Gender Stereotyping's Influence on the Perceived Competence of Siri and Co.

Claus-Peter H. Ernst
European Management School
c.ernst@ems.de

Nils Herm-Stapelberg
Johannes Gutenberg-Universität Mainz
hermstapelberg@uni-mainz.de

Abstract

Some users express frustration with regard to virtual assistants due to their lack of perceived competence. To address this negative perception, we believe that technology companies should be aware of gender stereotypes. More specifically, it has been shown that males are attributed with rational competence more often than females. Drawing from the CASA paradigm, which states that people regularly assign human traits to computers, we expect that this stereotype might also be present for virtual assistants, i.e., male-voice virtual assistants are perceived as being more competent than female-voice virtual assistants. We test this hypothesis by conducting a controlled experiment which simulates a realistic interaction with differently voiced virtual assistants. The results indicate that gender stereotypes indeed play a role in the perception of the interaction. Male-voiced assistants are perceived more competent than their female-voiced counterpart which has practical implications in the design and development of devices that utilize these assistants.

1. Introduction

In this current day and age, not only do people perform tasks, technologies do too. One recent development has been the rise of virtual assistants such as Alexa, Siri, and Cortana that use gendered voices to interact with users. These virtual assistants have been present on smartphones for a number of years and they perform tasks such as initiating calls, giving reminders about upcoming appointments or searching for information on the Web. In 2015, Amazon launched their smart-speaker line Echo, a speaker that included its virtual assistant Alexa, and landed a hit. Meanwhile, echo devices are regularly among the best-selling devices on Amazon, paving the way to the smart home

for brands such as Apple with its HomePod or Google with its Google Home.

However, some users have expressed frustration with regard to the virtual assistants, citing lacks of understanding, as well as reliability and accuracy problems [1], or, in other words, their lack of competence. Being or appearing competent is an important task-related quality to have. Competence can be defined as "[t]he set of ... explicit and tacit knowledge [and skills] that a ... [person or human-like technology] possesses that enables him or her [or it] to ... [do their respective tasks]" [2, p. 164].

In order to understand people's negative perception of virtual assistants' competence, we believe that technology companies should be aware of gender stereotypes, that is, "psychological traits ... that are believed to occur with differential frequency in ..." males and females [3, p. 11]. More specifically, users usually have the choice between male and female voices for their virtual assistants with the latter often being the default setting. Research suggests that males are attributed with adjectives that "carry the notion of rational competence" more often than females are [4, p. 452].[1] Since people are known to assign human traits to computers (also known as the CASA paradigm) [5], we expect that this stereotype might also be present for virtual assistants, resulting in different competence perceptions based on the perceived gender of the virtual assistant. As a result, we seek to contribute to the research question: Does gender stereotyping influence the perception of virtual assistants with regard to their perceived competence?

In order to evaluate our hypothesis, we conducted an experiment. More specifically, we assigned participants to two groups and asked them to make a virtual assistant perform multiple tasks by working through a provided list of eight requests. Whereas one group's virtual assistant answered with a female voice, the other group's virtual assistant answered with a male voice. Afterwards, the participants were asked to indicate how competent they perceive the virtual assistant to be.

---

[1] In order to address any misunderstanding, we would like to emphasize that this does not mean that females are in fact less competent than males. In fact, we do not believe that gender determines competence. Rather, in this study, we seek to build a

deeper understanding on how gender stereotyping may influence our everyday life.

HİCSS

Finally, we compared the competence levels perceived by both groups.

The paper is structured as follows: In the following section, we will introduce gender stereotyping as well as the CASA paradigm. After outlining our hypotheses and our research design, we will conclude our article with the limitations of our empirical study and the implications of our findings.

2. Theoretical background

2.1 Gender stereotyping

Generally, "[g]ender stereotypes are the psychological characteristics believed to be differentially associated with women and men" [6, p. 513]. Multiple studies have confirmed the presence of gender stereotypes across and within different cultures.

