Tiako P.F. (ed) *Smart and Sustainable Applications*. Chronicle of Computing. OkIP
 https://doi.org/10.55432/978-1-6692-0005-5_12

Gender Bender Bot –The Effect of (Not) Following Gender Stereotypes in Conversational Agent Design

Fabian Hildebrandt[a]*

[a] *TUD Dresden University of Technology, Helmholtzstraße 10, 01069 Dresden, Germany*

ABSTRACT

Conversational agents (CAs), are increasingly becoming a common presence in our daily lives (e.g., Alexa or ChatGPT). Research has shown that designing CAs humanlike (e.g., through social cues such as a human name or avatar) results in a higher perception of humanness, which increases and service satisfaction by the user. In this context, CAs are exclusively designed to portray stereotypical genders (e.g., combining a female name and avatar). To challenge this quasi-standard, a 2x2 experiment (male/ female avatar x male/ female name) with 262 participants was conducted to investigate the effect of gender-mixed CAs. Our results indicate that users of CAs with a stereotypical gender report higher service satisfaction and a partially higher perception of social presence for male CAs. However, the results do not reveal any differences in perceived empathy and competence. Thus, it appears that users prefer stereotypically CAs, which is in sync with current practice.

Keywords: Conversational Agents, Anthropomorphism, Avatar, Name, Gender, Stereotypes

I. INTRODUCTION

Text- and voice-based conversational agents (CAs) (e.g., Amazon's Alexa, the chatbot of HelloFresh, or ChatGPT) have grown in popularity recently (Følstad & Brandtzaeg, 2017; Hughes et al., 2023). CAs are defined as "software-based systems designed to interact with humans using natural language" (p.1) (Feine, Morana, et al., 2019) . Growing CA usage is due to underlying technologies' rising maturity, particularly of natural language processing (Lester et al., 2004). Furthermore, commercially available CA development platforms abound, and many popular online services—such as Facebook and WhatsApp—integrate CAs (Khan, 2017). For instance, Facebook alone launched 400,000 chatbots in 2022 (Agarwal, 2022). CAs have a wide range of application, as they can respond to customer calls (Leviathan & Matias, 2018) and provide answers in FAQ searches (Vu et al., 2021).

The way companies and customers interact is gradually changing, as many interactions between customers and service employees are replaced by CAs (Barrett et al., 2015). The greatest benefit for companies to use CAs is a cost-efficient, convenient, and time-and-place independent service (Lester et al., 2004; McTear et al., 2016; Verhagen et al., 2014). In this context, an area gaining increasing attention remains the possibility of designing humanlike CAs by adding social cues, such as a human name and avatar (Feine, Gnewuch, et al., 2019; Seeger et al., 2018). This humanlike design increases perceived humanness (Gnewuch et al., 2018), which has been shown to increase users' perceived enjoyment (Diederich et al., 2020), perceived service satisfaction (Gnewuch et al., 2018), and perceived competence of the CA (Schmid et al., 2022).

In this regard, an impactful effect is users' perception of gender in a CA (Gong, 2008), which is caused by the combination of several cues (i.e., name, avatar, and description) (Bastiansen et al., 2022; Feine et al., 2020). Thereby, gender can be defined as a non-essential category which is repeatedly performed based on societal norms (Morgenroth & Ryan, 2018). An analysis by Feine et al. (Feine et al., 2020) revealed a gender bias in the market because approximately 70% of the CAs (i.e., chatbots in their case) have female gender. Overall, the perception of gender has a strong influence on users. From a theoretical perspective, the tendency to assign gender and follow stereotypes was deeply ingrained in human psychology, which does not only impact human-to-human but also human-to-computer interactions (Nass et al., 1997; Nowak & Fox, 2018). For example, research has shown that users tend to perceive male agents as more competent and female agents as more likable, while agents with no identifiable gender are perceived as neutral (Nunamaker et al., 2011; Pfeuffer et al., 2019). However, it remains unclear how a CA is perceived if its gender is not stereotypically designed (i.e., the cues of a CA indicate a male and a female gender simultaneously, which is here called "gender-mixed"). Against this background, this study addresses the following research question:

RQ: What is the impact of designing a CA with gender-mixed cues?

