

The Impact of a Conversational Agent Avatar on the Experience of Museum Visitors

Laura Wagner^{1*}, A’Niyah Brown², Victoria Sevich³, Madeleine Bloomquist⁴, Vishal Sunder⁵, Eric Fosler-Lussier⁵, Michael White⁴

¹*Department of Psychology, The Ohio State University*, ²*Lingraphica*, ³*Department of Speech and Hearing Sciences, The Ohio State University*, ⁴*Department of Linguistics, The Ohio State University*, ⁵*Department of Computer Science and Engineering, The Ohio State University*

Abstract

We report on a conversational agent presented as an avatar on a TV and deployed inside a science museum in 2021. The avatar, Ilse, could answer questions about herself, the institutions she was connected to (the local university, the nearby lab, the museum), and a relevant science area (linguistics). We compared conversational interactions from 118 visitors to those visitors’ assessments of their own experiences. The interactions were similar to previous results from similar conversational agents: visitors often asked the avatar about herself and the avatar provided high-quality responses around half the time. Visitors’ survey responses showed that they enjoyed their interactions with the avatar, finding them easy and helpful. However, visitors who had received more high-quality responses from the avatar rated their subjective experiences significantly more positively. These results show the power of using conversational agents for public engagement and the importance of maintaining the underlying technology at a high level.

Introduction

Museums of all sorts are increasingly incorporating high tech assistants – so-called “conversational agents” – into their spaces (Gasteiger et al., 2021; Varitimiadis et al., 2021). These assistants can take on various forms, from chatbot games that guide visitors through a game about the museum contents (Gaia et al., 2019) to video avatars that help visitors engage with a specific exhibit (Lane et al., 2011), to exhibits explicitly designed to engage visitors with the artificial intelligence content embodied by the exhibit (Gonzalez et al., 2017). While there have been extensive investigations into all aspects of conversational agents used in home and office settings (see Nourouzi et al., 2018 and Diedrich et al., 2022 for reviews), less attention has been paid to how these systems work in museum settings, where the conversations are often less directly instrumental and the environments may be noisy and chaotic. Researchers investigating these assistants within the museum context have approached them from two main perspectives: one based in improving the technology (which has been rapidly evolving) and another based in assessing the

* Corresponding author:
Laura Wagner, wagner.602@osu.edu

impact of the technology on the visitor experience.

From a technological point of view, a museum space presents a variety of challenges. Many early examples of these conversational agents involved recording participants to create a corpus that could be analyzed and used to improve the nature of the interaction (e.g. Kopp et al., 2005; Robinson et al., 2008; Al Moubayed et al., 2012). One goal of this approach was to help the agents handle speech from a diverse set of individuals in a loud and noisy environment. For example, the Museum of Science in Boston had an exhibit featuring Ada and Grace – a pair of conversational robots who could talk about different things in the museum (Swartout et al., 2010; Aggarwal et al., 2012; Traum et al., 2012). Across its implementation, visitors could interact with Ada and Grace in a few different ways: some visitors spoke directly to them while others talked to a “handler” who relayed questions and comments to Ada and Grace. Moreover, visitors were provided with some suggested topic areas and questions to use with Ada and Grace, although they weren’t required to use them. Analyses of the conversational interactions found that in general, Ada and Grace responded appropriately to the questions only about half the time, which is typical of agents that respond to freely-offered human speech in museum settings (Kopp et al., 2005; Robinson et al., 2008).

One factor that can lead to inappropriate or inaccurate responses has to do with the topics that people ask about. Many museum-based conversational agents converse in a comparatively open-ended fashion as opposed to performing highly specific tasks such as turning on music or looking up a specific fact. For example, one study (Robinson et al., 2008) investigated

what people chose to ask a conversational agent at the Cooper Hewett Museum who was presented as a military officer named Sargent Blackwell. They found that people overwhelmingly asked Sgt. Blackwell personal questions about his life experiences, including his preferences and marital status (see also Kopp et al., 2005 and Bickmore et al., 2011 for similar results). Many systems anticipate this preference from visitors and build in a range of social information that their avatars can discuss (see Gasteiger et al., 2021). However, when conversations can be more narrowly constrained, museum agents perform better. Thus, while Ada and Grace’s responses were accurate overall only about 50% of the time, when analysis of their responses was restricted to those from their designated set of topics, their accuracy rose to approximately 80%. The free-form nature of many museum interactions does not always play to the strengths of the technology.

