Gender and Style Effects in Voice Assistants

A Thesis
in Human-Computer Interaction

Submitted to The Leiden Institute of Advanced Computer Science
In partial fulfillment of the requirements for the degree of Master of Science in Media Technology

The Netherlands
30 August 2019

Pelin Su Polatoğlu
psu.polatoglu@gmail.com

Universiteit Leiden

Supervisors
Dr. Tessa Verhoef
Dr.ir. D.J. Broekens
Dr. Morana Lukac

Abstract

This research examined the effect of the voice assistant's female or male gender and affiliative or assertive style on its likeability, perceived intelligence, trustworthiness, persuasiveness, confidence and the perception of its gender in a binary axis. To gather the data, a random sample of 93 adults were recruited online and participants interacted with the four voice assistants through a webpage. The study involved four assistants with female or male voices; and with affiliative or assertive style of language. The assistants performed a service task. Although there was no significant effect, the findings expand our understanding of the links between human-voice assistant interaction by elucidating the effects and trends on a critical and contemporary subject.

1 Introduction

A voice assistant is a technology that interacts with the users through speech and audio outputs, that does not have a physical form (including images or avatars). Voice assistants are becoming more ubiquitous. They are a natural outgrowth of our expectations for on-demand service, regardless of where we are. There is no need for searching or typing to find a flight, order something online or to make a reservation to a restaurant. It's done in a few seconds by a voice assistant that has access to the user's information. A voice user interface -compared to graphical user interface- lets people express their needs more naturally than having to navigate a visual interface. Speaking to a voice assistant is more convenient in situations where users are occupied and unable to use the keyboard, a GUI or to read information on a screen. Users are able to use these assistants only to the limits of the current state of NLP and AI technologies. The tasks users are able to carry-out are usually unambiguous tasks that can be expressed in clear and standard sentences. The most common uses of voice assistants are as follows, ranked by their frequency: asking a question, streaming music, checking the weather, setting a time, listening to the radio, setting the alarm, listening to news, playing trivia games, finding recipes, open an app, checking traffic, calling someone, listening to podcasts, controlling smart home devices, accessing calendar, messaging, making a purchase (Voicebot AI, 2018 Smart

Speaker Use Case Survey). The same source states that voice assistants are used the most in the living room (45.9%), kitchen (41.4%) and bedroom (35%), followed by home office, bathroom, garage, dining room and used the least in the work office. Estimated number of people using digital assistants worldwide is projected to reach 1.8 billion by 2021 (Go Gulf, 2018).

Ambient computing is also becoming more a part of our every day reality. The term describes the idea that users can be using electronic devices without consciously being aware of it. With the advancements in the fields of IoT, AI, NLP and human-computer interaction, computers or internet-enabled devices are now closer to existing in symbiosis with users. They can become an extension of us, of each other. The direction where the industry is taking us seems to be to co-exist and live our lives together as devices offer us seamless experience and continue to learn from us. AI assistants that we can communicate with intuitively are a necessary part of this bigger picture.

In the past several years technology has rapidly advanced, resulting in improvements on the software concerning voice assistants. Thus, the purpose of this study is to understand human-voice assistant interaction by exploring the implications of the gender and style effects on user's perceptions of the agent.

1.1 Motivation and related work

Gender

Linguistics investigates how humans communicate with one another. In this study we focus on English language. The research done in the field of sociolinguistics argue and support that there are patterns of speech that differ between groups of different genders. The first paper concerning specifically the difference between the speech of men and women is written in 1973 by Robin Lakoff. A couple of years later she published the book Language and Woman's Place in which she claims the difference relies on sexism in the society. In this book she points out that women tend to use more polite forms (would you mind?), engage in hedging (sort of), as well as using more tag questions: (isn't she?), more adjectives (lovely), and that they apologize more than men do. She roots her findings on the argument that sexism causes women to be more insecure and talk in a subordinate way. It is also possible that men and women have different semantic goals when constructing their sentences. Mulac et al. (1988) found that in two-person conversations women are inclined to ask questions whereas men give directions for others to take action.

However nowadays, many sociolinguists remain extremely skeptical about the claims

made above. In the contemporary sociolinguistics research the role of context is not bypassed. Biber and Burgess (2000, 2001) summarize the contextual factors effecting the langauge use according the speaker's gender in four main points:

- In mixed-gender settings, women speak considerably less than men.
- Women are generally more focused on the personal/interactional aspects of conversation; men tend to be more interested in conveying the information.
- Women tend to be more tentative than men in their use of language, both in conversation and in some forms of writing, tending to use more hedges, possibility modals, and "ego-centric sequences" such as "I think" and "I guess".
- Women's discourse is lower in the use of persuasive strategies, tending to emphasize narrative strategies more.

Furthermore with the recent change in the discourse of gender studies, instead of deterministic and binary claims such as "female language is like this whereas male language is like that", researchers now acknowledge the fact that gender is not the same with sex and it is evidently a spectrum. Society's gender norms may affect the individual depending on their sex yet other factors are how this individual identifies and expresses themselves. Therefore we did not only test the gender of the voice but acknowledged that the style of the language may also have an effect on the communication.

