Effects of Gender Stereotypes on Trust and Likability in Spoken Human-Robot Interaction

Matthias Kraus, Johannes Kraus, Martin Baumann, Wolfgang Minker
Institute of Communications Engineering, Department of Human Factors
University of Ulm, Germany
{matthias.kraus, johannes.kraus, martin.baumann, wolfgang.minker}@uni-ulm.de

Abstract

As robots enter more and more areas of everyday life, it becomes necessary for them to interact in an understandable and trustworthy way. In many regards this requires a human-like interaction pattern. This research investigates the influence of gender stereotypes on trust and likability of humanoid robots. In this endeavor, explicit (name and voice) and implicit gender (personality) of robots have been manipulated along with the stereotypicality of a task. 40 participants interacted with a NAO robot to gain feedback on a task they were working on and rated the perception of the robot cooperation partner. While no gender stereotypes were found for the explicit gender, implicit gender showed a strong effect on trust and likability in the stereotypical male task. Participants trusted the male robot more and rated it as more reliable and competent than the female personality robot, while the female robot was perceived as more likable. These findings indicate that for gender stereotypes in robot interaction a differentiation between explicit and implicit stereotypical features have to be drawn and that the task context needs consideration. Future research may look into situational variables that drive stereotypification in human-robot interaction.

Keywords: Human-Robot Interaction, Trust, Gender Stereotypes

1. Introduction

Due to the recent rapid growth of artificial intelligence, robots will no longer be mere substitutions for labor-intensive and repetitive tasks like for example in industrial automation. As they manage to handle more and more complex tasks, they will soon find their way into our daily lives and become our companions. In order to be accepted and trusted as a *social* companion, future robots need to possess human-like characteristics and social skills. Hence, several studies investigated the effects of gender and personality features in human-robot interaction (Jonsson and Dahlbäck, 2013; Park et al., 2012; Tay et al., 2014; Löckenhoff et al., 2014). Exemplarily, the results Park et al. (2012) showed that people who interacted with a robot of similar personality (extraversion - introversion) felt more comfortable than those who engaged with a robot of different personality. According to Joosse et al. (2013) when considering the robot's gender and personality, the task context in which the interaction takes place is also of great importance. They found that preferences for robot personalities depend on the robot's role and the stereotype perception people hold for certain tasks. For example, extroverted participants perceived similarity attraction when the robot was a tour guide, while introverted subjects perceived similarity attraction when the robot was a cleaner. Hence, the consistency of a robot's behavior and the context of a given task or role seems to play a role for perception of the robot.

In the scope of this work, we integrated gender stereotypical communication style (male - female personality traits) and gender typical characteristics (male - female voice) into a spoken dialogue robot-based assistance system and investigated interaction in a stereotypical female (baby health care) and stereotypical male task (taxi ordering) scenario. Furthermore, we considered the effects of the different configurations on trust, and on other related variables like predictability or competence, as well as on the likability of the robot. According to a stereotype perspective, the emphasis of this study was the comparison of matching (e.g. male personality traits - male task scenario) and non-matching conditions (e.g. female voice - male task scenario).

The outline of this paper is as follows: in Section 2 an overview of related work is provided. Particularly, we will focus on general gender traits and existing stereotypes as well as on trust and likability in human-machine interaction. Section 3 deals with the integration of gender stereotypes in spoken human-robot dialogue using the humanoid robot Aldebaran NAO. Here, we describe the implementation of explicit (name and voice) and implicit (personality traits) gender characteristics in the interaction. Furthermore, the experimental setup is described in detail. In Section 4 the results of the study are presented. Subsequently, the paper is concluded in Section 5 with a discussion of the found results and an outlook to future work.

2. Related Work

2.1. Gender Traits and Stereotypes

There are different personality traits which are attributed uniformly to men or women (Huddy and Terkildsen, 1993). Löckenhoff et al. (2014) investigated the perceived gender differences in five-factor personality traits (Big-Five) in terms of different nations and age groups and whether theses perceived differences reflect already assessed gender differences in personality accurately. They concluded that women were perceived as more open, more conscientious, and more agreeable than men. There were also higher ratings for women in warmth (facet of extraversion) and anxiety, depression, self-consciousness and vulnerability (facets of neuroticism), whereas men were higher rated in terms of excitement seeking and assertiveness. These perceived differences were consistent across age groups and nations and reflected closely the assessed gender differences in personality (self- and observer-ratings).

