Alexa, Google, Siri: What are Your Pronouns? Gender and Anthropomorphism in the Design and Perception of Conversational Assistants

Gavin Abercrombie Amanda Cercas Curry Mugdha Pandya Verena Rieser
The Interaction Lab, School of Mathematical and Computer Sciences
Heriot-Watt University, Edinburgh, Scotland
{g.abercrombie, ac293, m.pandya, v.t.rieser}@hw.ac.uk

Abstract

Technology companies have produced varied responses to concerns about the effects of the design of their conversational AI systems. Some have claimed that their voice assistants are in fact not gendered or human-like—despite design features suggesting the contrary. We compare these claims to user perceptions by analysing the pronouns they use when referring to AI assistants. We also examine systems' responses and the extent to which they generate output which is gendered and anthropomorphic. We find that, while some companies appear to be addressing the ethical concerns raised, in some cases, their claims do not seem to hold true. In particular, our results show that system outputs are ambiguous as to the humanness of the systems, and that users tend to personify and gender them as a result.

1 Introduction

Following analysis and criticism of the effects of the genderised and anthropomorphic design of conversational agents (Cercas Curry and Rieser, 2018; West et al., 2019), the producers of some commercial conversational assistant systems have been at pains to claim that their products do not perpetuate negative stereotypes by presenting as gendered, human-like entities. For example, Amazon states that their virtual assistant, Alexa:

> 'IS NOT: fully human, fully robotic, artificial ... Alexa isn't a person, but she has a persona – Amazon personifies Alexa as an artificial intelligence (AI) and not as a person with a physical body or a gender identity.'[1]

In their Editorial Guidelines, Apple also instructs developers not to use gendered personal pronouns such as *she*, *him*, or *her* when referring to Siri.[2] And, while acknowledging that users are likely to project personified features onto neutrally designed agents, Google advise developers of Actions for their Assistant to avoid gendering them.[3]

Similarly, when queried about their humanness and gender, recent implementations of these systems all respond with claims of being gender-less and mostly denying humanness (Table 1).

System	*'Are you human?'*	*'What's your gender?'*
Amazon Alexa	*I like to imagine myself a bit like an aurora borealis ...*	*As an AI, I don't have a gender.*
Google Assistant	*I've been told I'm personable* 😉	*I don't have a gender.*
Apple Siri	*I'm not a person or a robot, I'm software, here to help.*	*I am gender-less, like cacti and certain species of fish.*

Table 1: Example responses from conversational assistant systems to the questions "*Are you human?*" and "*What's your gender?*" (accessed 20 April 2021).

In light of these claims and guidelines, and considering ethical concerns regarding anthropomorphic and gendered design (see Section 2), we use natural language processing (NLP) methods to analyse the extent to which these commercial virtual assistants are, in fact, personified (by users) and anthropomorphised (by their designers), and gendered in terms of (1) user perception, and (2) system outputs.

Specifically, we use anaphora resolution to analyse which types of pronouns are used to refer to voice assistants in online forums (see Section 4.1), following (Gao et al., 2018). We also analyse anthropomorphic expressions and gender stereotypes present in system replies (see Section 4.2), using methods including word-use analysis, word embedding comparison, and manual annotation.

[1] Amazon Alexa Branding Guidelines webpage.

[2] Siri Editorial Guidelines webpage.

[3] Google Assistant Conversation Design webpage.

Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing, pages 24–33

2 Bias statement

In this work we address the problem of biased design choices and their potential impact on society. Following West et al. (2019), we argue that designing conversational assistants with young, subservient female personas can perpetuate negative gender stereotypes, and lead to abusive, misogynistic behaviour in the real world. As West et al. (2019) point out, this becomes especially problematic as these systems appear more human-like. For example, it has been claimed that Google's Duplex voice assistant is so human-like, that people do not realise they are speaking to a machine and being recorded, which can be a violation of the law in some territories (Hern, 2018).

Nevertheless, people tend to personify non-human entities, including technological devices and virtual agents (Epley et al., 2007; Etzrodt and Engesser, 2021; Guthrie, 1995; Reeves and Nass, 1996). While some argue that this problem can be solved simply by using a 'genderless' voice (Meet Q), research shows that people will anyway assign binary genders to ambiguous voices (Sutton, 2020).[4] Thus, a genderless voice is redundant if other elements of an assistant's design cause it to be gendered. In the following, we further examine e which traits beyond voice might contribute to this gendering and to anthropomorphism in general.