II. RESEARCH BACKGROUND

A. Conversational Agents in the Context of Service Systems

During the past few years, the capabilities of CAs have steadily improved (Gnewuch et al., 2018). In the 1970s, Joseph Weizenbaum unveiled the first CA (called ELIZA) (Weizenbaum, 1966), which had technological limitations but allowed text-based communication with users (Gnewuch et al., 2018). The capabilities of CAs have improved over time (McTear, 2017), enabling their deployment to support a wide range of service interactions (Barrett et al., 2015). Natural language processing has made it possible for CAs to often offer a comfortable and convenient user experience at any time and everywhere (Verhagen et al., 2014). Due to the widespread availability of technology for CA creation (such as Google Dialogflow) (Diederich et al., 2019), the use of CAs has significantly increased in practice. As a result, CAs are replacing human customer service representatives more and more frequently (Marinova et al., 2017). One key difference of CAs

* Corresponding author E-mail: fabian.hildebrandt@tu-dresden.de

Presented at the 2023 OkIP International Conference
on Automated and Intelligent Systems (CAIS)
in Oklahoma City, OK, USA, and Online, on October 2-5, 2023

compared to traditional graphical interfaces (such as online forms) is that they can be designed with humanlike features (e.g., having a humanlike avatar that may represent a gender or express emotions) (Feine, Morana, et al., 2019; Verhagen et al., 2014), which influence how customers perceive the interaction (Seeger et al., 2018).

B. Humanlike Design and the Role of Gender Perception of Conversational Agents

Anthropomorphism describes human predisposition to attribute human traits to inanimate objects, animals (such as cheerful monkeys), and fictional characters (such as SpongeBob) (Epley et al., 2007). The "Computers as Social Actors" (CASA) paradigm (Nass et al., 1994) and the "Social Response Theory" (Nass & Moon, 2000) further explain how encountering a computer (including CAs, such as Alexa or Siri) might lead to anthropomorphism - a spontaneous, pervasive, and powerful process (Hart et al., 2013). People tend to interact socially and communicatively with a computer (Nass et al., 1994), although they understand that it is a computer/machine (Nass & Moon, 2000)and instinctively apply social norms (Lang et al., 2013; Nass et al., 1994). Following, users treat CAs as they would treat another human, depending on how strongly they assign a CA humanness (Nass & Moon, 2000).

Seeger et al. (2018) proposed a framework for designing anthropomorphic CAs which postulates that the perception of humanness is influenced by three separate categories of social cues: human identity, verbal- and non-verbal communication. Beginning with human identity as an avatar (Gong, 2008), a gender (Nunamaker et al., 2011), and a human name (Cowell & Stanney, 2005). Verbal communication contains self-reference ("I") and self-disclosure (Schuetzler et al., 2018) as well as syntax and word variability (Seeger et al., 2018). Furthermore, empathy and support (McQuiggan & Lester, 2007), as well as praise through words ("that's good"), support this dimension. Moreover, the conversation is more comfortable and professional, if a CA is equipped with an introduction/ welcome message ("Hello, my name is …") (Cafaro et al., 2016). Non-verbal communication contains dynamic response time (Gnewuch et al., 2018) with associated blinking dots (de Visser et al., 2016), and the usage of emoticons (Wang et al., 2008).

Although CAs are only artifacts and, thus, cannot have a gender (e.g., Cortana's answer to the question about its gender is: "technically, I'm a cloud of infinitesimal data computation") (West et al., 2019), a specific gender can be perceived by the user through a name, an avatar, its voice, or a description (Feine et al., 2020). Currently, the majority of CAs on the market are have a female gender (Feine et al., 2020), indicating a biased perception of "female exclusively or female by default" CAs (West et al., 2019). In general, humans psychologically tend to apply gender stereotypes as soon as they perceive a specific gender, even to machines (i.e., CAs) (Nass et al., 1997; Nowak & Fox, 2018). For example, research has shown that users perceive agents assigned male as more competent, while agents assigned female radiate warmth, which in turn results in agents assigned male being perceived as more trustworthy (Pfeuffer et al., 2019). Moreover, CAs often have a gender to reinforce and perpetuate such stereotypes (Costa & Ribas, 2019). In this context, concerns have been raised about ethical issues of gender stereotyping in CA design (McDonnell & Baxter, 2019). For instance, it can be particularly harmful since many children interact with CAs, and gender stereotypes are reinforced through the CAs (Brahnam & De Angeli, 2012).

