Blair, A. and Foster, M. E. (2023) Real-World Evaluation of a University Guidance and Information Robot. In: 15th International Conference on Social Robotics (ICSR 2023), Doha, Qatar, 3-7 Dec 2023, pp. 193-203. ISBN 9789819987184 (doi: 10.1007/978-981-99-8718-4_17)

This is the author version of the work. There may be differences between this version and the published version. You are advised to consult the published version if you wish to cite from it: https://doi.org/10.1007/978-981-99-8718-4_17

https://eprints.gla.ac.uk/307944/

Deposited on 16 October 2023

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

Real-world evaluation of a university guidance and information robot

Andrew Blair[0000−0002−0453−1381] and Mary Ellen Foster[0000−0002−1228−7657]

School of Computing Science, University of Glasgow
a.blair.2@research.gla.ac.uk,maryellen.foster@glasgow.ac.uk

Abstract. We have developed a social robot to assist an existing support team in a large, recently-built university building designed for learning and teaching. Over the course of a week-long, supervised deployment, we collected long form questionnaire results (N = 59) on attitudes and feelings towards the robot from students and staff. We observed an overall positive response to the robot, but with a wide variety of specific opinions. We describe the limitations and challenges we found with the real-world deployment and outline next steps to allow an unsupervised deployment of the robot as part of the university's wider service delivery strategy.

Keywords: Human-robot interaction · Field studies.

1 Introduction

In recent years, social robots have been used in a wide range of public spaces, including shopping centres [6], hotels [14] and airports [8]. They offer the potential to add novel methods of service delivery to an organisation's repertoire, as well as providing an additional opportunity for public engagement.

At our university, a new learning and teaching building has recently been opened with capacity for over 2500 students. Information Services (IS) supports users of this building by deploying a support team to roam the building and assist with any queries users may have. However, this team are not present for the entire opening hours of the building, so an opportunity exists to develop a social robot to assist these building users as an additional service delivery tool for IS.

In a previous paper [1], we describe how the requirements for the robot system were developed and give a technical description of the implemented system, which combines the Pepper robot with the RASA [13] open-source chatbot framework, with an external microphone for speech recognition. In the current paper, we present the design and results of a week-long study where the robot was deployed in the aforementioned building and the university library. Hundreds of students and staff interacted with the robot, and we used a number of qualitative and quantitative measures to gather their opinions including a long-form questionnaire completed by 59 participants. We present the results on all of these measures and suggest necessary modifications for a potential future unsupervised deployment of the robot in a public space. The study results have been shared with IS senior management and will influence future service delivery strategies at the university.

2 Deployment Overview

We deployed the robot for five consecutive days, four in the learning and teaching building and one in the University library, for a total of 30 hours. We placed the robot at the three main entrances of the learning and teaching building (Figure 1), and on the main floor of the library. At least one researcher was present with the robot at all times. After each interaction, all users were prompted to rate the robot via a single-item Likert scale, and interested participants were also invited to fill in a longer questionnaire. In the following sections, we discuss the study outcomes from several perspectives. In Section 3 we discuss the system performance and the responses to the initial Likert scale; in Section 4, we discuss the user responses to the longer questionnaire; while in Section 5, we briefly describe other behaviours that were observed during the deployment, specifically focussing on the challenges presented by this real-world deployment setting.

