Keep an Eye on the Task! How Gender Typicality of Tasks Influence Human–Robot Interactions

Dieta Kuchenbrandt · Markus Häring ·
Jessica Eichberg · Friederike Eyssel ·
Elisabeth André

Abstract In an experiment, we tested whether the gender typicality of a human–robot interaction (HRI) task would affect the users' performance during HRI and the users' evaluation, acceptance and anthropomorphism of the robot. $N = 73$ participants (38 females and 35 males) performed either a stereotypically male or a stereotypically female task while being instructed by either a 'male' or a 'female' robot. Results revealed that gender typicality of the task significantly affected our dependent measures: More errors occurred when participants collaborated with the robot in the context of a stereotypically female work domain. Moreover, when participants performed a typically female task with the robot they were less willing to accept help from the robot in a future task and they anthropomorphized the robot to a lower extent. These effects were independent of robot and participant gender. Our findings demonstrate that the gender typicality of HRI tasks substantially influences HRI as well as humans' perceptions and acceptance of a robot.

D. Kuchenbrandt (✉) · F. Eyssel
Center of Excellence in Cognitive Interaction Technology (CITEC),
University of Bielefeld, Inspiration 1, 33619 Bielefeld, Germany
e-mail: dkuchenb@cit-ec.uni-bielefeld.de

F. Eyssel
e-mail: feyssel@cit-ec.uni-bielefeld.de

M. Häring · J. Eichberg · E. André
Human Centered Multimedia, Institute of Computer Science,
Augsburg University, Augsburg, Germany
e-mail: haering@informatik.uni-augsburg.de

J. Eichberg
e-mail: jessica.eichberg@student.uni-augsburg.de

E. André
e-mail: andre@informatik.uni-augsburg.de

1 Introduction

Robots have become increasingly popular in various application areas. On the one hand, this is due to the rapidly growing technological development in the field of robotics. On the other hand, robotic technologies promise to be highly beneficial for humans by releasing them from tasks that might be dangerous, tedious, hard to handle or highly monotone. If humans would actually get the opportunity to own a robot that could provide assistance for any given task they deemed suitable, how should their robot ideally look like? Which name should it have? Would people prefer a male or a female robot prototype? Importantly, what kind of work should their robot get done? Most people might now think about a robot that could clean the house or do the dishes, a robot that could provide help on tasks we often find annoying or time-consuming. Indeed, numerous already existing and newly developed robots are supposed to assist people with tasks such as doing daily household chores (Human Support Robot HSR, Toyota Motor Corp.), collecting and delivering commodities or food (e.g., Movaid [1]), serving as a social companion and communication partner (e.g., iCat [2]), and providing medical services (e.g., looking after patients; Robina, Toyota Motor Corp.). This also seems to be in line with humans' actual preferences regarding robot applications. Enz et al. [3] have demonstrated that individuals mostly prefer robots in a nursing context compared to other potential social robot roles (e.g., robots as teachers). Interestingly, a closer look at these application domains reveals that they are often closely associated with prescriptive societal gender roles. That is, these tasks can be categorized in terms of 'stereotypically female' versus 'stereotypically male' tasks (e.g.[4, 5]). More specifically, many robot applications seem to be associated with traditionally female work domains (e.g., housework, health care). Does this gender stereotypicality of robot applications

influence how humans interact with, perceive, and evaluate a specific robot prototype? To date, we are not aware of any study that has investigated this question. In the present research, we will address this gap in the literature.

Gender is one of the most salient and omnipresent social categories in human societies that affects virtually every aspect in our every-day live [6]. To a large extent, gender determines people's social roles, occupations, relationships, and opportunities [7]. Importantly, our own gender as well as the gender of others influences how we think about and interact with each other. Interestingly, this influence is not restricted to human–human social interactions. Gender also largely impacts how humans perceive and interact with non-human agents, such as computers [8,9] or robots. In the context of human–robot interaction (HRI) gender effects were investigated mostly from two different angles: The alleged robot gender (as indicated, e.g., by its appearance, behavior, or name) and the user's gender.

From a user perspective, robots often do not appear gender-neutral, but instead they are perceived in terms of male or female prototypes. To illustrate, Eyssel and Hegel [10] have demonstrated that a visual cue, such as a robot's hair length led to differential ascriptions of stereotypically male (e.g., competence) and female traits (e.g., warmth). The alleged robot gender also affected the perceived suitability of the robot for stereotypically male versus female tasks. Likewise, vocal cues (i.e., a male or female voice) indicate a robot's gender [11] and this, in turn, can affect the user's reactions toward the robot. Moreover, Powers et al. [12] have shown that people use knowledge about gender roles when interacting with a gendered robot. The robot's gender was here indicated with a male versus female voice plus different colors (grey vs. pink) of the robot's lips. In this study, participants elaborated less on a typically female topic (i.e., dating norms) when talking to an ostensibly female robot than when talking to a 'male' robot.

Besides robot gender, user gender impacts the users' evaluative reactions toward robots, whereas findings are inconsistent across studies. To illustrate, research by Siegel et al. [13] has shown that users evaluate a robot of the *opposite* gender more positively than a same-gender robot; they even tend to behave more positively toward robots of the opposite gender. In contrast, Eyssel et al. [11] have demonstrated that participants perceived a *same*-gender robot significantly more positive and felt psychologically closer to it compared to the opposite-gender robot. To date, however, it is unclear, which factors account for these inconsistencies across studies and which variables might moderate whether humans prefer same- or opposite-gender robots.

