# MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

## „Trusting Chatbots: The influence of anthropomorphism, gender bias and social impact framing"

verfasst von / submitted by

Yuan Lin Qiu

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien, 2019 / Vienna, 2019

| | |
|---|---|
| Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet: | UA 066 840 |
| Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet: | Masterstudium Psychologie UG2002 |
| Betreut von / Supervisor: | Prof. Dipl.-Psych. Dr. Arnd Florack |

## Table of Contents

## List of Figures

## List of Tables

## Acknowledgments

I would like to thank Univ. Prof. Dr. Arnd Florack for supervising and supporting my master's thesis through answering all my questions and giving me feedback.

I would also like to thank Simona Haasova, MSc for all the advice and active support of my efforts throughout the process, from conceptualizing the experiment to writing the thesis.

Finally, thank you to my family and fiancé, for the continued support and encouragement. and for making this opportunity possible.

Abstract

This study aimed to investigate whether user trust in anthropomorphized chatbots is affected by the chatbot's gender and whether this effect is moderated by communicating a social impact frame emphasizing the social impact of services. Stereotypes and gender role inconsistency can often make people biased against female professionals of certain sectors such as financial investment. Prior research has shown that a social impact frame could increase the perceived warmth of a female actor as well as the congruity of the expected female stereotype with their actual perceived role. We expected these patterns to apply to human-chatbot relationships as well, as gender stereotypes have been shown to appear in anthropomorphized objects. Investment advisor chatbots were developed for this study: These first introduced users to a fictional investment application and then recommended an investment service based on personal data provided by the user. Participants interacted with chatbots that had 1) either a male or female name and avatar, and 2) either did or did not have a social impact frame that presented an optional charitable service. Data on reported trust and willingness to follow the chatbot's recommendation as well as perceived warmth and competence was collected from 116 participants to analyze the effects of gender and social impact framing. Though the results showed that the social impact frame had a marginally significant effect in increasing perceived warmth of the chatbot and it had a marginally significant interaction with chatbot gender to affect user trust, no other predictions were confirmed.

*Keywords:* Chatbots, anthropomorphism, social impact frame, stereotypes, warmth, competence, trust

# Trusting Chatbots: The influence of anthropomorphism, gender bias and social impact framing

With the growing sophistication of Artificial Intelligence (AI), virtual conversational agents, or chatbots, are increasingly being used by companies as part of their marketing strategy. According to BI Intelligence of Business Insider (2016), 80% of businesses reported that they already used or planned to use chatbots by 2020. While research on chatbot design and trust is still very limited, marketers often design chatbots to be humanlike, using features like gendered names or anthropomorphic avatars. This follows key theories of anthropomorphism, which suggest that when people perceive an object to be human-like, or anthropomorphized, they tend to have stronger feelings of trust due to a sense of social relatedness (Nass, Steuer, & Tauber, 1994; Waytz, Heafner, & Epley, 2014). Like Amazon's Alexa or Apple's Siri, chatbots are computer programs automated to reply to requests using natural language. But unlike these popular home assistants, most chatbots today are actually disembodied, interacting with customers via text on websites or mobile applications with nothing more than an avatar and a name to represent a person. Since they are cheap to implement and intuitive to use, chatbots often play an important role in a company's customer relations strategy. For example, Bank of America's customers can use a chatbot named Erica in their mobile application to not only check balances and make transactions, but also receive personalized money-saving advice. But how much do customers actually trust female chatbots like Erica to handle their financial tasks? After all, the finance industry is particularly vulnerable to gender bias and female professionals are often underestimated: In a 2016 report by KPMG, 79% of those surveyed working in financial fields reported the belief that it is more difficult for women to succeed. Prior research has indicated that interactions with anthropomorphized computers with even the most basic gender cues can carry over gender bias (Nass, Moon, & Green, 1997). It is therefore possible, that a financial services chatbot anthropomorphized as female would face the same gender bias as if it were human, potentially harming customer trust. The goal of our research is to offer a counterpoint to past theories of anthropomorphism and suggest that in certain situations, anthropomorphizing may actually impede user trust. To do this, we wanted to show that anthropomorphized chatbots can carry over gender bias from human-human interactions, giving a female financial advisor chatbot a disadvantage over a male one. Additionally, using a strategy previously shown to reduce the negative effects of gender bias against female professionals, we want to show that by reframing a chatbot's responses to emphasize social impact, we will reduce the bias against a female financial advisor chatbot. This study would not only provide large societal relevance in terms of human-human interaction and the ever-growing field of human-computer interaction, it would also serve marketing professionals by guiding them in their customer engagement strategies.

**Anthropomorphism and Trusting in Chatbots**

Anthropomorphism is the natural human tendency to attribute humanlike traits such as intentions, emotions or motivations to non-human agents (Epley, Waytz, & Cacioppo, 2007). It is what makes us talk about our pets the same way we would describe the personality of a friend, or even think of a box of tissues as social and agentic as a human when it says “bless you” after someone sneezes (Jia, Wu, Jung, Shapiro, & Sundar, 2013). Still, in the natural environment, there is a distinct boundary between what is human and not human, where there are visual characteristics, traits, behaviors, and abilities that are uniquely human, and there are also behavioral and visual properties that distinguish between living and inanimate things (Heider, 2004; Sosna, 1991). The digital world is not so clear-cut, where humans may be represented by non-anthropomorphic images, and computer programs may be represented by anthropomorphic images, show intelligent capacity, and even follow rules of social interaction (Nowak, 2004).

Brand managers and marketers often seek to encourage and facilitate our tendency to anthropomorphize: Past research has shown that by applying anthropomorphic images and distinct humanlike personalities to products and brands, companies may be able to affect their perceived credibility (Keller 2001) and foster relationships (Fournier 1998). Furthermore, anthropomorphic cues in a computer, such as a humanlike personality (Kim & Sundar, 2011) and humanoid morphology, are known to promote socialness in human-technology interaction (Sundar, Jia, Waddell, & Huang 2015). This is in-line with the Computers as Social Actors (CASA) paradigm (Nass, Steuer, & Tauber, 1994), a prominent conceptual basis behind research looking at computers, anthropomorphism and trust, which supports the theory that anthropomorphism increases user trust in computer agents (Waytz, Heafner, & Epley, 2014; Gong, 2008; Visser, Monfort, & McKendrick, 2016).

Marketers are eager to use chatbots as a way to create an anthropomorphized representative, thereby helping anthropomorphize their brands with the aim of promoting trust. This trust determines the extent an object serves as a source of social influence on the user and the user’s willingness to follow its advice (Waytz, Cacioppo, & Epley, 2010). According to Seeger and Heinzl (2018), user trust is a prerequisite for successful adoption of chatbots, and the successful design of chatbots "is contingent upon an understanding of users' expectations and perceptions in order to assure that users are willing to rely on these agents" (p. 130). Although the CASA paradigm suggests that anthropomorphism is linked to promoting trust, there has been limited research looking more closely at the process of anthropomorphism and the expectations and perceptions at play when users interact with an anthropomorphized chatbot. As Reeves and Nass (1997) argue, the human brain has not evolved to respond to modern media technologies in a way that distinguishes between human-human and human-computer communication. So it stands to reason that there may still be important aspects of chatbot interactions that need to be considered and explored.

Despite limited research, chatbots have become a popular emerging digital marketing tool. Rather than requiring users to dig through pages of text in websites for information and giving rigid commands through buttons and search boxes, companies can utilize chatbots as virtual companions that have been programed to assist users through text messaging – a preferred method for the current generation to communicate with their friends and families. The more sophisticated chatbots can utilize artificial intelligence to mimic human conversation, but even the most basic pre-scripted chatbots are able to follow social conventions by offering a greeting or a thank you. Simulating the interface of popular messaging applications, chatbots often first greet the user with a small introduction and there is often an area for an avatar, or a representative picture like that of a Facebook profile picture. This is especially useful for marketers who wish to have an anthropomorphic digital spokesperson representing their brand. For example, the American train company, Amtrak, uses a chatbot they named "Julie" (in Figure 1 on page 10), which has a photo-realistic photo of a woman, greets customers as a friendly virtual assistant, and uses first-person pronouns when referring to itself:

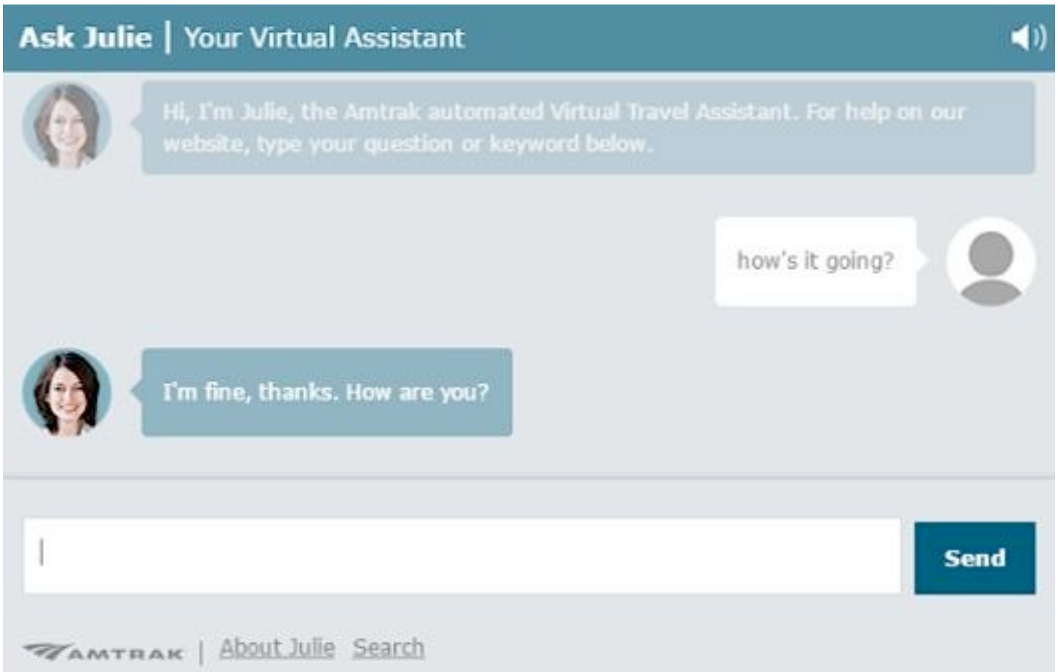


*Figure 1.* Screenshot of Amtrak's virtual assistant *Julie.* Retrieved May 13, 2019, https://www.amtrak.com/home.html.,

**Perceiving Gender in Anthropomorphism**

According to past research, anthropomorphism activates social schemas from human-human interactions in the user's mind as they attribute human features to non-human objects like computers, such as age and gender, and make judgments of credibility and likeability (Reeves & Nass, 1997; Sundar, Jia, Waddell, & Huang, p.48, 2015; Nowak, p.94, 2015). According to Sundar, et al. (2015), one social category that is especially relevant when an entity has been perceived as anthropomorphic, is gender, so that the entity is assigned as either male or female. Furthermore, the process of assigning an entity a gender category is done without even perceiving the corporeal body, which means that the attribution is based on stereotypes associated with gender as a social category and not physical differences (Balsamo, 1995; Nowak, 2003). This natural need for the categorization of gender is so strong that users interacting with computers with androgynous avatars feel more uncertain, and seem to dislike the ones that are difficult to categorize in terms of gender (Nowak & Rauh, 2008).