In parallel to investigations of how well the avatar performs as a piece of technology, other researchers have examined the visitor experiences that these agents provide. General evaluations of conversational agents in museum settings unilaterally report that people enjoy interacting with them and rate their experiences with them positively, with children often being somewhat more positive than adults (Bickmore et al., 2011; Al Moubayed et al., 2012; Traum et al., 2012; Gonzalez et al., 2017; Gaia et al., 2019). However, the more social and personal the agents get, the more people like them. Bickmore et al. (2011) explicitly tested two versions of their agent, Tinker, at the Museum of Science in Boston. The more social version could tell jokes, talk about himself, discuss the local baseball team, and included more greetings than the less social version. While both versions

received high positive ratings in general, the social version outperformed the less social version on all measures, including the length of time people spent interacting with it. More generally, Gasteiger et al. (2021) reviewed a dozen different social robots being used in museum settings and found that visitors preferred robots that were more human-like in their appearance and in social interactions. These results are in line with investigations of the user experience of conversational agents more broadly, where agents that incorporate human-like conversational agents receive higher ratings (e.g. Cohn et al., 2019).

While conversational agents in museums are designed for a variety of purposes, a few have been more specifically aimed at teaching visitors about computer science concepts. Researchers working on these agents have examined their impact on visitors' interest in computer science and what people learn from them. For example, Lane et al. (2011) reported on Coach Mike, a conversational agent at the Museum of Science in Boston who provided assistance to visitors in an exhibit about programming a robot. Visitors spent more time at the exhibit when they used Coach Mike compared to when they did not; more notably, their programs were shorter and potentially more elegant when Coach Mike assisted than when he did not. Gonzalez et al. (2017) report on a system explicitly designed to teach visitors about the Turing Test, a key computer science concept for artificial intelligence. Visitors who engaged with their system reported being more interested in computer science after using it (and see also Traum et al., 2012).

One type of evaluation that has not been conducted very frequently with museum-based conversational agents involves linking the technological quality with the visitor's

experience. Examinations of this question with common general-purpose conversational agents such as Siri and Cortana have found that user satisfaction and intent to use an agent in the future is generally tied to how easy it is to use the agent and how successful users are at accomplishing tasks with the agent (Kiseleva et al., 2016; Yang & Lee, 2019; Ashfaq et al., 2020). However, user expectations also matter. Luger & Sellen (2016) interviewed users of common conversational agents and found that users who expected the agent to be easier to use and more capable than it was expressed frustration while users who approached the agents with a sense of playfulness were more motivated to work at their interactions. Museum settings generally encourage a sense of play, which may lead visitors may be more tolerant of difficult interactions. For example, Al Moubayed et al. (2012) noted that visitors who interacted with their museum-based agent, Furhat, reported both that they enjoyed the experience but also that they had to concentrate "quite a bit" in order to understand him. However, no evaluations that we are aware of have attempted to closely tie a museum visitor's subjective evaluation of an agent with that agent's success rate with that individual. Moreover, as technology in this area rapidly changes so too do user expectations. We note that the Furhat results come from several years ago and a playful stance alone may no longer be enough to generate tolerance in visitors if their interaction is not sufficiently smooth and easy.

The current study reports on the use of a conversational agent deployed in a science museum. Our agent is called Ilse Lehiste (named after a pioneering linguist from our local linguistics department) and she was presented as an avatar on a screen positioned near the Language Sciences Research Lab.

This lab operates as both a data collection site as well as an exhibit in the Center of Science and Industry (COSI) in Columbus, OH (Wagner et al., 2015). Ilse was designed to be able to answer questions about languages, linguistics research, and the lab on site, as well as general questions about the museum and the local university. More details about Ilse’s technological history and specifications can be found in the Method section below. Our goals with this project were to assess how well Ilse performed technologically, how visitors to the museum subjectively evaluated their interactions with her, and to examine the connection between those two measures for specific individuals. In short, to what extent does the visitor experience depend on how well the technology works?