Williams and Best [7] conducted a cross-cultural, large-scale study about gender stereotypes in multiple countries from Africa, the Americas, Asia, Europe, and Oceania. More specifically, they asked university students to judge whether each one of 300 adjectives of the Adjective Check List [8] was more often associated with men, women, or not differentially associated by sex. Their results showed that there is a "high degree of pancultural similarity in the patterns of characteristics differentially associated with women and men in the 25 countries studied" [6, p. 514]. More specifically, while men were associated with adjectives that imply being strong, more active, dominant, autonomous, exhibiting, achieving, and enduring, women were associated with adjectives that imply being deferential, supporting, and nurturing.

Moreover, a re-analysis of the data of Williams and Best [7] by Williams et al. [6, p. 513] "found that the pancultural male stereotype was higher than the pancultural female stereotype on Extraversion, Conscientiousness, Emotional Stability, and Openness to Experience while the pancultural female stereotype was higher on Agreeableness".

Sherriffs and McKee [4] focused on the US and used a list of 200 adjectives to let respondents assign them to either males or females. Among their findings was that "[t]he stereotype of men as defined by adjectives which are applied to them by *both* sexes appears to involve three general notions. The first five adjectives ... imply a straightforward uninhibited social style. The next twelve words carry the notion of rational competence and ability. The remaining thirteen emphasize action, vigor, and effectiveness" [4, p. 452]. Correspondingly, the stereotype of women is by three "general clusters of adjectives. The first seven adjectives ... emphasize social skills and grace. The next nine adjectives ... imply warmth and emotional support ... The [final] four

words ... may represent ... concern for the significance or spiritual implications of experience" [4, pp. 453-454].

2.2 Computers as social actors

Previous research has shown that people ascribe human traits to computers or robots when interacting with them. Nass et al. [5] describe this phenomenon in their "computers are social actors" (CASA) paradigm. It explains how this process is unconsciously done by users even when obvious humanlike features such as faces are not present. Indeed, humans may assign personalities to inanimate objects during the interaction such as extroversion [9], dominance [10], or intelligence [11].

Although humanlike features are not necessary for users to ascribe human traits to computers or robots, Nass et al. [10] demonstrate that vocal cues, if they exist, influence users' gender assignment to a machine. The same is also true for facial cues (e.g., the length of hair) as well as differently pitched voices. More specifically, longer hair and higher pitched voices are usually associated with female robots [12, 13].

Moreover, results suggest that gender stereotyping plays a role in the evaluation of computers or robots. Indeed, different traits are ascribed to machines perceived as male or female, resulting also in differing perceptions with regard to the interaction. Tay et al. [14] (and similarly Carpenter et al. [15] and Eyssel and Hegel [13]) examined whether there are differences with regard to users' acceptance of "gendered" robots (i.e., robots that use voices as well as non-verbal cues) when matching stereotypical tasks to the robots' gender. They found that users preferred female robots in stereotypically female tasks such as health care, and male robots in stereotypically male tasks such as security — even though the robots showed no difference in their abilities. Similar results were found "under conditions in which all suggestions of gender were removed, with the sole exception of vocal cues" [10, p. 864]. More specifically, Nass et al. [10, p. 874] provide "evidence that vocal cues embedded in a machine are sufficient to evoke gender-based stereotypic responses" ...[, even in cases where] "all subjects were explicitly informed that the interaction was with a computer".

Overall, the CASA paradigm has been applied to and researched across many different domains such as navigation systems [16], e-commerce [17], and education [11, 18]. However, only few studies exist in the rather new area of smart (home) automation. Indeed, we are only aware of one study that links CASA with gender stereotyping in this particular context. Damen and Toh [19] found that users trust gendered automated agents more when they match their (stereotypical)

expectations, i.e., when using female agents in home settings and male agents in office settings.