In addition, research has shown that gender not only influences humans' evaluative reactions but also their anthropomorphic inferences about robots. Anthropomorphism denotes the psychological process of attributing humanlike properties, characteristics, and mental states to nonhuman agents [14]. Importantly, Epley et al. [14] have proposed that anthropomorphism represents an effective strategy to increase the ability to make sense of a nonhuman agent's actions and thus to reduce the uncertainty that is often associated with human-nonhuman interactions. It "can enable a sense of efficacy with these agents, a sense that actually increases one's apparent competence interacting with these agents" [14, p. 879]. According to this reasoning, anthropomorphism can be considered as an important factor contributing to the quality of HRI. Crucially, previous research has tested the role of gender with regard to anthropomorphism. Schermerhorn et al. [15], for instance, have investigated gender differences in the perception of robots and found that males compared to females showed a stronger tendency to perceive a robot as humanlike. However, other research findings indicate an opposite tendency. Siino and Hinds [16] investigated how male and female hospital workers conceptualized a mobile autonomous robot in their workplace. In their study, male workers perceived the robot as a controllable machine, whereas females tended to see the robot more like a human male acting with agency. Furthermore, Eyssel et al. [11] have demonstrated that user and robot gender together influenced the degree to which humans anthropomorphize a robot: participants attributed more mental capacities to a *same*-gender robot than to the opposite-gender robot.

The findings reviewed here illustrate that robot and user gender elicit complex effects in HRI that cannot easily be put in a nutshell. In addition, not only features of the robot and the user characterize HRIs, but also features of the task the user and the robot collaborate on. As outlined above, the different types of tasks a robot performs might be perceived as being either stereotypically male or stereotypically female (see [10]). Thus, just like robot and user gender, the perception of gender typicality of a task could shape the users' mental model about a specific robot and therefore could impact how people would perceive, evaluate, and interact with the robot on the respective task. This notion is supported by Nass and colleagues. Their research in the field of human-computer interactions [9] has shown that a gender match between task and computer features resulted in more positive perceptions of the computer. In one of their studies, a male-voiced tutor computer was judged as more informative about issues related to technology compared to a female-voiced tutor, whereas the female-voiced tutor computer was perceived as more informative about topics related to interpersonal relationships compared to the male-voiced one. Further evidence that task features along with user and robot features differentially influence HRI and perceptions of a robot comes from Mutlu et al. [17]. In their experiment, males and females played an interactive video game with a robot, and they did so either in a cooperative or in a competitive way. The results showed that men based their

evaluation of the robot to a large extent on the different features of the tasks, whereas women were more influenced by the characteristics of the robot. In a different set of studies [18], participants found a robot more suitable for a task when the degree of the robot's humanlikeness matched the degree of sociability required by the task. Accordingly, task characteristics indeed influence humans' perceptions of a robot and HRI quality. However, previous research has not yet considered the gender typicality of different tasks as an important aspect in HRI, despite the fact that many domains of robot applications are closely associated with societal gender roles.

The aim of the present exploratory experiment was to investigate the impact of gender stereotypicality of an HRI task on humans' task performance during HRI, on their evaluation and acceptance of the robot, and on anthropomorphization of the robot. Moreover, we also tested the interplay of this gender factor with user and robot gender. Given the novelty of this research topic and the inconsistent findings regarding user and robot gender effects in HRI across studies, we make no specific assumptions about the differential effects of gender typicality of the task as well as of user and robot gender on the dependent measures.

2 Method

2.1 Participants and Design

$N = 73$ German participants (38 females, 35 males) with a mean age of 25.04 years ($SD = 4.34$; age ranged from 19 to 42 years) took part in our experiment[1]. They were randomly assigned to one of four experimental conditions that resulted from a 2 (*gender typicality of task:* Male vs. female) $\times$ 2 (*robot gender:* Male vs. female) between-subjects factorial design: accordingly, together with an allegedly male or female robot participants had to solve a task that constituted either a stereotypically male or a stereotypically female task. As the participants' gender also was treated as an experimental factor, the full experimental design was a 2 (*gender typicality of task:* Male vs. female) $\times$ 2 (*robot gender:* Male vs. female) $\times$ 2 (*participant gender:* Male vs. female) design resulting in eight experimental groups. Each experimental group consisted of eight to 10 participants.

Fig. 1 The robot shows the participant a possible position for the selected item in the sewing kit, using speech, gaze and pointing gestures

2.2 Procedure

Participants were tested individually in a laboratory at Augsburg University. They were seated in front of a Microsoft Surface[2] touch-screen table opposite to the robot NAO (Academic Edition V3.2, Aldebaran Robotics, see Fig. 1). On the touch-screen, different items (either sewing accessories or tools) and a container (either a sewing box or a toolbox, see Fig. 2a, b) were depicted[3]. Initially, the experimenter briefly introduced the participants to the robot and mentioned the robot's name to indicate robot gender (i.e., NERO or NERA). Participants were then informed that they would work on a sorting task together with the robot and that the robot would instruct them on the task. The robot operated fully autonomously during the experiment (for details see [20]). After a short tutorial with two sample trials, participants completed 15 critical trials of the sorting task. On average, the interaction between participant and robot lasted for approximately 10 min. Subsequently, the experimenter asked participants to complete several computerized questionnaires that contained our dependent measures. Finally, participants were reimbursed, debriefed and dismissed.