According to Sundar, et al. (2015), users do not seem to differentiate between computers and humans in perceptions of behaviors and would evaluate computers with the same criteria as for humans, as long as there is a minimum threshold of social potential (Nass & Moon, 2000; Reeves & Nass, 1997). It can therefore be argued that users of Bank of America's Erica or Amtrak's Julie can very easily perceive them as female and thus project the same expectations and biases on them as if they were human, preferring them for certain tasks but rejecting them for others. Nass, Moon and Green (1997) showed in their study that even with all suggestions of gender removed except for a male and female voice, participants rated the male-voiced computer more positively than the female-voiced computer, even though they were aware that they were speaking to a computer. This not only shows a parallel in the tendency to use human-human biases in human-object interactions, but also that merely audio cues are sufficient to evoke complex schemas like gender stereotypes.

**Gender Bias and Trust**

In human-human social interactions, the perception and categorization of gender, as any other social categorization, brings many assumptions and expectations on a person. These often lead to stereotypes and biases, which contribute to the reason why certain professional industries are more favorable for one gender than the other. This is likely due to prevailing stereotypes that men are more competent, competitive and decisive, while women are warmer, kinder and more concerned about others (Heilman, p.658, 2001; Williams & Best, 1990). According to Fiske, Cuddy, Glick and Xu (2002), stereotypes typically follow two prevalent dimensions in the stereotype content model: Warmth and competence. Assessments of social behavior along these dimensions offer a functional value by helping people predict another person's intent and their capability to pursue it. Furthermore, Fiske et al. (2002) showed that the two dimensions are often mutually exclusive: A person is frequently perceived as either high in warmth and low in competence, or vise versa, and rarely are both jointly high or low. Therefore, the male stereotype content model is typically high competence, but low warmth, and the female stereotype content model consists of high warmth, but low competence. Stereotypes can be perceived in brands as well: Aaker, Vohs, and Mogilner (2010) showed in their study that for-profit organizations are perceived as being more competent than non-profit organizations and therefore more male-stereotype-congruent, while non-profits are perceived as warmer and more female-stereotype-congruent. According to Heilman (1983) and Eagly and Karau (2002), bias results from gender stereotypes when a perceived "lack of fit" or role incongruity occurs between the expected norm and the actual behavior. So when a person has the choice between a female or male investment advisor, they may prefer the male advisor because it conforms with their stereotypical expectations.

Similarly, we argue that when a person interacts with a humanlike chatbot performing the role of a financial advisor as part of the services of a for-profit investment company, how the chatbot is evaluated is dependent on whether it is presented as male or female. This is because with the perception of gender comes gender role expectations, and if the user perceives the chatbot as female, they would expect the female chatbot to conform to human female stereotypes. Since in our case, the chatbot is performing the role of an (male-stereotype-congruent) investment advisor, this would elicit a perceived "lack-of-fit", leading to a bias against the female chatbot and a poorer evaluation than the male chatbot. To define the evaluation of the chatbot, we decided to look at trust: According to Moorman, Deshpande and Zaltman (p. 82, 1993), trust is the willingness to rely on an exchange partner in whom one has confidence. In the world of business-to-consumer e-commerce, trust is the core of consumer relationships and has a large influence on consumer activities, promoting participation and service adoption (Ganesan,1994; Kim & Prabhakar, 2000). We believe that trust would not only be an apt reflection of how the financial advisors chatbots are evaluated, but it would also provide particular relevance for companies who either use or are interested in using chatbots to drive and support sales:

**Hypothesis 1:** The female financial advisor chatbot would be perceived as warmer and less competent than the male chatbot.

**Hypothesis 2:** The female financial advisor chatbot would be less trusted than the male chatbot.

## Minimizing Gender Bias with Social Impact Framing

Prior research (Aaker, Garbinsky, & Vohs, 2010; Lee & Huang, 2018) supports the idea that stereotype congruity is vital for positive social consequences. In Lee and Huang's 2008 study, when female and male entrepreneurs gave presentation pitches about their startup ventures, female business entrepreneurs were evaluated less favorably than their male counterparts. This gender gap exists potentially because of the perceived lack-of-fit between the expected gender role of the "warm" female and the more male-stereotypical, "competent" role of business leadership. Lee and Huang (2018) were able to eliminate this gender bias using their method of adding a social impact frame in the entrepreneurs' presentations. Social impact framing emphasizes an organization's care for the social-economic environment and highlights consequent benefits to societal welfare (Ghimire & Pimbert, 2013; Goodstein & Polasky, 2013). Researchers have previously argued that this process elicits attributions of warmth to not only the organization, but also the role of the individual communicating the frame (Aaker et al., 2010, Lee & Huang, 2018) and consequently, to the individual themselves (Macrae & Bodenhausen, 2001). The increased perception of warmth for the female entrepreneur and their business venture (and therefore their role) resulting from the social impact frame would elicit a better-perceived fit and potentially decrease the gender bias against them. Lee and Huang's 2018 study results supported this theory: When female participants added a social impact frame highlighting the social-welfare aspects of their business venture in their presentations, they were able to receive higher perceived warmth as well as higher evaluations, making them more equally evaluated as their male counterparts.

In the same way that Lee and Huang (2018) were able to use social impact framing to manipulate gender role congruity and mitigate gender bias for female business entrepreneurs, we expected the same effects for female financial advisor chatbots as well: Users with female advisor chatbots without a social impact frame would experience a lack-of-fit between the female stereotype and the stereotypically male role of a financial advisor of a for-profit business. However, users with female chatbots with an additional social impact frame introducing the social benefits of their company's services would not only see the chatbots as warmer, but they would also see the organization as warmer and more fitting for the female stereotype. This means that the potential gender bias against the female chatbots in their evaluations would be reduced and they would receive higher trust ratings. In light of prior findings and the results of the Lee and Huang study in 2018, our study is a conceptual replication of their study. Our hypotheses are as follows:

**Hypothesis 3a:** Financial advisor chatbots that use the social impact frame would be perceived as warmer than chatbots that do not use the frame.

**Hypothesis 3b:** The fictional investment company represented by the financial advisor chatbots that use the social impact frame would be perceived as warmer than chatbots that do not use the frame.

**Hypothesis 4:** Financial advisor chatbots that use the social impact frame would be more trusted than chatbots that do not use the frame.

**Hypothesis 5:** Due to the stronger perception of warmth in the company and the female chatbot, this leads to a perceived fit in evaluating the female chatbots. Therefore, the female chatbot receives a bigger boost in user trust so that there is no longer a significant difference in the evaluations between the male and female chatbots.

## The Study

This study was part of a larger study, which compared the evaluation of anthropomorphized chatbots (with gendered names and human avatars), to low-anthropomorphized chatbots, which used the company logo as the avatar and only referred to themselves as “chat assistance programs”. The larger study looked at social impact framing as well, but it had a broader focus looking at differences between high- versus low-anthropomorphized chatbots, examining the expectations and perceptions of humans and robots in terms of warmth and competence, as well as the impact on the company they represent.

The present study however, focused mainly on gender stereotypes and compared the evaluations of male and female chatbots within the user group that used anthropomorphized chatbots. The goal of this study was to look at the effects of chatbot gender on user perceptions and trust in a for-profit, male stereotypical, context, versus trust in an investment context with a warmer, more female stereotypical, social impact framing that highlights societal benefits of the service.

## Method

### Participants

Participants were recruited through the University of Vienna’s laboratory credit system that allows bachelor’s students of psychology to receive credits in exchange for participating in experiments. Participation in the study lasted approxmately 30 minutes in a computer room provided by the University’s psychology department. The data collection lasted one week in December 2018. A small preliminary study was conducted to receive qualitative feedback to improve the design of the chatbot and dialogue.

An a priori power analysis using G*Power software indicated that we needed to have a total sample size of 122 subjects to have 95% power for detecting a medium-large sized effect (f= .33) as indicated by the Lee and Huang 2018 study when employing the traditional .05 criterion of statistical significance. There were 116 participants in this study, 88 females and 28 males ranging from 18 to 35 years of age ($M_{age}$ = 21.2 years). The sample consisted entirely of psychology bachelor’s students at the University of Vienna, who completed the study to receive credits as part of the study program’s lab participation requirement. Data from two participants were excluded from the study: One due to an incorrect response to the gender manipulation check (reported male chatbot as female) and another exceeded the maximum allowable answer to the investment decision task, asking them for the amount they would be willing to invest (€6,000 on a range from €0 to €1,000). Furthermore, we also collected the following demographic data (Table 1) for our control variables:

*Table 1. Participant Control Variables (high-anthropomorphism conditions only)*

| | *Min* | *Max* | *M* | *SD* |
|---|---|---|---|---|
| Comfort with New Tech | 1 | 7 | 4.81 | 1.65 |
| Familiarity Chatbots | 1 | 7 | 3.39 | 1.92 |
| Familiarity Investment Applications | 1 | 6 | 1.51 | 1.15 |
| Investment Experience | 1 | 4 | 1,47 | .86 |
| Smartphone Use (Hours Weekly) | 3 | 80 | 20,78 | 13,01 |
| Computer Use (Hours Weekly) | 1 | 60 | 19,66 | 14,98 |

*Note. M* = Mean. *SD* = Standard Deviation. Items for Comfort with New Tech, Familiarity Chatbots, Familiarity Investment Applications, and Investment Experience range from 1 (low) to 7 (high).