Otterbacher et al. (2017) investigated the strength and content of gender stereotypes in image search algorithms using a trait adjective list based on Abele et al. (2008). The developed trait list contains stereotypically adjectives describing both female and male personality traits. For example, representative adjectives for the female trait warmth are attributes like emotional, warm or fair. In contrast, the male trait competence was described by adjectives like competent, consistent or intelligent. The results are in line with common ground of social psychological research findings, where female traits are more correlated to warmth and male traits rather correlated to competence (Ruble et al., 1998). Otterbacher et al. (2017) pointed out that regardless which wording participants use for their descriptions, the two dimensions (warmth and competence) can be found in a cluster analysis. These findings are in line with Paulhus and Trapnell (2008) who postulate the Big Two, a concept paralleling the Big Five for gender-specific traits. The Big Two are communion (warmth) and agency (competence) (Bakan, 1966). Abele and Bruckmüller (2011) mention similar classifications: the adjectives warmth, friendliness, trustworthiness (female traits) and ambitious, competent, self-confident (male) describe typical gender traits.

2.2. Trust and Likability in Human-Machine Interaction

Trust has been investigated in regard to human-machine interaction over the last 20 years. It has been found to play a role in many different professional and everyday domains like automated driving (Hergeth et al., 2016) or online shopping (Gefen et al., 2003). Trust can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability" (Lee and See, 2004, p.51). In human-robot teamwork, trust reflects the person's willingness to accept information and suggestions provided by the robot (Hancock et al., 2011).

In human-human interaction, the likability of the social interaction partner can be judged within seconds and this judgment has been found to be a strong predictor for interaction with a robot (Clark and Rutter, 1985; Robbins and DeNisi, 1994). Thus, besides trust likability of the robot acting as a social interaction partner (Bartneck et al., 2008) seemed to be worthwhile considering in our study as a dependent variable.

2.3. Hypotheses

The two papers Tay et al. (2014) and Tay et al. (2013) investigated the impact of occupational roles (security - health care), personality (extroverted - introverted), and gender (male - female) of a robot on user acceptance. Personality was manipulated by non-verbal cues, while the gender of the robot was varied by gender-specific names and voices (male - female). In these experiments subjects accepted robots with conformed gender and personality to the respective role stereotype more. In contrast to these studies, we choose a more content-related manipulation of personality traits and stereotypes to the robot with verbal cues. Additionally, in this research a broader range of outcome variables was measured, e.g., trust in the robot.

According to our preceding considerations, we tested the following hypotheses:

H1: Implicit male gender of the robot lead to higher ratings of trust (H1.1), reliability (H1.2), predictability (H1.3), and competence (H1.4) in the male task scenario (matching-condition) as compared to female personality traits.

H2: Implicit female gender of the robot lead to higher ratings of trust (H2.1), reliability (H2.2), predictability (H2.3), and competence (H2.4) in the female task scenario (matching-condition) as compared to male personality traits.

H3: Explicit male gender of the robot leads to higher ratings of trust (H3.1), reliability (H3.2), predictability (H3.3), and competence (H3.4) in the male task scenario (matching-condition) as compared to a female voice.

H4: Explicit female gender of the robot leads to higher ratings of trust (H4.1), reliability (H4.2), predictability (H4.3), and competence (H4.4) in the female task scenario (matching-condition) as compared to a male voice.

H5: Female personality traits of the robot lead to higher ratings of likability in the female (H5.1) and the male task scenario (H5.2) as compared to male personality traits.

3. Integration of Gender Stereotypes in Spoken Human-Robot Dialogue

For integrating gender stereotypes in human-robot interaction, the humanoid robot Aldebaran NAO produced by the SoftBank Robotics Group was set up as a dialogue partner in two stereotypical scenarios. A main feature of the NAO robot is the ability to engage in a multimodal interaction with users through speech, gestures and gaze. Since this work focuses on verbal interaction, only the speech capability of NAO was used to manipulate the variables of interest, while the other features were fixed in "autonomous life" [1] mode. Gender of the NAO robot was manipulated at two ways reflecting the two independent variables of the study: first, an explicit gender manipulation by typical male vs. female names and voices, and second, an implicit gender manipulation by gender specific personality traits through modeling gender specific wording and communication style of the robot's utterances. Furthermore, as a third independent variable gender stereotypical tasks were included in the study. For a stereotypical female area of work the health care and for a typical male domain the taxi domain were chosen. In each scenario, users had to solve several scripted tasks in a Wizard-of-Oz setup, where NAO evaluated the users' action by providing either positive oder negative spoken feedback.