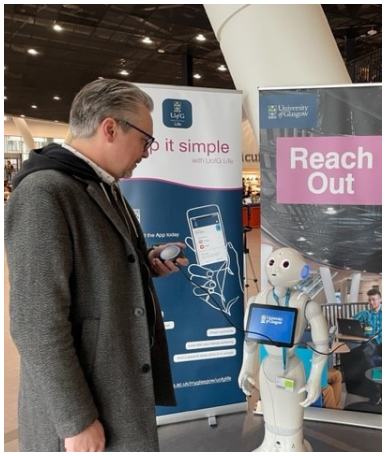

Fig. 1. A user interacting with the robot

3 System performance and conversation ratings

351 interactions occurred during the deployment, with 323 of them over one turn in length; the one-turn conversations generally represented early failure of one or more system components. The designed conversation length was five turns, and this was also the most common interaction length; some conversations were slightly longer, generally representing users who asked more than one question, while the longest conversations (up to 21 turns) were mainly due to a series of speech-recognition errors. Some utterances were misclassified, but users generally repeated themselves when necessary and most conversations were ultimately successful. From the interactions, 179 ratings were received on the Likert scale, with a mean of 3.47 (0.77) on a scale of 0 to 4, showing that the general impression of the robot was positive despite minor errors (Figure 2).

4 Questionnaire responses

59 participants completed the long questionnaire: 27 who identified as male, 27 as female, 4 as non-binary, and one participant who declined to disclose their gender. The majority of the participants (n=32) had never interacted with a robot

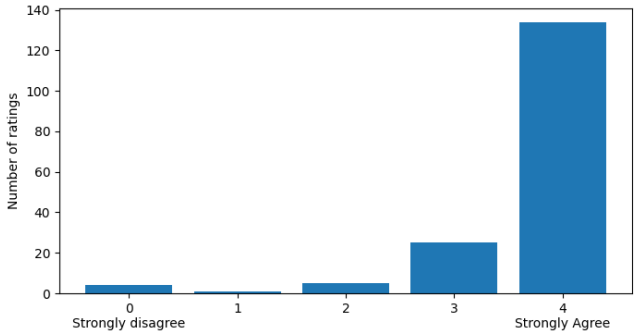

Fig. 2. Distribution of Likert ratings

before, but a large minority had (n=27). 12 participants were staff and 47 were students. The 12 staff were mainly building staff who had heard about the robot's presence in their work, but also included lecturers and project managers from other departments, representing a range of seniority levels. Of the 47 students, the majority were undergraduate students (n=33), with the remainder being postgraduate taught (n=13) and one postgraduate research student, and their fields of study varied widely, including both STEM and non-STEM subjects.

4.1 Likeability

The first section of the questionnaire was based on the Likeability section of the SASSI scale [9], which was developed to evaluate speech-based systems. Users were asked to rate the robot on a number of parameters on a scale of 0 to 4, with 0 representing "Strongly Disagree" and 4 representing "Strongly Agree".

As shown in Figure 3, on all questions, the user sentiment was greater than "Neutral" (2.0); indeed, for all but two questions, the mean was greater than "Agree" (3.0). The two questions with less positive responses, *It is clear how to speak with Pepper* and *I felt in control of the interaction with Pepper*, suggest that there was some confusion around how to interact with the robot. This matches what we observed during the deployment: users would look to the researchers for guidance on how to talk to the robot, even after being given instructions by the robot once the facial recognition system triggered.

4.2 Interaction factors

The next section of the questionnaire asked users to rate the robot on a number of factors, each to do with specific decisions that were made in the system implementation. For each question, the responses were given on a scale of 0 to 4, with 0 representing "Strongly Disliked" and 4 representing "Really Liked". A section of free-text was also provided to allow users to expand their answers if they wished to do so.