2.3 Human–Robot Interaction Task

2.3.1 Experimental Set-up

On the touch-screen table, participants were presented with a container that included 10 compartments that were of small,
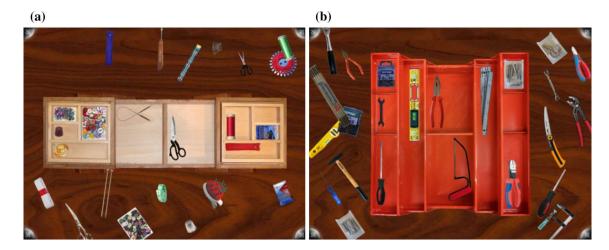

Fig. 2 Pictures of the experimental set-up: **a** a sewing box with sewing accessories, **b** a tool box with tools

medium, and large size. Moreover, participants were presented with nine items that were already sorted into the different compartments of the container. Fifteen further items were distributed around the container. These items had to be sorted into different compartments of the container. All but three items could be grouped into nine object categories (e.g., different types of scissors or water levels). Each category was represented by at least two items. Importantly, one item of each category was already stored in the container (see Fig. 2a, b). This was done in order to give participants guidance as to where the remaining objects could be stored. In addition, three remaining unsorted items did not belong to any of the nine categories. Figures 2a, b depict the set-up of the sorting task.

2.3.2 Instructions

In each of the 15 trials, participants received two instructions from the robot: the first instruction was the *selection instruction* that concerned the choice of the object (e.g., 'Pick the small scissors.'). This instruction included a specific description of the respective item (e.g., name, size, color if applicable). Moreover, following the procedure by Ishiguro et al. [21], each selection instruction was accompanied by a pointing gesture and a gaze toward the object. The participant was then supposed to select the respective item by tapping on the touch screen. The robot verbally confirmed a correct choice (e.g., 'This is correct.'). In case of a wrong choice, the robot provided respective feedback (e.g., 'No, this is the wrong object.') and repeated the instruction.

The second instruction in each trial concerned the *position instruction*. This instruction specified the target position for each item (e.g., 'Put it in the upper right small compartment.'). Similar to the selection instructions, the position instructions were accompanied by gaze and pointing behavior, as depicted in Fig. 1. However, to make the interac-

tion more realistic, natural and interesting [22], participants not always received correct or optimal instructions by the robot. That is, the robot used three different types of *position instructions*: in six of 15 trials, the robot gave *optimal position instructions* asking participants to put an item into a compartment that already contained an object of the same category. In six further trials, participants received *suboptimal position instructions*. That is, they were instructed to sort an item into a compartment that did not already contain an item of the same category, although an exemplar of the same object category was depicted in one of the other container's compartments. In three trials, the robot gave *wrong position instructions* and asked participants to put an item into a compartment that was too small to accommodate the chosen item. Thus, participants were obliged to choose an alternative compartment to store the specific item.

After each position instruction the robot commented on the participants' behavior with a short feedback. The robot used up to ten different utterances differentiating between three different types of participant behavior. First, when participants followed the robot's optimal or suboptimal position instructions, the robot uttered, e.g., 'Right, this is the correct compartment.' or 'Very good, that was correct.'. Second, in case participants did not follow the robot's optimal or suboptimal instructions but instead chose a correct alternative, the robot confirmed the participants' correct choice (e.g., 'Right, this fits as well.' or 'Yes, this is correct, too.'). Third, in case participants chose an incorrect target position for the item (i.e., a compartment that was too small for the item), the robot corrected the participants with utterances such as 'No, you need to find a bigger compartment.' or 'Unfortunately the compartment is too small.' and repeated the original position instruction.

The trials were realized in a fixed randomized order. The sequence of optimal, suboptimal, and wrong instructions was identical for all experimental conditions.

2.4 Experimental Manipulation

2.4.1 Gender Typicality of the Task

Participants were confronted with one of two different types of tasks. Participants were either asked to sort different tools into a toolbox. This represented the typically male task condition. In the female task condition, in contrast, participants were asked to sort sewing equipment into a sewing box. Accordingly, the two tasks were equal regarding the specific actions required from the participants (i.e., sorting various objects into different compartments). Gender typicality, however, was based on the kind of objects that had to be sorted and that are differentially associated with gender stereotypes. The tasks were identified as suitable for our experimental purpose based on the following pretest: $N = 30$ participants (12 male, 18 female, age: $M = 23.17, SD = 4.58$) were asked to rate 19 different tasks on bipolar 7-point scales regarding the tasks' complexity (-3 = simple, $+3$ = complex), the intellectual demands (-3 = low, $+3$ = high), and the gender typicality (-3 = female, $+3$ = male). Sorting sewing equipment into a sewing box was identified as typically female ($M = -1.67, SD = 1.16, t(29) = -26.88, p < .001$, tested against 0, the neutral midpoint of the scale) whilst sorting tools into a toolkit was rated as being typically male ($M = 1.80, SD = 1.06, t(29) = -11.33, p < .001$). Importantly, the two tasks did not differ from each other with redard to intellectual demands ($t < 1$) and complexity ($t < 1$). This was important in order to assure that the potential effects of gender typicality are not confounded with other factors such as task difficulty or complexity.