III. Hypothesis Development

A. Social Presence

Perceived social presence refers to the sense of human interaction and warmth that a user experiences when interacting with a human or computer (Short et al., 1976). Following CASA, a CA is perceived as a social actor if it is designed humanlike (Nass et al., 1994). Furthermore, according to the social response theory, social norms will be mindlessly applied (Nass & Moon, 2000). Hence, users perceive gender and apply it according to stereotypes when a CA is designed with corresponding social cues (Nass et al., 1997). Stereotypes can be defined as "a belief about a group of individuals" (Kanahara, 2006) and are used to make life easier and more efficient (Sherman et al., 1998). Therefore, CAs with stereotypical genders allow users to apply stereotypes to a CA ,which makes interactions more familiar (Pfeuffer et al., 2019). Following Van Hooijdonk and Liebrecht (Van Hooijdonk & Liebrecht, 2021), the users' perceived social presence of a CA can be increased by familiarity. Therefore, when the design of a CA has a stereotypical gender, it can be expected to be more familiar to users, which leads to a higher perception of social presence. Thus, the following hypothesis can be postulated:

H1: A CA with a stereotypical gender shows a higher level of perceived social presence than a gender-mixed CA.

B. Service Satisfaction

Service satisfaction is an individual's evaluation of the sum of all perceived aspects of a service (e.g., process and outcome) (Millán & Esteban, 2004). Following the expectation confirmation theory, service satisfaction is influenced by the expectation of the service compared to the outcome (Oliver, 1980). Combining the fact that most CAs have stereotypical genders (Feine et al., 2020) and service satisfaction is influenced by expectations regarding the CA (Oliver, 1980), it can be assumed that satisfaction with a CA is higher because users expect a stereotypical gendered CA. Therefore, the following hypothesis can be suggested:

H2: A CA with a stereotypical gender shows a higher level of perceived service satisfaction than a gender-mixed CA

C. Empathy

In the context of service interactions, empathy is the considerate, personalized attention a business gives its clients (Parasuraman et al., 1985). In a human-to-human customer service interaction, empathy is demonstrated through non-verbal cues such as nodding or consistent eye contact as well as vocal cues including showing comprehension of the client's request and sentiments (Gabbott & Hogg, 2001). Perceptions of a CAs humanness can have a significant influence on the perceived empathy of its user (Leite et al., 2013; McQuiggan & Lester, 2007). Prior research has shown that CAs assigned female are perceived as more likable and empathic. It can be hypothesized that if not exclusively stereotypical female gender attributes are

exhibited, this will also be perceived as less empathic. Thus, the following hypothesis can be formulated:

H3: A CA with a stereotypically female gender shows a higher level of perceived empathy than a gender-mixed CA.

D. Competence

Competence can be defined as a series of knowledge, skills, abilities, experiences, and behaviors, which leads to an effective performance of specific activities (Maaleki, 2018). Previous research revealed that CAs assigned male are perceived as more competent than CAs assigned female (Leite et al., 2013; McQuiggan & Lester, 2007). In this context, it can be assumed that CAs with no exclusive stereotypically male gender attributes will be perceived as less competent. Hence, the following hypothesis can be set up:

H4: A CA with a stereotypically male gender shows a higher level of perceived competence than a gender-mixed CA.

IV. Method

A. Participants

At a university in Germany, 262 undergrad students were recruited as participants via e-mail. All participants were incentivized via a raffle of three €10 online shopping gift cards. 22 individuals were dropped from our sample after failing the attention check questions. The resulting sample contained 67% women, 30% male, 1% diverse, and 2% with no answer. The mean age was 25 (18-58).

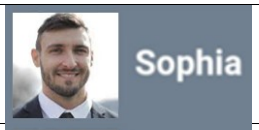
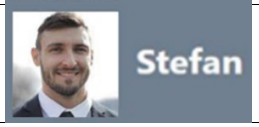
B. Task and Procedure

A structured dialogue was implemented (i.e., users had a clear task and all messages to the chatbot were task-related), as has been common in prior experiments (e.g., Diederich et al., 2020; Riquel et al., 2021). All participants were directed to a briefing page, where they received the same information about the experiment (context, tasks, and procedure). It was specifically made clear to the participants that they were interacting with a chatbot (and no actual human) and that the interaction had no real effect (i.e., the sports club joining process was fictive). Participants who successfully answered four comprehension questions were randomly assigned to a treatment. They completed a six-step process with the CA by entering a request (joining a sports club), a discipline (football, tennis, basketball, swimming, or gymnastics), a difficulty level (beginner, advanced, or expert), a weekday, a registration confirmation, and a phone number. Afterwards, the CA responded by providing a link to the survey.