**Preliminary Analysis**

Pearson correlations were conducted to assess the relationship between the collected control variables (Table 1) and all of the dependent variables (chatbot warmth and competence, company warmth, trusting beliefs, investment decision). It revealed no significant correlations ($p<.01$), and therefore no co-variates were used in the model.

**Design**

This study followed a 2 (male vs. female chatbot) x 2 (social impact framing vs. no framing) between-subject design. The independent factors were chatbot gender and the presence of a social impact frame in the chatbot dialogue. To look at the effects of gender bias and social impact framing, we used user-reported warmth and competence in the chatbot and the fictional company as dependent variables. To examine the effects on trust, we measured user-reported trust as well as the reported amount of money that they would be willing to invest (in the investment package suggested by the chatbot).

As part of the larger study, the conditions were randomized, meaning that participants were randomly assigned to one of six chatbots: A male chatbot, a female chatbot, a genderless (low-anthropomorphism) chatbot, a male chatbot with a social impact frame, a female chatbot with a social impact frame, and a genderless chatbot with a social impact frame. For this portion of the study, we do not look at low-anthropomorphized chatbots, which is why we used data from users of the low-anthropomorphism chatbots for our manipulation check only and not for hypothesis testing.

**Procedure**

All participants were tested on a desktop computer in the computer laboratory of the Psychology department of University of Vienna. After being given a short introduction and signing an informed consent page on Unipark, participants interacted with their randomly assigned chatbot for a session lasting around 10 minutes. Other than the addition of social impact framing for some of the conditions, all chatbots had the same script so that all participants went through the same content and length of interaction. In the interaction, the chatbots introduced themselves, the company and the investment package options. They then asked the participants to provide some personal information and then service suggested an option for them.

At the end of the chatbot interaction session, participants were instructed by the chatbot to return to the Unipark page to the trust behavioral task, where they were asked how much money (out of €1,000) they would be willing to invest in the option the chatbot had suggested for them. Afterwards, the participants completed the questionnaire portion of the study lasting around 10 to 15 minutes. Throughout the experiment, German was the operational language.

**Materials**

**Chat Interface.** Disembodied (text messenger) chatbots were developed using the Landbot.io service to play the role of an investment advisor for a fictional mobile investment application. The chatbots function through a pre-written dialogue script uploaded into the Landbot.io chatbot creation application. The chatbots are automated to pull up pre-scripted statements and questions in their programmed sequence, and users are able to reply to the chatbots using pre-scripted buttons.

The chatbots first introduced themselves as the assistant of a fictional investment company, 360Invest. It also explained the services of 360Invest's mobile investment management application and gave some basic information on how investment works. The chatbots explained to the participants the three different investment packages available to choose from in the investment application. These varied in degrees of risk and were named: Plan S (low risk), Plan M (medium risk), and Plan L (high risk). Afterwards, the chatbots asked participants questions about their investment goals, age, income, and marital status and pretended to use their answers to recommend them one investment package out of the three options. In actuality, all participants received the same recommendation of Plan M, and this was revealed in the debrief.

**Chatbot Gender and Anthropomorphism Manipulation.** The Landbot.io service also allows choosing various characteristics for the design of the chatbot, which allowed manipulation of the chatbot's perceived gender. To increase the perception of anthropomorphism and manipulate the perception of male or female gender in the chatbot, one chatbot was designed with a popular male name "Paul" (Figure 2 on page 18) with an avatar with a photo-realistic stock image of a man. Another chatbot was designed with a popular female name "Anna" (Figure 3 on page 18) with an avatar with a photo of a woman. The use of human names and gender categorization has been shown to increase anthropomorphism (Araujo, 2018; Nowak, 2015). The low-anthropomorphism chatbots had only the logo of 360Invest as the avatar and introduced itself as a "chat assistance program". Furthermore, to make the Paul and Anna chatbots appear more natural and humanlike, their chat response times were programmed so that the time that it took for them to write their reply reflected the length of their message, while the low-anthropomorphism had always a constant and minimal delay in their reply.

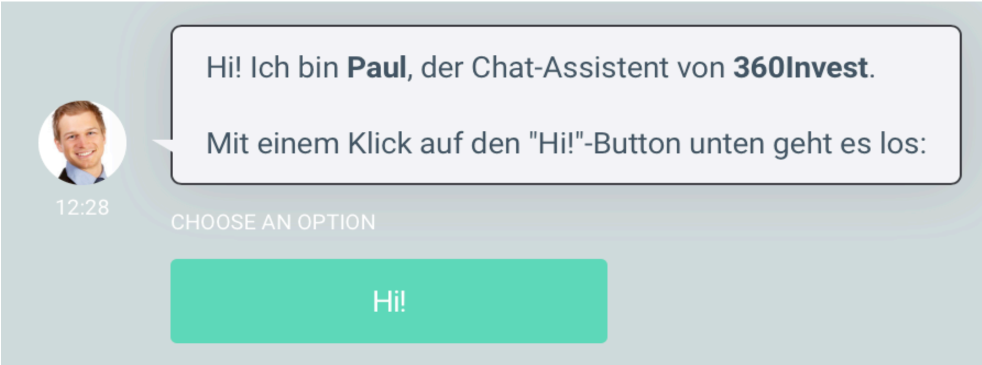


*Figure 2.* Screenshot of "Paul" chatbots greeting. Retrieved May 13, 2019, https://landbot.io/u/H-125328-LAIM8ZWB25R9NN42/index.html.

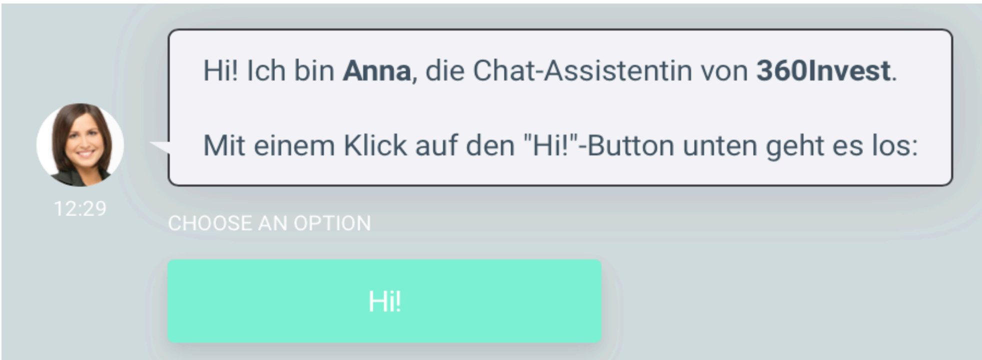


*Figure 3.* Screenshot of "Anna" chatbot's greeting. Retrieved May 13, 2019, https://landbot.io/u/H-125337-CI4OM0EX239IUE5N/index.html.

**Social Impact Frame.** In-line with the Lee and Huang (2018) study, we tested the social impact framing effect using an additional version of the male and female chatbots. This means one set of male and female chatbots interacted with users by describing the services of a fictional for-profit investment company named "360Invest", while another set had the same dialogue but with additional statements presenting a social impact frame. With the social impact frame, the chatbot introduces the user to the fictional company's partnership with a real, locally based, non-profit organization, Caritas, explaining that there is an option to donate a portion of investments to the charity during service signup. Below in Figure 4 on page 17 is a snapshot of the female chatbot with the social impact frame in its dialogue highlighted in red.


*Figure 4.* Screenshot of "Anna" chatbot with the social impact frame. Retrieved May 13, 2019, https://landbot.io/u/H-125352-L6MC5HEQ60IXG0F9/index.html.

**Measures**

After the interaction session with their chatbot, participants were redirected to Unipark and first given a decision task to measure their behavioral trust. The question was: "Imagine you have €1,000 available to invest, how much would you be willing to pay into the 360Invest plan that the chat assistant recommended to you?" Participants were then able to indicate a number in Euros in a blank box, from €0 (lowest) on. The next measure was an adaptation of the Trusting Beliefs Scale (TBS) from McKnight, Choudhury and Kacmar (2002), where participants reported the degree to which they trust the chatbot by rating six statements on a 5-point Likert scale from 1 *strongly disagree* to 5 *strongly agree*. Afterwards, items were adapted from Fiske et al. (2002) to measure participants' perceptions of warmth and competence in the chatbots. For warmth, participants were asked to indicate the extent to which they would describe the chatbot as "warm", "friendly", "good natured", and "likeable" by rating each item on a 5-point Likert scale from 1 *disagree* to 5 *agree*. For competence, the items were: "competent", "capable", efficient", and intelligent". To measure the perception of anthropomorphism, a 7-point Semantic Differential Scale was adapted from Powers and Kiesler (2006), where participants rated the chatbot based on three sets of items: Human vs. machine-like, natural vs. unnatural, lifelike vs. artificial. The scale ranges from 1 (for high anthropomorphism) to 7 (for low anthropomorphism). Finally, participants were asked for demographic information where they were also asked to report their technology-use habits and investment experience.

**Manipulation check**

To check whether the manipulation of anthropomorphism was successful, we reverse-coded the data from the anthropomorphism scale so that 1 indicates low-anthropomorphism and 7 indicates high, and used average participant response scores. Participants reported overall a slightly higher than the middle value of anthropomorphism in both the male and female chatbots ($M = 3.95$, $SD = 1.21$). Additionally, data from the groups with the anthropomorphized, gendered chatbots were compared with data of groups with the low-anthropomorphism chatbots (with no names or human avatars) using an independent samples *t*-test for each of the three items (Table 2).

*Table 2. Anthropomorphism Manipulation Check Using t-test Comparison of Means*

| | *Low-Anthropomorphized (n=118)* | | *Anthropomorphized (n=116)* | | |
|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *t*-test |
| Humanlike | 3.64 | 1.55 | 3.81 | 1.53 | -2.96** |
| Natural | 3.97 | 1.47 | 3.96 | 1.41 | ns |
| Lifelike | 3.58 | 1.34 | 4.08 | 1.30 | ns |

**p< .01

*Note.* *M* = Mean. *SD* = Standard Deviation. Items range from 1 low to 7 high (reverse-coded).