2.4.2 Gender of the Robot

In order to manipulate the alleged gender of the robot, we varied two aspects of the robot: Its name and its voice. Previous research has demonstrated that the name of a robot is a suitable indicator of the robot's alleged membership in social groups (e.g., ethnicity) [23]. In line with this, we assume that giving the robot a male versus female name is a suitable manipulation of the robot's alleged gender. In addition, related work in the field of human–computer interaction and HRI yielded that voice is another strong indicator of a computer's or robot's perceived gender, respectively [9,11]. To identify a male versus female name, $N = 30$ participants (17 male, 13 female, age: $M = 28.80, SD = 5.32$) were asked in a pretest to rate 12 names with regard to the names' gender typicality. On 7-point Likert scales, participants indicated how feminine (7) or masculine (1) they perceived the respective name. Based on this, the names NERA ($M = 6.57, SD = 0.63$) and NERO ($M = 1.23, SD = 0.50$) were chosen. Importantly, both names differed significantly from the neutral midpoint of the scale (4) in the expected direction ($t(29) = 22.46, p < .001$ for NERA; $t(29) = -30.07, p < .001$ for NERO) and were thus used to indicate female versus male gender of the robot, respectively. During the experimenter's instructions at the beginning of the study, the name of the robot has been mentioned repeatedly. In addition, the 'male' robot NERO spoke with a typically male voice (low frequency), whereas the 'female' robot NERA has been equipped with a more female-type voice (high frequency). The robotic voices have been generated by the robot's Text-To-Speech system (Acapela Mobility 7.0) and were selected in accordance with pretest results: $N = 41$ participants (18 male, 23 female, age: $M = 26.12, SD = 1.15$) listened to a neutral sentence ("According to my watch, it is now quarter past three", adapted from [11]) uttered with six different voices. For each voice, participants had to decide whether the respective voice was a male or a female one. Based on this, two voices were chosen: one 'male' voice that 98 % of the participants perceived as male, and one 'female' voice that 61 % perceived as female.

2.5 Dependent Measures

2.5.1 Manipulation Check

As a manipulation check, we asked participants to indicate on a 7-point Likert scale (endpoints: 1 = 'more female' vs. 7 = 'more male') whether they perceived the robot as being more female or more male. A value below 4 would indicate that the robot was perceived as female whereas a value above 4 would show that participants perceived the robot as male.

To ensure that the two types of tasks were equally demanding for the participants depending on gender typicality of the task, participants had to indicate on a 7-point Likert scale how difficult they perceived the task. Higher values indicate a greater level of task difficulty.

2.5.2 Performance During HRI

We used three different behavioral indicators for participants' performance during HRI. First, we measured the duration of each of the 15 trials, resulting in an *average duration per trial* (in seconds). Second, we counted how often participants chose a compartment that did not fit the size of the object. This was considered as a position error, resulting on an *average error rate per trial*. These two indicators represent the measures of objective task performance. That is, the longer the average duration per trial, and the higher the average number of errors per trial, the lower the quality of task performance.

As a third performance measure, we used the number of alternative compartment solutions participants have chosen, that is, the number of times participants did not follow the robot's position instructions. However, after receiv-

ing wrong position instructions from the robot, participants were obliged to choose an alternative compartment. Thus, we only considered the number of alternatives after optimal and suboptimal position instructions. Accordingly, we calculated the *average number of chosen compartment alternatives per trail* after optimal and suboptimal instructions. This measure is used as an indicator of compliance with the robot.

2.5.3 Evaluation and Acceptance of the Robot

With two items, participants rated the *robot's task-related competence* ('The robot knew exactly what I had to do in this task.', 'The robot was well informed about the task.'). The endpoints of the scales were 1 = 'not at all' and 7 = 'very much'. These two items formed a reliable index of task competence of the robot, $\alpha = .83$.

To measure *robot acceptance*, we asked participants to indicate how willing they would be to accept help from the robot on a possible future task. The endpoints of the scale ranged from 1 = 'not at all' to 7 = 'very much'.

2.5.4 Anthropomorphization of the Robot

In order to measure anthropomorphic inferences about NAO, we used two different measures. First, at a more general level, we asked participants to what extent they perceived the robot as a machinelike versus humanlike entity (*humanlikeness*). Participants indicated their impression on a 7-point Likert scale (1 = machinelike, 7 = humanlike), with higher values indicating more perceived humanlikeness. Second, adapting a scale developed by Gray et al. ([24], see also [11,25]), we measured the extent to which participants attributed mind to the robot (*mind attribution*). Participants rated NAO with regard to 25 mental capacities (e.g., the capacity to feel pain, to have desires, intentions, or to make plans). Items were summed and avaraged and formed a reliable index ($\alpha = .93$). Higher values reflect a greater extent of mind attribution to the robot.

3 Results

3.1 Manipulation Check

We first tested whether participants recognized the alleged gender of the robot. Results of a t-test revealed that in the female robot condition the robot was correctly identified as female ($M = 2.94, SD = 1.59$), whereas in the male robot condition the robot was correctly identified as male[4] ($M = 5.97, SD = 1.16$), $t(70) = 9.26, p < .001, d = 2.18$.

---

[4] Note that the endpoints of the 7-point Likert scale were 1 = more female versus 7 = more male. That is, values below 4 indicate that the robot was perceived as more female, whereas values above 4 show that participants perceived the robot as more male.

Moreover, to make sure that participants perceived both sorting tasks as equally demanding, we conducted a t-test comparing the typically female and the typically male task. Results indicate a marginally significant difference between the typically female task ($M = 1.32, SD = 0.55$) and the typically male task ($M = 1.57, SD = 0.60$), $t(70) = 1.85, p = .07, d = 0.43$, indicating that the female task was perceived as slightly less difficult than the male task. In order to control statistically for a potential influence of the perceived task difficulty, this variable has been used as a covariate in all subsequent analyses. In no case, however, the task difficulty turned out to be a significant covariate, all $ps > .15$.