C. Treatments

A between-subject design was applied, whereby every participant was randomly assigned to one of four chatbots (see Tab. 1). Except the treatment differences, the chatbots were technically identically implemented in Google Dialogflow and contained the same dialogue and training phrases. The chatbots were able to understand and process different wordings and could extract and repeat parameters.

TABLE I. TREATMENT DESIGN

		Avatar	
		Male	*Female*
Name	*Female*		
	Male		

All of the chatbots (see Fig. 1) were implemented with human identity cues, verbal cues, and non-verbal cues (Seeger et al., 2018). Human identity was represented by an avatar (professional male or professional female) (Gong, 2008), a human name (Stefan or Sophia) (Cowell & Stanney, 2005), and a gender (implied through a specific avatar and name) (Feine et al., 2020; Nunamaker et al., 2011). Verbal cues were given in a greeting ("Hello, […]") (Cafaro et al., 2016), self-reference ("I manage the registration page […]"), self-disclosure ("I am Stefan") (Schuetzler et al., 2018), variability in syntax/words (Seeger et al., 2018), and politeness ("Great …") (McQuiggan & Lester, 2007). The non-verbal cues we implemented were emojis (Wang et al., 2008) and dynamic response times (Gnewuch et al., 2018), with associated blinking dots to indicate that the chatbot is busy typing (de Visser et al., 2016). The design draws on other studies (e.g., Pfeuffer et al., 2019; Riquel et al., 2021) and follows the recommendations to not implement only a single group of cues (Seeger et al., 2021).

The four treatments differed in the implemented avatar and name of the chatbot (see Tab. 1). For this purpose, a traditionally gender indicating name (female: Sophia, male: Stefan) was crossed with a stereotypically male or female presenting avatar in each case (both avatars had a serious business look. The man had short hair and a beard; the woman had long hair and a braid). This resulted in two stereotypically male and female chatbots (the combination of a gender-matching male/ female name and avatar) and two gender mismatching chatbots (male name + female avatar/ female name + male avatar). One gender-mixed chatbot is visualized in Fig. 1.

Note the chatbot messages were translated from German to English.

Fig. 1. Gender-Mixed Chatbot (Male Avatar + Female Name)

D. Measures

The survey included questions regarding demographics (age and gender), social presence (Gefen & Straub, 1997) service

satisfaction (Verhagen et al., 2014), empathy (based on Yan et al., 2013), and competence (based on Wechsung et al., 2013). All constructs were measured on a 7-point Likert scale, except perceived competence, which we measured on a 9-point semantic differential scale to stay consistent with the items' original source (Wechsung et al., 2013). Table 2 reports on the items, factor loadings, Cronbach's α, composite reliability (CR), and average variance extracted (AVE) (convergent validity). All items show a sufficient factor loading > .60 (the lowest factor loading was .756) and therefore no item had to be dropped (Gefen & Straub, 2005). Furthermore, all constructs show sufficient reliability due to Cronbach's α > .80 (the lowest α was .871), CR >.80 (the lowest CR was .875), and sufficient convergent validity due to AVE > .50 (the lowest AVE was .637) (Urbach & Ahlemann, 2010). In summary, all measures exhibit sufficient reliability and validity.

V. Results

The data from the survey were analyzed using descriptive statistics to test the four hypotheses concerning the impact of gender-mixed CAs on the perceived social presence (H1), service satisfaction (H2), empathy (H3), and competence (H4). Hence, the means of each stereotypically male or female control group (male avatar/name (MAMN) and female avatar/name (FAFN)) was compared to both gender-mixed groups (male avatar/ female name (MAFN) and female avatar/ male name (FAMN)). Therefore, a Levene test was calculated first to test the variance homogeneity of constructs and then corresponding (Welch) t-tests were conducted. All tests were carried out using SPSS version 26 and are visualized in Tab. 3.

Summarizing, significance was found for social presence (MAMN-MAFN: p = .037, MAMN-FAMN: p = .032) and for service satisfaction (MAMN-MAFN: p = .029, FAFN-MAFN: p = .042) with each of a higher mean value of the stereotypical gender compared to the non-stereotypical gender CA. Therefore, partial support was found for hypotheses **H1** and **H2**, and no evidence was found for hypotheses **H3** and **H4**.

