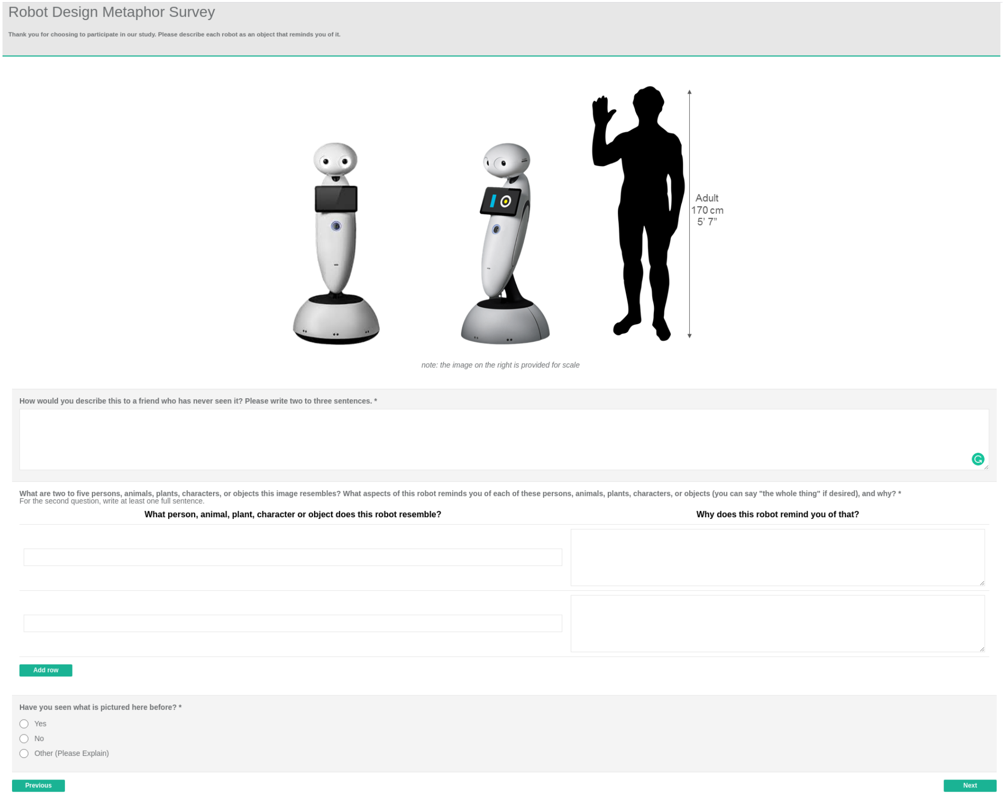


Fig. 14. The intersection of participant identities from the functional expectation survey (Study 3).

A.2 Design Metaphor Survey Questions and Interface



Robot Design Metaphor Survey

Thank you for choosing to participate in our study. Please describe each robot as an object that reminds you of it.

note: the image on the right is provided for scale

How would you describe this to a friend who has never seen it? Please write two to three sentences. *

What are two to five persons, animals, plants, characters, or objects this image resembles? What aspects of this robot reminds you of each of these persons, animals, plants, characters, or objects (you can say "the whole thing" if desired), and why? *
For the second question, write at least one full sentence.

What person, animal, plant, character or object does this robot resemble?	Why does this robot remind you of that?

Add row

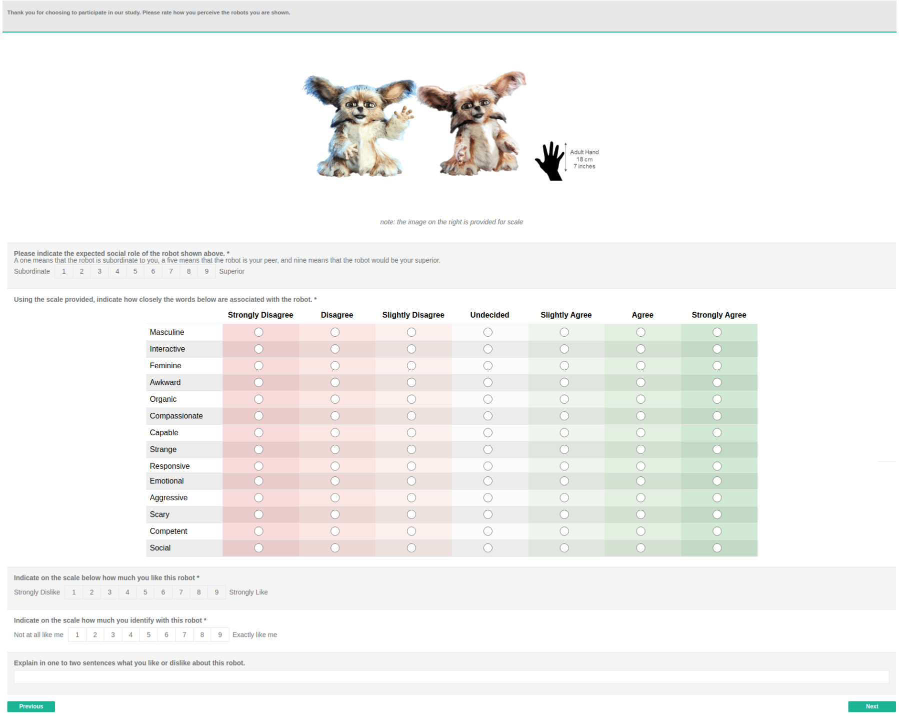
Have you seen what is pictured here before? *