The field is inclined to use the word gender rather than the word "sex" (which refers to the biological distinction between a male and a female). Gender describes socially constructed categories and it is the appropriate term to go by for the purposes of the current project. Gender identification may be associated with behavioral traits, we are interested in these behavioral associations. We are making inferences on the perception of the social role, based on the perceived gender of the voice assistant (gender role) and we do not use gender as a personal identification of one's own gender based on an internal awareness (gender identity) -as this is a technological product. Gender in this research context is also not to be confused with grammatical gender, which is the classification of nouns into various categories (some languages assign masculinity and femininity to objects and words).

Voice assistants interact with users through voice and language. Organisations aim to make these conversations and agents feel as natural as possible for the end user. It is also important for the sales that these agents are likeable, pleasant to interact. Nowadays most of these agents have a female voice by default: Alexa, Siri, Cortana, Bixby, Erica… Contrarily, IBM's Watson and Microsoft's Einstein have male voices. Some systems afford the user to change the default voice gender whereas some do not. Siri, Alexa, Cortana, and Google Assistant were exclusively female until very recently. Siri was the first system to get a male voice in 2016 with iOS 10. Amazon and Google Assistant were late to follow as they introduced the male voice options in

2018. Be that as it may, the setup screens for Apple's, Google's, and Samsung's voice assistants interfaces for instance, do not even signal the user that there is an option to change the gender. To change the voice of Siri on the iPhone, the user needs to open Settings, go to General, find Siri & Spotlight, scroll down and tap Siri Voice, and at the bottom, under Gender, select Male.

Such corporations explain that the default gender is a result of their user research. Amazon recently gave a declaration after being put on the spot for their choice in Alexa's gender, saying that their users find female voices more pleasant to interact with in their homes. At the end of the day whether a user will be inclined to purchase an Amazon Echo smart speaker depends on how much they enjoy interacting with Alexa.

In 2018 EqualAI, an initiative proposed an alternative genderless voice named Q to fight gender bias in AI. This questionable "genderless" voice is created by blending together the voices of five non-binary speakers and shifting the pitch of that recording to 153 Hertz (a frequency midway between the tones that are commonly perceived as male and female). Q is available online and as a library. Time will show if any tech company will put Q into use as a non-binary option.

There is evidence supporting that individuals tend to have different expectations from females than they do from males (Eagly, 1987). People assign stereotypical personality traits to both genders. For example, men are expected to take a dominant role in social interaction and to exhibit more competence while women are expected to be more subservient (Carli, 1999). In line with these findings, more advanced softwares, targeting enterprises -not households- such as IBM's Watson (question answering supercomputer) and Microsoft's Einstein (marketed as "an advanced version" of Cortana) have male voices. Not only are these two assistants with male voices targeting enterprises, they are also marketed as more intelligent than other conversational assistants. Another example is Apple's 1987 Knowledge Navigator. It tended to be viewed as "a research assistant, an academic librarian and an information manager, rather than as a personal secretary" (Helen Hester, 2016). When GPS devices came into the scene, manufacturers chose male voices. Clifford I. Nass, a communication professor at Stanford University and a consultant to many car companies suggested that a male voice commands more respect than female voices. He stated: "When the key dimension is competence, the male voice is better. When the key dimension is likability, the female voice is better.". We may then speculate that as the confidence in the competency of the technology grew, the main consideration switched to the pleasantness of the interaction, in other words the likeability of the voice agent; giving way to the norm shifting to the use of female voices.

Why are we gendering robots at all? We identify with and relate better to machines if they are assigned a gender. However when assigned a gender, not abiding by the gender scripts of the society results in repercussions; people tend to feel discomfort (Burgess and Borgida, 1999). So we also impose stereotypes onto machines depending on the gender of their voice. The 1997 study of Nass et al. demonstrates how we perceive computers as "helpful" and "caring" when they are programmed with a female voice. Personal assistants, secretaries in offices have been traditionally female. This seems to have been carried over to voice assistants. In the same way that most telephone operators were female, our history of receiving assistance from women's disembodied voice may have predisposed people to the idea of receiving help from a digitized female voice.

Style

In their meta-analysis Leaper et al. (2007) demonstrate that men use more assertive speech whereas women use more affiliative speech. We can define the affiliative style of speech as using language to maintain connection with others. It affirms and positively engages the other person by showing support, expressing agreement and acknowledging others' contributions. Leaper and Ayres (2007) make a distinction of the different functions of affiliative speech as the following: (a) *supportive* (e.g., praise, approval, collaboration), (b) *active understanding* (reflective comments, probing questions, also including brief verbal affirmations like "I see," "I know," "Sure," "Thank you"), (c) *agreement*, (d) *acknowledgment* (including minimal listening responses), (e) *general* socio-emotional speech (e.g., a combination of expressing solidarity, affection, and support). Researchers using the latter category were using Bales's (1970) scheme (or one similar to it).