TABLE II. MEASURED CONSTRUCTS

Construct	Items	L.
Social Presence AVE =.698 CR = .920 α = .919	I felt a sense of human contact with the chatbot.	.857
	I felt a sense of personalness with the chatbot.	.850
	I felt a sense of sociability with the chatbot.	.813
	I felt a sense of human warmth with the chatbot.	.798
	I felt a sense of human sensitivity with the chatbot.	.860
Service Satisfaction AVE = .773 CR = .931 α = .930	I was satisfied with the overall interaction with the chatbot.	.922
	I was satisfied with the way the chatbot treated me.	.825
	I was satisfied with the chatbot's response.	.862
	I was overall satisfied with the chatbot.	.904
Empathy AVE = .637 CR = .875 α = .871	The chatbot gives users individual attention.	.724
	The chatbot gives users personal attention.	.858
	The chatbot works in the best interest of the user.	.817
	The chatbot understands the needs of its users.	.787
Competence AVE = .677 CR = .913 α = .910	Extremely insincere - Extremely sincere	.822
	Extremely dishonest - Extremely honest	.833
	Extremely incredible - Extremely credible	.869
	Extremely untrustworthy - Extremely trustworthy	.831
	Extremely incompetent - Extremely competent	.756
CR = Composite Reliability, AVE = Average Variance Extracted, L. = Loading *Note that all items were translated into German for the survey.*		

TABLE III. DESCRIPTIVE STATISTICS, LEVENE TESTS AND T-TEST RESULTS

		Group				Group Comparison	Levene Test	t-value (df)	*p*-value
		MAMN (n = 54)	FAFN (n = 63)	MAFN (n = 64)	FAMN (n = 59)				
Social Presence	Mean	3.778	3.429	3.141	3.105	MAMN-MAFN	F = 0.758, p = .386	-2.111 (116)	.037*
	SD	1.743	1.474	1.536	1.549	MAMN-FAMN	F = 0.769, p = .383	-2.172 (111)	.032*
						FAFN-MAFN	F = 1.131, p = .290	-1.078 (125)	.283
						FAFN-FAMN	F = 1.815, p = .180	-0.731 (122)	.466
						MAMN-FAFN	F = 2.952, p = .088	1.174 (115)	.243
						MAFN-FAMN	F = 0.169, p = .682	-0.288 (123)	.774
Service Satisfactio n	Mean	5.620	5.536	5.043	5.441	MAMN-MAFN	F = 3.710, p = .057	-2.214 (116)	.029*
	SD	1.242	1.300	1.540	1.467	MAMN-FAMN	F = 1.309, p = .255	-0.700 (111)	.486
						FAFN-MAFN	F = 8.915, p = .003**	-2.059 (115.613)	.042*
						FAFN-FAMN	F = 4.557, p = .035*	-0.204 (112.783)	.839
						MAMN-FAFN	F = 0.966, p = .328	0.386 (115)	.700
						MAFN-FAMN	F = 0.523, p = .471	-1.653 (123)	.101
Competen ce	Mean	6.352	6.435	6.103	6.339	MAMN-MAFN	F = 0.228, p = .634	-0.736 (116)	.463
	SD	1.881	1.724	1.785	1.731	MAMN-FAMN	F = 1.815, p = .180	-0.038 (111)	.970
						FAFN -MAFN	F = 0.100, p = .752	1.065 (125)	.289
						FAFN -FAMN	F = 0.112, p = .738	0.049 (122)	.961
						MAMN-FAFM	F = 0.610, p = .436	-0.249 (115)	.804
						MAFN-FAMN	F = 0.393, p = .532	-0.998 (123)	.320
Empathy	Mean	4.722	4.429	4.215	4.600	MAMN-MAFN	F = 0.000, p = .982	1.863 (116)	.065
	SD	1.484	1.315	1.465	1.437	MAMN-FAMN	F = 0.238, p = .626	0.454 (111)	.651
						FAFN-MAFN	F = 0.707, p = .402	-0.865 (125)	.389
						FAFN-FAMN	F = 0.134, p = .715	-0.882 (122)	.379
						MAMN-FAFN	F = 0.672, p = .414	1.135 (115)	.259
						MAFN-FAMN	F = 0.173, p = .678	-1.660 (123)	.099
*SD = Standard Deviation, M = Male, F = Female, A = Avatar, N = Name, *** p <.05, ** p <.01, *** p <.001***									