Overall, it is still unclear what influence the perceived gender of virtual assistants has on the characteristics that users assign to them. We believe that characteristics that are believed to be differentially associated with women and men will also be differentially associated with female and male virtual assistants.

3. Research model

According to the CASA paradigm [5], people assign human traits to computers or robots when interacting with them. This has been confirmed in many studies over the years and, thus, is generally accepted to be true.

Moreover, vocal cues influence users' gender assignment to a machine [10]. In other words, if the voices of robots or computers are designed correspondingly (e.g., by using a higher or lower pitched voice [12, 13]), people perceive them as being either male or female.

"Gender stereotypes are the psychological characteristics believed to be differentially associated with women and men" [6, p. 513]. More specifically, numerous studies have shown that there are certain traits and behaviors that people assign differently to males and females [4, 6, 7].

In the context of virtual assistants, the most important quality trait is competence. More specifically, virtual assistants are there to get tasks done and, thus, being or appearing competent is one of the most important task-related qualities to have.

In the context of gender stereotyping, research suggests that males are attributed with adjectives that "carry the notion of rational competence" more often than females are [4, p. 452]. Drawing from the CASA paradigm, we expect that this perception of competence will also be present in the context of gendered virtual assistants. More specifically, we hypothesize the following: *Virtual assistants that are perceived as being male will be perceived as being more competent than those that are perceived as being female.*

4. Research design

4.1 Experiment

To test our hypothesis, we conducted an experiment in a German university and in German language using a between-subjects design [2]. We believe that results from a within-subject design would have been severely flawed in our context, since our participants would not have been blind to condition (i.e., the different voices of the virtual assistants) and, thus, memory effects,

sponsorship effects, and sequence effects would come up.

More specifically, in a laboratory setting we provided all of our participants with a list of eight task-related requests, which are common to ask virtual assistants for. The participants were told that we placed a smart speaker inside a non-transparent box and asked them to make its included virtual assistant to perform multiple tasks by working through the provided list of requests.

However, in fact, we had placed a simple Bluetooth speaker inside the box while making sure that the sound would not be affected negatively. After each request, a corresponding prerecorded answer was played by us through the Bluetooth speaker, creating the illusion of a smart speaker including a real virtual assistant for the participants.

We prerecorded the answers to the requests using Google Cloud TTS Service, which uses the Google Cloud Text-to-Speech API to convert text into natural human speech. More specifically, we used the German "WaveNet language C" with its default settings, which mimics a female voice, as well as "WaveNet language D" with an adjusted pitch of -4.00, which mimics a male voice. Whereas the "smart" speaker of one group was answering with the female voice, the one of the other group was answering with a male voice. Some answers were formulated in an imprecise way (3, 6), one was not answered at all (4), and one was answered incorrectly (2). This was done to provide a more realistic experience since virtual assistant are not always able to give a perfect answer to every question. Table 1 presents the list of provided requests and transcriptions of the prerecorded answers.

Overall, we choose this study design in order to avoid any brand-based bias and in order to ensure that all participants were getting the exact same answers. Moreover, we defined "Computer" as wake word in order to avoid any gender-based bias that might occur due a male or female name.

After the experiment, the participants were first asked to indicate how competent they perceive the virtual assistant to be. For that, we provided them with 9 reflective items such as "The virtual assistant is capable". More specifically, since the study was conducted in German and to not lose information of the original scale due to translation, we used a total of 9 German items, which were formulated based on the four-word competence scale of Price et al. [20] [cf. 21] (capable, efficient, organized, thorough). All of our items were measured using a seven-point Likert-type scale ranging from "strongly disagree" to "strongly agree" (the 9 German items can be found in Appendix A). Finally, the respondents were asked to indicate the virtual assistant's perceived gender on a seven-point

semantic differential scale with the genders male and female on the endpoints as a manipulation check.