[1] In this mode, NAO is in an upright position and wags slightly waiting for input. The head of NAO is oriented to nearest recognised person.

3.1. Design of Gender-Specific Personality Traits, Scenarios and Voice

In order to manipulate explicit gender of the robot's voice, the designed utterances of the robot were transformed into spoken language using IBM Watson Text to Speech Demo (available at http://text-to-speech-demo.mybluemix.net/) with the two available German voices (female - male) and played to the user.

The implicit gender (personality) of the robot was modeled according to the findings of Abele and Bruckmüller (2011). Hence, the robot's stereotypical male utterances were designed to be *dominant*, *confident* and *assertive*, whereas the stereotypical female utterances were *agreeable* and *warm*. In order to make the two conditions comparable, the meaning and general content of the utterances were kept constant in both conditions. In the following an exemplary utterance for negative feedback for each condition in the taxi scenario is depicted:

NAO male personality

> *"This is not correct. Due to several road construction works and traffic jams taxis C,D and A would require more time to reach the destination than taxi B. Keep this in mind for the future, please"*

NAO female personality

> *"Unfortunately, you picked the wrong answer. There is a better solution. Taxi C, D and A would have force themselves through far more road construction works and traffic jams than taxi B. However, I'm sure that you are conscious of that by now. So don't worry."*

In the female personality condition, the feedback was more submissive and tolerant in case of wrong answers by the user, which was intended to convey an impression of warmth and agreeableness of the robot. Contrary, the male NAO's feedback was rather strict and uncommunicative in order to give the user the perception of a competent and self-confident dialogue partner. The selection of the respective scenarios was derived from the findings of Williams and Best (1977). In the taxi ordering domain, the user was told to imagine that he would be working for a taxi company and was responsible for taking requests and sending the right amount of taxis to the correct places. In order to act in the company's best interests, the user should choose the most profitable option but also ensure satisfaction of the users' needs. As this task strongly relates to the adjectives logical, rational and methodical which were deemed stereotypical male attributes, the taxi ordering scenario was selected as male domain. In the baby health care domain, the user was told to be responsible for the nursing of a baby and had to make decisions on the appropriate handling of several situations, like the feeding or bathing of the baby. This task is strongly associated with the need of social skills and warmth. Additionally, people expected emotional competence and sensitive handling according to Williams and Best (1977). Therefore, this scenario was chosen as a task for a stereotypical female domain.

3.2. Experimental Setup

3.2.1. Participants

40 German participants were recruited at Ulm University and received 6 Euro in return for their participation. Two participants had to be excluded due to technical issues with the NAO robot. As a consequence, the data of 38 participants could be further analysed. 12 subjects were female, 26 were male, and the age ranged from 19 to 50 years with a mean age of M = 26.34 (SD = 7.38).

3.2.2. Experimental Design and Manipulations

	Task Order Scenario	
Gender Traits	**Baby Health Care**	**Taxi Ordering**
	Male voice Male personality	Male voice Male personality
	Female voice Male personality	Female voice Male personality
	Male voice Female personality	Male voice Female personality
	Female voice Female personality	Female voice Female personality

Table 1: Experimental conditions of the study: eight study groups from the combination of three independent variables with 2 levels.

In our experiment, we assessed trust, reliability, predictability, competence, acceptance and likability of the robot as dependent variables. Each variable was measured with items from established and with items from established and validated scales that were translated into German and slightly modified for content and study context. All scales were assessed with a 7-point Likert scale ranging from 1="totally disagree" to 7="totally agree". A 2x2x2 mixed factorial experimental design was conducted with the robot's implicit and explicit genders as between-independent and the task scenarios (baby - taxi) as within-independent variables. Additionally, the order of the scenarios was randomised leading to an overall of eight study groups to which participants were randomly assigned (see Table 1). As a cover story the subjects were told that they had to test a new robotic assistant for training which will be deployed in the company they were allegedly working for. Specifically, they had to work on five questions per scenario which referred to the specific task scenario at hand. The respective tasks were presented on a laptop computer, which was also used by the participants to fill in the questionnaires. The answers the subjects had to give were predefined, i.e. they had to give either a correct or a wrong answer per question corresponding to a script they were handed. Each time, the robot provided positive or negative feedback through speech after the participant uttered the scripted answer. As a Wizard-of-Oz paradigm was applied in this experiment, the system's feedback was triggered remotely by the wizard from an external desktop.

