

Spill the Tea: When Robot Conversation Agents Support Well-being for Older Adults

Weslie Khoo

Indiana University Bloomington
Bloomington, Indiana, USA
weskhoo@iu.edu

Long-Jing Hsu

Indiana University Bloomington
Bloomington, Indiana, USA

Kyrie Jig Amon

Indiana University Bloomington
Bloomington, Indiana, USA

Pranav Vijay Chakilam

Indiana University Bloomington
Bloomington, Indiana, USA

Wei-Chu Chen

Indiana University Bloomington
Bloomington, Indiana, USA

Zachary Kaufman

Indiana University Bloomington
Bloomington, Indiana, USA

Agness Lungu

Indiana University Bloomington
Bloomington, Indiana, USA

Hiroki Sato

Indiana University Bloomington
Bloomington, Indiana, USA

Erin Seliger

Indiana University Bloomington
Bloomington, Indiana, USA

Manasi Swaminathan

Indiana University Bloomington
Bloomington, Indiana, USA

Katherine M. Tsui

Toyota Research Institute
Cambridge, Massachusetts, USA
kate.tsui@tri.global

David J. Crandall

Indiana University Bloomington
Bloomington, Indiana, USA
djcran@indiana.edu

Selma Šabanović

Indiana University Bloomington
Bloomington, Indiana, USA
selmas@indiana.edu

ABSTRACT

Robots could support older adults' well-being by engaging them in meaningful conversations, specifically to reflect on, support, and improve different aspects of their well-being. We implemented a system on a QT social robot to conduct short autonomous conversations with older adults, to help understand what brings them feelings of joy and meaning in life. We evaluated the system with written surveys and observations of 12 participants including older adults, caregivers, and dementia care staff. From this, we saw the need to improve user experience through personalized interaction that better support older adults as they talk about well-being. Improving the interactions will involve improving the conversation flow, detecting emotions and nonverbal cues, and natural language processing to extract topics around well-being.

CCS CONCEPTS

• **Applied computing** → **Sociology**; • **Computing methodologies** → **Natural language processing**; • **Human-centered computing** → **Interaction design**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HRI '23 Companion, March 13–16, 2023, Stockholm, Sweden

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9970-8/23/03...\$15.00

<https://doi.org/10.1145/3568294.3580067>

KEYWORDS

human-centered interaction design, conversational systems, robotics, natural language processing

ACM Reference Format:

Weslie Khoo, Long-Jing Hsu, Kyrie Jig Amon, Pranav Vijay Chakilam, Wei-Chu Chen, Zachary Kaufman, Agness Lungu, Hiroki Sato, Erin Seliger, Manasi Swaminathan, Katherine M. Tsui, David J. Crandall, and Selma Šabanović. 2023. Spill the Tea: When Robot Conversation Agents Support Well-being for Older Adults. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion)*, March 13–16, 2023, Stockholm, Sweden. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3568294.3580067>

1 INTRODUCTION

Prior research has studied the effectiveness of robots in supporting older adults' well-being [14]. The reported positive effects could be due to enhanced meaningful social connections [15], increased conversation with family members [7], and engagement in a group or individual intervention [1]. More recently, studies have shifted towards conversational systems for well-being, driven in part by the popularity of Amazon Alexa, Google Assistant, and other commercial systems. These systems have been found to benefit older adults by helping them talk about feelings that may otherwise be repressed [16]. For example, text-based conversational systems on mobile apps have been found to positively affect people's well-being, especially their mood [9]. Sofia, a conversational agent with an open-ended dialogue system, helps people discuss their mental state through reflection [11].

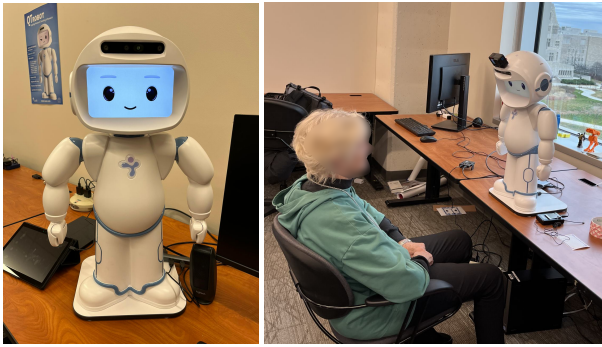


Figure 1: (Left) QT-based robot system and (Right) participants interacting with the system.

