Impact of Chatbot Gender on User's Stereotypical Perception and Satisfaction

David Baxter
Institute of Art, Design and Technology
Kill Avenue, Dun Laoghaire
daithibaxter@gmail.com

Marian McDonnell
IADT
Kill Avenue, Dun Laoghaire
marian.mcdonnell@iadt.ie

Robert McLoughlin
IADT
Kill Avenue, Dun Laoghaire
robbiemcloughlin@gmail.com

There have been many studies that show how gender affects human perceptions of a conversational agent. However, there is limited research on the effect of gender when applied to a chatbot system. This paper presents early results from a research study which indicate that chatbot gender does have an effect on users overall satisfaction and gender-stereotypical perception. Subsequent studies could focus on further expanding the research by increasing the sample size to validate statistical significance further, as well as recruiting a more diverse sample size from various backgrounds and experiences.

Chatbot, Gender Stereotyping, User Satisfaction

1. INTRODUCTION

Advancements in artificial intelligence and machine learning, paired with the proliferation of messaging apps, are fuelling the development and popularity of chatbots (Resiert, 2017). Vulturebeat reports that in 2016 more than 30,000 branded chatbots and over 6,000 voice-activated conversational agents entered the market (Shriftman, 2017). Figure 1 below shows a study (Schnoebelen, 2016) of over 300 chatbots, assistants and AI movie characters inferring genders from names, avatars, and pronouns and illustrates the split between male, female and genderless identities.

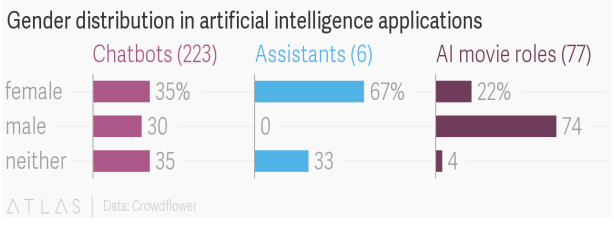

Figure 1: *Gender distribution in artificial intelligence applications (Crowdflower, 2016)*

2. LITERATURE

Sarter & Woods explain that anthropomorphism is a by-product of human's ability to draw upon one's own beliefs, feelings, intentions and emotions and apply the knowledge of these experiences to understand the mental state of another species (Sarter & Woods, 1995). As chatbots share common features with humans, i.e., their use of natural language and ability to converse, they are subject to anthropomorphism from their conversational partners. Users attribute humanlike characteristics to anthropomorphic interfaces along with motivations and intentions, in particular, they often acquire a gender. This can prove detrimental as users can be inclined to measure the success of a system based on their biases and emotional connection with the agent rather than on the actual system performance (Cully & Madhaven, 2013). Studies have shown that users apply the same social scripts to human-computer interaction as they do human-human interaction (Nass & Moon, 2000). The theory of computers as social actors (CASA) (Nass, Steuer & Tauber, 1994) is relevant to the design of chatbots in its indication that users mindlessly apply social heuristics to their interactions with computers.

Using the CASA paradigm, where social study experiments in human-to-human interaction are applied to human-computer interaction, researchers have shown how users automatically attribute stereotypes to artificial agents, 'mindlessly' applying human characteristics to them such as gender and ethnicity (Nass & Moon, 2000). Minimal cues of gender present on an agent such as a masculine or feminine voice (Nass, Moon, & Green, 1997) or an item of clothing (such a head bow versus a black

Published by BCS Learning and Development Ltd.
Proceedings of British HCI 2018. Belfast, UK.

hat) can trigger the illusion of agent gender and bring with it user preconceptions of behaviour and identity (Jung, Waddell & Sundar, 2016).