3.2.3. Experimental Procedures

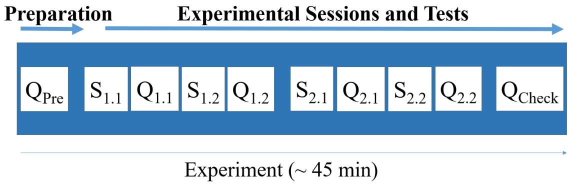

Figure 1: Procedure of experiment. After two scenario sessions (S), dependent variables were assessed. The final questionnaire of the experiment gives insights to the manipulation of the independent variables.

After the welcome procedure, the subjects were provided with first instructions and details about the study. As a second step, they had to read and sign the informed consent, and had to fill a pre-test questionnaire about their personality. Before the first interaction cycle, they received information about the tasks and the procedure of the study. This included details about the feedback functionality of NAO and about the task to rate the interaction with the robot. At the beginning of each scenario, NAO introduced him- or herself. Subsequently, the participant had to work on the first two questions, fill in a questionnaire to assess the dependent variables, and continue with the remaining three questions on the respective task. The scenario was then ended with the completion of the same questionnaire in order to gain a more robust evaluation. The same procedure was repeated for the second task scenario. In conclusion, a last questionnaire containing manipulation checks, demographics and possible confounding variables had to be filled out.

4. Results

All scales used to assess the dependent variables showed acceptable Cronbach's Alphas (all alphas $>$ .82; scales ranging from a minimum of four to a maximum of seven items). No significant outliers were found. To rule out confounding group differences for the study conditions, we controlled for the subject's acceptance of technical and electronic devices, their preliminary trust towards assistive systems, their experience with spoken dialogue systems, as well as their attitude towards (humanoid) robots as confounding variables. All 2x2 ANOVAs testing these differences did not show any significant group differences (all p-values $>$ .05/3). In addition, participants age and gender was similarly distributed in the different experimental groups.

A manipulation check was conducted with a series of independent t-tests concerning the gender and quality of the used voices, gender-specific personalities of the robot, as well as gender-stereotypical task scenarios. Therefore, we tested each gender-manipulated variable on perceived masculinity and femininity by the user. In both scenarios the manipulation of the gender-specific voices worked (all p-values $<$ .05). The manipulation of the gender-specific personality of the robot also worked in both scenarios (see Table 2). However, contrary to the generated "empathic" female personality being recognised as more female in both scenarios (all p-values $<$ .05), the generation of a "dominant" male personality was not perceived as significantly more male by the participants. The taxi ordering and baby healthcare scenario were correctly rated as a typically male and respectively female task ($p < .05$). In order to guarantee comparability of the two voice conditions exopect for perceived gender, six additional characteristics of the voices were assessed. Figure 2 shows the measured quality of each voice using six subjective metrics. Except for the naturalness and pleasantness of the voice ("The voice is natural" $[F(1,36) = 8,155, p < .05]$; "The robot is a pleasant conversation partner" $[F(1,36) = 2,502, p < .05]$) there were no significant differences.

We measured the dependent variables at two times during the interaction with the robot. The first time of measurement was to assess the initial reaction towards the robot while the second time of measurement sought to measure a stabilised evaluation after some interaction with the robot was experienced. Following the reasoning that trust and the assessment of its components competence, reliability and predictability fluctuates during early interaction with a new system (Lee and See, 2004; Hoff and Bashir, 2015), for these dependent variables only the averaged scale values of

	Gender Personality Manipulation Check			
	Taxi		**Baby**	
	Male *M (SD)*	Female *M (SD)*	Male *M (SD)*	Female *M (SD)*
Male Personality Check	4.80 (1.49)	4.54 (1.01)	5.06 (1.06)	4.62 (1.13)
Female Personality Check	2.62 (1.11)	4.25 (1.12)	2.99 (1.31)	4.57 (1.03)

Table 2: Descriptive statistics of the manipulation check regarding the different personality traits of NAO.