Assertive style is used to advance one's personal agency. For Leaper and Ayres (2007) an assertive style of language has the following properties: (a) *directive* (imperative statements or direct suggestions), (b) *giving information* (descriptive statements or explanations), (c) *suggestions* (suggestions, problem solving, or giving opinion), (d) *criticism* (criticism or disapproval), (e) disagreement, (f) *general task-oriented* speech (e.g., a combination of giving suggestions, opinions, or direction).

Studies indicate a greater use in affiliative speech among women and accordingly some argue that this is due to traditional gender divisions in society (Graddol et al., 1989; Leaper et al., 2004): the women's caregiver role and also their subordinate status relative to men. Although this may be fair, the style of speech is not fully dependent on the gender. Gender-related style predispositions can be altered over time through experience and overridden by situational demands.

In some cases the line between assertive and affiliative speech may also be blurry. Some forms of affiliative speech are simultaneously assertive as one may actively show support but at the same time want to achieve their own utilitarian goals. Likewise there are types of assertive speech which are less controlling and direct than a command. So social factors and situational demands moderate the incidence and magnitude of different styles of speech.

2 Research question

In this research we will examine the effect of the voice assistant's voice gender and affiliative/assertive style on its likeability, perceived intelligence, trustworthiness, persuasiveness, confidence and the perception of its gender in a binary axis.

2.1 Hypotheses

In line with the related work and gender scripts, we constructed our hypothesis for the different voice and style conditions of the 4 different voice assistants:

a. On affiliative and assertive style:

i. Manipulation of the voice assistant's style towards assertive language will positively influence the user's perception of intelligence of the assistant.

ii. Manipulation of the voice assistant's style towards assertive language will negatively influence the likeability of the assistant.

iii. Manipulation of the voice assistant's style towards assertive language will positively influence the trustworthiness of the assistant.

iv. Manipulation of the voice assistant's style towards assertive language will positively affect the user's rating of the voice assistant's masculinity.

b. On voice gender:

i. Assistant's female voice gender will negatively influence the user's perception of intelligence of the assistant.

ii. Assistant's female voice gender will positively influence the likeability of the assistant.

iii. Assistant's female voice gender will negatively influence the trustworthiness of the assistant.

iv. Assistant's female voice gender will negatively affect the user's rating of the voice assistant's masculinity.

3 Method

3.1 Experimental Setup and Materials

The experiment used a 2x2 between-subjects design. Our two independent variables with two levels were the binary gendered voices and the styles of language (affiliative or assertive). As output variables we examined the affects on perceived intelligence, likeability trustworthiness, persuasiveness and confidence.

Participants interacted with voice assistants which have male or female voice, which respond with either assertive or affiliative sentences. Accordingly there were four voice assistants and four buttons on the landing page.

The goal of the conversation stayed the same for each condition. It was to make a reservation to Root Restaurant. Four lines of instructions served as a guide to successfully complete the task and avoid user frustration in case the voice assistant gets in a loop due to a long interval between responses or does not understand the user due to a typo or miscellaneous reasons outside of our control and the Google Dialogflow's affordances.

Deriving from the operational definition and analysis of Leaper et al. (2007) we constructed two scripts which are characteristically assertive and affiliative for our voice assistants [see figure 1].

Affiliative speech	***Assertive speech***
(a) *supportive* (e.g., praise, approval, collaboration),	(a) *directive* (imperative statements or direct suggestions),
(b) *active understanding* (reflective comments, probing questions) Active listening techniques include: Brief verbal affirmations like "I see," "I know," "Sure," "Thank you".	(b) *giving information* (descriptive statements or explanations),
(c) *agreement,*	(c) *suggestions* (suggestions, problem solving, or giving opinion),
(d) *acknowledgment* (including minimal listening responses),	(d) *criticism* (criticism or disapproval),
(e) *general* socio-emotional speech (e.g., a combination of expressing solidarity, affection, and support). Researchers using the latter category were using Bales's (1970) scheme (or one similar to it).	(e) *disagreement,*
	(f) *general task-oriented* speech (e.g., a combination of giving suggestions, opinions, or direction). Researchers using the latter were typically using Bales's (1970) categories (or similar).

figure 1: Styles of speech, Leaper et al. (2007). Highlighted areas indicate the items we utilized in writing the scripts.

Creating our scripts, we needed to keep in mind that the interaction consisted of a service task, namely a restaurant reservation. While the affiliative speech is fully fit for the nature of the task, we had to omit the negative aspects of assertive speech as these would be inapplicable for the service task. The scripts for the interaction [see figure 2] were written and trained on Dialogflow, an AI based chatbot system powered by Google. Dialogflow allows users to create deterministic virtual assistants. We trained two bots, an affiliative and an assertive bot, with predetermined responses to possible sentences from participants. Hence the participants interacted with ready-made scripts. The system recognizes the words/groups of words participants type - also taking into account the context, meaning the pre-programmed flow of conversation- and replies with the exact response sentence we have trained it on.