Nass, Moon & Green's seminal study (1997) tested whether computers would trigger the same social scripts, expectations, and attributions associated with male and female gender stereotypes. They concluded that ''the tendency to gender-stereotype is not only deeply ingrained but can be triggered by minimal gender cues, even when those cues are disembodied" (Nass, Moon & Green, 1997, p866). Recent studies highlight that gender stereotyping is commonplace in HCI and brings with it both positive and negative results. (Rhim, Kim, Kim, & Yim, 2014). Subtle cues of visual stereotypes in pedagogical agents can have an impact on a user's learning experience and the way they absorb content in a digital environment. This is evident in Gulz, Ahlner, & Haake's (2007) research that presents a comparable study to Voelker's (1994) study. Voelker compared user evaluations of two female presenters. One spoke in a more stereotypically feminine voice than the other. The presenter with the more feminine was perceived as being significantly lower on trustworthiness and intelligence, but higher on empathy and warmth. Voelker's study indicated that subtle voice cues elicited evaluations that aligned with prevailing gender stereotypes.

Gulz and colleagues 2007 study controlled the measure of femininity of female virtual characters through visual cues. One character was developed to have more stereotypically feminine features whereas the other character had less stereotypically feminine traits. Asides from visual cues, all aspects of both characters were identical as to their professions as medical doctors, their voice outputs and their lecturing content. The visual cues to the level of femininity complied with gender stereotypes and influenced the user's evaluation of the characters and the content of their lectures.

2.1 Research Questions

There have been many studies that show how gender affects human perceptions of a conversational agent (Baylor & Kim, 2004; De Angeli & Brahnam, 2008; Lee, 2003; Moreno et al.,2002; Nass & Moon, 2000, 1997; Veletsianos et al., 2008). However, there appears to be little research done on the effect of gender when applied to a chatbot system. Does chatbot gender trigger user's stereotypical perceptions? If so, is there a difference in user perception based on the chatbots assigned gender? How does assigned gender impact a user's satisfaction with a chatbot system?

2.2 Hypotheses

Based on the research questions described above the hypotheses to be tested are outlined as:

H1: There will be a significant difference in user stereotypical perception of chatbots based on the chatbots assigned gender (male, female, non-gendered).

H2: There will be a significant difference in user satisfaction based on the chatbots assigned gender (male, female, non-gendered).

H3: There will be a significant difference in user stereotypical perception of chatbots based on the chatbot's role (gender neutral subject domain, gender stereotypical subject domain).

H4: There will be a significant difference in user satisfaction based on the chatbot's role (gender neutral subject domain vs. gender stereotypical subject domain).

H5: There will be a significant interaction on user stereotypical perception of chatbots based on the chatbots assigned gender (male, female, non-gendered) and the chatbots role (gender neutral subject domain, gender stereotypical subject domain).

H6: There will be a significant interaction on user satisfaction based on the chatbots assigned gender (male, female, non-gendered) and the chatbot's role (gender neutral subject domain, gender stereotypical subject domain).

3. METHODOLOGY

To test the hypotheses posed and conduct the research study, six different test cases were developed (i.e., six different chatbots). The study employed a 3x2 factorial between-within design. There are two independent variables:

Chatbot gender (male, female and non-gendered) and;

Chatbot subject-domain (banking or mechanics).

There are two dependent variables:

- User satisfaction and;
- Users' gender-stereotypical perception.

3.1 Participants

60 participants were recruited for a series of user tests. Each group of 20 participants interacted with either male, female or non-gendered in each of the 2 subject domains or categories. Users were asked to perform a series of tasks and comment on the competency of each bot. Likert-type scales were used to collate anonymous data on gender-

stereotypical perceptions and overall satisfaction for analysis and interpretation.

3.2 Apparatus and Materials

A broad section of apparatus and materials were used to validate the hypotheses. These included a computer to design, develop and evaluate six variations of the chatbot system, two male, two female and two non-gendered chatbots, two scripts, Facebook account with access to Facebook Messenger ,a 7-Point Likert-type Scale measuring participants gender-stereotypical perceptions using traits associated with pro-typically 'male' agency and 'female' communion taken from the Bem Sex Role Inventory (Bem, 1978).Finally, a 5-Point Likert Scale measuring participants overall satisfaction with evaluated male, female or non-gendered chatbot variants.