The analysis found a significant difference for only one (humanlike vs. machine-like) item out of three between the low- and high-anthropomorphism conditions in the expected direction. Differences from the natural vs. unnatural and lifelike vs. artificial items were not significant between the conditions. Furthermore, to check whether chatbot gender had an effect on perceived anthropomorphism, an independent samples *t*-test showed no significant difference in perceived anthropomorphism between the male ($M = 4.01$, $SD = 1.30$) and female chatbot ($M = 3.90$, $SD = 1.12$), $t(114) = -.46$, $p = .648$. This shows that the anthropomorphism manipulation worked the same for all conditions.

To check the gender manipulation and make sure that people were able to perceive the ascribed chatbot genders, we conducted a chi-squared test for independence. Fisher's Exact Test provided a 2-sided *p*-value of .000, which indicates significance in the participants' ability to identify that their chatbot as either female or male. Out of the participants who interacted with a female chatbot, 58 correctly reported that it was female and none reported it incorrectly as male. Out of the participants who interacted with a male chatbot, 58 correctly reported that it as male and 1 person reported it incorrectly as female–this person's data was excluded.

To check whether participants could understand the social impact frame, we used a manipulation check question: "Does 360Invest have a program that addresses societal needs?" Participants were able to answer, "yes" versus "no" or "don't know". Out of the 59 participants who interacted with a chatbot in the for-profit investment only condition with no social impact frame, 49 correctly reported either "no" or "don't know". Out of the 57 participants who interacted with a chatbot with a social impact frame, 43 correctly reported "yes". To check our manipulation, we conducted a chi-squared test for independence using the social impact frame condition as the independent variable, and manipulation check answers as the dependent variable. Fisher's Exact Test provided a 2-sided *p*-value of .000, which indicates a successful manipulation and significance in the participants' ability to identify whether their chatbot introduced a social impact program by the fictional investment company, 360Invest.

**Data Analysis**

The software program SPSS was used to analyze the data. A two-way ANOVA analysis was used to analyze the effect of chatbot gender and social impact framing on chatbot warmth and competence, company warmth and competence, and user trust.

## Results

**Hypothesis Testing**

A two-way ANOVA was used to test the hypotheses using the chatbot gender and the type of framing as the independent variables. To look at how chatbot gender and social impact framing affects the perception of the chatbots in terms of the gender stereotype content model (Fiske, et. al., 2002), data from the warmth and competence items were averaged for each trait and used as the dependent variables. To look at whether the social impact frame affected user perceptions of warmth in the company, we used data from the company warmth scale (also averaged). To look at how chatbot gender and social impact framing affects user trust in the chatbots, both the data from the Trusting Beliefs Scale and responses to the investment decision task were used as the dependent variables. Means and standard deviations for our dependent variables are summarized in Table 3.

*Table 3. Descriptive Statistics for Warmth, Competence and Trust Measures*

| | Male Chatbot (*N*=29) | | Male Chatbot +SIF (*N*=29) | | Female Chatbot (*N*=30) | | Female Chatbot +SIF (*N*=28) | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Chatbot Warmth | 2.91 | .78 | 3.38 | .77 | 3.22 | 1.05 | 3.38 | .76 |
| Chatbot Competence | 3.62 | .54 | 3.66 | .84 | 3.71 | .91 | 3.65 | .77 |
| Company Warmth | 2.75 | .73 | 2.94 | .86 | 2.82 | .98 | 3.05 | .85 |
| TBS | 2.91 | .78 | 3.28 | .81 | 3.22 | 1.05 | 2.94 | .86 |
| Investment Decision | 603.45 | 254.23 | 588.97 | 269.50 | 580.83 | 277.90 | 616.07 | 289.33 |

*Note.* SIF = Social Impact Frame. *M* = Mean. *SD* = Standard Deviation. Warmth and competence items range from 1 (low) to 5 (high). Trusting Believes Scale (TBS) items range from 1 (low) to 5 (high). Investment decision ranges from €0 to €1,000.

**Chatbot Warmth and Competence**

We expected that the effects of the gender of the chatbot on perceived warmth and competence would reflect prior research on gender bias in human-human interactions as well as human-machine interactions: Participants with the female chatbot would report overall higher warmth but lower competence than those with the male chatbot. However, the analysis revealed no significant main effect from chatbot gender on reported warmth for the chatbots

$F(1,112)=1.59$, $p= .210$. There was also no significant main effect of chatbot gender on reported competence, $F(1,112)=.09$, $p=.771$.

Based on prior research on social impact framing, particularly the Lee and Huang (2018) study, we also predicted that a social impact frame would increase the perception of warmth for both male and female chatbots. The analysis revealed a marginally significant main effect of social impact framing on chatbot warmth $F(1,112) = 2.87$, $p = .093$, $\eta p^2 = .025$. Interestingly, when the analysis is split by chatbot gender, the effect of the framing had a marginally significant main effect for only male chatbots, $F(1,112) = 2.87$, $p = .493$, showing that the mean ratings of perceived warmth of the male chatbot was higher with the social impact frame than without. The main effect of the framing was not significant for female chatbots, $F(1,56) = .43$, $p = .517$, but, looking at the mean scores in Table 3, we could speculate a small difference in scores in the expected direction, where the female chatbot could have been perceived with higher warmth with the frame than without (but this may be due to chance). The analysis did not reveal a significant Chatbot Gender × Social Impact Frame interaction $F(1,112) = .46$, $p = .498$.

Though we did not hypothesize the effect of adding a social impact frame on perceived competence of the chatbot, prior research has shown a tradeoff effect between warmth and competence (Judd, James-Hawkins, Yzerbyt, & Kashima, 2005), which would affect the evaluation of a person or entity. However, as an explorative analysis, the results showed no significant main effect of social impact framing on perceived competence for the chatbots $F(1,112) = .01$, $p = .939$. There was also no main effect of chatbot gender on perceived competence $F(1,112) = .09$, $p = .771$, and no Chatbot Gender × Social Impact Frame interaction effects $F(1,112) = .10$, $p = .753$.

Though the main effect of chatbot gender on chatbot warmth and competence was not significant, the mean warmth scores from the conditions without social impact framing show a small tendency in the expected direction (male chatbot lower than female chatbot). Interestingly, it can be seen in Figure 5 on page 24 that chatbot warmth scores indicate a tendency to not only increase across both gender conditions but also equalize in the presence of the social impact frame. However, the interaction between chatbot gender and social impact framing is not significant for chatbot warmth either $F(1,112) = .462$, $p = .498$, so this may be due to chance.

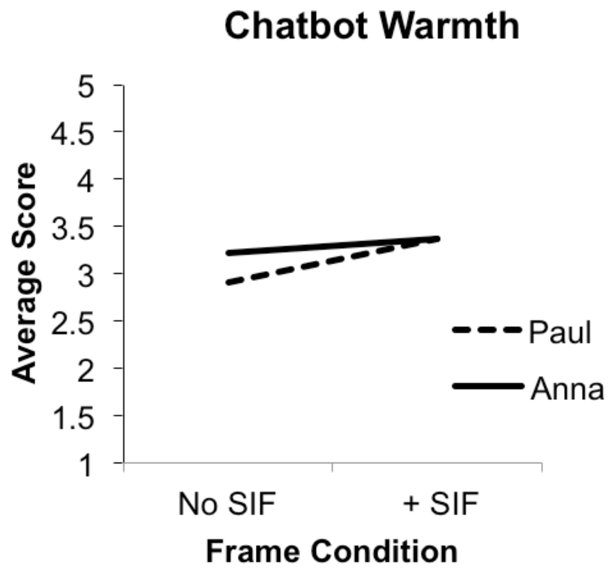


*Figure 5.* Chatbot warmth scores in framing and gender conditions: No social impact framing (No SIF) vs. with social impact framing (+ SIF); male chatbot (Paul) vs. female chatbot (Anna)

**Company Warmth**

An important mechanism behind social impact framing is that it emphasizes warm, societal benefits of company services, thereby raising not only the perceived warmth of the communicator, but also the company they represent. Whereas there would be a perceived lack-of-fit between the warm female stereotype and a colder, more expectedly competent for-profit business, in this way, a better-perceived fit is established with the special framing. Therefore, we checked whether participants with social impact framing in their chatbots also reported the fictional company, 360Invest, as warmer than those who did not have the frame. The analysis showed that the main effect of social impact framing on company warmth was not significant $F(1,112) = 1.78$, $p = .185$. But, It can be seen in Figure 6 on page 25, that the mean scores for company warmth indicate a possible minor tendency in the expected direction (higher company warmth for both chatbot genders with social impact frame than without). There was also no main effect of chatbot gender on perceived company warmth $F(1,112) = .32$, $p = .573$, and no Chatbot Gender × Social Impact Frame interaction effects $F(1,112) = .02$, $p = .883$.