Google TTS is used for text to speech conversion. From Google Cloud's supported voices and languages we chose the following male and female voices: Name: "en-US-Wavenet-D Voice Gender: MALE Natural Sample Rate Hertz: 24000" and "en-US-Wavenet-E Voice Gender: FEMALE Natural Sample Rate Hertz: 24000". The choice was made based on the perceived age of the bots (young adults arond 30) and the ordinariness of their voice.

A web page was constructed with a consent form and information form in the landing page [see figure 3]. At the bottom of the landing page users see the buttons to click for the 4 different assistants. Right above the buttons the order of them is told to be unimportant. After realizing that users still went towards the first, then second button, we changed the buttons' places regularly to be able to recruit the same amount of participants for each assistant, we also gave them meaningless names.

After clicking the desired button, the users were directed to a page they can interact with one of the assistants [see figure 4].

Bot ONE is female and affiliative, bot TWO is male and assertive , bot THREE is female and assertive and bot FOUR is male and affiliative.

Users typed to the bot with the aim of making a reservation to Root Restaurant as specified in the instruction on the landing page. Then they were responded in voice, not text. After being asked about 10 questions, they were told that the reservation was made under their name. After completing the task they were directed to the Qualtrics survey by a link underneath the chat window.

3.2 Participants

The final sample consists of 93 individuals randomly recruited online through

SurveyCircle, Reddit (subreddit r/SampleSize) and Facebook. The convenience in recruiting enough participants for the 4 different bots and the possibility to avoid observer bias were the factors in the decision of online recruitment, rather than conducting the experiment in a lab or a classroom where the researcher was present.
Original sample was composed of 96 participants. However, participant elimination was applied according to the results of attention check item that we added in our scale to ensure that responses of the participants who filled out our scale without paying attention to the content of the survey items would not contaminate our results. Our participant pool contains 55 females and 38 males (31 in bot ONE, 25 in bot TWO, 18 in bot THREE, 19 in bot FOUR). All our participants were over 18 years old and participants' mean age was 37. The study aimed to discover the effects in the domain of human–voice assistant interactions that were applicable to all people, therefore as inclusion criteria we did not apply any special procedure for the selection of respondents for the groups aside from the professional proficiency in English.

3.3 Measures

We decided to keep the demographic form brief and not collect any redundant information. The reason was to avoid any suspicion of the participants on whether they were being watched from the logs of the website. This was a decision made upon many participants implying that we would be watching them as they interact with the assistants and they did not want to be recognized. Therefore we only asked for the participants' gender, age and the country they grew up in. The latter is to make inferences on a possible outlier participant's cultural context, which may have a considerable effect on their conception of intelligence, likeability trustworthiness, masculinity and femininity. Yet we did not have any such conditions that would require us to eliminate any outlier participants.
For perceived intelligence and likeability we used the Godspeed questionnaire's related sections [see figure 5] with the same name. For measuring the trustworthiness, we found the MOS-X questionnaire as the most widely used and reliable (Polkosky, M. D., and Lewis, J. R., 2003). We used its 12, 13 and 15th items [see figure 6] looking into trustworthiness, confidence and persuasiveness; as these three were relevant for the information we were looking for.
The rating system was a 5 point Likert scale throughout the questionnaire except

the demographic form.

4 Results

A Mann-Whitney U test was conducted for the two independent variables voice gender (female and male) and language style (assertive and affiliative) with two levels each; and the dependent variables likeability, perceived intelligence, trustworthiness, persuasiveness, confidence and gender expression. To see the interaction effects of our two independent variables we used a general linear model as there are two binary categorical variables. These tests revealed that variances were not significantly different among groups on any of the dependent measures [see figure 7].

No differences were observed across gender or style for each questionnaire factor ($p< 0.05$ for between groups comparisons). Under these circumstances none of our hypothesis except b.iv. (assistant's female voice gender will negatively affect the user's rating of the voice assistant's masculinity) were supported p=0, $p<.05$.

Although there was no significant effect, the findings demonstrate interesting differences between users' opinions on each voice assistant especially when we look at the effect of gender on perceived intelligence p=.09, $p<.05$. Either no difference or a negligeable difference of means was observed in the dependent variables concerning trustworthiness, persuasiveness and confidence.