3.2 Performance During HRI

3.2.1 Duration of Task Completion

Results of a 2 (*gender typicality of task*: Male vs. female) × 2 (*robot gender*: Male vs. female) × 2 (*participant gender*: Male vs. female) analysis of covariance (ANCOVA) yielded no significant main effects on the duration of task completion (in seconds), all $ps > .15$. However, we obtained a significant *robot gender* by *participant gender* interaction effect, $F(1, 62) = 5.32, p = .02, \eta^2 = .08$. In order to inspect this pattern of results further, we conducted post-hoc t-tests. Moreover, to control for inflated type I error rates due to multiple testing, we tested against a Bonferroni-corrected significance level of $\alpha = .025$. Results show that female participants completed the task equally fast, regardless of whether they interacted with an ostensibly female or male robot ($M = 5.58 s, SD = 1.04$ vs. $M = 6.03 s, SD = 1.38$, respectively, see Fig. 3a), $t(36) = 1.01, p = .32, d = 0.33$. In contrast, male participants were faster in completing the task when they interacted with the male ($M = 4.96 s, SD = 0.88$) than with the female robot ($M = 5.85 s, SD = 1.31$), $t(33) = -2.37, p = .002, d = 0.80$.

3.2.2 Errors

The 2 (*gender typicality of task*: Male vs. female) × 2 (*robot gender*: Male vs. female) × 2 (*participant gender*: Male vs. female) ANCOVA revealed a significant main effect of *gender typicality of task*, $F(1, 62) = 7.67, p = .01, \eta^2 = .11$. As can be seen in Fig. 3b, c, participants chose a wrong compartment more often when working on a typically female ($M = 0.20, SD = 0.13$) than on a typically male task ($M = 0.12, SD = 0.09$). Furthermore, we obtained a significant effect of *participant gender*, $F(1, 62) = 4.61, p = .04, \eta^2 = .07$, showing that female participants made less errors in making a position decision ($M = 0.14, SD = 0.11$)
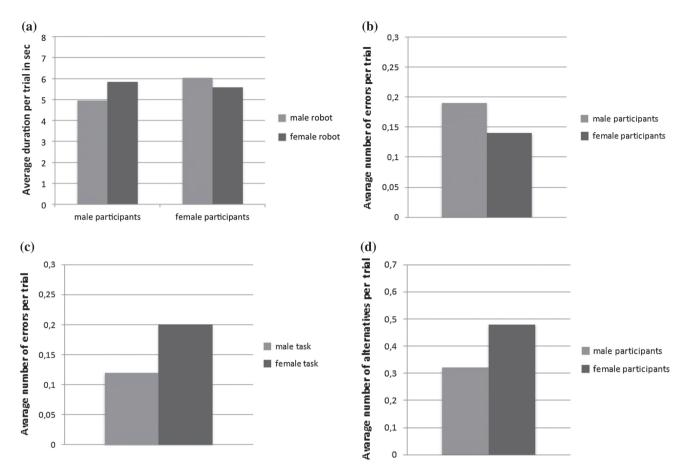
Fig. 3 **a** Average duration per trial as a function of participant and robot gender. **b** Average number of errors per trial as a function of participant gender. **c** Average number of errors per trial as a function of gender typicality of the task. **d** Average number of alternatives per trial as a function of participant gender

than males ($M = 0.19$, $SD = 0.12$). No other main or interaction effects reached statistical significance, all $ps > .19$.

3.2.3 Alternatives

Results of a 2 (*robot gender*: Male vs. female) × 2 (*gender typicality of task*: Male vs. female) × 2 (*participant gender*: Male vs. female) ANCOVA yielded a significant main effect of *participant gender*, $F(1, 62) = 13.01$, $p < .001$, $\eta^2 = .17$, indicating that female participants used alternative solutions more frequently instead of following the robot's instructions ($M = 0.48$, $SD = 0.17$) compared to male participants ($M = 0.32$, $SD = 0.20$, see Fig. 3d). No other significant effects were found, all $ps > .24$.

3.3 Evaluation and Acceptance of the Robot

3.3.1 Robot's Task-Related Competence

A 2 (*gender typicality of task*: Male vs. female) × 2 (*robot gender*: Male vs. female) × 2 (*participant gender*: Male vs.

female) ANCOVA revealed no main effects, all $ps > .16$, nor significant interaction effects, all $ps > .07$.

3.3.2 Robot Acceptance

Results of a 2 (*gender typicality of task*: Male vs. female) × 2 (*robot gender*: Male vs. female) × 2 (*participant gender*: Male vs. female) ANCOVA yielded a main effect of *gender typicality of task*, $F(1, 61) = 7.82$, $p = .01$, $\eta^2 = .11$. Accordingly, participants were more willing to accept help from the robot on a future task when they previously interacted with the robot on a typically male task ($M = 5.09$, $SD = 1.77$) than when they worked on a female task ($M = 3.69$, $SD = 1.91$, see Fig. 4). No other main or interaction effect was significant, all $ps > .16$.