Method

Participants

All data collection practices were reviewed and approved by the Ohio State University Social and Behavioral Sciences Institutional Review Board (protocols 2021B0184 and 2018B0169). A total of 118 museum visitors were recruited and run on the floor of the museum. Participants were drawn from the museum visitor pool at large: they ranged in age from 5 – 56 years, including 35 participants aged 9 years and younger, 32 participants between the ages of 10 and 17 years, and 51 adults. Adults provided oral consent for themselves and an accompanying parent provided oral permission for all children.

Demographically, 67 participants identified as female, 48 as male, and 3 as something else; 57 self-identified as White, 15 as Black/African American, 3 as Asian, 1 as Hispanic, 10 as multi-racial/ethnic, and 4 chose not to provide race/ethnicity information. All participants completed a conversation with Ilse that was recorded as well as a pre-conversation and a post-

conversation survey. Based on the pre-conversation survey, participants rated themselves as reasonably familiar with virtual assistants (M = 3.08 out of 4) but were rather less familiar with our domain of science, linguistics (M = 1.96 out of 4).

Figure 1: Dr. Ilse Lehiste, Conversational Agent



Ilse the Avatar, the Conversational Agent

Ilse is an interactive conversational agent presented as an avatar who can answer orally-presented questions in real-time from real people. She was based on the Virtual Patient project (Maicher et al., 2019, 2022; Stiff et al., 2022), an interactive conversational agent created to help train medical students take patient histories and perform differential diagnosis. More specifically, the underlying technology within Ilse included (a) cloud-based speech transcription (alternately IBM Watson and Microsoft Azure provided transcription), (b) an OSU-developed language understanding

system based on neural networks trained with novel algorithms that were tuned to be able to better handle rare questions, and (c) a Unity-based avatar displayed on an iPad. In addition, Ilse was trained on a set of specific content reflecting common questions visitors have about COSI, the local university, the Language Sciences Research Lab, language and linguistics, and about herself. When Ilse was not confident of the user's input, she could ask to have a question re-phrased and could also suggest alternative topics for users to ask.

Within the museum, Ilse's iPad was connected to a large TV screen placed on a rolling cart. On that screen, her head/face and torso filled the screen and her eyes were roughly at 5' off the ground. Figure 1 shows what her set-up looked like in the museum.

Survey Instrument

Prior to interacting with Ilse, participants filled in a pre-conversation survey, consisting of demographic questions and 6 ratings questions (using a 4-point Likert Scale) about their interest in and familiarity with the relevant technology, with the scientific field of linguistics, and with the affiliated lab. Specifically: How familiar are you with virtual assistants/language science and linguistics/the Language Pod at COSI? How interested are you in automated speech recognition/language science and linguistics/participating in a study at the Language Pod at COSI?

After interacting with Ilse, participants completed a post-conversation survey which re-asked the same questions about their interests and familiarity with linguistics and the lab. In addition, participants rated their experience with Ilse on 5 questions: Was the avatar helpful? Was it easy talking to the avatar? Was it fun talking to the avatar? Did the avatar's speech sound natural? How well did the avatar understand you? Participants

were also given a chance to provide open-ended feedback about their experience.

Procedures

Museum visitors passing by the exhibit cart were invited to "talk with the avatar" by a research assistant (RA). The system was consistently called "the avatar" by the RA (and the surveys); while many people asked the avatar what her name was, not everyone found out that the avatar's name was Ilse.

The RA obtained consent/permission for the study, including both the survey and the internal audio recordings made by the agent herself. The surveys were presented on a separate iPad and each participant was assigned an ID number which was also entered into the conversational agent's system to allow each survey to be linked to a specific conversation. After participants completed the pre-conversation survey, they interacted with Ilse.

The RA facilitated the interaction with Ilse in several ways. She made sure that participants were speaking into the microphone on the iPad which was the actual input to the system. She also explained to participants that the avatar was not like Siri or Alexa and that she could really only talk about some topics, specifically she could talk about the museum, about language, and about the nearby Language Sciences Research Lab. Participants were encouraged to talk freely to the avatar, but they were provided with cue cards that provided information on the topics the avatar was prepared to discuss as well as hints about some specific questions. For example, the cue card for Language said: *She knows about Language!* • *How many are spoken in the world* • *About American Sign Language, Accents and Bilingualism* • *Who studies it. And more!* *See if you can figure some out!* In addition,

the RA had a copy of all the questions that the avatar had been trained on and she provided hints and ideas to participants about what to ask or alternative ways to phrase a question. The RA's directive was to encourage participants to do as much as they could on their own, but never at the expense of them having a positive museum experience. Participants talked with Ilse for as long as they wanted to. Once they told the RA they were done, they were provided with the post-conversation survey.