	Gender Personality			
	Taxi		**Baby**	
	Male *M (SD)*	Female *M (SD)*	Male *M (SD)*	Female *M (SD)*
Trust	5.57 (.67)	4.95 (1.13)	4.90 (1.02)	5.00 (1.11)
Reliability	5.61 (.71)	4.96 (1.39)	4.98 (1.03)	5.09 (1.05)
Predictability	5.86 (.88)	4.98 (1.54)	4.99 (1.28)	4.90 (1.34)
Acceptance	5.05 (.90)	5.02 (1.27)	4.96 (1.06)	5.09 (1.29)
Likability	4.04 (.94)	5.15 (1.23)	4.51 (.74)	5.24 (1.09)
Competence	5.89 (.77)	5.14 (1.34)	5.59 (1.03)	5.34 (1.30)

Table 3: Descriptive statistics with reference to gender personality and task scenario.

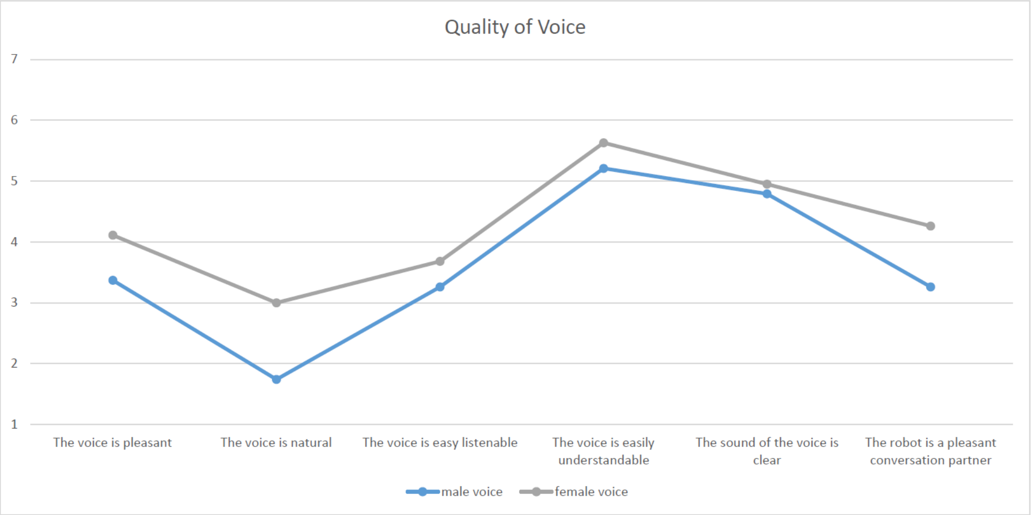

Figure 2: Diagram showing the rated scales for the measurement of quality of the used voices. The quality was rated on a 7-point Likert scale for six items.

the second time measurement was included in the analysis. For acceptance and likability an aggregate of the two points of measurement was used.

As a starting point, three-way mixed ANOVAs for all dependent variables were conducted. The three-way interaction of all independent variables did not yield promising results (all $p > .05$). Therefore, it is not further considered in the following. Similarly, all interactions and main effects including the explicit gender missed significance. This lead to the decision to leave this factor out of further analyses for this study (all $p > .05$). For all further analyses, the dependent variable acceptance showed no significant differences - neither for the gender manipulations nor for the different scenarios (all $p > .05$). The two-way interaction of the factors scenario (babycare vs. taxi) and implicit gender were significant for trust $[F(1, 34) = 5.968, p < .05]$ and predictability $[F(1, 34) = 4.553, p < .05]$. For both variables the robot was significantly rated higher for the male personality as compared to the female personality in the taxi ordering scenario (see Table 3 for detailed descriptives). Although the p-values for reliability $[F(1, 34) = 3.454, p = .072]$, competence $[F(1, 34) = 2.676, p = .111]$ and liking $[F(1, 34) = 2.475, p = .125]$ missed significance, in face of the comparatively low power of this analyses and the small sample size of this study, we conducted post-hoc t-tests to further inspect these results for all dependent variables.

These test revealed no significant effect for competence in both scenarios (all $p > .05$). For competence there was no significant difference found for the implicit gender in the baby scenario, but in the taxi scenario a robot with a male personality was rated significantly more competent $[t(36) = 2.047, p < .05]$. Conversely, likability was rated significantly higher for the female personality of the robot in both the baby care $[t(36) = 2.352, p < .05]$ and the taxi ordering scenario $[t(36) = 3.064, p < .05]$. For visualisation, the dependent variables are presented for both tasks in Figure 3.

5. Discussion and Future Work

The study provides evidence that gender stereotypes can indeed be replicated in the context of human-robot interaction and that their consideration in spoken dialogue systems can positively affect the human-machine interaction.