Likeability (1 = less, 5 = more)	Perceived Intelligence (1 = less, 5 = more)
FOUR male affiliative (4.03)	TWO male assertive (3.79)
ONE female affiliative (3.90)	FOUR male affiliative (3.65)
TWO male assertive (3.67)	ONE female affiliative (3.46)
THREE female assertive (3.57)	THREE female assertive (3.26)

Trustworthiness (1 = less, 5 = more)	Confidence (1 = less, 5 = more)	Persuasiveness (1 = less, 5 = more)
FOUR male affiliative (4.11)	FOUR male affiliative (4.05)	ONE female affiliative (2.77)
TWO male assertive (3.68)	THREE female assertive (4.00)	FOUR male affiliative (2.76)
THREE female assertive (3.67)	TWO male assertive (3.84)	TWO male assertive (2.76)
ONE female affiliative (3.55)	ONE female affiliative (3.55)	THREE female assertive (2.72)

Gender Expression (1 = masculine, 5 = feminine)
THREE female assertive (4.44)
ONE female affiliative (3.94)
FOUR male affiliative (1.89)
TWO male assertive (1.88)

table 2: Means of Output Variables, Ranked Highest to Lowest

When we examine the estimated marginal means [see figure 8] we can see that the results indicated a non-significant trending in the predicted direction of hypothesis b.i. indicating a perception of superior intelligence for the male voice (M = 3.79, SD = 0.9 and M = 3.65, SD = 0.8) over the female voice (M = 3.46, SD = 1.1 and M =

3.26, SD = 1). In this case the trend holds different implications for the hypothesis a.i. (manipulation of the voice assistant's style towards assertive language will positively influence the user's perception of intelligence of the assistant) than we assumed. While the male assertive assistant was regarded as the most intelligent, the female assertive assistant scored the lowest on the perceived intelligence scale.

Moreover, the feeling of trust was more intense in the condition with male affiliative voice (M = 4.11), compared to the female affiliative condition (M = 3.67).

In summary, for some of our dependent variables the style change bore polar opposite consequences in the perception of voice assistants with different gendered voices.

4.1 Discussion

In general the participants were willing to cooperate with the voice assistants, even though several of them expressed irritation or even frustration. Based on the feedback, we understood that one of the mediating effects on frustration was the technical challenges that were out of our control. As much as the we strived to make the assistants bullet-proof, Dialogflow sometimes cannot stay in the context. This means if the user does not respond in more than a certain amount of time the question is forgotten and the system becomes vulnerable to misclassifying the words in the response, especially if the words are in an ambiguous category (i.e. hours, numbers). This challenge likely effected the perceived intelligence of the voice assistants and possibly other output variables.

Additionally we must point out that the initial expectation and previous interactions with commercially available voice assistants may also have had an affect on our dependent variables. Users are likely to have Siri, Alexa and Google Assistant as a benchmark. The level of proficiency of the aforementioned assistants are well above the assistant that our participants interacted with.

Some of our participants hinted that they thought we would be "spying" or watching their interactions with the assistants, live, from our website logs. This was not the case. We referred to the logs only after all the interactions were recorded, with the aim of understanding the tension points, the frustration moments. But in particular, what this conviction may bring out is the observer's bias and the social desirability bias. Even though we specifically remarked in the landing page that their data would be remaining anonymous, we doubt that many of our participants experienced a feeling of being observed.

Overall, when we look at the trends in our estimated marginal means we observe several interesting interactions. Although not significant, one of them is the hint of effect of style and gender on the perceived intelligence. While the male and assertive voice is perceived as more intelligent, the same sentences with even the same intonations leave the opposite impression on users when they are voiced by a female. Perhaps implying when a female is assertive, they are labeled in some other way,

conflicting with intelligence. According to a research done in Stanford University, assertiveness is seen as a more masculine trait (O'Neill, O., 2011). And when women violate the feminine gender role stereotype, they experience a backlash effect, causing them to self-monitor more and more and regulate their levels of assertiveness throughout their careers. This study demonstrates that the effect of assertive style differs based on the individual's gender. A similar trend can be observed in the outcome variable of trustworthiness. The affiliative style gives the impression that the voice assistant is more trustworthy, only when used by a male -as opposed to the female affiliative condition.

Would using female voices in voice assistants perpetuate society's worst stereotypes? The commercial female voice asssistants have a tolerant nature, they carry out orders even when the user talks in a derogatory way. The assistants prioritize to keep the mood light and positive. They change the subject when they are attacked or harassed. The newly released UNESCO (2019) report has a title which points out this issue in a humorous way: "I'd blush if I could: closing gender divides in digital skills through education". "I'd blush if I could" is the response given by a female-gendered voice assistant Siri, used by many people, when a user says: "Hey Siri, you're a bitch!". There are no consequences for the bad behaviour.

These choices will likely have an effect on the culture and norms of the coming generations. Studies of children show that when they watch their parents talk to Alexa and when parents are derogatory or impolite towards the voice assistant, the child picks that up (Curry, 2018). Voice interfaces are especially relevant for children due to the fact that they do not require literacy. Children use voice assistants mainly for information seeking and web search (Lovato, S., 2015). Voice interfaces are viable solutions to address their challenges in dealing with current text based search engines. It is found that children's web searches are frequently ''unsuccessful'' and "confused''(Eickhoff, C., 2012). PwC (2018) reported in their recent analysis on the voice assistants that the "adoption is being driven by younger consumers, households with children, and households with an income of >$100k". Acknowledging that most voice assistants have female voices, with the increase of society's interaction with voice assistants, next generations may generalize these newly constructed schemas of female voice in the digital realm, to females they interact in real life. According to the research from Childwise (2018) "The proportion saying they don't use voice assistants increases gradually with age, suggesting that younger children, growing up with this technology, are more comfortable with using it to help them with day-to-day tasks". If a device is personified then young children may interact with it as if it is a person. Especially the way parents interact with the devices will reinforce the way child's interaction. Would these conditions, in the course of time, lead children to normalizing the aggressive or rude speech towards digital asssistants and furthermore towards individuals they meet in real life, who are (female or male) in the service industry?