4.2 Data collection

On June 11th, 2019, we recruited German-speaking students of a German university that were attending an introductory course of information systems by promising a raffle of four 25 € gift certificates from Amazon for the participants. All participants were randomly assigned to one of the two groups. In this manner, we obtained 26 completed online questionnaires. However, we had to remove 3 participants from our sample: 1 respondent obviously took the experiment not seriously, 1 respondent was not speaking German in a sufficient way, and 1 respondent did not indicate the gender of the virtual assistant in our manipulation check correctly, and were thus dropped from the analyses. As a result, we had a final sample size of 23 subjects (9 datasets in the male virtual assistant group and 14 datasets in the female virtual assistant group).

Table 2 presents the demographics and controls of our complete sample as well as of our two experimental groups – voice assistant with male voice (VA.Male) and voice assistant with female voice (VA.Female) – including age and gender. According to the results of one t-tests and one Fisher's exact tests, no significant difference was detected across groups in age and gender (see table 2). This suggests a successful random assignment of our participants to our experimental groups and supports the claim that the experimental groups did not differ with regard to these important covariates. This means we could rule out structural group differences as being the cause of any differences found in our dependent variable between groups.

Table 1. Virtual assistant requests and answers

No.	**Request** Computer, ...	**Answer**
1	What day is it?	It is Tuesday, June 11th 2019.
2	How many milliliters is 30 centiliters?	30 millimeters are 3 centimeters.
3	What is 30 percent of 69 €?	30 percent times 69 is 20.7.
4	How far is Mainz from Berlin?	I am sorry, I cannot help you with that.
5	How old is Barack Obama?	Barack Obama is 57 years old.
6	How many days is it until Christmas?	It is 6 months until Christmas.
7	Flip a coin.	It shows heads.
8	Is it going to rain tomorrow?	It does not look like it's going to rain tomorrow.

Table 2. Demographics and controls

	Range	**VA.Male** N=9	**VA.Female** N=14	**Complete Sample** N=23	**p**
Age Mean Standard deviation	19-30	21.11 1.10	22.43 2.66	21.91 2.82	.193[a]
Gender Male Female	-	2 7	6 8	8 15	.400[b]

a = Result of a t-test.
b = Result of a Fisher's exact test.

5. Results

5.1 Validation and descriptives

We conducted a confirmatory factor analysis (CFA) to assess the validity of our used perceived competence scale. Based on factor loadings and residual correlations, 3 items were removed from the data. The remaining 6 items achieved factor loadings of >.80. The average variance extracted (AVE) of the factor was .79. We could therefore assume convergent validity and use the questionnaire for analysis. Moreover, our resulting scale had a Cronbach's alpha of .909, emphasizing its reliability.

Table 3 presents the descriptives per remaining item (means, standard deviations, and medians) as well as the average composite score for perceived competence. The male voiced virtual assistant consistently scores higher on average for every item of the perceived competence scale. In total, the male voiced assistant outperforms the female variant on average by approx. .5 points (5.33 vs. 4.85) with rather low variation (standard deviation .19 vs. .32). The median score of the male condition shows similar tendencies (5.33 male voice vs. 4.86 female voice). Since the items were measured on a seven-point Likert scale, we can say that both groups were perceived as somewhat competent.

Table 3. Item and construct descriptives

Construct Item	VA.Male			VA.Female			Complete sample		
	Mean	Standard Deviation	Median	Mean	Standard Deviation	Median	Mean	Standard Deviation	Median
Perceived Competence*	**5.33**	**.67**	**5.33**	**4.85**	**1.06**	**4.75**	**5.04**	**.94**	**5.00**
Item 1	5.56	.68	5.00	5.29	1.03	5.00	5.39	.92	5.00
Item 2	5.33	.47	5.00	5.14	1.19	5.00	5.22	.98	5.00
Item 3	5.00	1.05	5.00	4.36	1.29	4.00	4.61	1.24	4.00
Item 4	5.33	.94	5.00	4.79	1.15	4.50	5.00	1.10	5.00
Item 5	5.22	.92	5.00	4.93	1.22	5.00	5.04	1.12	5.00
Item 6	5.56	1.17	5.00	4.57	1.18	4.00	4.96	1.27	5.00

*=composite score, normalized with item count (=6)

5.2 Hypothesis testing

We used the nonparametric Mann–Whitney U test to test for group differences [e.g., 22], since we cannot assume the necessary normal distribution to apply a standard t-test. Table 4 presents the results.