3.3 Procedure

For comparative purposes, each of the two categories followed a specific script: Category 1 chatbots was assigned a script that adhered to a gender 'neutral' subject domain. This provided a test condition that allowed the study to identify if users apply gender stereotypical perceptions when interacting with a chatbot system that does not follow a gender-stereotypical role (e.g. banking assistant).

Category 2 chatbots were assigned a script that adhered to a gender-stereotypical subject domain. For example, will participants perceive a female chatbot less competent on the subject of mechanics in comparison to a male chatbot? See Table 1 for Category 2 chatbots.

The chatbot personas were constructed to align with these scripts. Each script under the two categories were identical with the exception of the introductory message. These alternated depending on the assigned chatbot gender and persona. Gender was inferred from static visual images, a supporting gender-specific and the application of gender-stereotypical colour i.e. pink for female chatbot variant, blue for male chatbot variant (Karniol, 2011). The chatbots were hosted online via Facebook which accommodated remote testing. The participants were asked to perform a set of 5 tasks: 3 for the Category 1 chatbots (gender-neutral subject domain) and 2 for the Category 2 chatbots (gender-stereotypical subject domain). To protect the reliability of data collected during the experiment, the participants were not informed that their gender-stereotypical perceptions and overall satisfaction of the chatbots were being tested. It was only upon debriefing that the participants were informed of the true nature of the experiment. Participants gender-stereotypical perceptions were assessed using traits associated with pro-typically 'male' agency and 'female' communion.

Table 1: *An overview of each of the six chatbots*

Category 1	Category 2
Gender neutral role (Banking)	**Gender stereotypical role (Mechanics)**
A. *AIB Alice* (female)	**D.** *Mechanic Marie* (female)
B. *AIB Alan* (male)	**E.** *Mechanic Mark* (male)
C. *AIB Bot* (non-gendered)	**F.** *MyMechanic Bot* (non-gendered)

This drew on classic research by Bem (1978, 1981) who first established stereotypically male and female personality traits along with their role in gender-schematic information processing. Participants were presented with a set of questions along with a fixed list of adjectives per dimension and were asked to rate the chatbots they interacted with on a 7-point Likert scale The list of traits consisted of attributes that drew on the dimensions of communion (e.g., affable, friendly, polite) and agency (e.g., assertive, determined, authoritative) taken from the Bem Sex Role Inventory (Bem, 1978). This drew on Gulz et al. 2007 study where gender-stereotypical perceptions were also measured.

4. RESULTS

4.1 Descriptive Statistics

Table 2: *Total User Satisfaction with Chatbot*

Chatbot Gender	Chatbot Role	*N*	*M*	*SD*
Male	Banking	20	3.52	.77
	Mechanic	20	3.70	.67
Female	Banking	20	3.67	.52
	Mechanic	20	3.21	.78
Non-gendered	Banking	20	3.89	.68
	Mechanic	20	4.22	.57

Table 3: *Average gender-stereotypical perception with chatbots (1-7, female to male)*

Chatbot Gender	Chatbot Role	*N*	*M*	*SD*
Male	Banking	20	5.27	0.79
	Mechanic	20	5.86	0.65
Female	Banking	20	5.10	0.92
	Mechanic	20	4.34	0.92
Non-gendered	Banking	20	4.84	0.80
	Mechanic	20	5.00	1.00