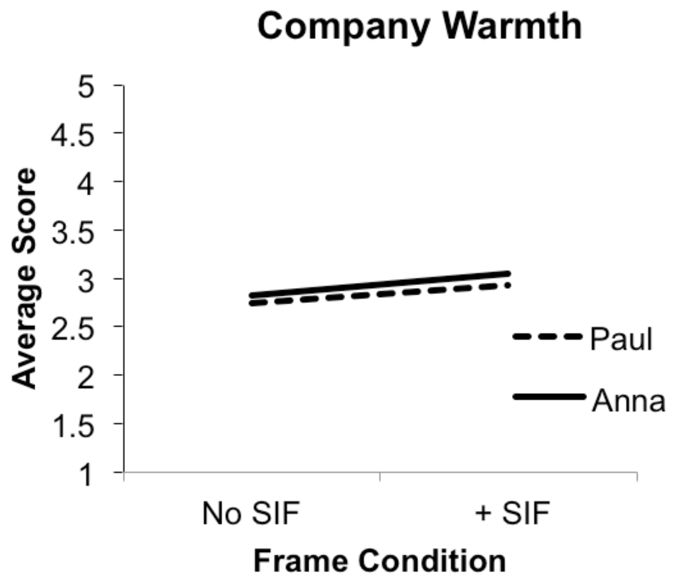


*Figure 6.* Company warmth scores in framing and gender conditions: No social impact framing (No SIF) vs. with social impact framing (+ SIF); male chatbot (Paul) vs. female chatbot (Anna)

**User-Reported Trust**

We predicted that the addition of the social impact framing would alleviate the gender bias against the female chatbot so as to increase its user trust ratings, making it more equal to the male chatbot in its trust evaluation. When splitting the data by the framing conditions, the analysis using data from the Trusting Beliefs Scale revealed no significant main effect from chatbot gender on reported trust in the chatbots without social impact framing $F(1,57) = 1.02$, $p$=.316. Interestingly however, the mean scores show possible a minor tendency for participants to report more trust for the female chatbot than the male one (Figure 7 on page 26), even though we expected that the female advisor chatbot would be less trusted than the male when presenting a solely for-profit service,

Looking at all conditions through the two-way ANOVA analysis, the data revealed that the main effect of social impact framing on reported trust was not significant $F(1,112) = .61$, $p = .435$), and neither was the main effect of chatbot gender $F(1,112) = .31$, $p = .578$). However, the interaction of the social impact frame and chatbot gender was marginally significant, $F(1,112) = 2.41$, $p = .066$, $\eta p^2 = .03$. This interaction was expected as we predicted that the addition of a social impact frame would moderate the effects of chatbot gender on user trust so that there would be no significant difference in trust evaluations between the male and female chatbots with framing. Interestingly, as seen in Figure 7 on page

26, the direction of the mean scores indicate a possibly different direction than expected between the type of framing and reported trust: Trust ratings for the male chatbot was higher with the presence of the frame, while trust for the female chatbot actually was lower, giving the male chatbot an advantage.

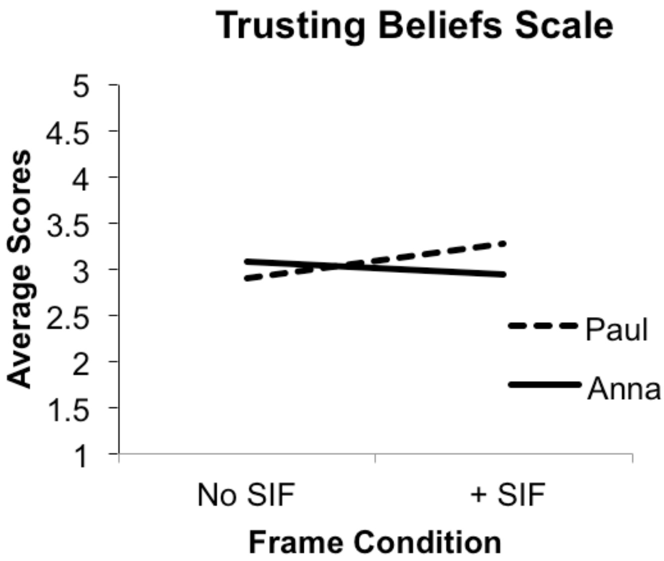


*Figure 7.* Reported trust and interaction between chatbot gender and framing: No social impact framing (No SIF) vs. with social impact framing (+ SIF); male chatbot (Paul) vs. female chatbot (Anna)

**Behavioral Trust Task**

Trust was also tested in terms of the participants' willingness to invest in the investment package suggested by their chatbot in an investment decision task. Since financial data may often have non-normal distributions, we first tested the investment responses for normality and homogeneity of variance using a Shapiro-Wilk test of normality, which showed that the response data for willingness to invest are not normally distributed ($p<0.01$). However, a visual inspection of the Q-Q plot and the histogram of the frequency distribution showed that a reasonable amount of normality can be assumed. Additionally, the Central Limit Theorem posits that sample sizes equal to or above 30 are sufficiently large enough to justify their use for analyses like ANOVA (Field, 2018). For these reasons, with our sample size of 116, we determined that it was not necessary to run a non-parametric test.

Again, when splitting the behavioral task data by the framing conditions, there was no significant main effect from chatbot gender on responses in the behavioral trust task in the non-social impact frame conditions $F(1,57) = 1.06$, $p$=.746, which means that there was no significant difference in reported investment amounts between users of the male and female chatbots without social impact framing. Looking at the data of both framing conditions, the analysis showed that there was neither a main effect from the social impact frame $F(1,112) = .04$, $p = .838$, nor was the main effect from chatbot gender significant $F(1,112) = .00$, $p = .965$. Additionally, the interaction of the social impact frame and chatbot gender was also not significant, $F(1,112) = 3.45$, $p = .625$. However, as can be seen in Figure 8, the reported investment values may show a potential minor tendency in the expected direction, where participants may have been willing to invest more for the male chatbot without the social impact framing. It also shows a potential interaction between framing and chatbot gender, where the female chatbot may have received an advantage over the male with the social impact frame.

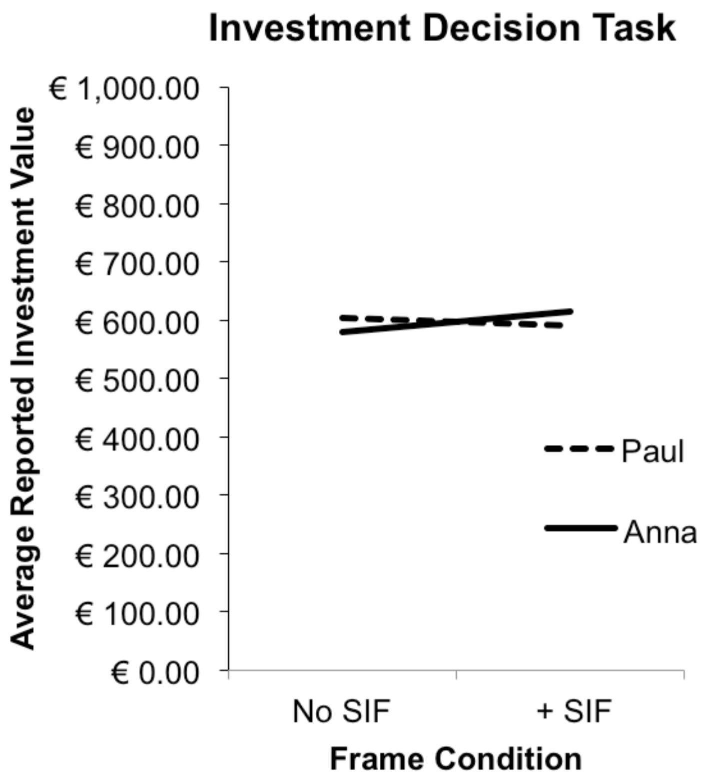


*Figure 8.* Behavioral trust and interaction between chatbot gender and framing: No social impact framing (No SIF) vs. with social impact framing (+ SIF); male chatbot (Paul) vs. female chatbot (Anna)

# Discussion

## Chatbot Gender

The goal of the study was to validate whether social patterns from human-human interactions arising from gender stereotypes could be passed on to chatbots when they are anthropomorphized and gendered. More specifically, we wanted to see whether a chatbot's gender could affect its perceived warmth, competence and user trust. We expected that, in congruence with the stereotype content model (Fiske, et. al., 2002), the female chatbot would be perceived with higher warmth, lower competence and lower trust than its male counterpart when they play the role of investment advisors.

Our data analysis could not support our predictions that male and female financial advisor chatbots would be evaluated according to human gender stereotypes in terms of warmth and competence in accordance with the stereotype content model (Fiske et al., 2002): There was no significant difference in chatbot warmth and competence ratings across the gender conditions. Still, the actual participant scores had some interesting (although not significant) outcomes: While the female chatbot on average may have been rated slightly warmer than the male chatbots, unexpectedly, they seemed also to be rated more competent. Furthermore, though there was no significant effect from gender, participants in the investment decision task showed a tendency in the expected direction with higher scores on average for the male chatbot than the female chatbot. While by contrast, participants actually seemed to have potentially reported more trust for the female financial advisor chatbot in the Trusting Beliefs Scale.

Furthermore, to the best of our knowledge, this is the first study looking at gender effects in disembodied (text-based) chatbots. Prior research looking at interactions with gendered computer programs has usually used embodied chatbots or avatars that incorporated other cues such as vocals (Nass, Moon, & Green, 1997) or moving images (Garau, Skater, Bee, & Sasse, 2001). Due to past theories suggesting that gender perception is easily established in interactions with anthropomorphized entities known to be bots and not human (Nowak, 2015), we expected that only gendered avatar and names would be enough as gender cues for our chatbots. However, the lack of significant data from looking at the effects of chatbot gender on warmth, competence and trust evaluation may then be an indication that the use of only a static photo and gendered name in a chatbot is not enough to elicit the same strength of gender stereotyping and bias.

Additionally, the anthropomorphism manipulation check showed that only one of the three items (humanlike vs. machine-like) showed a significant difference between the high- and low-anthropomorphism conditions, while data for the other two items (natural vs. unnatural and lifelike vs. artificial) was not significant. This potentially shows that something worked in the anthropomorphism manipulation, but something also did not. Therefore, the mixed reaction to the anthropomorphism manipulation may have been a contributor to the mixed results. This is a possible indicator that the users simply may have not paid much attention to the chatbot name and avatar (the main anthropomorphism and gender manipulations) and focused on the dialogue text instead. Other possible reasons for the lack of significant results are discussed in the limitations section.

**Social Impact Framing**

We wanted to see whether social impact framing, a method previously demonstrated by Lee and Huang (2018) to alleviate gender bias against female business leaders, could reduce bias against female financial advisor chatbots and increase their trust evaluations. We wanted to validate the complexity of chatbot interactions by showing that gender role congruity plays an important role in the evaluation of anthropomorphized chatbots. Though the results show that the social impact frame did allow marginally significant higher perceived warmth in the chatbots overall, they could not support the prediction that the frame would result in higher overall trust as well.