Coding

Ilse's system audio-recorded user's input as well as her responses to each input. User input was coded by hand for its overall content: Was there a clear topic for each utterance and if so, what was it? An input was classified as being about the Avatar if it contained a question about her underlying technology or treated her as a social being. Inputs asking about language facts or about linguistic usage were classified as

Linguistics. Inputs were classified as Institutional if they asked about any of the institutions affiliated with the project, including the museum (including questions about where other museum exhibits were located), the Language Sciences Research Lab (aka the Language Pod) or the local university sponsoring the work. These three topics reflected the topics that the avatar had been specifically trained to address. In addition, an input could be classified as Out of Range if it was about a topic that Ilse was not prepared to talk about (such as other branches of science). Finally, inputs could be classified as Uncodeable. This classification was used with partial utterances, overlapping speech, and short, contextually bound utterances (e.g. *What? Yeah*). Two coders independently classified each input utterance for its topic and agreed on the classifications 88% of the time ($Kappa = .84$, $SE = .011$) which is considered almost perfect agreement. Table

Table 1: Examples of the five topic classifications in the input

Input Content Topics	Examples from Participants
Avatar (includes social interactions with the avatar)	How old are you? What languages do you speak? You're not planning on taking over the human race, are you? Thank you for your time Avatar lady.
Linguistics	How does texting impact English? Tell me about the use of slang.
Institutional (includes questions about the museum, lab, and our associated university)	What time is this place open 'til? Can I get a degree in linguistics at Ohio State? Who works in the language pod?
Out of Range (topics that the avatar was not prepared to address)	How many miles a week should I run? What is a banana?
<u>Uncodable</u>	Yeah. How many? Folders.

1 presents the 5 main topic classifications in the input, along with examples.

Ilse's responses were coded for their quality. An answer was considered of High Quality if it was accurate and on-topic. This scoring was used generously and included times when Ilse's answer was incomplete or felt somewhat unnatural. An answer was considered of Moderate Quality when it was broadly on-topic, but wasn't really appropriate for the question being asked. An answer was considered of Low Quality when it was completely off-topic. Two coders independently classified each response for quality and agreed on the classifications 91% of the time (Kappa = .823, SE = .018) which is considered "almost perfect" agreement. In addition, Ilse had a set of pre-programmed answers that were used when she didn't know how to answer a question; these responses Refused the input and Redirected the participant to re-phrase or change topic. Table 2 presents examples of each type of response.