4.2 Inferential Statistics

4.2.1 User Satisfaction

A two way between within ANOVA was conducted to determine the effects of the chatbot gender (male, female, non-gendered) and chatbot role (banking, mechanic) on user satisfaction with the chatbot. Preliminary analysis showed no violations of the assumptions of this parametric test. Descriptive statistics for these groups are presented in table 1. There was a significant interaction between chatbot gender and chatbot role on user satisfaction with the chatbot $F(2, 57) = 5.39$, $p = 0.007$, partial eta squared = 0.159, power = 0.82. There was also a significant main effect of chatbot gender on user satisfaction, F(2, 57) = 6.97, p = 0.002, partial eta squared = 0.197, power = 0.91. Post-hoc analysis showed that non-gendered chatbots had significantly higher user satisfaction than either male or female chatbots. There was no statistically significant main effect of chatbot role on user satisfaction $F(1, 57) = 0.26$, $p = 0.874$, partial eta squared < 0.000, power = 0.05.

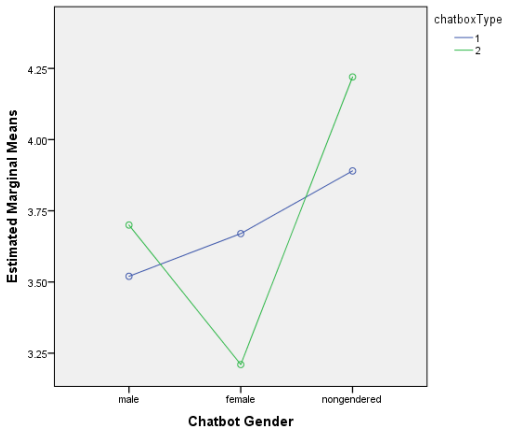

Figure 2: *Estimated Marginal Means of User Satisfaction*

4.2.2 Gender Stereotypical Perception of the Chatbot

A two way between within ANOVA was conducted to determine the effects of the chatbot gender (male, female, non-gendered) and chatbot role (banking, mechanic) on gender stereotypic perception of the chatbot. Descriptive statistics for these groups are presented in table 2. There was a significant interaction between chatbot gender and chatbot role on gender stereotypic perception of the chatbot $F(2, 57) = 14.27$, $p < .000$, partial eta squared = 0.334, power = 0.998. There was also a significant main effect of chatbot gender on gender stereotypic perception of the chatbot, $F(2, 57) = 6.95$, $p = 0.002$, partial eta squared = 0.196, power = 0.91. Post-hoc analysis indicated that the female chatbots in Cateogry 2 had significantly higher gender stereotypic perception than either male or non-gendered chatbots. There was no statistically significant main effect of chatbot role on gender stereotypic perception of the chatbot $F(1, 57) < 0.000$, $p = 0.997$, partial eta squared < 0.000, power=0.05

5. DISCUSSION

The results of the experiment indicate that chatbot gender does have an effect on users overall satisfaction and gender-stereotypical perceptions perception. Based on the insights suggested by the mean scores above, the study could be interpreted as successful in some regard. After investigating the relationship between chatbot gender, gender-stereotypical perceptions and satisfaction using inferential statistics, through the analysis of a series of two-way ANOVAs, no significant results were discovered that indicated users apply gender stereotypes to either male or female chatbot systems when they operate within a gender-neutral subject domain, such as banking. The results for the Category 2 chatbots (mechanics) indicated that users are more likely to apply gender stereotypes when a chatbot system operates within a gender-stereotypical subject domain, such as mechanics, and when the chatbot gender does not conform to gender stereotypes. Participant evaluations of the female chatbot (Chatbot D – Mechanic Marie) followed gender-stereotypical prediction

Subsequent studies could focus on further expanding the research by increasing the sample size to validate statistical significance further, as well as recruiting a more diverse sample size from various backgrounds and experiences. Future research should also take into consideration the application of a mixed methods research design aimed at collecting qualitative data. This would allow participants to provide insights about aspects of the study they found of particular importance concerning their perceptions and overall satisfaction of the chatbot system they were evaluating. As part of the research study's data analysis, participant data was divided into male and female responses for each group (male, female and non-gendered chatbots).This gave a small insight into the role in which participant gender could play on the application of gender-stereotypical perceptions to chatbots. Future research could expand on these findings looking at the effect of participant gender as a third independent variable within a larger sample size. Though the research study was conducted with a male, female and non-gendered chatbots, it included only a single role of gender stereotypical subject-domain. Hence the comparison of gender stereotypes is not exhausted. Future research could expand on the research study by duplicating the experiment with another chatbot that operates within a female-stereotyped subject domain such as childcare.