Furthermore, though the data analysis showed a marginally significant interaction between social impact framing and chatbot gender on user-reported trust, actual scores on the Trusting Beliefs Scale for the male chatbot unexpectedly show potentially higher ratings with the addition of the frame. However, trust for the female chatbot was potentially lower, with the gender gap staying about the same. Interestingly, the results from the behavioral trust measure, the investment decision task, did not produce a significant result for the interaction, but the means (Table 3) showed that scores were more consistent with predictions. However, it is of note here that though chatbot gender did not have a significant relationship with the investment decision data, the real scores potentially indicate that the gender gap may have actually widened with the addition of a social impact frame: While it may have benefited the female chatbot as predicated, it also may have given the male chatbot a disadvantage.

A vital mechanism behind the social impact framing's ability to reduce bias against females in a male dominated field is that it increases the perceived warmth of the company that they represent (and thereby their role), which could lead to a better-perceived fit for a female chatbot representing an investment company as well. This is why we wanted to check whether the social impact frame was able to increase perceptions of warmth of our chatbots' fictional company, 360Invest, as well. Though not the main effect was not significant, the actual scores indicate a potential relationship in the expected direction (with slightly higher warmth with the framing) – but this could be due to chance. Since the analysis did not show a significant main effect of the social impact framing on company warmth, this may have been a potential reason behind the unexpected results for user-trust evaluations. It is also interesting to note here, that in the larger study that also looks at low-anthropomorphized chatbots, an independent *t*-test did confirm a significant main effect of social impact framing on company warmth $t(116)=-2.12$, $p = 0.036$, where the groups with social impact framing reported higher company warmth ($M$=2.8, $SD$=.74) than groups without the frame ($M$=2.56, $SD$=.72). This indicates that a larger sample may help provide significant results.

## Complexity of Chatbot Interaction

The inconsistencies in the data may be the result of errors in manipulation and measurement, which are discussed further in the limitations section below. Despite the inconsistencies, the results show that the manipulation of the social impact frame had an effect in changing perceptions of the chatbots, and made an impact (along with gender perceptions) in the chatbot evaluations. This supports the broader argument that when it comes to promoting trust, interactions with anthropomorphized chatbots are not so straightforward and require more careful consideration for the chatbot characteristics and evaluation context.

## Limitations and Future Research

**Anthropomorphism Manipulation.** The ratings on the anthropomorphism scale for the chatbots were not very high, so it is possible that participants still perceived them as just computer programs. During the evaluation portion of the experiment, the questions in the Unipark questionnaire always referred to the chatbots as "das Chat-Assistenzprogramm" so that the German wording could remain gender-neutral. Therefore, it is possible that this constant reminder that the chatbots are only computer programs led to the wearing off of any gender effects. For this reason, it would be beneficial to use English wording with an English-speaking population to avoid the anchoring of gendered nouns inherent to the German language.

In a study by Aroujo (2018), the researcher compared evaluations of anthropomorphic and non-anthropomorphic chatbots: Using informal language for the anthropomorphic chatbot such as “hello” and “goodbye” versus “start” and “quit” for the non-anthropomorphic chatbot, he was able to successfully manipulate anthropomorphism using different dialogical cues between the conditions. We decided not to use this type of manipulation in our study so as to prevent unintended influences from the dialogue content on the perception of chatbot warmth.

Additionally, prior studies looking at chatbots were usually more sophisticated, using artificial intelligence to power their chat-interface rather than a pre-scripted dialogue (Shawar, & Atwell, 2007). Due to time and financial constraints of the study, we could not develop AI chatbots on our own and resolved to use the pre-scripting framework provided by Landbot.io. Therefore, it is possible that the rigid pre-scripted dialogue may have contributed to the less than optimally perceived anthropomorphism in our chatbots.

**Gender Manipulation.** Stereotype biases tend to have a stronger effect on evaluations when people perceive a comparison group at the same time (Biernat, & Manis, 1994). Because participants only interacted with either a male or a female chatbot and were asked to only rate the chatbot that they interacted with, it is possible that participants did not experience much of a gender contrast in their mental representations, and evaluated their chatbots based on different standards than they would if they were asked to compare a male and female chatbot.

**Measuring Trust.** One of the surprising results from the analysis was that there is no significant difference in perceived competence in any of the conditions, so it seems as if trust was only evaluated based on perceived warmth. Past research has shown that judgments of warmth are made more readily and quickly than competence, and make a greater impact on overall attitude towards others (Wojciszke, Bazinska, & Jaworski, 1998; Wojciszke & Abele, 2008). This, along with the fact that the items in the Trusting Beliefs Scale are rather broad and abstract, may have led to the unexpected outcomes in reported trust. Rather than evaluating the chatbots as financial advisors, it is possible that participants had evaluated them as customer assistants, a more female stereotypical and socially conscious role (Gustavsson and Czarniawska, 2004), which would also explain why TBS data for the female chatbot without the social impact frame was on average more higher than the male’s despite the financial context. For future research, the wording for Trusting Beliefs Scale and competence items should be better specified to match the context.

Furthermore, Fiske et al. (2002) and Judd et al. (2005) have shown that judgments of warmth and competence are often negatively correlated, where a surplus of one implies a deficit on the other, so that people tend to be seen as either competent but cold, or warm but incompetent. Therefore, it is possible that with the marginally significant boost in warmth from the social impact frame, when the female chatbot spoke about financial investment, perceptions of lack of fit became stronger so that it decreased its reported trust.

The invest decision task data actually fits the predicted direction much better than the data from the participant-reported trust. It could be an indication that it is the better trust measure here, especially because the language is much more specific to the context. However, the effects were not significant, and this could be the result of an unintended anchoring effect: We noticed in the data that €800 was the most frequently reported answer to the investment decision task. A possible reason is that when the participants were interacting with the chatbots, the chatbots explained the fees of the services and mentioned that €800 is the minimum to start investing. Even though the questionnaire did not refer to this or mention this anywhere else, people may have remembered it and used it as an anchor. We would suggest that future research using a similar testing method should avoid potential anchoring effects like this in the chatbots dialogue.

**Financial Resources.** Due to limited resources, we were only able to use one stock photo of a male and a female person each that looked similar to each other. A study with more resources should use a pool of similar stock photos of male and female persons to randomize in the experiment to prevent any unintended effects from the pictures themselves. Limited financial resources also meant that there was never real money involved in this study and we could only ask the participants to imagine the investment decision task. Past research has indicated that real monetary rewards tend to have stronger effects than nominally equivalent hypothetical rewards (Camerer, Hogarth, Budescu, & Eckel, 1999; Hertwig & Ortmann, 2000; Smith & Walker, 1993). This may have also contributed to the lack of significance in the investment decision results.

In addition to the improvements in methodology as suggested above, this study opens up potential areas to be explored further in terms of chatbot gender perception and trust. Our study mainly focused on the efficacy of chatbot design in terms of using anthropomorphism and gender, and whether user-chatbot trust could be established. However, it would be interesting to measure user trust in the brand that the chatbot represents as well. After all, the end goal of using chatbots as part of customer support and marketing strategy is to improve user-brand relationship. Furthermore, it would also be interesting to use an eye-tracking study to see what particular features the user pay attention to in a chatbot dialogue box. As discussed earlier, the lack of significant gender effects suggests that it is possible that users might not attend much to the avatar or chatbot name and focus mainly on the dialogue text instead. In the future, a study with a larger sample that has more equally distributed gender would be able to explore whether there are differences in how male and female users perceive male and female chatbots. Lastly, we would highly suggest conducting a replication study using the English language, as this language, in contrast to German, does not use gendered declensions with nouns. Because the participants in our study were German native speakers and because the German word for robot is masculine, the participants may have used this as a mental anchor with the chatbots, impeding any gender effects in their evaluations.

**Relevance**

It is important to note that the purpose of this experiment was not to recommend that chatbot designers perpetuate gender stereotypes. Rather, the first purpose is to show that social biases can be transferred to human-computer interactions via anthropomorphism even in text-based interactions. Another goal was to put into question the widely held belief that evoking anthropomorphism would always help build human-computer trust. Though mixed, the results from our study show that chatbot designers must treat user interactions with more mindfulness and delicacy as they would an interaction with another human if they want to avoid confusing or alienating their customers.

## References


Aggarwal, P., & Mcgill, A. L. (2007). Is That Car Smiling at Me? Schema Congruity as a Basis for Evaluating Anthropomorphized Products. *Journal of Consumer Research*, *34*(4), 468–479. doi: 10.1086/518544

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. doi: 10.1016/j.chb.2018.03.051

Balsamo, A. (1995). Forms of Technological Embodiment: Reading the Body in Contemporary Culture. *Cyberspace/Cyberbodies/Cyberpunk: Cultures of Technological Embodiment*, 215–238. doi: 10.4135/9781446250198.n13

BI Intelligence (2016, December 14). 80% of businesses want chatbots by 2020. *Business Insider.* Retrieved from https://www.businessinsider.de/80-of-businesses-want-chatbots-by-2020-2016-12?r=US&IR=T

Biernat, M., & Manis, M. (1994). Shifting standards and stereotype-based judgments. *Journal of Personality and Social Psychology*, *66*(1), 5–20. doi: 10.1037//0022-3514.66.1.5

Camerer, C. F., Hogarth, R. M., Budescu, D. V., & Eckel, C. (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Elicitation of Preferences*, 7–48. doi: 10.1007/978-94-017-1406-8_2

Cowell, A. J., & Stanney, K. M. (2003). Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents. *Intelligent Virtual Agents Lecture Notes in Computer Science*, 301–309. doi: 10.1007/978-3-540-39396-2_50

Eagly, A. H., & Karau, S. J. (2002). Role congruity theory of prejudice toward female leaders. *Psychological Review*, *109*(3), 573–598. doi: 10.1037//0033-295x.109.3.573

Field, A. P. (2018). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (5th ed.). London: Sage.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878–902. doi: 10.1037//0022-3514.82.6.878

Fournier, S. (1998). Consumers and Their Brands: Developing Relationship Theory in Consumer Research. *Journal of Consumer Research*, *24*(4), 343–353. doi: 10.1086/209515

Ganesan, S. (1994). Determinants of Long-Term Orientation in Buyer-Seller Relationships. *Journal of Marketing*, *58*(2), 1. doi: 10.2307/1252265

Garau, M., Slater, M., Bee, S., & Sasse, M. A. (2001). The impact of eye gaze on communication using humanoid avatars. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI 01*. doi: 10.1145/365024.365121

Ghimire, K. B. (2013). Social Change and Conservation. doi: 10.4324/9781315066189.