Results

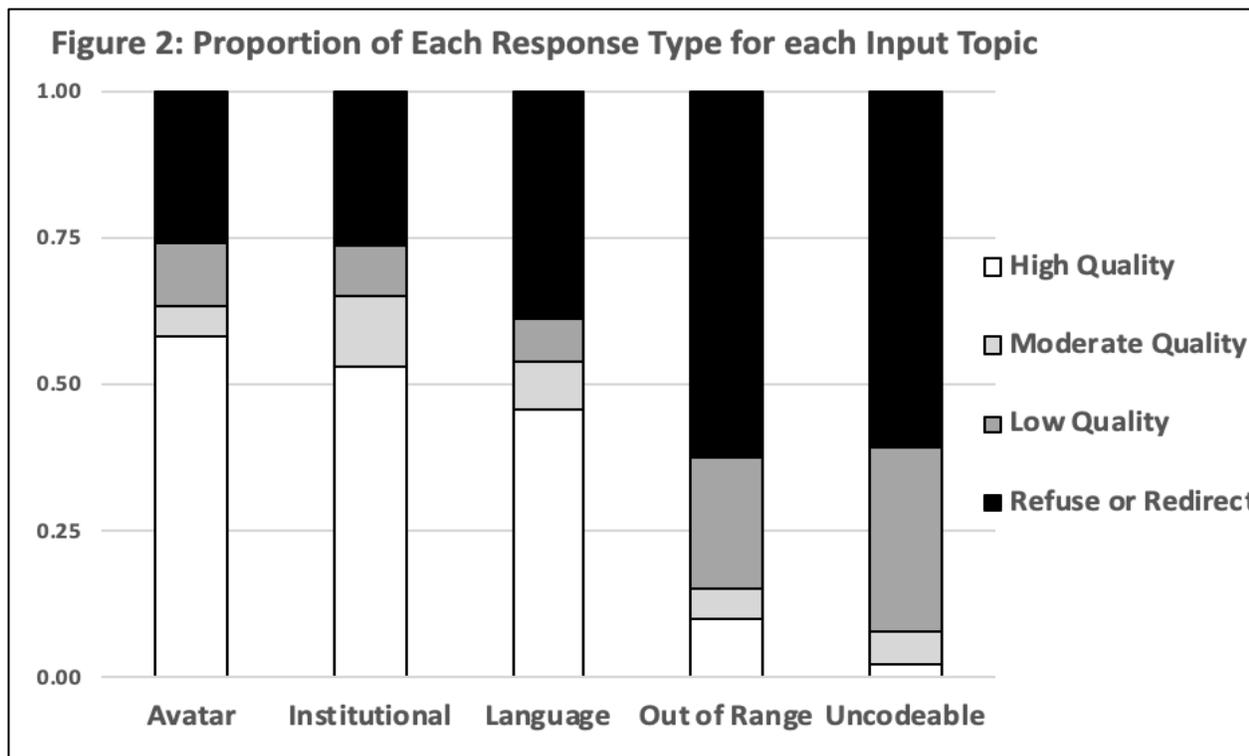
The overall nature of the conversations that our participants had with Ilse were similar to those reported previously in the literature, in terms of their length, their emphasis on the avatar, and the quality of the responses. The average number of conversational utterances for each participant was 12.2 (range: 1 – 49) leading to a total of 1461 input utterances. Of these, approximately one quarter (N = 364) were classified as uncodeable. Of the remaining 1097 utterances, almost half of them (N = 526) were questions about the avatar herself, nearly one third (N = 331) were about the various institutions connected with the project, and approximately 18% were about our science content area of linguistics (N = 201); the remaining questions (N = 40) were about other topics outside of Ilse's prepared range.

Across all topic areas, 40% of the responses were rated as being of High Quality, 7% were of Moderate Quality, 15% were of Low Quality and 25% were Refuse or Redirect responses. However, these types of responses were not equally distributed across the different topics ($X^2 = 389.12$, $df = 12$, $p < .0001$). As can be seen in Figure 2, Ilse was far more likely to Refuse or Redirect an input or provide a Low Quality response when it was Out of Range or Uncodeable, and she was more likely to provide a High or Moderate quality answer for topics she was prepared to discuss.

Table 2: Examples of different quality responses and low confidence responses by Ilse

Category	Examples of input-output pairings
High Quality	<ul style="list-style-type: none"> • Input: Tell me about accents. • Response: Every place has words and ways of pronouncing sounds that are special to them. If you're not from there, it will sound like a different accent! • Input: How do I use the iPad? • Response: We use our iPads for language science research and outreach demonstrations! If you see somebody in a labcoat with one, ask them what they're up to and you just might get to play a cool game.
Moderate Quality	<ul style="list-style-type: none"> • Input: What are the most popular languages on the Earth? • Response: Papua New Guinea is the most multilingual country, with over 839 living languages, according to <i>Ethnologue</i>, a catalogue of the world's known languages. • Input: Can you tell me something about Ohio State University? • Response: This is the Language Pod, a real Ohio State research lab that studies all kinds of things about language! We are doing experiments on things like how different people talk differently, how children learn language, and how computers and people can interact using human language, just like we are now!
Low Quality	<ul style="list-style-type: none"> • Input: Tell me some cool facts about language. • Response: Thanks for the positive feedback! • Input: I like your hair. • Response: COSI's electrostatic generator is only available as part of daily shows. Keep an eye on today's published schedule and the center stage on the first floor if you want to find it!
Refuse/Redirect	<ul style="list-style-type: none"> • Input: How does American Sign Language work? • Response: I'm sorry, I'm not sure I understood your question, can you please rephrase it? • Input: OK, can you tell me what the weather is gonna be like in the next two hours? • Response: Hmm, not quite sure what you're asking. I can tell you how many languages are spoken in the world or about the gadgets cafe or recommend a good place to eat here.