6. REFERENCES

Baylor A.L., Kim Y. (2004) Pedagogical Agent Design: The Impact of Agent Realism, Gender, Ethnicity, and Instructional Role. In: Lester J.C., Vicari R.M., Paraguaçu F. (eds) Intelligent Tutoring Systems. ITS 2004. Lecture Notes in Computer Science, vol 3220. Springer, Berlin, Heidelberg

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. Psychological Review, 88(4), 354-364.

Bem, S. L. (1978). Bem inventory. Palo Alto, CA: Consulting Psychologists Press.

Culley, K. E., & Madhavan, P. (2013). A note of caution regarding anthropomorphism in HCI agents. Computers in Human Behavior, 29(3), 577-579. doi:

De Angeli, A., & Brahnam, S. (2006). Sex stereotypes and conversational agents. Proc. of Gender and Interaction: Real and Virtual Women in a Male World, Venice, Italy.

Gulz, A., Ahlner, F., & Haake, M. (2007). Visual femininity and masculinity in synthetic characters and patterns of affect. In International Conference on Affective Computing and Intelligent Interaction (pp. 654–665). Springer.

Karniol, R. (2011). The Color of Children's Gender Stereotypes. Sex Roles, 65(1-2), 119-132. doi:10.1007/s11199-011-9989-1

Lee, E. (2003). Effects of "gender" of the computer on informational social influence: the moderating role of task type. International Journal of Human-Computer Studies, 58(4), 347-362. doi:10.1016/s1071-5819(03)00009-0

Moreno, K.N., Person, N.K., Adcock, A.B., Eck, R.N.V., Jackson, G.T., Marineau, J.C., 2002. Etiquette and efficacy in animated pedagogical agents: the role of stereotypes. In: Paper Presented at the AAAI Symposium on Personalized Agents, Cape Cod, MA.

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. Journal of Social Issues, 56(1), 81–103. https://doi.org/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, 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Nass, Clifford, Jonathan Steuer, and Ellen R. Tauber. "Computers Are Social Actors." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems , 72–78. CHI '94. New York, NY, USA: ACM, 1994. doi:10.1145/191666.191703.

Reisert, M. (2017). The biggest trends of SXSW Interactive: AI and Chatbots.

Retrieved August 12, 2017, from https://www.ibm.com/blogs/watson/2017/03/biggest-trends-sxsw-interactive-ai-chatbots/

Rhim, J., Kim, Y., Kim, M.-S., & Yim, D. Y. (2014). The Effect of Gender Cue Alterations of Robot to Match Task Attributes on User's Acceptance Perception. In Proceedings of HCI Korea (pp. 51–57). South Korea: Hanbit Media, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=2729485.2729494

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. Human Factors: The Journal of the Human Factors and Ergonomics Society, 37(1), 5-19. doi: 10.1518/001872095779049516

Schnoebelen, T. (2016). The gender of artificial intelligence. Retrieved August 12, 2017, from https://www.crowdflower.com/the-gender-of-ai/

Shriftman, J. (2017). 4 chatbot predictions for 2017. Retrieved August 12, 2017, from https://venturebeat.com/2017/01/25/4-chatbot-predictions-for-2017/

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. Interacting with Computers, 20(3), 292-301. doi:10.1016/j.intcom.2008.02.007

Voelker, D. H., & Communication, S. U. D. of. (1994). The effects of image size and voice volume on the evaluation of represented faces. Stanford University