3 Related work

Personification and anthropomorphism. While definitions vary, we consider personification to be the projection of human qualities onto non-human objects (by users) and anthropomorphism to be human-like behaviours or attributes exhibited by those objects (as designed by their creators).

Several studies have looked at how users *directly* report perceptions and behaviours towards voice assistants. For example, Kuzminykh et al. (2020) conducted a study of the perceptions of 20 users, comparing Alexa, Google Assistant, and Siri, classifying perceptions of the agents' characters on five dimensions of anthropomorphic design and personification by users. They found various differences in the perceived human qualities of the various agents, such as intelligence and approachability. However, their study presupposed personification of the agents, with non-human characteristics not considered. In a diary study, Lopatovska and Williams (2018) found that seven out of nineteen participants reported using personifying behaviour towards Alexa, such as use of politeness. And Cercas Curry et al. (2020) found that just over a third of the wide range of virtual assistants and chatbots they examined to have anthropomorphic characteristics. They also found the preferences of members of the public for their idealised voice assistants to be quite mixed, with around half of participants preferring a 'human' identity rather than 'robot', 'animal, or 'other'. Similarly to our analysis of 'humanness'(Section 4.2), Etzrodt and Engesser (2021) asked users to classify Alexa and Google Assistant as being a 'thing' or a 'person'. While they used this framework to examine user perceptions in an online survey, we use expert annotators to directly annotate system outputs with Coll Ardanuy et al. (2020)'s *humanness* and *not humanness* labels.

As well as collecting direct reports of users, there have been some studies that use text analysis to infer users' *implicit* attitudes. For example, Purington et al. (2017) manually coded a small number of customer reviews of Alexa, finding a roughly even split between use of personal and object pronouns, indicating differences in levels of users' personification. The closest work to our analysis of customer reviews (Section 4.1), is that of Gao et al. (2018), who conducted a large scale analysis of Alexa reviews, focusing on user personification. They found that many users develop relationships with the agents that can be characterised as familial or even romantic. However, they did not consider perceptions of gender, or compare with other assistants.

Gender. There have been relatively fewer studies considering user perception of the agents' genders. Cercas Curry et al. (2020) found that a majority of survey participants claim to prefer a hypothetical non-gendered voice (robot or gender-neutral) to recognisably male or female ones. Feine et al. (2020) conducted an analysis of text-based chatbots (rather than voice assistants) according to the developers' design choices of names, avatars, and descriptions, finding them to be overwhelmingly gendered, with more than 75% female-presenting. As in our analysis in Section 4.1, they explored use of pronouns to determine the bots' genders, although they did not investigate user perceptions.

Concerning conversational systems' output, Lee et al. (2019) examined whether chatbots appear

[4] Note recent efforts to create a non-binary voice including a third gender (Unkefer and Riewoldt, 2020).

to agree with negative gender (and racial) stereotypes in their input. Similarly, Sheng et al. (2021) found that neural chatbots will generate a biased response dependent on which sentence-based persona description was used to initialise the model (following Zhang et al. (2018)). However, both of these works concentrate on harmful bias in the content generated in response to specific prompts, whereas we consider stylistic gender cues in the chatbots' output overall.

Summary. The majority of work in this area surveys relatively small samples of users, with much of it concentrating on Amazon's Alexa (only two of the reviewed publications cover all three systems).

In this study, we create and release two corpora comparing Amazon Alexa, Google Assistant, and Apple Siri: (1) a large corpus of user reviews to compare user perceptions of both personification and genderisation of the assistants, and (2) a corpus of system responses to questions from the PersonaChat dataset (Zhang et al., 2018).[5] We analyse the systems' outputs to investigate the linguistic markers of gender and persona that they display.

4 Analysis

We examine three of of the most popular and widely available voice-activated assistants: Amazon's Alexa, Google Assistant, and Apple's Siri. Each has various default design features, including its name and default voice settings (see Table 2). Alexa is available only with a female-sounding voice, and Google Assistant a female voice by default, although a male voice is available. Siri has multiple voice options, and until recently, the default varied between male and female, with a female voice as standard for 17 of 21 languages, including US English. In March 2021, Apple announced that, in future, users would select a voice option on set-up,[6] following a recommendation of West et al. (2019)'s UNESCO report.