3.4 Anthropomorphization of the Robot

3.4.1 Humanlikeness

A 2 (*gender typicality of task*: Male vs. female) × 2 (*robot gender*: Male vs. female) × 2 (*participant gender*: Male
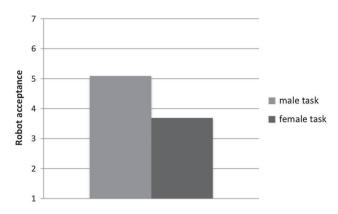
Fig. 4 Acceptance of the robot as a function of gender typicality of the task

vs. female) ANCOVA revealed a main effect of *gender typicality of task*, $F(1, 62) = 4.92$, $p = .03$, $\eta^2 = .07$, demonstrating that when the robot was the instructor for a female task participants perceived it to be more machinelike ($M = 2.25$, $SD = 1.32$). In contrast, when the robot gave instructions on the male task, it has been perceived more as a humanlike entity ($M = 3.16$, $SD = 1.75$). No other main or interaction effects reached statistical significance, all *ps* > .15.
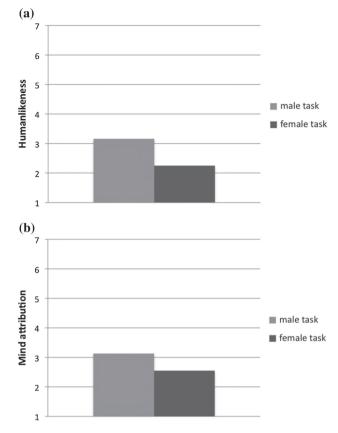
3.4.2 Mind Attribution

Similar effects were found for mind attribution. The 2 (*gender typicality of task*: Male vs. female) $\times$ 2 (*robot gender*: Male vs. female) $\times$ 2 (*participant gender*: Male vs. female) ANCOVA revealed a main effect of *gender typicality of task*, $F(1, 62) = 4.98$, $p = .03$, $\eta^2 = .07$. Within the context of a male task, participants attributed more mental capacities to the robot ($M = 3.13$, $SD = 0.96$) than within the context of a female task ($M = 2.55$, $SD = 0.93$). The findings are depicted in Fig. 5a, b.

4 Discussion

Take a second and think back to the issues we raised in the introduction of this article. If humans could own a robot for their personal use, for what kind of tasks should the robot ideally provide assistance? According to our findings, there is still no simple answer to this question.

In the present experiment, female and male participants performed a stereotypically female or stereotypically male task while interacting with an ostensibly female or male robot. Accordingly, we tested the effects of gender typicality of an HRI task, robot gender and user gender on participants' performance during HRI, on their evaluation and acceptance of the target robot and on anthropomorphic inferences about

Fig. 5 **a** Perceived humanlikeness of the robot as a function of gender typicality of the task. **b** Mind attribution to the robot as a function gender typicality of the task

the robot. Results show that the gender typicality of the HRI task significantly influenced how well participants performed the task, how they evaluated and perceived the robot: when being instructed on a typically female versus male task, participants made more errors, were less willing to accept help from the robot on a future task, and anthropomorphized the robot to a lesser extent. In addition, participant gender influenced the task performance. Females compared to male participants made less errors and were less compliant with the robot, that is, they more often chose alternative compartments. Males, in contrast, were faster when performing the task than females, while this was only true when they interacted with a same-gender robot.

With our experiment, we extended the previous li-terature on gender effects in HRI with respect to se-veral aspects: We were the first to test the effects of such *gender typicality of an HRI task* on HRI. With this, we demonstrated that gender typicality of the task that the user and the robot completed together had substantial impact on the outcomes of the HRI. Particularly, on a variety of measures ranging from objective performance measures to participants' evaluation and perception of the robot, we found that working on a male task was superior compared to a human–robot collaboration on

a typically female task. Interacting with a robot in the context of a typically female 'work domain' thus resulted in less optimal outcomes than working on a male task. It should be noted that participants perceived the female sewing box task as slightly less demanding than the male toolbox task, thus, the differences in perceived task difficulty could account for the poorer outcomes in the female task condition. However, the present results contradict this notion: even after statistically controlling for the level of perceived task difficulty, the typically female task showed—compared to the typically male one—detrimental effects on HRI and the perception of the robot. This suggests that the gender-specific stereotypes that are attached to the different kinds of tasks might have influenced how successfully participants dealt with the robot and the task. Interestingly, many robots are developed to provide assistance on every-day tasks that are generally perceived as being typically female (e.g., providing assistance in the household). Future research thus needs to address possible measures to counteract the difficulties of human–robot collaborations in female domains that were identified in the present study. Moreover, prospective research should focus on long-term effects of a task's gender typicality on HRI, as our findings cannot reveal whether also long-term interactions between robots and humans would be negatively affected in the context of stereotypically female robot applications.

In addition, our results add to previous findings that have shown that male and female users react differently toward robots (e.g., [11,26]). In the present study, female participants more frequently worked autonomously on the task and made their own choices instead of following the instructions of the robot compared to male participants. This result indicates that females were less compliant with the robot than males. This could be based on reduced trust in the robot or on a lesser extent of willingness to cooperate with the robot. However, this assumption needs to be tested in future studies. One might further argue that this gender difference could explain the fact that females made less position errors than males as female participants generally followed less often the robot's (wrong) instructions. In fact, the two measures of errors and alternatives were significantly correlated with each other ($r = .63$). However, when investigating task and participant gender effects on errors while statistically controlling for chosen alternatives, we still found a significant influence of participants' gender on position errors, at least in the male work domain (females vs. males make less errors in the male work domain). In addition, when testing gender effects on chosen alternatives while controlling for average errors per trial, we still observe that females compared to males choose more alternatives. Thus, the participant gender effect with regard to errors cannot be completely explained by the fact that females generally were less influenced by the robot and vice versa. Moreover, with respect to participant

gender we found that male participants seemed to be more efficient (i.e., faster) when collaborating with a same-gender than with the opposite-gender robot whereas for women the robot's gender did not influence their velocity in performing the task. Interestingly, this is in line with previous research [13] showing that men are more reactive to a robot's gender cues than women.