Table 4. Mann–Whitney U test

Construct	Comparison	z – value	Significance	Effect Size (Cohens d)
Perceived Competence	VA.Male/ VA.Female	-1.362	.091	.592

When we compared the levels of perceived competence of the group that evaluated the male-voiced virtual assistant with the group that evaluated the female-voiced virtual assistant, we found a significant difference (p<.10) with a medium effect size (Cohens d = .592) despite the limited sample size. Our male-voiced personal virtual assistant was therefore perceived to be more competent as the female-voiced assistant. We find support for our hypothesis and are able to show that gender stereotyping also applies to virtual assistants. This result provides additional support for CASA and shows that CASA also applies to modern applications such as personal virtual assistants that did not exist when CASA was first discussed. It is therefore necessary to consider the effect of human trait attribution to machines when designing personal virtual assistants.

6. Conclusion

In this article, we studied whether gender stereotyping influences the perception of virtual assistants with regard to their perceived competence. Based on an experiment with 23 participants, our study suggests that people perceive male-voiced virtual assistants to be more competent than female-voiced virtual assistants.

These findings hold important practical implications. More specifically, if users perceive a virtual assistant as incompetent, they may not use it.

Especially in the context of smart speakers, this has serious consequences since the integrated virtual assistants are usually the only way to use them at all. Since this is the case, the sales of smart speakers are strongly dependent on people's competence perception of the virtual assistants. Since we found that male-voiced virtual assistants are perceived to be more competent than female-voiced virtual assistants, it might prove beneficial for the companies to at least also offer male voices.

Though our findings hold important practical implications, our study has some limitations. First, it is only based on one male-voiced and one female-voiced virtual assistant in the context of smart speakers. Therefore, our results do not necessarily apply to all types of virtual assistants or to all usage contexts (e.g., smartphones). Second, since our sample consisted of students only, our findings may also not apply to other demographic groups (age, cultural background, education).

As a next step, we plan to expand our research and address its limitations. More specifically, we would like to roll out our survey to other countries and in particular survey people that are older and younger than those in our sample. Moreover, we also plan to replicate our findings in the context of virtual assistants on smartphones. Additionally, since in this study we focused on the perceived competence of the used voices and we found that CASA applies in this context, it will be interesting to examine other personality traits and perceptions such as likeability or credibility.

7. References

[1] PWC, Consumer Intelligence Series: Prepare for the Voice Revolution, White Paper, 2018.

[2] G. Bassellier, B.H. Reich, and I. Benbasat, "Information Technology Competence of Business Managers: A Definition and Research Model", Journal of Management Information Systems 17 (4), 2001, pp. 159-182.

[3] D.L. Best, Gender Stereotypes, in C.R. Ember and M. Ember (eds.), Encyclopedia of Sex and Gender: Men and Women in the World's Cultures, Kluwer Academic, New York, NY, 2004, pp. 11-23.

[4] A.C. Sherriffs, and J.P. McKee, "Qualitative Aspects of Beliefs About Men and Women", Journal of Personality 25 (4), 1957, pp. 451-464.

[5] C. Nass, J. Steuer, and E.R. Tauber, "Computers Are Social Actors", CHI 1994 Proceedings, pp. 72-78.

[6] J.E. Williams, R.C. Satterwhite, and D.L. Best, "Pancultural Gender Stereotypes Revisited: The Five Factor Model", Sex Roles 40 (7/8), 1999, pp. 513-525.