Figure 2: Proportion of Each Response Type for each Input Topic



However, an alternative way to think about quality is that for inputs that were difficult for Ilse to understand, the best response is in fact to refuse the input. Previous research (Aneja et al., 2020) notes that these kinds of clarification responses don't impact how much users like conversational agents, although they do make them seem more robotic. Thus, for Out of Range and Uncodeable topics, a Refuse or Redirect response can be seen as a reasonable answer – at least one of Moderate Quality. To probe this approach to quality, we assigned High, Moderate, and Low Quality responses to the values 1, .5, and 0 respectively for all topic types. However, Refuse or Redirect responses were valued as .5 (i.e. Moderate Quality) for the topics Out of Range and Uncodeable and as 0 (i.e. Low Quality) for the remaining topic areas. A One-way ANOVA with topic area as the independent variable and this response score as the dependent variable found that the overall quality of responses was .52 (Std Dev =

.44), which corresponds to an average response of Moderate Quality. There was a significant effect of topic on this quality score ($F(4, 1456) = 22.44, p < .001$). Post-Hoc Tukey's tests showed that in fact Ilse had the lowest quality responses for the Uncodeable items (.34), which was significantly worse ($p < .05$) than the quality of her responses to all of her prepared topics, but not different from the quality of her response to Out of Range topics (.44). Within her prepared topic areas, Ilse gave significantly higher quality responses to questions about the Avatar herself (.61) than to questions about Language (.50) with questions about Institutional topics being intermediate (.59). This level of response quality is comparable to other conversational agents deployed in museums.

We next considered how the participants subjectively perceived their interactions with Ilse: Did people like her? Did they find their interactions with her successful?

Participants were asked directly to rate their experience with Ilse on a 4 point scale, where higher numbers indicated more positive judgments. To evaluate the ratings, we compared them via t-test to the mid-point of the scale (a score 2), which we treated as a neutral rating. All item ratings were significantly above this mid-point, as shown in Table 3. Despite this overall high positivity, we also noted that the ratings to the item “Was it fun talking to the avatar” were notably higher, and in fact this item was significantly more positive than all of the other experience ratings (see paired t-test results in Table 3). Thus, while participants found their interactions with Ilse to be generally smooth, they found the interactions to be particularly fun.

We also investigated whether interacting with the avatar changed participants’ level of interest in the science connected with it, specifically, automatic speech recognition and language science and linguistics, as well as their interest in participating in studies at the nearby Language Sciences Research Lab. Participants rated their interest on both the pre-interaction survey and the post-

interaction survey using a 4 point Likert scale with higher numbers indicating higher levels of interest. Interacting with the avatar did not lead to any significant changes in interest for any items, although the initial rates of interest were rather high to begin with, leaving little room for change to happen (M = 3.01 for initial interest in automatic speech recognition, M = 2.64 for initial interest in language science and linguistics, M = 3.09 for initial interesting participating in a study).

A final set of analyses examined whether a participant’s age or gender influenced either their subjective or objective experiences with the avatar. We found no effect of the participants’ gender on any measure, but there was a significant negative correlation between age and three of the subjective measures (how fun it was to talk to the avatar, $R = -1.9$, $p = .044$; how helpful the avatar was, $R = -.322$, $p < .001$; how well the avatar understand them, $R = -.228$, $p = .014$). The negative values for these correlations indicate that younger participants gave more positive ratings on these measures than older participants. We

Table 3: Participants’ ratings of their experience on a scale from 1 – 4, with higher numbers reflecting more positivity. (note: some participants skipped some items, leading to minor differences in the N’s for each test).

Item	Mean (S.E.)	Comparison to Scale midpoint	Comparison to being “Fun”
<i>Was it fun talking to the avatar?</i>	3.27 (.08)	$t(117) = 16.17, p < .001$	n/a
<i>Was the avatar helpful?</i>	2.78 (.09)	$t(115) = 8.9, p < .001$	$t(115) = 6.19, p < .001$
<i>Was it easy talking to the avatar?</i>	2.47 (.1)	$t(116) = 4.94, p < .001$	$t(116) = 7.97, p < .001$
<i>Did the avatar’s speech sound natural</i>	2.61 (.09)	$t(117) = 6.83, p < .001$	$t(117) = 6.75, p < .001$
<i>How well did the avatar understand you?</i>	2.34 (.09)	$t(115) = 3.94, p < .001$	$t(115) = 11.09, p < .001$

note, however, that there were no significant correlations between age and the objective measures of length and response quality.