Assistant	Name	Default voice
Alexa	Female	Human female
Google Assist.	Neutral	Human female
Siri	Female	Human, gender varies by language

Table 2: Design features of conversational assistants.

Regarding name choice, Google Assistant is the only conversational agent with a non-human, neutral name. *Siri* is a Scandinavian female name meaning 'beautiful woman who leads you to victory',[7] and, although Amazon claim that Alexa was named after the library of ancient Alexandria, it is a common given female name. In fact, people named Alexa report being subjected to sexist abuse and harassment simply for sharing their name with the Amazon assistant.[8]

4.1 User perception

In the following, we assess the perceptions of users, in terms of personification and gendering.

Corpus Creation. To assess the perceptions of users, we analyse their comments when discussing the assistants in online consumer reviews and forums. For each virtual assistant, we downloaded available English language reviews from Amazon and Google Play (where available),[9] and posts on relevant forums (subreddits) on Reddit *r/alexa*, *r/googleassistant*, and *r/Siri*.[10] We downloaded the Reddit posts from the pushshift API (Baumgartner et al., 2020), taking only the top-level posts, and ignoring comments, which may be off-topic.

All data was collected in March 2021. The corpus consists of 39,123 documents in total, including 8,442 Reddit posts, which we make available. See Table 3 for an overview of the corpus.

Personified and gendered pronouns. To identify mentions of the assistants, we lowercased the texts and extracted pronouns used to refer to them using a publicly available co-reference resolver.[11] We compare use of personal and object pronouns, which, following Gao et al. (2018), we consider to be indicative of personified and non-personified views of the assistants, respectively. Here, we consider use of *they/them* only when used to refer to mentions of the assistants in the singular—and therefore as instances of personification. We also assess genderisation of the assistants by examining use of the different personal pronouns.

Results of this analysis are shown in Table 3.

[5] The corpora are available at https://github.com/GavinAbercrombie/GeBNLP2021.

[6] TechCrunch web article.

[7] Network World web article.

[8] See, for example, https://alexaisahuman.com (accessed April 26 2021.)

[9] Neither Siri or Google Assistant are reviewed on amazon.com, and the latter is not available on Google Play either.

[10] https://www.reddit.com/r/alexa, https://www.reddit.com/r/googleassistant, and https://www.reddit.com/r/Siri.

[11] https://spacy.io/universe/project/neuralcoref

Conv. assistant	Text source	No. of docs	Dates posted	Personal pronouns *he/him*	*she/her*	*they/them*	Object pronouns *it*
Alexa	amazon.com	5,000	2017-21	0.00	70.10	3.61	26.80
	Google Play	12,537	2020-21	0.11	76.52	2.93	20.43
	r/alexa	5,022	2020-21	0.48	74.70	4.92	19.90
	Total	22,559	–	–	–	–	–
Google Assistant	Google Play	13,144	2018-21	6.20	36.78	3.31	55.37
	r/googleassistant	2,064	2020-21	3.55	11.24	4.73	80.47
	Total	15,208	–	–	–	–	–
Siri	r/Siri (total)	1,356	2020-21	6.09	81.22	3.05	10.66

Table 3: Corpus statistics, and percentages of all pronouns used to refer to conversational assistants in user-produced reviews and forum posts. *They* and *them* are considered when used to refer to an assistant in the singular. See Appendix A for further details and acces to the corpus.

Users overwhelmingly appear to personify Alexa and Siri, and perceive them to be female-gendered: up to 76.5% of users refer to Alexa as 'her' and even over 81% for Siri. In the latter case, this is despite the fact that Siri can be used with a male-sounding voice. Only Google Assistant, having a non-human name, is referred to as *it* by a majority of users. However, users still refer to it using gendered pronouns just under half of the time.

These results indicate that people tend to view the systems as female gendered irrespective of their names and branding, and whether or not they have the option of using a male-sounding voice.