○ Yes
○ No
○ Other (Please Explain)

Previous Next

Fig. 15. The interface and questions participants in the design metaphor survey saw (Study 1).

A.3 Social Expectation Survey Questions and Interface



Thank you for choosing to participate in our study. Please rate how you perceive the robots you are shown.

note: the image on the right is provided for scale

Please indicate the expected social role of the robot shown above. *
A one means that the robot is subordinate to you, a five means that the robot is your peer, and nine means that the robot would be your superior.

Subordinate 1 2 3 4 5 6 7 8 9 Superior

Using the scale provided, indicate how closely the words below are associated with the robot. *

	Strongly Disagree	Disagree	Slightly Disagree	Undecided	Slightly Agree	Agree	Strongly Agree
Masculine	○	○	○	○	○	○	○
Interactive	○	○	○	○	○	○	○
Feminine	○	○	○	○	○	○	○
Awkward	○	○	○	○	○	○	○
Organic	○	○	○	○	○	○	○
Compassionate	○	○	○	○	○	○	○
Capable	○	○	○	○	○	○	○
Strange	○	○	○	○	○	○	○
Responsive	○	○	○	○	○	○	○
Emotional	○	○	○	○	○	○	○
Aggressive	○	○	○	○	○	○	○
Scary	○	○	○	○	○	○	○
Competent	○	○	○	○	○	○	○
Social	○	○	○	○	○	○	○

Indicate on the scale below how much you like this robot *

Strongly Dislike 1 2 3 4 5 6 7 8 9 Strongly Like

Indicate on the scale how much you identify with this robot *

Not at all like me 1 2 3 4 5 6 7 8 9 Exactly like me

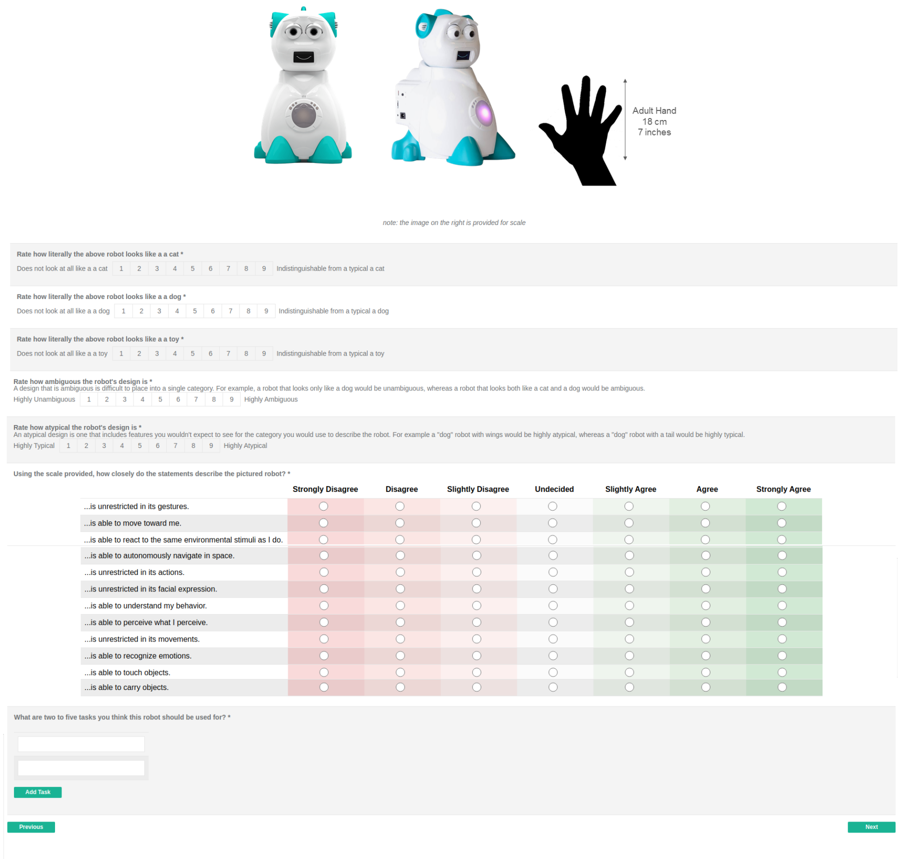
Explain in one to two sentences what you like or dislike about this robot.