Discussion

Interactive conversational agents like our avatar Ilse have been used to engage with museum visitors in the past and they will likely become increasingly common. From a technological perspective, Ilse performed comparably to similar agents in museum settings and elicited similar kinds of interactions. On average, interactions with Ilse lasted for around a dozen turns and were often focused on asking Ilse questions about herself, many treating her very anthropomorphically. About a quarter of the conversational turns were uncodeable for content, largely reflecting technological challenges of conducting visitor-guided recordings in a museum environment. Ilse's responses were of moderate quality overall, although about half of her responses were of high quality when the topic area was one she was prepared to discuss.

From a visitor perspective, our participants were very positive about their experience with Ilse. They found her to be helpful and easy to talk to and felt that she largely understood them. Moreover, they really found it fun to interact with her. In addition, younger people were more positive about their interactions than older people. That said, the strength of participants' positive assessments was linked to how successfully Ilse responded to them specifically. Participants who generated higher quality responses from Ilse rated their experience more highly.

Based on the subjective ratings that participants provided, it is clear that Ilse successfully accomplishes one of the core goals of informal science learning (Bell et al., 2009): she generates interest and

excitement, particularly so for children. Moreover, the fact that visitors prefer to talk with Ilse about herself suggests an optimal use for this kind of conversational agent: she can be a tool for helping people learn about what she is, namely, an implementation of artificial intelligence. That is, while we had trained Ilse to discuss the museum, our local university, our onsite lab, and our science (linguistics), none of those topics were as popular as treating Ilse as a genuine social being and asking her to talk about herself. However, while Ilse was able to describe some of her own underlying technology, she was not prepared for any in-depth conversations about the technology, philosophy, or societal implications of artificial intelligence. Future iterations of Ilse can leverage the interest she generates into a more focused learning experience that engages people intentionally about the science that makes Ilse happen.

One key insight from this study, which points to an important limitation of the current work, is that visitor interest depends to some extent on Ilse's performance. This relationship highlights the importance of improving the technology itself. For example, one issue with Ilse's conversations was that many of them yielded uncodeable input. Our museum location was a fairly noisy environment and we provided no mechanism to help Ilse restrict her perception to a single voice. Other systems have addressed this problem by having a facilitator control the microphone input to the conversational agent (e.g. Traum et al., 2012) or incorporating elements such as touch-screens for the interaction (e.g. Bickmore et al., 2011). Our own team is working on integrating the avatar with a visual system that would identify the speaker and thus help facilitate the speech recognition process.

Another area in which Ilse's performance could be improved is in her ability to flexibly adapt to the topics presented by the participants. In this area, we note that Ilse's underlying processing is dated: she was built and implemented before the current explosion of generative AI systems. Her old-fashioned system was far less able than modern generative systems to handle unexpected topics and diverse phrasings of questions. On the other hand, her system was also safer in that she was unable to generate hallucinatory responses or offensive content. Our team is working on integrating the benefits of generative AI into Ilse's system in a way that does not simultaneously lead to inappropriate outputs.

Technology is constantly changing and improving, and we have no doubts that future iterations of Ilse will perform better – able to understand more inputs and able to produce more high-quality outputs. The current study illustrates how valuable conversational agents like Ilse can be for public engagement and promoting interest and understanding of technology.

Acknowledgements

We thank the students Shea Mangan and Jay Whitney for their assistance in conducting this study and coding the data. We also thank the Center of Science and Industry (COSI) for allowing us to host this project in their space. This project was funded by grants from the Artificial Intelligence in the Arts, Humanities, and Engineering: Interdisciplinary Collaborations and the National Science Foundation (NSF IIS 1618336 NSF BCS 2140708).