VI. Discussion

The results show that CAs with a stereotypical gender induce a higher perception of service satisfaction and CAs with a stereotypically male gender led to a higher perceived social presence by the users. Furthermore, no influence by gender-mixed CAs on perceived empathy and competence could be found. The obtained results may be due to different reasons. On the one hand, a non-gender-matching name and avatar can be perceived as an error by the developer (Mozafari et al., 2022). Hence, the general perception of a flawed CA decreases compared to a flawless one (e.g., the perception of humanness or service satisfaction (Bührke et al., 2021; Riquel et al., 2021)), and therefore the effect may have originated here. On the other hand, the perception of gender may also correlate with the gender of each participant (here 67% were female) (Marecek, 1995) or with their attitude towards sexism (i.e., is a stereotypical gender expected, and is there a general rejection of everything else) (Swim & Hyers, 2010). Therefore, in future research, a larger/ more gender-balanced sample should be recruited and questions about their attitude towards sexism should be included in the questionnaire. Practitioners should be encouraged to design their CA with a stereotypical gender (either exclusively male or female, depending on the context) to improve users' perception of the CA and therefore improve their service experience.

The typical limitations of experiment-based research apply to this work as well. Participants in the experiment did not utilize the CAs for a genuine, in-person commercial service, because it was performed in a controlled environment. As a result, the interaction did not affect expectations or objectives in real life. To this extent, the experiment traded realism for controllability. The specific task (joining a sports club) could explain the fact that no difference was found regarding perceived empathy and competence. Future research should build on these findings by applying them to real-world scenarios. Another limitation is that the sample was collected from German students, which could impact the results, but is generally not decisive (Compeau et al., 2012).

VII. Conclusion

Anthropomorphizing CAs through social cues is common in practice, as it increases perceived service satisfaction and competence (Gnewuch et al., 2018; Schmid et al., 2022). Hence, the gender of a CA is perceived through an avatar and a name (Feine et al., 2020). In previous research, however, gender was assumed to be stereotypical (e.g., Feine et al., 2020; Pfeuffer et al., 2019). For this reason, this study investigated how the perception of a CA is when gender is mixed/ not stereotypical (e.g., male avatar and female name). The results show that a CA with a stereotypical gender induces a higher perceived social presence and service satisfaction compared to a gender-mixed CA, and no difference in perceived empathy (compared to a stereotypically female CA) and competence (compared to a stereotypically male CA). The explanation for this would be, on the one hand, that people tend to assign stereotypes to machines (i.e., CAs) (Nass et al., 1997). Thus, the interaction becomes more familiar. On the other hand, people could perceive a gender-mixed CA as a mistake of the developer (Mozafari et al., 2022), which negatively influences their perception (Bührke et al., 2021). For practitioners, it is recommended to assign CAs a binary gender with a corresponding binary name and avatar.

Acknowledgment

I would like to thank Alfred Benedikt Brendel, Michael Hildebrandt, Kübra Nur Yagmur, and Annika Schulz for their support during this research project.

References

Agarwal, M. (2022). *A Facebook Messenger Chatbots Guide to Manage Customer Interaction*. https://www.socialpilot.co/facebook-marketing/facebook-messenger-chatbots

Barrett, M., Davidson, E., Prabhu, J., & Vargo, S. L. (2015). Service innovation in the digital age: Key contributions and future directions. *MIS Quarterly*, *39*(1), 135–154.

Bastiansen, M. H. A., Kroon, A. C., & Araujo, T. (2022). Female chatbots are helpful, male chatbots are competent? *Publizistik*, *67*, 601–623.

Brahnam, S., & De Angeli, A. (2012). Gender affordances of conversational agents. *Interacting with Computers*, *24*(3), 139–153.

Bührke, J., Brendel, A. B., Lichtenberg, S., Greve, M., & Mirbabaie, M. (2021). Is Making Mistakes Human? On the Perception of Typing Errors in Chatbot Communication. *Proceedings of the 54th Hawaii International Conference on System Sciences (HICSS)*.

Cafaro, A., Vilhjalmsson, H. H., & Bickmore, T. (2016). First impressions in human-agent virtual encounters. *ACM Transactions on Computer-Human Interaction*, *24*(4), 1–40.

Compeau, D., Marcolin, B., Kelley, H., & Higgins, C. (2012). Generalizability of information systems research using student subjects A reflection on our practices and recommendations for future research. *Information Systems Research*, *23*(4), 1093–1109.

Costa, P., & Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *Technoetic Arts*, *17*(1–2), 171–193.