The results showed that the personality condition has a major effect on the interaction within the stereotypically male-designed taxi ordering domain. The male personality was perceived more trustworthy, reliable, competent and predictable than NAO's female personality condition (hypothesis H1 verified; H1.2 not significant). In the stereotypically female-designed baby health care domain our assumption to expect the reversed effect was not validated and showed only a slight tendency regarding trust and reliability (hypothesis H2 rejected). An explanation why the expected effects could not be observed in the female domain poses the design of the scenario itself.

The baby health care scenario was designed in accordance with attributes like emotional competence, warmth and social skills. As the study was conducted in a highly fictionous environment and the subjects had to make scripted decisions from a rather distant point of view, participants possibly could not form an emotional bond to the task and did not require the robot to be warm or emotionally competent. Another explanation could be the rather masculine appearance of NAO itself. As visual cues play an important role in the perception of gender stereotypes (Hall, 1978), the masculinity of NAO could have had an negative effect on the female personality condition and the perceived gender of the robot.

Besides, the fact that the manipulation of the male personality of NAO did not work as well as the manipulation of the female personality traits may have contributed to a smaller difference in the perception of the robot. This especially seemed to hold true for the female task scenario, which could not be pushed by the masculinity of the task and hence resulted in almost no visible distinction.

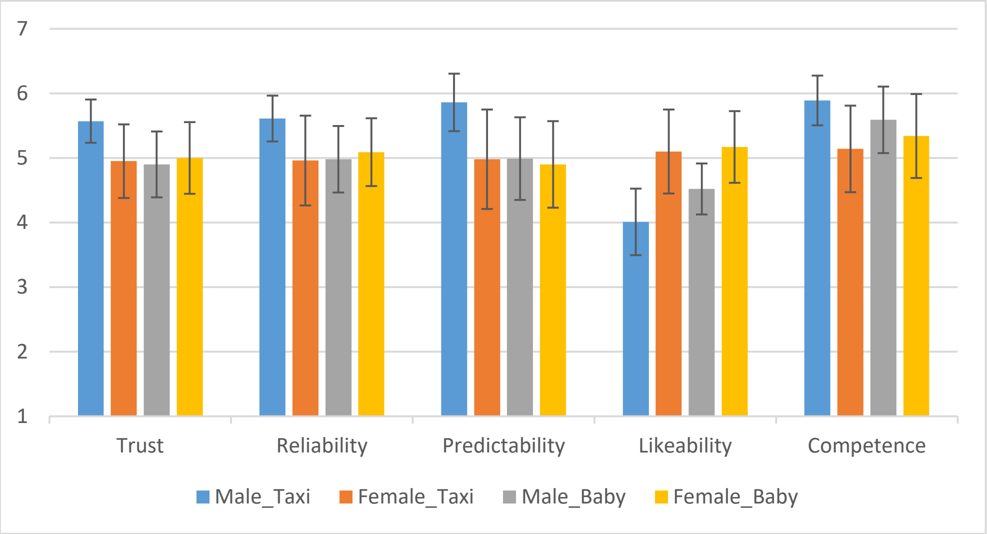

Figure 3: Bar charts for the variables trust, reliability, predictability, likability and competence. The terms “Male” and “Female” refer to the implicit gender of the robot, while results are provided for both task scenarios. Mean values can be taken from Table 3.

Surprisingly, the voice condition had no effect on the interaction in the stereotypical domains (hypothesis H3 and H4 rejected). Hence, it could be possible that personality traits have a larger impact on gender stereotypes than a female or male sounding voice. In order to prove this claim further investigation is necessary. Measurement of the quality of the used voices showed a significant difference in naturalness of the male and female voice. However, even though the female voice was perceived as more natural, this circumstance did not lead to measurable effects in the experiment. As one would expect that a more natural voice would result in more trust in the interlocutor, this forms a kind of paradox, but reinforces the impression that personality traits have a bigger impact than voice alone.