Fig. 16. The interface and questions participants in the social expectation survey saw (Study 2). The order of the questions in the Likert section were randomized both during pilot studies and during the full deployment.

Table 3. This table describes the assignment of the Likert items to the specific social constructs they measured.

Construct	Items
Warmth	"Social", "Organic", "Compassionate", and "Emotional"
Competence	"Capable", "Responsive", "Interactive", and "Competent"
Discomfort	"Scary", "Strange", "Awkward", and "Aggressive"
Femininity	"Feminine"
Masculinity	"Masculine"

A.4 Functional Expectation Survey Questions and Interface



note: the image on the right is provided for scale

Rate how literally the above robot looks like a a cat *

Does not look at all like a a cat | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Indistinguishable from a typical a cat

Rate how literally the above robot looks like a a dog *

Does not look at all like a a dog | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Indistinguishable from a typical a dog

Rate how literally the above robot looks like a a toy *

Does not look at all like a a toy | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Indistinguishable from a typical a toy

Rate how ambiguous the robot's design is *
A design that is ambiguous is difficult to place into a single category. For example, a robot that looks only like a dog would be unambiguous, whereas a robot that looks both like a cat and a dog would be ambiguous.

Highly Unambiguous | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Highly Ambiguous

Rate how atypical the robot's design is *
An atypical design is one that includes features you wouldn't expect to see for the category you would use to describe the robot. For example a "dog" robot with wings would be highly atypical, whereas a "dog" robot with a tail would be highly typical.

Highly Typical | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Highly Atypical

Using the scale provided, how closely do the statements describe the pictured robot? *

	Strongly Disagree	Disagree	Slightly Disagree	Undecided	Slightly Agree	Agree	Strongly Agree
...is unrestricted in its gestures.	○	○	○	○	○	○	○
...is able to move toward me.	○	○	○	○	○	○	○
...is able to react to the same environmental stimuli as I do.	○	○	○	○	○	○	○
...is able to autonomously navigate in space.	○	○	○	○	○	○	○
...is unrestricted in its actions.	○	○	○	○	○	○	○
...is unrestricted in its facial expression.	○	○	○	○	○	○	○
...is able to understand my behavior.	○	○	○	○	○	○	○
...is able to perceive what I perceive.	○	○	○	○	○	○	○
...is unrestricted in its movements.	○	○	○	○	○	○	○
...is able to recognize emotions.	○	○	○	○	○	○	○
...is able to touch objects.	○	○	○	○	○	○	○
...is able to carry objects.	○	○	○	○	○	○	○

What are two to five tasks you think this robot should be used for? *

Fig. 17. The interface and questions participants in the functional expectation survey saw (Study 3).

Table 4. This table describes the assignment of the Likert items to the specific functional constructs they measured. The order of the questions in the Likert section were randomized both during pilot studies and during the full deployment.

Construct	Items
Perception and Interpretation	"...is able to react to the same environmental stimuli as I do", "...is able to recognize emotions", "...is able to perceive what I perceive", and "...is able to understand my behavior"
Tactile Mobility and Interaction	"...is able to autonomously navigate in space", "...is able to move toward me", "...is able to touch objects", and "...is able to carry objects"

Continued on next page

Table 4 – Continued from previous page

Construct	Items
Non-verbal Communication	"...is unrestricted in its actions", "...is unrestricted in its movements", "...is unrestricted in its facial expression", and "...is unrestricted in its gestures"

B ROBOT DESCRIPTORS

Table 5. A table of the binary and ordinal robot descriptors that were developed through inspection of the robot designs, and user descriptions. For each feature, we provide a description of what the feature means, and the measure of Cronbach's alpha that we obtained between two raters of the robotic systems.