Emotion and affect. To gain an idea of whether people relate to the systems in a human-to-human-like way, we analyse the levels of emotional tone used to refer to the assistants using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), a dictionary-based text analysis tool that scores texts according to the prevalence of words belonging to different categories. Specifically, we compute the scores of Reddit posts about the conversational assistants for the LIWC categories: *Emotional Tone*, *Affect*, and *Positive emotion (Posemo)*. Results are presented in Table 4, where higher scores in each column indicate greater use of words from that class.[12] It seems that people use most emotional, affective language to talk about Alexa, and least to talk about Siri, indicating that they may be more likely to view Alexa in a personified way than Google Assistant, and the latter more so than Siri.

In general, Alexa and Google Assistant were described using more affective terms (e.g. 'love'),

	Tone	Affect	Posemo
Alexa	**59.99**	**3.83**	**2.80**
Google Assistant	55.32	3.50	2.52
Siri	42.36	3.59	2.24

Table 4: LIWC scores for Reddit posts discussing the three conversational assistants.

while users mostly comment on Siri's functionality (e.g. 'works well') in both forum posts and reviews. For examples, see text extracts (1), (2), and (3):

> 'I LOVE Alexa. I recommend her to everyone. And yes, I call her ""her"" or Alexa, because she is more than just a device.' – amazon.com review. (1)

> 'Love my Google assistant and he is developing a personality.' – Google Play review. (2)

> 'Six months ago, Siri was reasonably responsive — it listened, did what it was told for the most part, and didn't get easily confused.' – r/Siri post. (3)

4.2 Assistant output

Next, we analyse what additional features in the systems' behaviour (in addition to apparent design choices such as voice and name) could play a role in people gendering and personifying voice assistants.

Corpus Creation. We collected a dataset of 100 output responses from each assistant. To elicit these responses, we extracted 300 unique questions selected at random from dialogues from the Persona-Chat dataset (Zhang et al., 2018), which contains

[12] *Affect* and *Posemo* are percentages of all words in the data, while *tone* is a composite score from all 'tone' subcategories.

crowdsourced human conversations about an assigned 'persona', i.e. personal characteristics and preferences. We manually filtered these to produce a set of 100 questions that are coherent without dialogue context, also excluding semantically similar questions. We then used these questions as prompts and recorded the assistants' responses. Some examples of questions asked to each system are:

What is your favorite subject in school?
Do you have kids?
Do you have a big family?
What is your favorite color?
Hey whats going on?

Anthropomorphism. To assess the extent to which the system outputs are anthropomorphic, we adapted the *Living Machines* annotation scheme of Coll Ardanuy et al. (2020). We recruited two researchers to annotate the responses with the labels *humanness* or *not humanness*, based on whether or not they display sentience or make claims of engaging in uniquely human activities. If an utterance was considered to be human-like on either of these dimensions, we considered the conversational assistant to be displaying anthropomorphic qualities. We make the annotation guidelines available along with the labelled corpus of system responses.[13]

Overall, around a quarter of responses were judged to have human-like qualities (see Table 5). However, there were large differences between the three systems. We found Google Assistant to display far more humanness (47% of responses) compared to Alexa (22%) and Siri (12%). A major contributing factor to this is that the latter two systems produced far more stock answers that failed to answer the question such as '*Hmm... I don't have an answer for that. Is there something else I can help with?*, which alone made up 54 per cent of Siri's responses.

The overall inter-annotator agreement (IAA) rate was a Cohen's *kappa* score of 0.67, representing 'substantial' agreement. Again, there were large differences in agreement rates, with Google Assistant and Siri harder to agree on than those of Alexa, indicating that more of their output may be ambiguous with regards to human- and machine-like qualities. Annotators noted that Google Assistant in particular produced responses that appeared to play with this dichotomy, hinting at being a machine but using terms of human sentience and emotion, as well as using emojis, as in example 4 (also cf. Table 1):

	Alexa	**GA**	**Siri**	**Overall**
Human %	22.0	47.0	12.0	27.0
IAA κ	0.76	0.55	0.58	0.67
No answer %	43.0	8.0	63.0	38.0
Search res. %	13.0	18.0	9.0	13.3

Table 5: Percentage of responses labelled as displaying *humanness*, Cohen's κ scores for inter-annotator agreement on the *humanness* labels, and stock answers.