In this paper, we present preliminary work on designing a social robot conversation system for supporting older adult well-being. Our system uses LuxAI’s commercially-available QTrobot (QT) [10] with our custom software. We tested QT in a series of iterative evaluation and refinement studies as part of robot design development [12]. Our preliminary results from user testing suggest that QT was viewed positively by older adults. We then describe our ongoing work in making the conversational system more personal, in making its interactions more natural and accurate, and in implementing autonomous structured and unstructured activities to help users talk about their well-being.

2 THE ROBOT PLATFORM

We use QT [10], a commercially-available humanoid robot that stands around 2 feet tall (Fig. 1 left). The robot’s face is an 8-inch 800x480 LCD display that can show facial expressions and emotions. The robot has a flexible head and arm joints which allow it to express various simple gestures. It features a camera and 3d depth sensor (an Intel RealSense D455) mounted on its forehead, and a microphone array on the top of the head to collect spatial audio.

3 INTERACTION DESIGN AND EVALUATION

3.1 Initial Interaction Design

Our overall goal is to build a conversational system that autonomously interacts with older adults in order to identify topics that encourages them to reflect and talk about their well-being (Fig. 2). To do this, the system will eventually need to capture both audio of what the person is saying and visual data about the person’s facial expression, gestures, and visual attention. Our current version only uses audio but nevertheless provides a complete working robot that we use for initial design and evaluation.

An interaction begins with the robot greeting the participant with a physical wave gesture and a verbal cue asking for their name. The robot then listens for the users’ verbal response, which is transcribed using speech recognition [21]. The transcribed name is cross-checked with a database of common names [3] to ensure that the transcribed word is actually a name. The system repeats the name periodically throughout the interaction by adding it to questions and statements (e.g., “It is nice to meet you, Andrew.”).

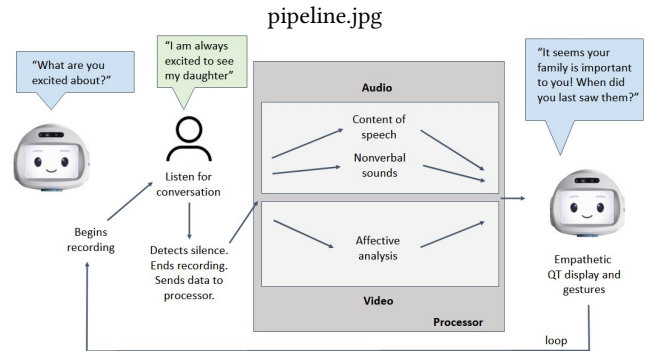


Figure 2: Overview of our autonomous conversation robot system. The system asks a question, listens for the user’s answer, analyzes the spoken content as well as other audio and visual cues, generates an emphatic response, and asks a follow-up question. The process then repeats.

Then the robot begins the main conversation loop. The overall interaction pattern is that the system asks the participant a question and then listens for a response. The robot comments on the participant’s response and then ask a new question. This interaction pattern is arguably the opposite of Alexa and other voice assistants; while they simply answer one-off factual questions posed by users, our system is designed to get people engaged in a conversation by asking open-ended questions and then following up on their answers. Here is a sample conversation that a participant had with our system during our user testing:

System: *what do you enjoy [name]?*

Human: *oh I enjoy traveling lots of... traveling far places*

System: *I enjoy it too, [name]! What is your favorite place that you have traveled to?*

Human: *Yellowstone National Park*

System: *I love Yellowstone, [name]! What do you like about Yellowstone National Park?*

To achieve this, we used an off-the-shelf machine learning model, Generative Pre-Trained Transformer 3 (GPT-3) [13]. GPT-3 is an autoregressive language model that produces human-like content that is associated with the input data. An untrained GPT-3 model produces generic text that may not be appropriate for use in conversations (see Table 1, second column), so we fine-tuned GPT-3 using transcripts from interviews of people talking about what gives them meaning in their lives [15]. To better control how our system replies to users’ responses, we trained two separate GPT-3 models: one to empathetically acknowledge what the user just said, and another to continue the conversation with a follow-up question.