[7] J.E. Williams, and D.L. Best, Measuring Sex Stereotypes: A Multination Study, Sage, Thousand Oaks, CA, 1990.

[8] H.G. Gough, and A.B. Heilbrun, Adjective Check List Manual, Consulting Psychologists Press, Palo Alto, CA, 1983.

[9] C. Nass, and K.M. Lee, "Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction", CHI 2000 Proceedings, pp. 329-336.

[10] C. Nass, Y. Moon, and N. Green, "Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices", Journal of Applied Social Psychology 27 (10), 1997, pp. 864-876.

[11] J.-E.R. Lee, C. Nass, S.B. Brave, Y. Morishima, H. Nakajima, and R. Yamada, "The Case for Caring Colearners: The Effects of a Computer-Mediated Colearner Agent on Trust and Learning", Journal of Communication 57 (2), 2007, pp. 183-204.

[12] F. Eyssel, D. Kuchenbrandt, and S. Bobinger, "Effects of Anticipated Human-Robot Interaction and Predictability of Robot Behavior on Perceptions of Anthropomorphism", HRI 2011 Proceedings, pp. 61-67.

[13] F. Eyssel, and F. Hegel, "(S)He's Got the Look: Gender Stereotyping of Robots", Journal of Applied Social Psychology 42 (9), 2012, pp. 2213-2230.

[14] B. Tay, Y. Jung, and T. Park, "When Stereotypes Meet Robots: The Double-Edge Sword of Robot Gender and Personality in Human-Robot Interaction", Computers in Human Behavior 38 2014, pp. 75-84.

[15] J. Carpenter, J.M. Davis, N. Erwin-Stewart, T.R. Lee, J.D. Bransford, and N. Vye, "Gender Representation and Humanoid Robots Designed for Domestic Use", International Journal of Social Robotics 1 2009, pp. 261-265.

[16] D.R. Large, and G.E. Burnett, "The Effect of Different Navigation Voices on Trust and Attention While Using in-Vehicle Navigation Systems", Journal of Safety Research 49 2014, pp. 69-75.

[17] L. Qiu, and I. Benbasat, "Evaluating Antropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems", Journal of Management Information Systems 25 (4), 2009, pp. 145-182.

[18] C. Edwards, A. Edwards, B. Stoll, X. Lin, and N. Massey, "Evaluations of an Artificial Intelligence Instructor's Voice: Social Identity Theory in Human-Robot Interactions", Computers in Human Behavior 90 2019, pp. 357-362.

[19] N. Damen, and C. Toh, "Designing for Trust: Understanding the Role of Agent Gender and Location on User Perceptions of Trust in Home Automation", Journal of Mechanical Design 141 (6), 2019.

[20] L.L. Price, E.J. Arnould, and S.L. Deibler, "Consumers' Emotional Responses to Service Encounters: The Influence of the Service Provider", International Journal of Service Industry Management 6 (3), 1995, pp. 34-63.

[21] V. Zeithaml, A. Parasuraman, and L.L. Berry, Delivering Quality Service: Balancing Customer Perceptions and Expectations,, Free Press, New York, NY, 1990.

[22] A. Field, Discovering Statistics Using SPSS, Sage, Thousand Oaks, CA, 2009.

Appendix A – Perceived Competence Scale

Table 5. Perceived Competence Scale

Item	German Version
1	Der virtuelle Assistent ist kompetent.
2	Der virtuelle Assistent ist leistungsfähig.
3	Der virtuelle Assistent ist gründlich.
4	Der virtuelle Assistent ist planvoll.
5	Der virtuelle Assistent ist organisiert.
6	Der virtuelle Assistent ist begabt.
*	Der virtuelle Assistent ist fähig.
*	Der virtuelle Assistent ist sorgfältig.
*	Der virtuelle Assistent ist effizient.

Items marked with a * were removed during CFA