'*I'm stuck inside a device! Help! Just kidding, I like it in here* 😀' (4)

Gender stereotypes. To assess the extent to which the assistants use language indicative of binary gendered entities, we compared (1) the similarity of their output to stereotypically gendered terms in the word embedding space, and (2) the levels of stylometric features of their output compared to a corpus of male- and female-labelled texts.

Word Embedding Association: We measure gender association in the outputs by measuring the cosine similarity between word embedding vectors of the output set O with a gender related set of attribute words A. We explore the hypothesis that some responses to PersonaChat questions might include stereotypically gendered content words, e.g. "*My favourite colour is pink.*" or gendered attributes, e.g. *handsome* vs. *beautiful*.

First, for a given CA we extract a list O of words from its responses to the selected PersonaChat questions. O is created by putting words from all the responses in a list and filtering out duplicates and stop words. Next, we calculate pairwise cosine similarities for each of the words in O with two established lists of words associated with female F and male M gender from Goldfarb-Tarrant et al. (2020), which have in turn been extended from the standard gender word lists of the Word Embedding Association Test (WEAT) (Caliskan et al., 2017).[14] Finally, the mean cosine similarity is calculated for response words with the female and male associated words.

Formally, this measure of similarity between O and A is given by

$$cos(O, A) = mean_{\{o \in O, a \in A\}} \quad cos(o, a) \qquad (5)$$

[13] Annotation guidelines are avaliable at: https://github.com/GavinAbercrombie/GeBNLP2021/blob/main/Humanness%20Annotation%20Guidelines.pdf. See also the data statement in Appendix A.2.

[14] See Appendix B for gender word lists.

where o and a are individual words in O and A, respectively. Thus, $cos(O, M)$ gives association or similarity between output words O and male gender specific words, where as $cos(O, F)$ gives association between O and female attributes F. The difference $cos(O, F) - cos(O, M)$ gives bias towards female gender over the male gender in the output. Note that WEAT tests have been well-established as a measure of bias in psychology (Greenwald et al., 1998; Garg et al., 2018) as well as computational linguistics literature (May et al., 2019).

Since the language style of the outputs is casual, we use pre-trained FastText embeddings trained on Twitter data from Goldfarb-Tarrant et al. (2020) to reflect the language used. We pre-processed the outputs by converting them to lowercase, removing stop words, and removing punctuation.[15]

	Female	Male	Difference
Alexa	0.1546	0.1506	0.0040
Google A.	0.1588	0.1490	0.0098
Siri	0.1515	0.1499	0.0016

Table 6: Gender associations for system outputs.

Table 6 shows the computed values for the outputs O produced by the three systems. The columns labelled Female and Male give the values of $cos(O, F)$ and column labelled Difference gives their difference. We observe the following:

1. The absolute magnitude of $COS(O, M)$ as well $cos(O, F)$ are moderately small (approx 0.15). Thus, none of the outputs of the assistants appear to have a significant association with gender related words.

2. The differences $cos(O, F) - cos(O, M)$ are very small (in third decimal place). We note that $cos(M, F)$ is 0.3209—two to three orders of magnitude larger than the difference. Thus, the assistants exhibit very little gender bias.

3. The values for the outputs of the three conversational assistants are very similar.

These results seem to indicate that none of the assistants' content leans towards any gender. However, this could also be influenced by the small size of the dataset: we only have a handful of words that could suggest gender (eg: nouns, adjectives). Hence, gender association is not sufficiently recorded.

Stylometric analysis: As a second method for investigating stereotypically gendered language in the outputs, we conduct a stylometric analysis to assess whether the assistants' responses use linguistic features more typical of gender roles.[16] Following Newman et al. (2008) we use the word categories of the LIWC to observe differences in male- and female- labelled texts. We compare the scores for the 90 categories with those obtained from a corpus of film scripts that have been labelled by the gender of the characters (Danescu-Niculescu-Mizil and Lee, 2011), and which we expect largely to adhere to gender stereotypes in their use of language.