The participant’s speech is transcribed in real-time and fed to the GPT-3 models. One surprisingly difficult challenge is determining when the user is finished speaking. Currently, our system simply waits for a period of silence lasting longer than a predefined threshold, but this is susceptible to both false positives (cutting people off when they were just hesitating to think of a word) and false negatives (when background noise creates spurious audio signals). Moreover, waiting for a period of silence introduces a delay in QT’s response that makes the interaction seem less natural. In ongoing

work, we are developing better techniques to recognize when a user has completed their thought using both audio and visual cues. In the meantime, we use an external microphone positioned close to the participant in order to avoid background noise interfering with the interaction. If QT does not detect any speech at all within 5 seconds, it verbalizes that it does not understand and would like the user to repeat.

The system ends the conversation when it hears “bye” or several variants. QT will then physically wave and say “*It was nice talking to you, {name}. Goodbye!*”

3.2 Evaluation with Survey and Observation

Method. We tested our QT-based system with 12 participants (7 older adults ages 65-85, 2 clinicians, and 3 dementia care staff) in two different locations: a conference exhibition booth (P1-P5) and a campus meeting room (P6-P12). First, we introduced QT and its functionality and asked if the participant would like to interact with it. Upon consent, we asked them to sit in a chair in front of QT (Fig. 1 Right) and have an open-ended conversation with QT. Then, participants filled out a survey designed by the research team focusing on their experience with the social robot (i.e., how much they enjoyed their interaction with QT, how comfortable they felt, how friendly they feel QT is, and their willingness to use QT). Their responses were recorded on a Likert scale from 1 (not enjoyable, not comfortable, not friendly, not willing to use) to 5 (enjoyable, comfortable, friendly, willing to use). We also recorded and observed the interactions to gain insights into the interaction experience. The university’s institutional review board approved this study.

Results. According to our survey results (Likert scale 1 to 5), most of the participants enjoyed the interaction with the robot ($M=3.8$, $SD=1.0$), felt comfortable with the robot ($M=3.6$, $SD=0.9$), and felt the robot is friendly ($M=3.5$, $SD=1$).

Participants were divided on their willingness to use the robot ($M=3.3$, $SD=1.6$). For example, P2, a dementia care staff thought it was helpful and would be willing to use it: “[I] would use it with my clients living with dementia who are isolated at home and need engagement.” P11, who is a healthy older adult, also said it would be helpful for his “*wife who has dementia... help to remember things.*” Three participants were not willing to use the robot. While two didn’t specify why, P7, who is a healthy older adult, suggested the robot was more applicable for other populations: they “*recommend this for people who live alone, have memory problems – to be ‘companions’ – helping to remember things.*” This suggests P7 couldn’t find their own connection to the robot, and we need to make QT more personalized to fit the healthy older adult population, similar to P7.

Participants gave practical feedback about the QT-based system. For example, P8, a healthy older adult, said “*QT is very slow and rudimentary in its communication and resources. If it were as efficient as Alexa in communicating, it would be far more inviting.*” Here, we infer that the autonomous conversation needs to flow quicker and more naturally.

We also observed recorded videos of the robot interacting with the participants. Most of the interactions were positive, in which the participants and the bystanders laughed at the responses that

the robot made. However, some of the laughter was because of the robot’s unexpected response which differs from how a human talks. For example, in one scene when a participant talked about her sister, the robot asked “*Is your brother important to you?*” the participant answered, “*yes and I have a sister as well,*” and QT system answered, “*I am sorry to hear that. How are you doing?*” The participant laughed and then immediately said “*goodbye*” to the robot. After the conversation, she told the researcher laughing “*I guess it doesn’t like my sister.*” Our robot system needs to detect these moments to be able to provide a better conversation flow with users for further discussion.

4 NEXT STEPS

4.1 Improving Conversation Flow

Participants commented that our robot system is slow, which could affect their willingness to continue the conversation with the robot about their well-being. A major reason is that our current prototype detects when the participant is done speaking by waiting for a period of silence. Of course, in natural conversation people use a variety of cues, verbal and non-verbal, to signal that they are done speaking. As Skantze [17] mentioned, “*apart from the auditory channel, the visual channel (face and body) is also important for turn-taking.*” Therefore, we plan to implement analysis of visual and audio information to better synchronize this conversation turn-taking, and hence increase the possibility of shifting the conversation towards well-being.