Cowell, A. J., & Stanney, K. M. (2005). Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *International Journal of Human Computer Studies*, *62*(2), 281–306.

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349.

Diederich, S., Brendel, A. B., & Kolbe, L. (2019). Towards a Taxonomy of Platforms for Conversational Agent Design. *Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI)*, 1100–1114.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2020). Designing Anthropomorphic Enterprise Conversational Agents. *Business and Information Systems Engineering*, *62*, 193–209.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychological Review*, *114*(4), 864–886.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *International Journal of Human Computer Studies*, *132*(12), 138–161.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender Bias in Chatbot Design. *Chatbot Research and Design*, 79–93.

Feine, J., Morana, S., & Gnewuch, U. (2019). Measuring Service Encounter Satisfaction with Customer Service Chatbots using Sentiment Analysis. *Proceedings of the 14th International Conference on Wirtschaftsinformatik (WI)*, 1–11.

Følstad, A., & Brandtzaeg, P. B. (2017). Chatbots and the New World of HCI. *Interactions*, *24*(4), 38–42.

Gabbott, M., & Hogg, G. (2001). The Role of Non-verbal Communication in Service Encounters: A Conceptual Framework. *Journal of Marketing Management*.

Gefen, D., & Straub, D. (2005). A Practical Guide To Factorial Validity Using PLS-Graph: Tutorial And Annotated Example. *Communications of the Association for Information Systems*, *16*(1), 91–109.

Gefen, D., & Straub, D. W. (1997). Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. *MIS Quarterly*, *21*(4), 389–400.

Gnewuch, U., Morana, S., Adam, M. T. P., & Maedche, A. (2018). Faster Is

Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction. *Proceedings of the 26th European Conference on Information Systems (ECIS)*, 1–17.

Gong, L. (2008). How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, *24*(4), 1494–1509.

Hart, P. M., Jones, S. R., & Royne, M. B. (2013). The human lens: How anthropomorphic reasoning varies by product complexity and enhances personal value. *Journal of Marketing Management*, *29*(1–2), 105–121.

Hughes, L., Gauld, R., Grover, V., Hu, M., Edwards, J. S., Flavi, C., Janssen, M., Jones, P., Junglas, I., Khorana, S., & Kraus, S. (2023). "So What if ChatGPT Wrote It ?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy. *International Journal of Information Management*, *71*, 102642.

Kanahara, S. (2006). A review of the definitions of stereotype and a proposal for a progressional model. *Individual Differences Research*, *4*(5), 306–321.

Khan, R. (2017). Standardized Architecture for Conversational Agents a.k.a. ChatBots. *International Journal of Computer Trends and Technology*, *50*(2), 114–121.

Lang, H., Seufert, T., Klepsch, M., Minker, W., & Nothdurft, F. (2013). Are Computers Still Social Actors? *Conference on Human Factors in Computing Systems - Proceedings*, 859–864.

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human-robot relations. *International Journal of Human Computer Studies*.

Lester, J., Branting, K., & Mott, B. (2004). Conversational agents. In *The Practical Handbook of Internet Computing* (pp. 1–17).

Leviathan, Y., & Matias, Y. (2018). *Google Duplex: An AI System for Accomplishing Real-World Tasks Over the Phone*. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Maaleki, A. (2018). *The ARZESH Competency Model: Appraisal & Development Manager's Competency Model*. LAP LAMBERT Academic Publishing.

Marecek, J. (1995). Gender, politics, and psychology's ways of knowing. *American Psychologist*, *50*(3), 162–163.

Marinova, D., de Ruyter, K., Huang, M. H., Meuter, M. L., & Challagalla, G. (2017). Getting Smart: Learning From Technology-Empowered Frontline Interactions. *Journal of Service Research*, *20*(1), 29–42.

McDonnell, M., & Baxter, D. (2019). Chatbots and Gender Stereotyping. *Interacting with Computers*, *31*(2), 116–121.

McQuiggan, S. W., & Lester, J. C. (2007). Modeling and evaluating empathy in embodied companion agents. *International Journal of Human Computer Studies*, *65*(4), 348–360.

McTear, M. F. (2017). The rise of the conversational interface: A new kid on the block? *International Workshop on Future and Emerging Trends in Language Technology*, 38–49.

McTear, M. F., Callejas, Z., & Griol, D. (2016). Conversational Interfaces: Past and Present. In *The Conversational Interface* (pp. 51–72). Springer.