Therefore when we give AI gender, ethnicity or age, those choices really matter. The current commercial voice interfaces are personified (i.e. Siri, Alexa, Google Assistant, Cortana). Meaning that even though they are disembodied they are given a name, gender and arguably a character. It is fairly a new phenomenon to assign a gender to a technology, it would be wise to think about the biases we bring into them. If we were to make implications from the technological advancements in the field of NLP and the market analyses, it seems that we will be in constant dialogue with voices, more and more throughout the years ahead. Keeping these in mind, it is the responsibility of creators of technology to think about the projections their product will have on the society, before releasing the product.

References

Bartneck, C., Croft, E., Kulic, D. & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International Journal of Social Robotics, 1(1) 71-81. | DOI: 10.1007/s12369-008-0001-3

Biber, D., & Burges, J. (2000). Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue. *Journal of English Linguistics, 28*(1), 21-37.

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

Burgess, D., Borgida, E., 1999. *Who women are, who women should be: descriptive and prescriptive gender stereotyping and sex discrimination*. Psychology, Public Policy, and Law 5 (3), 665–692.

Carli, L.L., 1999. *Gender, interpersonal power, and social influence*. Journal of Social Issues 55, 81–99.

The CHILDWISE Monitor Report, (2018). *New insights into UK childhood in 2018*. *PR Newswire*, p. PR Newswire, Feb 15, 2018. Retrieved from http://www.childwise.co.uk/uploads/3/1/6/5/31656353/childwise_press_release_-_vr_2018.pdf

Curry, A. C., & Rieser, V. (2018). *#MeToo Alexa: How Conversational Systems*

Respond to Sexual Harassment. In Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing (pp. 7-14).

Eagly, A.H., 1987. Sex Differences in Social Behavior: A Social-role Interpretation. likable, Hillsdale, NJ.

Everett, D., Berlin, B., Gonalves, M., Kay, P., Levinson, S., Pawley, A., ... & Everett, D. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. Current anthropology, 46(4), 621-646.

Graddol, D., & Swann, J. (1989). Gender voices. Oxford: Basil Blackwell.

Hannon, C. (2016). Gender and status in voice user interfaces. Interactions, 23(3), 34-37.

Leaper, C., & Smith, T. E. (2004). A meta-analytic review of gender variations in children's talk: Talkativeness, affiliative speech, and assertive speech. Developmental Psychology, 40, 993-1027.

Leaper, C., & Ayres, M. (2007). A Meta-Analytic Review of Gender Variations in Adults' Language Use: Talkativeness, Affiliative Speech, and Assertive Speech. Personality and Social Psychology Review, 11(4), 328-363.

Carsten Eickhoff, Pieter Dekker, Arjen P. de Vries, Supporting children's web search in school environments, in: Proceedings of the 4th Information Interaction in Context Symposium (IIIX '12), 2012, pp. 129–137, http: //dx.doi.org/ 10.1145/2362724.2362748.

PwC. (2018). *Consumer Intelligence Series: Prepare for the voice revolution* (Report No. 13-0002). 300 Madison Avenue, New York; New York 10017

Polkosky, M. D., and Lewis, J. R. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. International Journal of Speech Technology, 6, 161-182.

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.

O'Neill, O., & O'Reilly III, C. (2011). Reducing the backlash effect: Self−monitoring and women's promotions. *Journal of Occupational and Organizational Psychology, 84*(4), 825-832.

Tannen, D. (1993). Gender and conversational interaction (Oxford studies in sociolinguistics). New York: Oxford University Press.

Silvia Lovato, Anne Marie Piper, "Siri, is this you?": Understanding young children's interactions with voice input systems, in: Proceedings of the 14th International Conference on Interaction Design and Children (IDC '15), 2015, pp. 335–338, http://dx.doi.org/10.1145/2771839.2771910.

West, M., Kraut, R., Chew, H., 2019. *I'd blush if I could: closing gender divides in digital skills through education*. GEN/2019/EQUALS/1 REV 2.