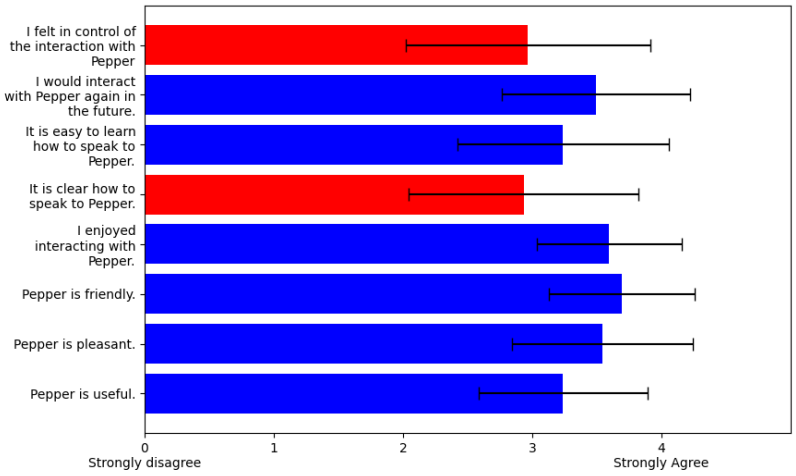

Fig. 3. Participant responses to the Likeability section of the survey, 0 = Strongly Disagree, 4 = Strongly Agree (mean and standard deviation)

How did you feel about the gestures, such as hand and arm movements? The responses were positive with a mean of 3.29 (0.74), with only one participant actively disliking the gestures. Despite the overall positive response, many people had less positive comments: for example, in its default mode, Pepper regularly flexes its hands, which some participants found "creepy,""unsettling," or "scary", and one felt like it was "going to fight [them]".

How did you feel about your face being tracked or followed as you interacted with the robot? There were mixed feelings surrounding this question. Overall, there was still a positive response with a mean of 2.76 (0.86); however, several students found it disturbing, with variations of the following quote:

> initial reaction [...] that it was a bit creepy with how its head would [move] as you walked past it. but once the interaction started it was adorable how she would look at you [...]

How did you feel about the robot starting the conversation with you? Mean response 3.08 (0.85).
Do you prefer the idea of you starting the conversation with the robot? Mean response 2.75 (0.86).
We will discuss these questions together. The high mean of the first question shows that in general the implemented *proactivity* (where the robot initiated a conversation as soon as a person was detected) was well received. However, the high standard deviation shows that there is no clear consensus. Where participants

responded with "Really liked" to the first question, they would often respond "Strongly disliked" to the second question; similarly, neutral respondents to the first question would tend to respond neutrally to the second.

How would you describe Pepper? The final question in this section asked users to describe Pepper in free text. The three most common responses to this question were "cute", "helpful" and "friendly". These are all common descriptions for Pepper, and are a positive sign that the robot design has aided the acceptance of the robot within the space. "Childlike" was also used, with two different connotations: most respondents who said this thought this made the robot appear unthreatening and therefore appealing to interact with, but some also found problems with this presentation, for example that because it appeared like a child that they were uncomfortable being prompted to touch it.

4.3 Feedback on QR codes

For external directions and Helpdesk articles, the system displayed a QR code on Pepper's tablet that would take them to the student mobile app for directions or to the IT Helpdesk website. 48 out of the 59 participants followed this user journey in their interaction, so the questionnaire included an item specifically addressing this part of the interaction.

Was it useful having the information on your phone rather than being on the tablet of the robot? On the same 0 to 4 scale as above, the mean result across the 48 participants was 3.1 (0.87), showing that the majority of participants agreed that having the information on their devices was a good addition to the system. However, some raised various issues with the QR code. One participant, despite answering the above question with "Strongly agree", suggested it would be:

> ...useful to both display the direction on the screen and with the qr code so the robot can be used without a smartphone...

The QR code was used to allow students to follow directions or read long-form Helpdesk articles even after walking away from the robot, but it is clear from this and other responses that the reason for this design choice could be made clearer.

4.4 Out of hours support

Would you prefer Pepper for out of hours support or a tablet? 42 respondents said Pepper, 14 said the Tablet and 3 suggested they would want both. For this question, we compared the pattern of responses to user opinions of the robot as expressed on the main survey, using Welch's T test. As we can see from Table 1, people's attitudes towards the robot's behaviour and the interaction that they had with the robot appeared to affect their willingness to use and accept the robot. We note an especially strong correlation between their attitudes towards the face-tracking and whether or not they would want to use the robot for out of hours support. One student even said they would be "terrified" if they saw the robot at night unannounced.