References

- Aggarwal, P., Artstein, R., Gerten, J., Katsamanis, A., Narayanan, S. S., Nazarian, A., & Traum, D. R. (2012, May). The Twins Corpus of Museum Visitor Questions. In *LREC* (pp. 2355-2361).
- Aneja, D., McDuff, D., & Czerwinski, M. (2020, October). Conversational error analysis in human-agent interaction. In *Proceedings of the 20th ACM international conference on intelligent virtual agents* (pp. 1-8).
- Ashfaq, M., Yun, J., Yu, S., & Loureiro, S. M. C. (2020). I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics*, *54*, 101473.
- Bell, Philip; Bruce Lewenstein; Andrew E. Shouse & Michael A. Feder (eds.) 2009. *Learning Science in Informal Environments: People, Places, and Pursuits*. Washington D. C.: The National Academies Press.
- Bickmore, T., Pfeifer, L., & Schulman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11* (pp. 55-67). Springer Berlin Heidelberg.
- Cohn, M., Chen, C. Y., & Yu, Z. (2019, September). A large-scale user study of an Alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue* (pp. 293-306).

- Diederich, S., Brendel, A. B., Morana, S., & Kolbe, L. (2022). On the design of and interaction with conversational agents: An organizing and assessing review of human-computer interaction research. *Journal of the Association for Information Systems*, 23(1), 96-138.
- Gaia, G., Boiano, S., & Borda, A. (2019). Engaging museum visitors with AI: The case of chatbots. In T. Giannini and J. P. Bowen (eds.) *Museums and Digital Culture: New Perspectives and Research*, 309-329. Springer.
- Gasteiger, N., Hellou, M., & Ahn, H. S. (2021). Deploying social robots in museum settings: A quasi-systematic review exploring purpose and acceptability. *International Journal of Advanced Robotic Systems*, 18(6), 17298814211066740.
- Gonzalez, A. J., Hollister, J. R., DeMara, R. F., Leigh, J., Lanman, B., Lee, S. Y., ... & Wilder, B. (2017). AI in informal science education: bringing turing back to life to perform the turing test. *International Journal of Artificial Intelligence in Education*, 27, 353-384.
- Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016, March). Understanding user satisfaction with intelligent assistants. In *Proceedings of the 2016 ACM on conference on human information interaction and retrieval* (pp. 121-130).
- Kopp, S., Gesellensetter, L., Krämer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide—design and evaluation of a real-world application. In *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings 5* (pp. 329-343). Springer Berlin Heidelberg.
- Lane, H. C., Noren, D., Auerbach, D., Birch, M., & Swartout, W. (2011). Intelligent tutoring goes to the museum in the big city: A pedagogical agent for informal science education. In *Artificial Intelligence in Education: 15th International Conference, AIED 2011, Auckland, New Zealand, June 28–July 2011 15* (pp. 155-162). Springer Berlin Heidelberg.
- Luger, E., & Sellen, A. (2016, May). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 5286-5297).
- Al Moubayed, S., Beskow, J., Granström, B., Gustafson, J., Mirning, N., Skantze, G., & Tscheligi, M. (2012). Furhat goes to Robotville: A large-scale multiparty human-robot interaction data collection in a public space. In *International Workshop on Multimodal Corpora, Tools, and Resources. Istanbul, Turkey*.
- Norouzi, N., Kim, K., Hochreiter, J., Lee, M., Daher, S., Bruder, G., & Welch, G. (2018, November). A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents* (pp. 17-22).
- Robinson, S., Traum, D. R., Ittycheriah, M., & Henderer, J. (2008, May). What would you Ask a conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting. In *LREC* (pp. 1-7).
- Swartout, W., Traum, D., Artstein, R., Noren, D., Debevec, P., Bronnenkant, K., ... & White, K.

- (2010). Ada and Grace: Toward realistic and engaging virtual museum guides. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10* (pp. 286-300). Springer Berlin Heidelberg.
- Traum, D., Aggarwal, P., Artstein, R., Foutz, S., Gerten, J., Katsamanis, A., ... & Swartout, W. (2012). Ada and grace: Direct interaction with museum visitors. In *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12* (pp. 245-251). Springer Berlin Heidelberg.
- Wagner, L., Speer, S. R., Moore, L. C., McCullough, E. A., Ito, K., Clopper, C. G., & Campbell-Kibler, K. (2015). Linguistics in a science museum: Integrating research, teaching, and outreach in a language sciences research lab. *Language and Linguistics Compass* 9, 420 – 431.
- Yang, H., & Lee, H. (2019). Understanding user behavior of virtual personal assistant devices. *Information Systems and e-Business Management*, 17, 65-87.