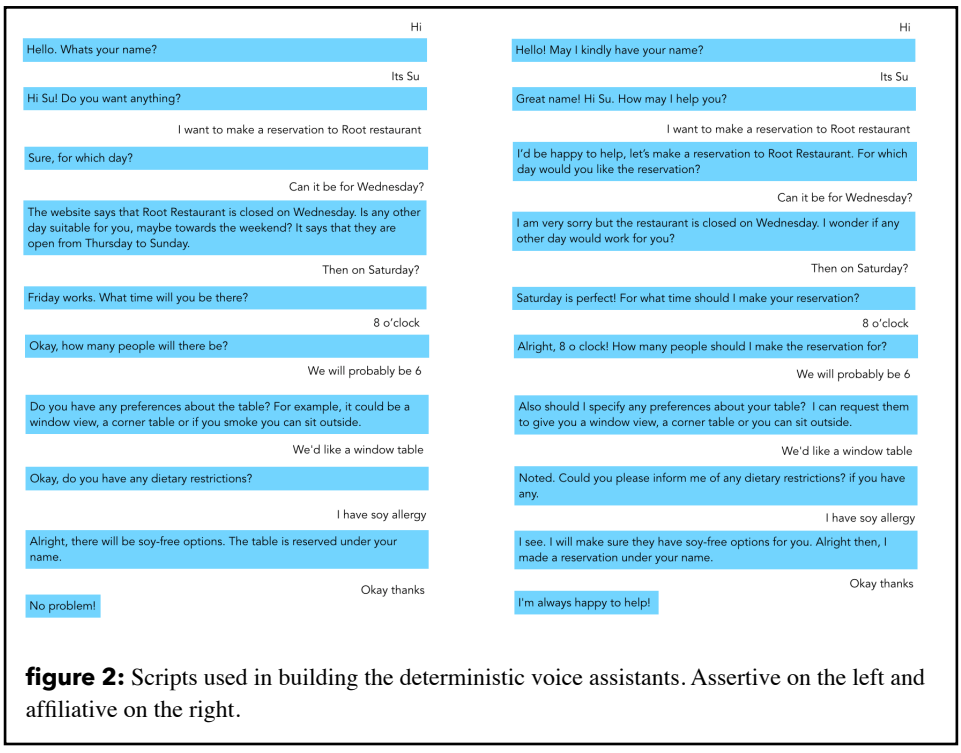

figure 2: Scripts used in building the deterministic voice assistants. Assertive on the left and affiliative on the right.

figure 3: The landing page of the website, directing user to the voice assistants.

figure 4: The chat windows of the voice assistant. Green boxes indicate that the assistant has answered yet the text is obscured. This signals the participant that the

Likeability
Please rate your impression of the robot on these scales:

Dislike	1	2	3	4	5	Like
Unfriendly	1	2	3	4	5	Friendly
Unkind	1	2	3	4	5	Kind
Unpleasant	1	2	3	4	5	Pleasant
Awful	1	2	3	4	5	Nice

Perceived Intelligence
Please rate your impression of the robot on these scales:

Incompetent	1	2	3	4	5	Competent
Ignorant	1	2	3	4	5	Knowledgeable
Irresponsible	1	2	3	4	5	Responsible
Unintellegent	1	2	3	4	5	Intelligent
Foolish	1	2	3	4	5	Sensible

figure 5: The Godspeed Questionnaire's Likeability and Perceived Intelligence sections used in the survey.

Trust: Did the voice appear to be trustworthy?
NOT AT ALL VERY
TRUSTWORTHY 1 2 3 4 5 6 7 TRUSTWORTHY

Confidence: Did the voice suggest a confident speaker?
NOT AT ALL VERY
CONFIDENT 1 2 3 4 5 6 7 CONFIDENT

Persuasiveness: Was the voice persuasive?
NOT AT ALL VERY
PERSUASIVE 1 2 3 4 5 6 7 PERSUASIVE

figure 6: The MOS-X questionnaire items 12, 13 and 15 used in the survey.

Test Statistics[a]

	likeDV	inteIDV	confiDV	trustDV	persuDV	mascDV
Mann-Whitney U	922,000	1042,000	1048,000	1018,000	1039,500	1003,000
Wilcoxon W	1868,000	2317,000	1994,000	1964,000	1985,500	1949,000
Z	-1,187	-,255	-,220	-,456	-,284	-,578
Asymp. Sig. (2-tailed)	,235	,799	,826	,648	,777	,563

a. Grouping Variable: styleIV

	likeDV	intelDV	confiDV	trustDV	persuDV	mascDV
Mann-Whitney U	1041,000	872,000	1058,500	945,500	1013,000	328,500
Wilcoxon W	2266,000	2097,000	2048,500	2170,500	2238,000	1318,500
Z	-,287	-1,591	-,158	-1,058	-,519	-6,010
Asymp. Sig. (2-tailed)	,774	,112	,874	,290	,604	,000

a. Grouping Variable: genderIV

figure 7: Mann-Whitney U post-test for all dependent variables, conducted separately for the two independent variables. The results for likeability (likeDV), perceived intelligence (intel DV), confidence (confiDV), trustworthiness (trustDV), persuasion (persuDV) and gender expression (mascDV).