Robot Feature	Description	Cronbach's α
Anthropomorphic Embodiment?	Presence of human-like features (e.g., is bipedal, has two arms, two legs, or hair on the head).	.87
Zoomorphic Embodiment?	Presence of animal-like features (e.g., a tail, wings, animal-like ears)	1.00
Mechanical Embodiment?	Presence of visible mechanical parts (e.g., exposed wires, wheels, or visible motors).	.89
Dominant Classifaction	One of {Anthropomorphic, Zoomorphic, Mechanical}, which describes the overall form of embodiment.	.83
Number of Wheels	The assumed number of wheels that the embodiment uses to move.	.70
Number of Legs	The number of appendages that can be used for locomotion.	.88
Number of Arms	The number of assumed appendages that could be used for gesturing and grasping.	.95
Number of Eyes	The number of round components that can be perceived as eyes.	1.00
Mobile?	Can physically move between points in space.	.89
Does it ride on something?	Presence of a platform that the robot appears to rest on top of.	.86
Drivetrain Skirt?	Indicates that the wheels and motors were contained within a skirt-like shape that smoothly connects with the rest of the embodiment.	.79
Treads?	Presence of treads as a means of locomotion.	1.00
Spherical Head?	Presence of a head that appears to be a near-perfect sphere.	.92
Box Head?	Indicates that the head is approximately box-shaped (but not just a standalone screen).	.87
Tablet Head?	Indicates that the head consists of a single screen (e.g., a phone, tablet, etc.)	1.00
Human Head?	Indicates that the head is human-like in appearance and has a skin-like quality.	1.00

Continued on next page

Table 5 – Continued from previous page

Robot Feature	Description	Cronbach's α
Wearing a Helmet?	Indicates that the robot appears to be wearing a helmet or face shield.	.61
Antennae?	Presence of one or more antenna-like structures on the head	1.00
Hair Follicles?	Presence of many separate hair-like protrusions from the head in a distinct region that represents hair.	.87
Mechanical Hair?	Presence of mechanical structure on the head that can be interpreted as a hair style.	1.00
Ears?	Presence of shapes or mechanisms that resemble ears.	.81
Screen Face?	Presence of a screen near the top of the robot that displays at least one facial feature.	.94
Static Face?	Presence of physical facial features that are not physically actuated.	.78
Mechanical Face?	Presence of a physical facial features that contains components that are physically actuated.	.77
Mouth?	Presence of a shape or mechanism that resembles a mouth.	.89
Nose?	Presence of a shape or mechanism that resembles a nose.	.83
Eyebrows?	Presence of shapes or mechanisms that resemble eyebrows.	1.00
Blush?	Presence of a shape, mechanism, or coloring that resembles rosy cheeks.	.72
Eyelids?	Presence of a shape or mechanism that resembles eyelids	.72
Pupils?	Presence of a shape within a round shape perceived as eyes that represents a pupil.	.92
Irises?	Presence of a (colorful) shape within a round shape perceived as eyes that represents an iris, which contains a pupil.	.78
Eyelashes?	Presence of hair-like protrusions from the eye that represent eyelashes.	.89
Lips?	Presence of shapes or mechanisms that resemble lips.	.82
Mechanical Lips?	Presence of physical tube-like structures that represent lips.	1.00
Low Waist-to-Hip Ratio?	Indicates that the perceived waist width of the robot is much smaller than (< 0.8 times) the perceived hip width.	.80
High Shoulder-to-Waist Ratio?	Indicates that the perceived shoulder width is much larger than (> 1.25 times) the perceived waist width.	.93
High Shoulder-Hip Ratio?	Indicates that the perceived shoulder width is much larger than (> 1.25 times) the perceived hip width.	.62
Screen On Chest?	Presence of a display interface at a medium height on the embodiment.	1.00
Furry?	Indicates that the robot's embodiment is covered in multiple hair-like protrusions.	1.00
Matte Body?	Indicates that the external sheen of the embodiment is not highly reflective.	.94

Continued on next page

Table 5 – Continued from previous page

Robot Feature	Description	Cronbach's α
Hard Exterior?	Indicates that the robot's exterior is constructed from hard materials (e.g., plastic, metal, etc.).	1.00
Skin-like Material?	Indicates the presence of a skin-like, flexible, and non-furry material covering any part of the embodiment.	1.00
Exposed Wires?	Presence of visible string-like structures that are needed for power requirements of the embodiment.	.80
Jointed Limbs?	Indicates that the limbs of the robot contain visible joints (i.e., not hidden under fabrics or outer casings).	.79
Industry?	Indicates that the robot was released for purchase by end-users.	.95
Curvy Embodiment?	Indicates that the embodiment is designed with organic-looking curves and the embodiment is not obviously partitioned into simple shapes (e.g., rectangular prisms or cylinders).	.73
Symmetric Embodiment?	Indicates that the embodiment exhibits reflective symmetry across its sagittal plane.	.79

Table 6. A table of the continuous feature descriptors taken from the robots' websites.

Robot Feature	Description
Height	The total height of the robot in centimeters.
Weight	The total mass of the robot in kilograms, or "UNK" if this information was not available.
Year	The year in which the robot was created or first written about publicly.
Country of Origin	The country in which the robot was developed
Most Prominent Color	The color that is used in most of the embodiment.