Table 1. Responses from users who prefer Pepper or a tablet for out of hours support

Question	Pepper	Tablet	p-value
It is easy to learn how to speak to Pepper	3.54	2.71	0.012
I would interact with Pepper again in the future.	3.69	2.85	0.010
How did you feel about the gestures, such as hand and arm movements	3.43	2.86	0.012
How did you feel about the robot starting the conversation with you	3.24	2.57	0.0105
How did you feel about your face being tracked or followed as you interacted with the robot	2.93	2.29	0.0046

4.5 Potential use cases

The final question concerned potential other use cases for the robot:

Would you like the robot to help you with anything else outside directions and the helpdesk? A number of participants referenced the ability of the robot to move, with one participant suggesting how it could be used to increase accessibility:

> people with disabilities. For example, if someone is blind, it could help them to find the elevator

As the target building is new, it is not an unreasonable suggestion; the building is fully accessible to wheelchair users, so it would be an ideal testbed for combining autonomous navigation and social robotics. Additionally, movement would allow us to further encapsulate the Reach Out roving model, whereby the ambassadors do not stay in one place and roam the building in an attempt to provide more visibility to building users. If this option is chosen, a different platform with better navigation capabilities than Pepper would be needed.

A common suggestion was to recommend events or activities around campus to students. Building from the QR code system discussed earlier, one participant suggested that Pepper could pass digital flyers for the events it would talk about via the QR codes such that the user could reference event details later. This would be straightforward to introduce, as the events data for the university is in a publicly accessible database.

Finally, several users spoke of just wanting to be able to chat to the robot. Whilst we explicitly discounted large language models (LLMs) for this experiment, as the goal was to provide answers to specific questions, the rise of services such as ChatGPT [12] has fuelled expectations of users to a much higher level. From a technical perspective it may be trivial to integrate an LLM, but the challenge would be controlling the output: for example, if the user asks about sensitive topics such as mental health, they have to be dealt with in the appropriate way which cannot be truly guaranteed with LLMs.

5 Observations

In addition to the formal questionnaire responses noted above, researchers also made notes of user behaviour during the deployment. We first describe general observations of people's reactions to encountering the robot, and then discuss specific incidents that exemplify the challenges that would be involved in deploying the robot more independently in this sort of real-world setting.

5.1 General observations

A number of students recognised Pepper from other deployments or from their studies. Pepper is used in Psychology courses at the University as a case study, and they were excited to see it in the real world. One student spoke of seeing the robot in various places in Japan but was surprised the interaction was via voice, as they had only ever interacted with Pepper via its tablet before.

One user was unable to interact with Pepper. They primarily communicate using sign language, and the robot neither understood or was able to respond with sign language. They usually were able to understand people by lip reading, but this is also not possible with Pepper. We made a design decision to not subtitle the robot or allow alternative input due to time constraints, but with future work accessibility would have to be a major consideration. It is also a limitation of Pepper to not be dexterous enough to perform sign language, but it could be developed to have it as an input from a user.

Along with all of our positive sentiment, we did see some strong negative sentiment and uncertainty. The following are direct quotes we observed from three students:

> oh my god what is that, why does it move
> do you aim to kill humans
> [walking by..] f*cking hell

This shows that despite making every effort, acceptance is fraught with challenges. Firstly, it shows the importance of making sure that responses are polite and appropriate. The bad publicity and therefore acceptance of the robot from stakeholders could be greatly affected if the robot responded inappropriately to these negative quotes.

The robot had an IS lanyard around its neck (Figure 1), which was noted by one student who said it made the robot seem "more official". They also suggested giving the robot a hat, as they felt it would make the robot stand out more and also make it "cuter". Clothing Pepper is a popular thing to do and has proven to have concrete effects on people's behaviour; when deployed as a shrine attendant in Japan and traditionally clothed, people would bow without prompting like they would with a human attendant [7].