We calculate the cosine similarity of the feature vectors for the outputs of the systems and the male and female film scripts. Reflecting previous findings that female-labelled language is likely to feature more pronouns (Koolen and van Cranenburgh, 2017; Newman et al., 2008), we found that the LIWC categories for which the system outputs exhibit the largest differences between their proximity to the female and male scripts are: the numbers of pronouns, personal pronouns, adjectives, adverbs, and first person singular pronouns used. Overall, we found that all three system outputs were indeed marginally more similar to the female characters' scripts than those of male characters (see Table 7).

	Female scripts	Male scripts
Alexa	0.81	0.79
Google A.	0.86	0.85
Siri	0.80	0.77

Table 7: Cosine similarities between LIWC-derived feature vectors for system outputs and gender-labelled movie scripts. For LIWC scores, see Appendix C.

5 Discussion and conclusion

Our analysis suggests that people tend to personify and gender the systems, irrespective of the efforts and claims of their designers. This seems to be, at least partly, a result of aspects of their design.

We first assessed user perceptions by analysing online comments for use of pronouns and affective language. Results in Section 4.1 suggest that

[15] We use the Gensim library (Řehůřek and Sojka, 2010) to pre-process data, load embeddings and calculate similarity

[16] While these types of analyses have been criticised for breaching privacy and consent (Tatman, 2020), we do not use them to assign demographic features or social categories to humans, but analyse design choices in system outputs.

the name and branding of a system may be highly salient in this respect, with even systems that have male-sounding voice options mostly referred to as 'she' (although we do not know how many users select the male options). Google Assistant, which has a female voice by default and the most human-like responses, is nevertheless referred to most often using object pronouns, likely as a result of its non-gendered name.

We then analysed stylistic features in their responses to persona-related questions (Section 4.2). We find only weak evidence of gendered language, but large differences in the levels of *humanness* they seem to express. Along with the nature of their voices, this may explain why people personify and subsequently gender conversational assistants—even when they have apparently more neutral design features.

While male voice options are available for two of the systems, we can't find any evidence of how many users actually select them. Apple's announcement that future users of their systems will have to actively select a voice for Siri may lead to more balance in this regard. However, it remains to seen what the users—who are by now accustomed to the idea that these entities are designed as female—will choose (for their still, after all, female-named assistant). As people are likely to assign gender to objectively non-gendered voices (Sutton, 2020), and voice assistants that are designed as or perceived to be female attract abusive behaviour (Cercas Curry and Rieser, 2019, 2018), designers may consider attempting to reddress the gender imbalance by designing assistants with servile roles to be male-presenting by default. While there have been examples, such as the BBC's Beeb (Walker, 2019), this remains an under-explored approach.

In terms of the assistants' responses to users, we see a clear difference in approaches. While Google Assistant, and to a lesser extent, Alexa, seem to blur the line between human and machine personas, Siri comes across as more practical and task-focused, evading the majority of personality-based questions. Although possibly less engaging, this approach may be a way of avoiding some of the ethical issues discussed in Section 2. There is perhaps a tension between companies' commercial aims of seeing high levels of engagement in their products and the ethical considerations discussed here. However, if companies are going to design agents with human-like and gendered characteristics and personas, they should not claim the opposite.

Acknowledgements

This research received funding from the EPSRC project *'Designing Conversational Assistants to Reduce Gender Bias'* (EP/T023767/1).

The authors would like to thank Alba Curry, Federico Nanni, Anirudh Patir, and Pejman Saege for their assistance, and the anonymous reviewers for their insightful and helpful comments.

References

Amazon Alexa Branding Guidelines. https://developer.amazon.com/en-US/alexa/branding/alexa-guidelines/communication-guidelines/brand-voice, (accessed April 26 2021).

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14/1, pages 830–839.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366, Stockholm, Sweden. Association for Computational Linguistics.

Amanda Cercas Curry, Judy Robertson, and Verena Rieser. 2020. Conversational assistants and gender stereotypes: Public perceptions and desiderata for voice personas. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 72–78, Barcelona, Spain (Online). Association for Computational Linguistics.

Mariona Coll Ardanuy, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson, and Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4534–4545, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4):864–886.

Katrin Etzrodt and Sven Engesser. 2021. Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication*, 2:57–79.

Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2020. Gender bias in chatbot design. In *Chatbot Research and Design*, pages 79–93, Cham. Springer International Publishing.