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. doi: 10.1016/j.chb.2007.05.007

Goodstein, E. S., & Polasky, S. (2013). *Economics and the environment*. Hoboken, NJ: John Wiley & Sons.

Gustavsson, E., & Czarniawska, B. (2004). Web Woman: The On-line Construction of Corporate and Gender Images. *Organization*, *11*(5), 651–670. doi: 10.1177/1350508404046455

Heider, F. (2004). Perceiving the Other Person. *Social Cognition*, 33–35. doi:10.4324/9780203496398-1

Heilman, M. E. (2001). Description and Prescription: How Gender Stereotypes Prevent Womens Ascent Up the Organizational Ladder. *Journal of Social Issues*, *57*(4), 657–674. doi: 10.1111/0022-4537.00234

Hertwig, R., & Ortmann, A. (2000). Experimental practices in economics: A challenge for psychologists? *PsycEXTRA Dataset*. doi: 10.1037/e683322011-032

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. *Journal of Personality and Social Psychology*, *89*(6), 899–913. doi: 10.1037/0022-3514.89.6.899

Keller, K. L. (2001). *Building customer-based brand equity: A blueprint for creating strong brands* (pp. 3-27). Cambridge, MA: Marketing Science Institute.

Kim, H.-S., & Sundar, S. S. (2011). Using interface cues in online health community boards to change impressions and encourage user contribution. *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI 11*. doi: 10.1145/1978942.1979028

Kim, K. & Prabhakar, B. (2000): Initial trust, perceived risk, and the adoption of internet banking, *Proceedings of the Twenty-First International Conference on Information Systems,* Brisbane, Australia.

KPMG (2016). *The Time is Now: Real Change, Real Impact, Seize the Moment: Women in Alternative Investments.* Retrieved from https://home.kpmg.com/us/en/home/insights/2016/12/women-in- alternative-investments-report.html

Lee, M., & Huang, L. (2018). Gender Bias, Social Impact Framing, and Evaluation of Entrepreneurial Ventures. *Organization Science*, *29*(1), 1–16. doi: 10.1287/orsc.2017.1172

Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, *92*(1), 239–255. doi: 10.1348/000712601162059

Mcknight, D. H., Choudhury, V., & Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *The Journal of Strategic Information Systems*, *11*(3-4), 297–323. doi: 10.1016/s0963-8687(02)00020-3

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors Affecting Trust in Market Research Relationships. *Journal of Marketing*, *57*(1), 81. doi: 10.2307/1252059

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Conference Companion on Human Factors in Computing Systems - CHI 94*. doi: 10.1145/259963.260288

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. doi: 10.1111/0022-4537.00153

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. doi: 10.1111/j.1559-1816.1997.tb00275.x

Nowak, K. L. (2003). Sex Categorization in Computer Mediated Communication (CMC): Exploring the Utopian Promise. *Media Psychology*, *5*(1), 83–103. doi: 10.1207/s1532785xmep0501_4

Nowak, K. L. (2015). Examining Perception and Identification in Avatar-mediated Interaction. *The Handbook of the Psychology of Communication Technology*, 87–114. doi: 10.1002/9781118426456.ch4

Nowak, K. L., & Rauh, C. (2008). Choose your "buddy icon" carefully: The influence of avatar androgyny, anthropomorphism and credibility in online interactions. *Computers in Human Behavior*, *24*(4), 1473–1493. doi: 10.1016/j.chb.2007.05.005

Powers, A., & Kiesler, S. (2006). The advisor robot. *Proceeding of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction - HRI 06*. doi: 10.1145/1121241.1121280

Reeves, B., & Nass, C. I. (1997). The media equation: How people treat computers, television, & new media like real people & places. *Computers & Mathematics with Applications*, *33*(5), 128. doi: 10.1016/s0898-1221(97)82929-x

Seeger, A.-M., & Heinzl, A. (2017). Human Versus Machine: Contingency Factors of Anthropomorphism as a Trust-Inducing Design Strategy for Conversational Agents. *Information Systems and Neuroscience Lecture Notes in Information Systems and Organisation*, 129–139. doi: 10.1007/978-3-319-67431-5_15

Shawar, B.A., & Atwell, E. (2007, January). Chatbots: Are they Really Useful? *LDV Forum, 22*, 29-49.

Smith, V. L., & Walker, J. M. (1993). Monetary Rewards And Decision Cost In Experimental Economics. *Economic Inquiry*, *31*(2), 245–261. doi: 10.1111/j.1465-7295.1993.tb00881.x

Sosna, M., & Sheehan, J. J. (1991). *The boundaries of humanity: humans, animals, machines*. Berkeley: Univ. of California Press.

Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a Theory of Interactive Media Effects (TIME). *The Handbook of the Psychology of Communication Technology*, 47–86. doi: 10.1002/9781118426456.ch3

Visser, E. J. D., 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. doi: 10.1037/xap0000092

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who Sees Human? *Perspectives on Psychological Science*, *5*(3), 219–232. doi: 10.1177/1745691610369336

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. doi: 10.1016/j.jesp.2014.01.005

Williams, J. E., & Best, D. L. (1990). *Measuring sex stereotypes: A multination study* (rev. ed.). Beverly Hills, CA: Sage Publications.

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the Dominance of Moral Categories in Impression Formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263. doi: 10.1177/01461672982412001

Wojciszke, B., & Abele, A. E. (2008). The primacy of communion over agency and its reversals in evaluations. *European Journal of Social Psychology*, *38*(7), 1139–1147. doi: 10.1002/ejsp.549

# Appendix A

## Unipark Instructions

*A1.* Unipark Welcome Message

Liebe Teilnehmerin, lieber Teilnehmer,

herzlichen Dank für Ihre Bereitschaft, an unserer Studie teilzunehmen. Diese Studie, in der wir uns mit virtueller Kommunikation beschäftigen, besteht aus zwei Teilen und dauert ca. 30 Minuten. Sie werden zur Durchführung ausschließlich den Computer verwenden.

Im ersten Teil werden Si emit einem Chat-Assistenzprogramm der Investmentfirma 360Invest interagieren, wobei Ihnen verschiedene Investmentpakete (Pläne) vorgestellt werden. Sie werden dann im Zuge der Beratung einige Fragen beantworten um das für Sir am besten geeignete Packet herauszufinden.

Bei der Interaktion mit dem Chat-Assistenzprogramm ist wichtig, dass Sie sich auf das Szenario einlassen und wahrheitsgemäß antworten. Ihre Angaben aus dem Chatbot werden nicht gespeichert.

Der zweite Teil der Studie ist ein Fragebogen, und es ist für uns notwendig, dass Sie alle Fragen beantworten. Wenn Sie sich bei einer Frager nicht sicher sind, kreuzen Sie bitten das Feld an, das am ehesten zutrifft.

Um die Kommunikation realistisch zu halten, versuchen Sie bitte sich vorzustellen, dass Sie 1000 Euro an Investitionskaiptal zur Verfügung haben.

*A2.* Informed Consent

Mit dem Klicken des "Continue"-Buttons, bestätigen Sie, die Einleitung gelesen zu haben, und willigen ein, an dieser Studie teilzunehmen:

Ich bin damit einverstanden, dass meine Angaben ausschließlich für wissenschaftliche Zwecke am Institut für Angewandte Psychologie: Arbeit, Bildung, Wirtschaft aufbewahrt und ausgewertet werden. Nach Beendigung des Forschungsvorhabens werden alle Daten gelöscht, die einen Bezug zu meiner Person erlauben.

Vielen Dank für Ihre Teilnahme an der Studie!

*A3.* Instructions

Liebe/r Teilnehmer/in,

Sie werden im nächsten Schritt mit dem Chat-Assistenzprogramm von 360Invest verbunden, dort werden Ihnen das Unternehmen und die Services vorgestellt und zum Abschluss erhalten Sie auf Basis Ihrer Angaben eine Investment-Empfehlung. Ihre Daten werden nicht gespeichert.

Bitte denken Sie im Verlauf der Studie daran, dass Sie davon ausgehen, 1000 Euro an Investitionskaputal zur Verfügung zu haben. Es geht um diesen Betragm, wenn Sie im Anschluss nach einer Einschätzung zu Ihrer Investition gefragt werden.

*A4.* Re-direction to Chatbot

Um mit dem Chat-Assistenzprogramm verbunden zu werden, klicken Sie bitte auf folgenden Link: *<link to randomly-assigned chatbot>*

# Appendix B

## Questionnaire

*B1.* Behavioral Investment Task

**Wenn Sie 1000 Euro Investitionskapital zur Verfügung hätten, wieviel wären Sie bereit, in den vom Chat-Assistenzprogramm vorgeschlagenen Plan zu investieren?**

Bitte geben Sie den Betrag als ganze Zahl in Euro an

Euro

*B2.* Trusting Beliefs Scale

Ich glaube, dass das Chat-Assistenzprogramm in meinem besten Interesse handeln würde.

Wenn ich Hilfe bräuchte, würde das Chat-Assistenzprogramm das Beste geben um mir zu helfen.

Das Chat-Assistenzprogramm ist an meinem Wohlergehen interessiert, nicht an seinem eigenen.

Das Chat-Assistenzprogramm ist im Umgang mit mir ehrlich.

Ich würde das Chat-Assistenzprogramm als aufrichtig bezeichnen.

Das Chat-Assistenzprogramm ist authentisch und ernstzunehmend.

**Bitte geben Sie an, inwieweit Sie der jeweiligen Aussage zustimmen.**

*B3.* Chatbot Competence Scale

Im nächsten Abschnitt möchten wir Ihre Einstellungen gegenüber dem Chat-Assistenzprogramm erheben. Bitte versuchen Sie, alle Abstufungen der vorgegebenen Skala für Ihre Einschätzung zu nutzen. Es gibt keine richtigen und falschen Antworten. Wir sind an Ihrer persönlichen Meinung interessiert.