5.2 Real world considerations

Dealing with abuse, of various forms, is an issue that social robots must be able to address in real-world deployments. We observed a number of concerning behaviours towards to the robot, even with staff present: one user threatened to hit the robot when it did not perform as expected, while multiple students attempted to flirt with it. Dealing with abuse [4], and specifically sexual harassment [3] has been explored in conversational agents. However, there is little research in how robots should deal with a physical threat in a real world environment [2]. The worry was echoed by some of the students and staff, with one student saying, "I hope people are not mean to Pepper" and another concerned that, if left unattended, "[Pepper] would get punched".

Another challenge is user preconceptions of speech recognition technology. Popular culture tries to show the humorous side of this with a sketch with a voice activated elevator[1] involving Scottish users; excerpts from this were also quoted when users were interacting with our robot. Some users were hesitant to interact with the robot saying, "it won't understand me". Especially at a university as diverse as Glasgow, with over 40% of the student body international, this is a significant challenge. Foster and Stuart-Smith's [5] research endorses our findings; they found that Scottish people expected to be able to understand the robot, but assumed it would struggle to understand their accent. The go-to response when Pepper responded with an error message was that it did not understand the user's speech, not that it was unable to match an intent and carry out a task. One participant even said "stop being racist robot". Care should be taken to generate appropriate errors, possibly repeating the phrase heard by the robot or displaying it on the tablet. However, this may come with the trade-off of making the user experience more frustrating with repetitive statements, so a balance must be struck.

We also observed "play-fighting" between students, causing concern to the other building users. It was stopped by a manager but raises a question of what a robot should do if supporting the building out of hours. Many would argue it is outside the domain of the robot, but if the robot is seen by users as truly working alongside staff then it should intervene. Other possibilities include Pepper simply recording the interaction. Some may argue this is a blatant privacy violation, but the building is already covered by CCTV, so Pepper would simply be adding another more versatile method of surveillance. Either way, it would be imperative to alert staff if Pepper was to witness such an incident.

Many students also asked the robot questions to vent their frustrations. One student asked "where can i find somewhere to study between 12 and 1" and when the robot responded that it could not find that room, the student responded "Of course, because there is nowhere". These types of interactions are where the personality of the robot becomes very relevant, as different people would answer the question in various ways. Völkel explored this and formalised an approach for developing personalities in conversational agents [15]. It once again raises the

[1] `https://www.bbc.co.uk/programmes/p00hbfjw`

importance of co-design with our users: university students present an interesting and one of the most wide ranging demographics of the general public, and the robot should be able to cater to each individual where possible.

Some students spoke of being reassured by the perceived anonymity the robot gave to a conversation; they felt they would not be judged for asking their questions. Whilst this is obviously a positive if it encourages students to ask for help, it also raises a larger problem of how we would deal if a student approached the robot for help with bullying, discrimination or harassment. Mbawa [10] presented an approach where the interaction was scored; if the score was deemed low, they were presented with self help resources, if it was high they were directed to a suicide helpline. Care would need to be taken actually deploying this in the real-world unsupervised, as in this experiment they used pre-defined scenarios for the majority of participants.

6 Conclusions and future work

We have developed a social robot for use in a large, newly-built teaching building at a university and deployed it alongside an existing human support team to respond to building user's queries. We evaluated the robot via a week-long field experiment using a range of subjective and objective measures. The robot generally performed well, with positive initial feedback. On the long questionnaire, we found a range of positive attitudes towards the acceptance of the robot, along with a number of constructive suggestions for future system enhancements. This is shown by all objective measures exceeding the neutral threshold. However, there were a significant number of participants who had negative responses to the robot of various forms; in future studies, we will likely incorporate items from the Negative Attitude to Robots Scale (NARS) [11] to better quantify this reaction.