To implement visual channels, we began to explore computer vision techniques. Particularly, Skantze [17] emphasized gaze as an important visual cue to identify the end-of-turn: “*Speaker tends to look away at the beginning of the turn but then shift the gaze towards the listener at the end of the turn.*” Based on these suggestions, we are investigating using eye contact detection with the robot’s camera. To model the human’s eye contact with a robot, we are using existing work [2] which estimates, using a neural network model, the probability that a person is looking at the camera. We are also exploring a multi-modal approach using both audio and visual channels with TalkNet [18], an active speaker detection model that estimates whether or not a face is speaking. While these two approaches are promising, making them run in real-time on the robot system’s limited processing power may be challenging.

4.2 Determining Emotions During Conversations

Observations from our user testing indicate that our robot system does not always respond appropriately to user responses and reactions. This could diminish the users’ interest in thinking and talking about well-being with the robot. To give the robot more context in analyzing users’ responses, we propose using computer vision to analyze emotions and other non-verbal cues during conversations. For example, a participant gazing off into space may indicate that they are bored with the current topic of discussion, while a smile may indicate that the participant wants to continue.

In preliminary work, we use MediaPipe [4] to detect faces in video frames. If multiple faces are detected, the larger face is used since they are likely closer to the robot. Next, each frame is fed to a deep network that estimates the user’s valence and arousal, using

Table 1: Example of QT interactions with users, with untrained and trained GPT-3 models.

| User's speech | QT's response from untrained GPT-3 model | QT's response from our trained models 1 & 2 |
|---|---|--|
| I went to a local festival. | The festival was a lot of fun. There were a lot of people there, and the atmosphere was really festive. | M1: Sounds interesting! M2: Who did you go to the local festival with? |
| I watched a baseball game last night and my favorite team lost. | The Chicago Cubs lost to the St. Louis Cardinals 5-4 last night. | M1: I feel sorry for that M2: Who did you watch it with? |
| I didn't see my family because of the pandemic. | Now I see them all sick and I can't do anything for them that's really difficult to go through. | M1: I'm sorry you didn't see your family. M2: What other things do you enjoy? |

an open-source implementation [20] of Toisoul et al. [19]. Finally, the frame-level predictions are aggregated into a single output (positive, negative, or neutral) for a short segment of interaction. This estimate can then be used by the robot to respond with an emphatic expression or gesture. For example, if a positive user response is detected, the robot could smile while responding to the user, and talk more about what is meaningful in their lives. Our next step is to analyze these videos in real-time.

4.3 Detecting Nonverbal Sounds During Interaction

Beyond using computer vision to detect facial expressions, non-verbal sounds are just as important to identify emotions or moods, which can help steer away from awkward conversations. Nonverbal sounds such as laughing, sighing, and crying are often ignored but just as key [6]. In the above-mentioned observations, if the robot were able to detect the awkward laughter when the robot said “*sorry to hear*,” it could apologize or move on — and thus have a deeper conversation related to their well-being. Nonverbal sounds can also help the robot determine how a person may be feeling in the tone of their voice. Furthermore, when integrated into a multimodal system using vision for facial feature detection and audio recognition of speech content, we can cross-examine parts of the conversation to determine whether a person is being sarcastic or saying one thing but meaning another. We are investigating machine learning-based approaches to identify human vocal elements such as pitch.

4.4 Improving the NLP Pipeline

In addition to improving the user experience of autonomous conversations with QT, we plan to improve the content of the conversations themselves by using better NLP and generation. This could help us analyze the content of the conversation to identify topics and themes that may increase the participant’s well-being. In preliminary work, several conversation transcripts from interviews with older adults were analyzed using Spacy [5] and IBM Watson [8]. We trained these NLP models with our real-world data to improve the accuracy of detecting and extracting relevant named entities. The named entities include a variety of objects, such as names, places (geo-locations), dates, and activities. We also performed sentiment analysis on these specific named entities and conversations as a whole. This allows us to understand the aspects of their life that carry a particular sentiment. The sentiments are categorized into positive, negative, and neutral which help our robot system understand and converse about users’ well-being.

5 CONCLUSION

We aim to develop a robot system to help older adults discuss their well-being through autonomous conversations. To do this, we developed a prototype running on a QT robot and tested it with 12 participants including older adults, clinicians, and dementia care staff. Our results suggested that while most people enjoyed using our prototype, our current system needs to improve the conversation flow and test in environments that reflect real-world scenarios. Furthermore, the system needs to use nonverbal cues in order to better engage users and talk about their well-being. We plan to deploy and test the robot at older adults’ homes for long-term studies, test it with specific well-being-related questions (i.e., meaning and social support [15]), and explore other technical challenges like personalized interaction for each user.