**Geben Sie bitte an, wie charakteristisch folgende Eigenschaften für das Chat-Assistenzprogramm sind:**

| | **Trifft gar nicht zu 1** | **2** | **3** | **4** | **Trifft sehr zu 5** |
|---|---|---|---|---|---|
| kompetent | | | | | |
| leistungsfähig | | | | | |
| effizient | | | | | |
| intelligent | | | | | |

*B4.* Chatbot Warmth Scale

**Geben Sie bitte an, wie charakteristisch folgende Eigenschaften für das Chat-Assistenzprogramm sind:**

| | Trifft gar nicht zu 1 | 2 | 3 | 4 | Trifft sehr zu 5 |
|---|---|---|---|---|---|
| herzlich | ○ | ○ | ○ | ○ | ○ |
| freundlich | ○ | ○ | ○ | ○ | ○ |
| gutmütig | ○ | ○ | ○ | ○ | ○ |
| sympathisch | ○ | ○ | ○ | ○ | ○ |

*B5.* Anthropomorphism Scale

**Bitte bewerten Sie das Chat-Assistenzprogramm, indem Sie den Regler verschieben.**

Durch einmaliges Klicken auf die graue Linie aktivieren Sie den Regler.

Menschlich — Maschinell

Natürlich — Unnatürlich

Lebensnah — Künstlich

*B6.* Chatbot Gender Manipulation Check

**Haben Sie das Chat-Assistenzprogramm als weiblich oder männlich wahrgenommen?**

Kreuzen Sie das aus Ihrer Sicht eher zutreffende an.

○ Weiblich

○ Männlich

*B7.* Company Competence Scale

Im nächsten Abschnitt möchten wir Ihre Einstellungen gegenüber Unternehmen 360Invest erheben. Bitte versuchen Sie, alle Abstufungen der vorgegebenen Skala für Ihre Einschätzung zu nutzen. Es gibt keine richtigen und falschen Antworten. Wir sind an Ihrer persönlichen Meinung interessiert.

**Geben Sie bitte an, wie charakteristisch folgende Eigenschaften für das Unternehmen 360Invest sind:**

| | Trifft gar nicht zu 1 | 2 | 3 | 4 | Trifft sehr zu 5 |
|---|---|---|---|---|---|
| kompetent | ○ | ○ | ○ | ○ | ○ |
| leistungsfähig | ○ | ○ | ○ | ○ | ○ |
| effizient | ○ | ○ | ○ | ○ | ○ |
| intelligent | ○ | ○ | ○ | ○ | ○ |

*B8.* Company Warmth Scale

**Geben Sie bitte an, wie charakteristisch folgende Eigenschaften für das Unternehmen 360Invest sind:**

| | Trifft gar nicht zu 1 | 2 | 3 | 4 | Trifft sehr zu 5 |
|---|---|---|---|---|---|
| herzlich | ○ | ○ | ○ | ○ | ○ |
| freundlich | ○ | ○ | ○ | ○ | ○ |
| gutmütig | ○ | ○ | ○ | ○ | ○ |
| sympathisch | ○ | ○ | ○ | ○ | ○ |

*B9.* Social Impact Frame Manipulation Check

**Verfügt 360invest über ein Programm zur Förderung gesellschaftlicher Bedürfnisse?**

Bitte wählen Sie die zutreffende Antwort.

○ Ja

○ Nein

○ Weiß Nicht

*B10*. Control Question: Comfort with new technology

**Wie wohl fühlen Sie sich insgesamt mit neuen Technologien?**

überhaupt nicht

ausgesprochen

*B11*. Control Questions: Comfort with chatbots and investment applications

**Wie vertraut sind Sie insgesamt mit der Nutzung von...**

Chatbots

Investmentapps

*B12*. Control Question: Experience with investment

**Wieviel Erfahrung haben Sie persönlich im Investment-Bereich?**

überhaupt keine ausgesprochen viel

*B13*. Demographics: Age

**Wie alt sind Sie?**

Bitte geben Sie Ihr Alter in vollen Jahren an:

[ ] Jahre

*B14*. Demographics: Participant Gender

**Geschlecht**

- ○ weiblich
- ○ männlich
- ○ anders

*B15*. Demographics: Education

**Höchste abgeschlossene Ausbildung:**

Pflichtschule
berufsbildende höhere Schule/Le
Matura/Abitur
Bachelor-Abschluss
Master-Abschluss

*B16*. Demographics: Hours per week on smartphone

**Wie viele Stunden nutzen Sie ungefähr Ihr Smartphone in der Woche?**

Bitte geben Sie die Dauer in Stunden pro Woche an.

[ ] Stunden

*B17*. Demographics: Hours per week on computer

**Wie viele Stunden nutzen Sie ungefähr Ihren Computer in der Woche?**

Bitte geben Sie die Dauer in Stunden pro Woche an.

[ ] Stunden

# Appendix C

## Debrief

Vielen Dank für Ihre Teilnahme!

Sie haben mit einem Investmentberatungs-Chatbot interagiert, der speziell für diese Studie ausgearbeitet wurde. Das Ziel unserer Studie ist, die Interaktion von KonsumentInnen mit Chatbots zu untersuchen. Dabei interessiert uns besonders die wahrgenommene Menschlichkeit (Anthropomorphismus), die Bereitschaft, dem Chatbot zu vertrauen und der Effekt von Geschlechtsverzerrungen (Biases) auf das Vertrauen. Die daraus abzuleitenden Implikationen sind nicht nur aus sozialpsychologischer Forschungsperspektive sehr interessant, sondern auch für Programmierer und Unternehmen, die Chatbots einsetzen möchten, von größter Bedeutung.

Es wurde gezeigt, das Vertrauen einen entscheidenden Faktor in der erfolgreichen Anwendung von Chatbots darstellt. Vorangegangene Forschung legt nahe, dass die Steigerung des Anthropomorphismus eines unbelebten Objekts Gefühle des Vertrauens beim Nutzer steigert.

Wir vermuten daher, dass gesteigerter Anthropomorphismus soziale Verzerrungen (Biases), die eigentlich bei zwischenmenschlicher Kommunikation auftreten, auch auf die Kommunikation von Mensch und Maschine überträgt. Dadurch kann unter Umständen das wahrgenommene Vertrauen abnehmen. Dem verbreiteten Rollenbild folgend, dass Frauen eher fürsorglich und um andere bemüht sind, wohingegen Männer eher kompetitives und entscheidungsfreudiges Verhalten zeigen, haben wir verschiedene Chatbots entwickelt.

Es gibt sowohl einen finanziell orientierten Chatbot, als auch einen, der die sozialen Auswirkungen des Investments in den Vordergrund stellt.

Mit den Daten der verschiedenen Fragebögen werden wir einige Gruppen hinsichtlich ihrer Tendenz, dem Chatbot zu vertrauen, miteinander vergleichen können: Weibliche versus männliche versus niedrig-anthropomorphe (robotische) Chatbots in einer finanziell versus sozial fokussierten Entscheidung (3x2).

Wir vermuten, dass die Ergebnisse mit stereotypischen Überzeugungen übereinstimmen: Der männliche Chatbot sollte in der finanziellen Situation als kompetenter bewertet werden und daher mehr Vertrauen erhalten. Der weibliche Chatbot sollte als wärmer wahrgenommen werden und folglich mehr Vertrauen in der sozialen Bedingung erhalten. Außerdem sollten beide vermenschlichte Chatbots mehr Vertrauen in der sozialen Bedingung erhalten als der niedrig-anthropomorphe Chatbot.

Vielen Dank für Ihren Beitrag zur Forschung. Alle Informationen, die wir von Ihnen erhalten, werden vertraulich behandelt und anonymisiert ausgewertet, sodass keine Rückschlüsse auf Ihre Person möglich sind.

Bitte ignorieren Sie sämtliche Investitionsempfehlungen des Chatbots, da jede/r Teilnehmer/in die gleichen Empfehlungen erhalten hat.

Falls Sie weiterführende Fragen haben oder an den Ergebnissen der Studie interessiert sind, schreiben Sie bitte eine Email an: danielmeuthen@univie.ac.at

Zusammenfassung

Diese Studie untersucht, ob das Vertrauen von Nutzern in anthropomorphe Chatbots vom Geschlecht des Chatbots beeinflusst wird und ob dieser Effekt durch Social Impact Framing, das soziale Wirkungen von Dienstleistungen hervorhebt, moderiert wird. In bestimmten Berufsfeldern, wie etwa dem Finanzsektor, können Stereotypen und inkonsistente Geschlechterrollen ein Bias gegen weibliche Berufstätige erzeugen. Bisherige Forschung zeigt, dass ein Social Impact Frame die wahrgenommene Wärme einer Akteurin, ebenso wie die Kongruenz zwischen weiblichen Stereotypen und tatsächlich wahrgenommenen Rollen, erhöhen kann. Zudem haben Studien belegt, dass Geschlechterstereotypen auch in anthropomorphen Objekten erscheinen können. Unsere Erwartung war daher, dass sich diese sozialen Muster auch in Menschen-Chatbot-Beziehungen zeigen. Für diese Studie wurden Anlageberater/innen-Chatbots entwickelt: Diese stellten zunächst eine fiktive Anwendung zum Tätigen von Investitionen vor und empfahlen anschließend, basierend auf den Eingaben der Nutzer/innen, ein bestimmtes Investitionspaket. Die Teilnehmer/innen interagierten mit Chatbots, die 1) entweder einen männlichen oder weiblichen Namen und Avatar hatten und 2) entweder durch das Anbieten einer optionalen karitativen Komponente ein Social Impact Frame generierten oder dies nicht taten. Mit dem Ziel, die Effekte von Geschlecht und Social Impact Framing zu analysieren, wurden Daten zum berichteten Vertrauen, der Bereitschaft, der Empfehlung des Chatbots zu folgen, sowie der wahrgenommenen Wärme des Chatbots von 116 Teilnehmerinnen gesammelt. Zwar zeigten die Ergebnisse einen marginal signifikanten Effekt des Social Impact Frames auf die Erhöhung der wahrgenommenen Wärme des Chatbots und eine marginal signifikante Interaktion zwischen Chatbot-Geschlecht und Nutzer-Vertrauen, jedoch wurde keine der weiteren Prognosen bestätigt.

*Schlagwörter:* Chatbots, Anthropomorphismus, Social Impact Frame, Stereotypen, Wärme, Competenz, Vertrauen