5 Critical Remarks and Future Directions

Even though the present findings are promising and hopefully stimulate further investiagtions of gender effects in HRI, some limitations and future directions of this research shall be discussed.

First, the reported experiment applied a simple and to some extent artificial HRI task, which could decrease the ecological validity of our findings. At the same time, however, for specific user groups (e.g., users with special needs) such instructed sorting tasks might represent an actual and useful robot application. Furthermore, the gender typicality of the task was only operationalized by a contextual manipulation, that is, via the different objects (tools vs. sewing accessories) that had to be sorted. From a methodological point of view, this was necessary in order to avoid confounding of our effects with other factors such as task difficulty or complexity. To illustrate, washing the dishes as a typically female task not only differs from repairing a car (as a typically male task) in terms of gender typicality. These tasks also differ regarding their intellectual and physical demands and with regard to their complexity. However, our manipulation also reflects the fact that the gender typicality of tasks is often contextually determined (e.g., washing the dishes vs. washing a car). Moreover, as we even found effects of this rather unobtrusive manipulation, this supports even more the assumption that gender typicality of HRI tasks influence HRI outcomes. Nonetheless, future work needs to investigate more complex and ecologically valid tasks whose gender typicality is not only defined by the objects involved in the task. This is particularly important in order to test how generalizable the present findings are.

Second, one could critically argue that the measure of position errors as one of our dependent variables does not measure errors by the human users but only assesses whether humans "blindly" followed the robot's wrong (3 out of 15) instructions or not. Consequently, the error measure might be confounded with participants' commitment to the robot's instructions. However, while following the robot's wrong instructions indeed can be seen as an indicator of commitment, such commitment at the same time is erroneous behavior. To illustrate, in other contexts, for instance in the production sector, such highly committed

behavior can have severe impact. When humans do not notice errors made by robots or follow a robot's wrong instructions due to high commitment, this could not only cost high amounts of money but could also be threatening and even dangerous for humans. In other words, following the robot's wrong instructions constitutes a "real error" (note: participants were free to not follow the robot's instructions). Nonetheless, when interpreting the significant difference between the typically female and male work domain with regard to errors (on average 3 errors in female task vs. 1.8 errors in male task), we need to be aware that this difference partly reflects the fact that participants followed more often the robot's erroneous instructions when working on a stereotypically female versus male task.

Third, the sample size in the present experiment was quite small given the complex experimental design, resulting in a low probability for detecting small effects of our experimental manipulations. In fact, according to power analyses [19], a sample size of 130 up to 780 participants would have been necessary to detect medium to small effects (f ranging between .10 and .25) given a high power of $1 - \beta = .80$, and a significance level of $\alpha = .05$. For HRI studies, however, such sample sizes are impractical and almost unattainable. As a consequence, however, some gender effects in the present study might remain undetected, particularly if these were small effects. However, the present study was not generally underpowered and medium to large effects could be statistically detected (all η^2 were between .07 and .17 and thus were medium or large effects, see [27]), pointing also to the practical relevance of the presented findings.

Finally, as mostly students took part in our experiment, the age range and the professional background of our sample was to some extent homogeneous. In order to draw general conclusions of our findings, gender effects need to be tested with different groups of participants. For instance, as traditional gender roles and stereotypes changed over the last decades, age could influence the endorsement of such stereotypes. Accordingly, older participants might react differently towards gendered tasks than younger participants. Future studies should therefore consider comparing different participant groups and test whether socio-demographic variables such as age or professions moderate gender effects in HRI.

6 Conclusion

Our findings emphasize that besides taking into account mental models users have about gendered robots (see [10, 11,26]), we need to consider social roles and attributes that are related to traditionally male and female work domains when developing and designing robot systems.

Many robot applications are related to societal gender roles. The present findings yielded initial evidence for the fact that such 'gendered tasks' substantially influence how users perform during an HRI and how they perceive a robot's humanlikeness. Thus, developers and researchers should particularly pay attention to those applications that traditionally have been occupied by women. Finally, as gender is one of the most basic social categories in human social perception that is also applied to robots and to robot applications, and as gender seems to elicit quite complex effects in HRI, more theory-driven testing is needed to approach gender issues in robotics systematically and to integrate the various existing research findings.

References

1. Dario P, Guglielmelli E, Laschi C, Teti G (1999) MOVAID: a personal robot in everyday life of disabled and elderly people. Technol. Disabil. 10:77–93
2. Heerink M, Krose B, Evers V, Wielinga B (2006) The influence of a robot's social abilities on acceptance by elderly users. In: Proceedings of the 15th IEEE international symposium on robot and human interactive communication, pp 521–526
3. Enz S, Diruf M, Spielhagen C, Zoll C, Vargas P (2011) The social role of robots in the future—explorative measurement of hopes and fears. Int J Soc Robot 3(3):263–271
4. Berk RA, Berk SF (1979) Labor and leisure at home: content and organization of houshold day. Sage, San Francisco
5. Thompson L, Walker AJ (1989) Gender in families: women and men in marriage, work, and parenthood. J Marriage Fam 51:845–871
6. Harper M, Schoeman WJ (2003) Influences of gender as a basic-level category in person perception on the gender belief system. Sex Roles 49:517–526
7. Bussey K, Bandura A (1999) Social cognitive theory of gender development and differentiation. Psychol Rev 106:676–713
8. Nass C, Moon Y (2000) Machines and mindlessness: social responses to computers. J Soc Issues 56:81–103
9. Nass C, Moon Y, Green N (1997) Are computers gender-neutral? Gender stereotypic responses to computers. J Appl Soc Psychol 27(10):864–876
10. Eyssel F, Hegel F (2012) (S)he's got the look: gender-stereotyping of social robots. J Appl Soc Psychol 42:2213–2230
11. Eyssel F, Kuchenbrandt D, Hegel F, de Ruiter L (2012) Activating elicited agent knowledge: how robot and user features shape the perception of social robots. In: Proceedings of the 21st IEEE international symposium on robot and human interactive communication, pp 851–857
12. Powers A, Kramer ADI, Lim S, Kuo J, Lee S-L, Kiesler S (2005) Eliciting information from people with a gendered humanoid robot. In: Proceedings of the 14th IEEE international symposium on robot and human interactive communication, pp 158–163
13. Siegel M, Breazeal C, Norton MI (2009) Persuasive robotics: the influence of robot gender on human behavior. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 2563–2568

14. Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114:864–886
15. Schermerhorn P, Scheutz M, Crowell C (2008) Robot social presence and gender: do females view robots differently than males?. In: The Proceedings of the 3rd ACM/IEEE conference on human–robot interaction, pp 263–270
16. Siino R, Hinds P (2005) Robots, gender and sensemaking: sex segregations impact on workers making sense of a mobile autonomous robot. In: Proceedings of the IEEE international conference on robotics and automation (ICRA), Barcelona, Spain
17. Mutlu B, Osman S, Forlizzi J, Hodgins J, Kiesler S (2006) Perceptions of Asimo: task structure and user attributes as elements of human-robot interaction design. In: Extended abstracts of the 15th IEEE symposium on robot and human interactive, pp 351–352
18. Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human–robot cooperation. In: Proceedings of the 12th IEEE international symposium on robot and human interactive communication, pp 55–60
19. Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191
20. Häring M, Eichberg J, André E (2012) Studies on grounding with gaze and pointing gestures in human–robot-interaction. In: Social robotics. Lecture notes in computer science, vol 7621, pp 378–387
21. Ishiguro H, Ono T, Imai M, Maeda T, Nakatsu R, Kanda T (2001) Robovie: an interactive humanoid robot. Int J Ind Robot 28:498–503
22. Salem M, Eyssel F, Rohlfing K, Kopp, S, Joublin F (2011) Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In: Proceedings of the international conference of social robotics, pp 31–41
23. Eyssel F, Kuchenbrandt D (2012) Social categorization of social robots: anthropomorphism as a function of robot group membership. Br J Soc Psychol 51(4):724–731
24. Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. Science 315:619
25. Eyssel F, Kuchenbrandt D (2011) Manipulating anthropomorphic inferences about NAO: the role of situational and dispositional aspects of effectance motivation. In: Proceedings of the 20th IEEE international symposium in robot and human interactive communication (RO-MAN 2011), pp 467–472
26. Crowell CR, Scheutz M, Schermerhorn P, Villano M (2009) Gendered voice and robot entities: perceptions and reactions of male and female subjects. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems, pp 3735–3741
27. Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum Associates, Hillsdale

Dieta Kuchenbrandt is a post-doctoral researcher in the Group "Gender and Emotion in Cognitive Interaction Technology" at the Center of Excellence in Cognitive Interaction Technology (CITEC) in Bielefeld. She earned her Masters degree from the University of Greifswald (Germany) in 2006 and received her Ph.D. in 2010. In her research, she is interested in intergroup processes both in human–human and human–robot interactions as well as in interventions to reduce prejudice towards various social groups.

Markus Häring graduated as a Master of Science in Informatics and Multimedia from Augsburg University, Germany in 2010. As a student researcher he worked on dialog modeling for virtual characters in the EU project DynaLearn. Afterwards, he started his Ph.D. at the lab for Human Centered Multimedia with a focus on collaborative human–robot interaction. His research interests also include usability engineering, multimodal interaction, and social robotics.

Jessica Eichberg graduated in 2011 as a Master of Science in Informatics and Multimedia from Augsburg University, Germany. In her Master thesis and as a student researcher at the lab for Human Centered Multimedia she worked on concepts and technical solutions for multimodal human–robot interaction. In April 2012 she started as an Internal Tool Developer at SAP in Irland and since April 2013 she is a Behavior Architect for Special Education at the robot manufacturer Aldebaran Robotics in Paris.

Friederike Eyssel is assistant professor and head of the Group "Gender and Emotion in Cognitive Interaction Technology" at the Center of Excellence in Cognitive Interaction Technology (CITEC) in Bielefeld. She earned her Masters Degree in psychology from University of Heidelberg (Germany) in 2004. She received her Ph.D. in psychology from University of Bielefeld (Germany) in 2007. She has held interim professorships for social psychology at the University of Muenster, the Technical University of Dortmund, and the University of Cologne. She is interested in various research topics ranging from social robotics and intelligent social agents to attitudes, sexual violence, sexism, dehumanization, and prejudice reduction.

Elisabeth André is a full professor of computer science at Augsburg University and chair of the Laboratory for Human Centered Multimedia, Augsburg University, Augsburg, Germany. Earlier, she worked as a principal researcher at DFKI GmbH, where she led various academic and industrial projects in the area of intelligent user interfaces. In summer 2007, she was nominated a Fellow of the Alcatel-Lucent Foundation for Communications Research. In 2010, she was elected a member of the prestigious German Academy of Sciences Leopoldina and the Academy of Europe. Her research interests include affective computing, intelligent multimodal interfaces, and embodied agents.