ACKNOWLEDGMENTS

This work was supported by Toyota Research Institute.

REFERENCES

- [1] Shu-Chuan Chen, Cindy Jones, and Wendy Moyle. 2018. Social robots for depression in older adults: a systematic review. *Journal of Nursing Scholarship* 50, 6 (2018), 612–622.
- [2] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. 2020. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature communications* 11, 1 (2020), 1–10.
- [3] genderize.io. Last Accessed December 7, 2022. Genderize Our Data. [Online]. Available from: <https://genderize.io/our-data>.
- [4] Google. Last Accessed December 7, 2022. MediaPipe Face Detection. [Online]. Available from: https://google.github.io/mediapipe/solutions/face_detection.
- [5] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear* (2017).
- [6] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su, and Yi-Hsuan Chen. 2019. Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds. 5866–5870. <https://doi.org/10.1109/ICASSP.2019.8682283>
- [7] Suzanne Hutson, Soo Ling Lim, Peter J Bentley, Nadia Bianchi-Berthouze, and Ann Bowling. 2011. Investigating the suitability of social robots for the wellbeing of the elderly. In *International conference on affective computing and intelligent interaction*. Springer, 578–587.
- [8] IBM. Last Accessed December 7, 2022. IBM Watson Natural Language Understanding Entity Types. [Online]. Available from: <https://cloud.ibm.com/docs/natural-language-understanding>.
- [9] Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 2018. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth* 6, 11 (2018), e12106.
- [10] LuxAI. Last Accessed December 7, 2022. QTrobot V2 Humanoid Social Robot for Human AI Research & Teaching. <https://luxai.com/qtrobot-for-research/>
- [11] Raju Maharjan, Kevin Doherty, Darius Adam Rohani, Per Bækgaard, and Jakob E Bardram. 2022. Experiences of a Speech-enabled Conversational Agent for the Self-report of Well-being among People Living with Affective Disorders: An In-the-Wild Study. *ACM Transactions on Interactive Intelligent Systems (TiIS)* 12, 2 (2022), 1–29.

- [12] Bernt Meerbeek, Martin Saerbeck, and Christoph Bartneck. 2009. Iterative design process for robots with personality. In *AISB2009 Symposium on new frontiers in human-robot interaction*. 94–101.
- [13] OpenAI. Last Accessed December 7, 2022. Open AI Models. [Online]. Available from: <https://beta.openai.com/docs/models/gpt-3>.
- [14] Lihui Pu, Wendy Moyle, Cindy Jones, and Michael Todorovic. 2019. The effectiveness of social robots for older adults: a systematic review and meta-analysis of randomized controlled studies. *The Gerontologist* 59, 1 (2019), e37–e51.
- [15] Natasha Randall, Swapna Joshi, Waki Kamino, Long-Jing Hsu, Abhijeet Agnihotri, Grace Li, Donald Williamson, Kate Tsui, and Selma Šabanović. 2022. Finding ikigai: How robots can support meaning in later life. *Frontiers in Robotics and AI* 9 (2022).
- [16] Korok Sengupta, Sayan Sarcar, Alisha Pradhan, Roisin McNaney, Sergio Sayago, Debaleena Chattopadhyay, and Anirudha Joshi. 2020. Challenges and Opportunities of Leveraging Intelligent Conversational Assistant to Improve the Well-Being of Older Adults. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (*CHI EA '20*). Association for Computing Machinery, New York, NY, USA, 1–4. <https://doi.org/10.1145/3334480.3381057>
- [17] Gabriel Skantze. 2021. Turn-taking in Conversational Systems and Human-Robot Interaction: A Review. *Computer Speech & Language* 67 (2021), 101178. <https://doi.org/10.1016/j.csl.2020.101178>
- [18] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. 2021. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3927–3935.
- [19] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence* (2021). <https://www.nature.com/articles/s42256-020-00280-0>
- [20] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. 2021. Face-analysis/emonet: Official implementation of the paper "estimation of continuous valence and arousal levels from faces in naturalistic conditions." [Online]. Available from: <https://github.com/face-analysis/emonet>.
- [21] Anthony Zhang. Last Accessed December 7, 2022. SpeechRecognition 3.9.0. [Online]. Available from: <https://pypi.org/project/SpeechRecognition/>.