It is important to note that the system, as deployed, had several limitations. Firstly, the data provided by IS was out-of-date or incomplete in places. We note we did not perform any pre-processing or modification of the data other than in the internal building directions, nor use customised information retrieval techniques to search the available data. Also, during the deployment, the robot was never unattended; this was primarily so we could observe user interactions and step in where necessary, but also due to deploying in a completely public building and the risk of harm to the robot. This likely affected some people's responses to the robots; they would instinctively look to staff rather than the robot in the first instance.

The user feedback has been shared with IS and is now being used to assist in making future decisions about novel service delivery methods within the university. Future work would include developing sophisticated social signal processing to recognise a user's intent to speak with the robot, using large language models to improve the variety and relatability of the system's responses, addressing the speech recognition challenges faced in the deployment location, integrating solutions for accessibility, and —overall— taking into account the suggested user enhancements to improve service delivery further.

References

1. Blair, A., Foster, M.E.: Development of a university guidance and information robot. In: Proceedings of HRI 2023 (2023). https://doi.org/10.1145/3568294.3580138
2. Brščić, D., Kidokoro, H., Suehiro, Y., Kanda, T.: Escaping from children's abuse of social robots. In: Proceedings of HRI 2015 (2015). https://doi.org/10.1145/2696454.2696468
3. Curry, A.C., Rieser, V.: #MeToo Alexa: how conversational systems respond to sexual harassment. In: Proceedings of the 2nd ACL workshop on Ethics in Natural Language Processing (2018). https://doi.org/10.18653/v1/W18-0802
4. Curry, A.C., Rieser, V.: A crowd-based evaluation of abuse response strategies in conversational agents. In: Proceedings of SigDial 2019 (2019). https://doi.org/10.18653/v1/W19-5942
5. Foster, M.E., Stuart-Smith, J.: Social robotics meets sociolinguistics: Investigating accent bias and social context in HRI. In: Proceedings of HRI 2023 (2023). https://doi.org/10.1145/3568294.3580063
6. Foster, M.E., et al.: MuMMER: Socially intelligent human-robot interaction in public spaces. In: Proceedings of AI-HRI 2019 (2019), http://arxiv.org/abs/1909.06749
7. Friedman, N., Love, K., LC, R., Sabin, J.E., Hoffman, G., Ju, W.: What robots need from clothing. In: Proceedings of DIS 2021 (2021). https://doi.org/10.1145/3461778.3462045
8. Furhat Robotics: (2018), https://furhatrobotics.com/press-releases/franny-frankfurt-airports-new-multilingual-robot-concierge-can-help-you-in-over-35-languages/
9. Hone, K.S., Graham, R.: Towards a tool for the subjective assessment of speech system interfaces (SASSI). Natural Language Engineering **6**(3-4), 287–303 (2000). https://doi.org/10.1017/S1351324900002497
10. Mbawa, S.Z.: How can a conversational agent (chatbot) be used to detect and prevent suicide based on recognisable suicide behaviours amongst young people with mental disorders? Masters thesis, University of Applied Sciences, Utrecht (2021)
11. Nomura, T., Suzuki, T., Kanda, T., Kato, K.: Measurement of negative attitudes toward robots. Interaction Studies **7**(3), 437–454 (2006). https://doi.org/10.1075/is.7.3.14nom
12. OpenAI: ChatGPT, OpenAI (2022), https://openai.com/blog/chatgpt/
13. Rasa Inc: Introduction to Rasa Open Source (2022), https://rasa.com/docs/rasa/
14. Stock, R.M., Merkle, M.: Can humanoid service robots perform better than service employees? A comparison of innovative behavior cues. In: Proceedings of the 51st Hawaii international conference on system sciences (2018)
15. Völkel, S.T., Schödel, R., Buschek, D., Stachl, C., Winterhalter, V., Bühner, M., Hussmann, H.: Developing a personality model for speech-based conversational agents using the psycholexical approach. In: Proceedings of CHI 2020 (2020). https://doi.org/10.1145/3313831.3376210