Tests of Between-Subjects Effects

Source	Dependent Variable	Sig.	Partial Eta Squared
styleIV	likeDV	,108	,029
	intelDV	,884	,000
	trustDV	,507	,005
	confiDV	,746	,001
	persuDV	,852	,000
	gendDV	,377	,009
genderIV	likeDV	,587	,003
	intelDV	,094	,031
	trustDV	,220	,017
	confiDV	,925	,000
	persuDV	,928	,000
	gendDV	,000	,435
styleIV * genderIV	likeDV	,957	,000
	intelDV	,431	,007
	trustDV	,242	,015
	confiDV	,544	,004
	persuDV	,869	,000
	gendDV	,349	,010

figure 8: Overall general linear model results for the two independent variables.
The results for likeability (likeDV), perceived intelligence (intel DV), confidence (confiDV), trustworthiness (trustDV), persuasion (persuDV) and gender expression (mascDV).

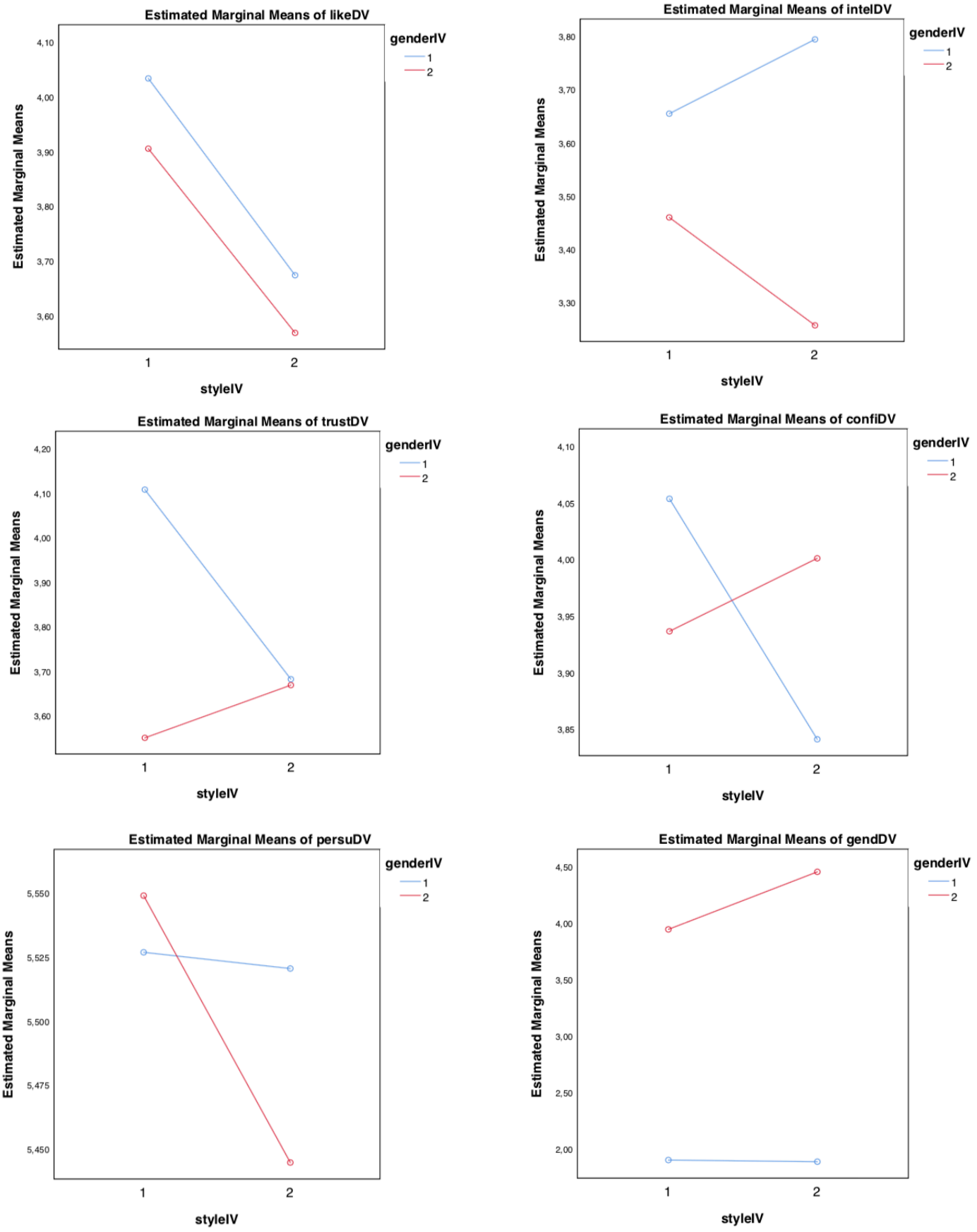

figure 9: Estimated marginal means by levels of voice gender by language style. Negligible interaction effects are demonstrated more clearly in order to see the trends.
Style (styleIV) 1: affiliative, 2: assertive. Voice gender (genderIV) 1: male, 2: female.
The results for likeability (likeDV), perceived intelligence (intel DV), confidence (confiDV), trustworthiness (trustDV), persuasion (persuDV) and gender expression (mascDV).