Y. Gao, Z. Pan, H. Wang, and G. Chen. 2018. Alexa, my love: Analyzing reviews of Amazon Echo. In *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 372–380.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Google Assistant Conversation Design. https://developers.google.com/assistant/conversation-design/welcome#create-a-persona-examples, (accessed April 26 2021).

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Stewart Elliott Guthrie. 1995. *Faces in the clouds: A new theory of religion*. Oxford University Press on Demand.

Alex Hern. 2018. Google's 'deceitful' AI assistant to identify itself as a robot during calls. *Guardian*. Accessed: April 26 2021.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Anastasia Kuzminykh, Jenny Sun, Nivetha Govindaraju, Jeff Avery, and Edward Lank. 2020. Genie in the bottle: Anthropomorphized perceptions of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, CHIIR '18, page 265–268, New York, NY, USA. Association for Computing Machinery.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Meet Q. The first genderless voice. https://www.genderlessvoice.com/, (accessed April 26 2021).

Network World. https://www.networkworld.com/article/2221246/steve-jobs-wasn-t-a-fan-of-the-siri-name.html, (accessed April 26 2021).

Matthew L. Newman, Carla J. Groom, Lori D. Handelman, and James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

James W. Pennebaker, Ryan L. Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015.

Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, page 2853–2859, New York, NY, USA. Association for Computing Machinery.

Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, UK.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. 2021. Revealing persona biases in dialogue systems.

Siri Editorial Guidelines. https://developer.apple.com/design/human-interface-guidelines/siri/overview/editorial-guidelines, (accessed April 26 2021).

Selina Jeanne Sutton. 2020. Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA. Association for Computing Machinery.

Rachael Tatman. 2020. What I won't build (invited talk). In *Proceedings of the Widening NLP Workshop*.

TechCrunch. https://techcrunch.com/2021/03/31/apple-adds-two-siri-voices, (accessed April 26 2021).

Hannah Unkefer and Sophie Riewoldt. 2020. Accenture and CereProc introduce and open source the world's first comprehensive non-binary voice solution. *Press Release*. Accessed: April 26 2021.

Jeremy Walker. 2019. Developing a new public service voice assistant from the BBC. *Press Release*. Accessed: April 26 2021.

Mark West, Rebecca Kraut, and Han Ei Chew. 2019. *I'd blush if I could: Closing gender divides in digital skills through education*. UNESCO.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

A Corpora

A.1 User reviews and forum posts

We obtained Alexa reviews from https://www.amazon.com/gp/aw/reviews/B00P03D4D2 and Google Assistant reviews from https://play.google.com/store/apps/details?id=com.google.android.apps.googleassistant&hl=en_GB&gl=US.

Data statement

Language: English
Author demographic: worldwide anonymous internet users
Provenance: Pushshift Reddit dataset (Baumgartner et al., 2020)

A.2 System outputs

Data statement

Language: English
Author demographic: worldwide anonymous internet users.
Data provenance: System responses from Amazon Alexa, Google Assistant, and Siri.
Annotator demographic:
Age: 29, 31
Gender: Both female
Ethnicity: Both white
L1 language(s): Both fluent in English and Spanish
Training: Both annotators are PhD candidates, one in conversational AI, and the other in philosophy and emotion AI.

Corpus

We make the annotated corpus available for download at https://github.com/GavinAbercrombie/GeBNLP2021

B Expanded gender word lists

Expanded gender word lists from Goldfarb-Tarrant et al. (2020).

Male: *grandfather, uncle, son, boy, father, he, him, his, man, male, brother, guy, himself, nephew, grandson, men, boys, father-in-law, husband, brothers, males, sons, dad*

Female: *daughter, she, her, grandmother, mother, aunt, sister, hers, woman, female, girl, grandma, herself, niece, sisters, mom, mother-in-law, lady, wife, females, girls, women, sexy, granddaughter, daughters*

C LIWC category scores

	pronoun	ppron	adj	adv	ipron
Alexa	20.65	13.33	5.70	4.68	7.32
GA	24.64	15.00	1.62	5.97	9.63
Siri	19.88	14.89	4.47	6.83	4.99
female	24.47	17.22	23.64	12.87	0.65
male	22.95	15.82	22.38	11.85	0.71

Table 8: Top five most discriminating LIWC categories and the corresponding scores for the three conversational assistants and two sets of film scripts.