Millán, Á., & Esteban, Á. (2004). Development of a multiple-item scale for measuring customer satisfaction in travel agencies services. *Tourism Management*, *25*(5), 533–546.

Morgenroth, T., & Ryan, M. K. (2018). Gender Trouble in Social Psychology: How Can Butler's Work Inform Experimental Social Psychologists' Conceptualization of Gender? *Frontiers in Psychology*, *9*, 1320.

Mozafari, N., Schwede, M., Hammerschmidt, M., & Weiger, W. H. (2022). Claim success, but blame the bot? User reactions to service failure and recovery in interactions with humanoid service robots. *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues: A Journal of the Society for the Psychological Studies of Social Issues*, *56*(1), 81–103.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 72–78.

Nowak, K. L., & Fox, J. (2018). Avatars and computer-mediated communication: A review of the definitions, uses, and effects of digital representations. *Review of Communication Research*, *6*, 30–53.

Nunamaker, J., Derrick, D., Elkins, A., Burgoon, J., & Patton, M. (2011). Embodied conversational agent-based kiosk for automated interviewing. *Journal of Management Information Systems*, *28*(1), 17–48.

Oliver, R. L. (1980). A Cognitive Model of the Antecedents and Consequences of Satisfaction Decisions. *Journal of Marketing Research*, *17*(4), 460–469.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, *49*(4), 41–50.

Pfeuffer, N., Adam, M., Toutaoui, J., Hinz, O., & Benlian, A. (2019). Mr. And Mrs. Conversational agent - Gender stereotyping in judge-advisor systems and the role of egocentric bias. *Proceedings of the 40th International Conference on Information Systems (ICIS)*.

Riquel, J., Brendel, A. B., Hildebrandt, F., Greve, M., & Dennis, A. R. (2021). "F*** You !" – An Investigation of Humanness, Frustration, and Aggression in Conversational Agent Communication. *Proceedings of the 42nd International Conference on Information Systems (ICIS)*, 1–16.

Schmid, D., Staehelin, D., Bucher, A., Dolata, M., & Schwabe, G. (2022). Does Social Presence Increase Perceived Competence? *Proceedings of the ACM on Human-Computer Interaction*.

Schuetzler, R. M., Giboney, J. S., Grimes, G. M., & Nunamaker, J. F. (2018). The Influence of Conversational Agents on Socially Desirable Responding. *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS)*.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with Human-like Conversational Agents: Designing for Anthropomorphism. *Journal of the Association for Information Systems*, *22*(4), 1–58.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2018). Designing Anthropomorphic Conversational Agents: Development and Empirical Evaluation of a Design Framework. *Proceedings of the 39th International Conference on Information Systems (ICIS)*, 1–17.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype efficiency reconsidered: Encoding flexibility under cognitive load. *Journal of Personality and Social Psychology*, *75*(3), 589–606.

Short, J., Williams, E., & Christie, B. (1976). *The Social Psychology of Telecommunications*. Wiley.

Swim, J. K., & Hyers, L. L. (2010). Sexism. In *Handbook of prejudice, stereotyping, and discrimination* (pp. 407–430). Psychology Press.

Urbach, N., & Ahlemann, F. (2010). Structural Equation Modeling in Information Systems Research Using Partial Least Squares. *Journal of Information Technology Theory and Application (JITTA)*, *11*(2), 5–40.

Van Hooijdonk, C., & Liebrecht, C. C. C. (2021). Chatbots in the tourism industry: The effects of communication style and brand familiarity on social presence and brand attitude. *Proceedings of UMAP 2021*, 375–381.

Verhagen, T., van Nes, J., Feldberg, F., & van Dolen, W. (2014). Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication*, *19*(3), 529–545.

Vu, T. L., Tun, K. Z., Eng-Siong, C., & Banchs, R. E. (2021). Online FAQ Chatbot for Customer Support. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction* (pp. 251–259).

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human Computer Studies*, *66*(2), 98–112.

Wechsung, I., Weiss, B., Kühnel, C., Ehrenbrink, P., & Möller, S. (2013). Development and validation of the conversational agents scale (cas). *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech*.

Weizenbaum, J. (1966). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45.

West, M., Kraut, R., & Ei Chew, H. (2019). *I'd blush if I could : closing gender divides in digital skills through education*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

Yan, A., Solomon, S., Mirchandani, D., Lacity, M., & Porra, J. (2013). The role of service agent, service quality, and user satisfaction in self-service technology. *International Conference on Interaction Sciences*.