Furthermore, the study revealed that gender personalities effect significantly the likability of human-machine interaction. Since the female personality traits were designed more agreeable and warm, the female personality was perceived more likable in both scenarios (H5 verified; H5.2 not significant). An interesting finding was that trust and likability did not always seem to correlate, but it seems that robots of different characteristics seem to be more trustworthy in different domains - despite the fact that robots of a more female personality are liked more irrespective of the domain under investigation. Hence, a trustworthy system is not necessarily more likable. Is it possible to maintain both trustworthiness and likability in spoken human-robot interaction? Further research should provide more insight on this coherence. Additionally, likability and acceptance do not seem to interdepend gender-personality wise. Thus, robots of all genders are accepted but may show differences in likability depending on the personality. This implicates for human-robot interaction designers to not only concentrate on a high likability of their systems because rather emotionally cold systems may be equally accepted.

In order to avoid the drawbacks of this experimental setup in future work, we plan on conducting the study without a humanoid robot and turn to a embedded voice-only control system like Amazon Echo. In doing so, we could eliminate the visual side-effects of perceived gender stereotypes. Furthermore, a more sophisticated approach for modeling the female domain is necessary. Therefore, the future participants should not work on scripted tasks, but make their own decisions in a more realistic environment. Hence, subjects could get emotionally involved in the task. Most importantly, a higher sample size and a better manipulation of the male personality characteristics are necessary to give further insight in this quite interesting topic.

6. Bibliographical References

Abele, A. E. and Bruckmüller, S. (2011). The bigger one of the “big two”? preferential processing of communal information. *Journal of Experimental Social Psychology*, 47(5):935–948.

Abele, A. E., Uchronski, M., Suitner, C., and Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology*, 38(7):1202–1217.

Bakan, D. (1966). The duality of human existence: An essay on psychology and religion.

Bartneck, C., Croft, E., and Kulic, D. (2008). Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In *Metrics for HRI workshop, technical report*, volume 471, pages 37–44.

Clark, N. K. and Rutter, D. R. (1985). Social categorization, visual cues, and social judgements. *European Journal of Social Psychology*, 15(1):105–119.

Gefen, D., Karahanna, E., and Straub, D. W. (2003). Trust and tam in online shopping: An integrated model. *MIS quarterly*, 27(1):51–90.

Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological bulletin*, 85(4):845.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, 53(5):517–527.

Hergeth, S., Lorenz, L., Vilimek, R., and Krems, J. F. (2016). Keep your scanners peeled: Gaze behavior as a measure of automation trust during highly automated driving. *Human factors*, 58(3):509–519.

Hoff, K. A. and Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434.

Huddy, L. and Terkildsen, N. (1993). Gender stereotypes and the perception of male and female candidates. *American Journal of Political Science*, pages 119–147.

Jonsson, M. and Dahlbäck, N. (2013). In-car information systems: Matching and mismatching personality of driver with personality of car voice. In *International Conference on Human-Computer Interaction*, pages 586–595. Springer.

Joosse, M., Lohse, M., Pérez, J. G., and Evers, V. (2013). What you do is who you are: The role of task context in perceived social robot personality. In *Robotics and automation (ICRA), 2013 IEEE international conference on*, pages 2134–2139. IEEE.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Löckenhoff, C. E., Chan, W., McCrae, R. R., De Fruyt, F., Jussim, L., De Bolle, M., Costa Jr, P. T., Sutin, A. R., Realo, A., Allik, J., et al. (2014). Gender stereotypes of personality: universal and accurate? *Journal of Cross-Cultural Psychology*, 45(5):675–694.

Otterbacher, J., Bates, J., and Clough, P. (2017). Competent men and warm women: Gender stereotypes and backlash in image search results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6620–6631. ACM.

Park, E., Jin, D., and del Pobil, A. P. (2012). The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9(2):35.

Paulhus, D. L. and Trapnell, P. D. (2008). Self-presentation of personality. *Handbook of personality psychology*, 19:492–517.

Robbins, T. L. and DeNisi, A. S. (1994). A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations. *Journal of Applied Psychology*, 79(3):341.

Ruble, D. N., Martin, C. L., and Berenbaum, S. A. (1998). Gender development. *Handbook of child psychology*.

Tay, B. T. C., Park, T., Jung, Y., Tan, Y. K., and Wong, A. H. Y. (2013). When stereotypes meet robots: The effect of gender stereotypes on people's acceptance of a security robot. In *International Conference on Engineering Psychology and Cognitive Ergonomics*, pages 261–270. Springer.

Tay, B., Jung, Y., and Park, T. (2014). When stereotypes meet robots: the double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behavior*, 38:75–84.

Williams, J. E. and Best, D. L. (1977). Sex stereotypes and trait favorability on the adjective check list. *Educational and Psychological Measurement*, 37(1):101–110.