Wizard of Oz: An Underestimated Method in the Field of Social Robotics

MEHRBOD MANAVI, University of Siegen, Germany
FELIX CARROS, University of Siegen, Germany and Fraunhofer Institute for Applied Information Technology, Germany
JULIA RENARDIAS, University of Siegen, Germany
SIMA AMIRKHANI, Bonn-Rhein-Sieg University of Applied Sciences, Germany
RAINER WIECHING, University of Siegen, Germany
VOLKER WULF, University of Siegen, Germany

In 2017, we designed a basic web dashboard to temporarily control Pepper's robot, which was in the process of implementing an Android remote controller app. Although the intention was to operate the robot's basic control, its design and function have led to its long-term use of this dashboard, even now. We quite underestimated the potential use of this dashboard. Often, it helps us bring the robot to life and make it functional, whereas the Choregraphe and smartphone remote app could not connect to it. Later, we often used the Wizard of Oz method to impress older adults in the care home and demonstrate the fun aspects of the robot to them. Nowadays, it is less frequent to use the dashboard and Wizard of Oz in general; still, it is a game-changer when new people see and interact with the robot. However, this raises some ethical considerations about manipulating the target groups to like the robot. Therefore, in this paper, we discuss the importance of ethical considerations in using the Wizard of Oz method and the design implications that lead to the long-term use of this dashboard.

CCS Concepts: • **Human-centered computing** → *Usability testing*; **Web-based interaction**; *Interface design prototyping*.

Additional Key Words and Phrases: Social Robot, Human Robot Interaction, HRI, Social Assistive Robot, SAR, Older Adults, User Experience, Usability Testing, Wizard of Oz

1 INTRODUCTION & RELATED WORK

The basic definition of Wizard of Oz (WoZ) technique involves connecting two systems so that the user can interact with a system that does not yet exist [15] or is not yet operational, while the experimenter, who observes the system, acts as a machine and executes the user's commands [10]. WoZ is one of the user testing approaches in Human-Computer Interaction, in which a human responds to the user's interaction with the system artifacts behind the scenes during the prototype test [15]. This approach is ideal for Human-Computer Interaction research, as it enables designers to study and evaluate the design before developing the complete prototype. This avoids finishing a fully developed prototype based on false user assumptions [9]. In addition, as Dow mentioned, the use of the WoZ has some limitations. It can be

Authors' addresses: Mehrbod Manavi, University of Siegen, Siegen, Germany, mehrbod.manavi@uni-siegen.de; Felix Carros, University of Siegen, Siegen, Germany, felix.carros@uni-siegen.de and Fraunhofer Institute for Applied Information Technology, Sankt Augustin, Germany, felix.carros@fit.fraunhofer.de; Julia Renardias, University of Siegen, Siegen, Germany, Julia.renardias@uni-siegen.de; Sima Amirkhani, Bonn-Rhein-Sieg University of Applied Sciences, Sankt Augustin, Germany, sima.amirkhani@h-brs.de; Rainer Wieching, University of Siegen, Siegen, Germany, rainer.wieching@uni-siegen.de; Volker Wulf, University of Siegen, Siegen, Germany, volker.wulf@uni-siegen.de.

This work is licensed under a Creative Commons Attribution 4.0 International License.

Mehrbod Manavi, Felix Carros, Julia Renardias, Sima Amirkhani, Rainer Wieching, Volker Wulf (2025). Wizard of Oz: An Underestimated Method in the Field of Social Robotics. In *Proceedings of the 12th International Conference on Communities & Technologies* (C&T 2025), July 21–23, Siegen, Germany. EUSSET. https://doi.org/10.48340/ct2025-1009

used by an operator whose role is to simulate certain system functions or intervene when the system is operational, but it may not be fully trusted.

Since 2017, we have actively used the Pepper robot in care homes to improve the healthy aging of older adults. The Pepper robot from Aldebaran has been widely used since 2015 [1] and remains one of the most popular social robots on the market. Despite the hold of production of this robot in 2020 [19], it is still in use in many research studies in different sectors [14], especially in care sectors with various stakeholders (see [6, 7, 16, 17]). However, as many studies have indicated [8, 21], its Natural Language Model system is not particularly effective. The lack of an updated Python library and a basic Linux operating system on the machine makes it challenging for developers to build a tremendous independent program that runs on the machine itself. One way to break through is by utilizing a server on the system, connecting the robot to the server via brokers, and then enhancing its capabilities; however, this needed to be tested and improved by the developers.

Here, the WoZ proved helpful in facilitating communication between the user and the robot. Notably, in interactions between the robot and children, we have observed that the robot struggles to capture the sound of small girls' voices. Even though they clearly pronounce the word, it is still impossible to understand them. Using third-party services, such as Google ASR and Microsoft Azure, could not solve this problem, causing frustration among small children, especially girls aged 6 and 8 [16]. The first design of the dashboard was inspired by a JavaScript sample on a Japanese forum [1], which motivated us to expand and build the web dashboard. The original project was basic and consisted mainly of Pepper's programming; it served only as a template for connecting Pepper via the NAOqi port. The idea of a dashboard hosting most of the useful functions required to control Pepper comes to mind. Some of these basic functions include Pepper movement, text-to-speech feature, autonomous life, showing a webpage, turning off, resetting, and other similar functions added to the JavaScript dashboard. This dashboard aimed to simulate the functionality of a remote app for the Pepper robot.

In this paper, we explored the influence of using the WoZ method on the development of a robotic control web-based dashboard. We especially focused on the design process, design considerations, and ethical implications of simulating autonomy in Human-Robot Interaction.

2 BACKGROUND

The basic definition of the WoZ is a system that, to the user interacting with it, appears to be fully autonomous, yet is actually operated by a human being [20]. This terminology originated from a children's novel called "The Wonderful Wizard of Oz" [3] written by Frank Baum in 1900. Then, in 1939, a movie with this name showed a wizard -Professor Marvel- controlling a machine behind a curtain [2], manipulating others to be a real wizard. An early example of operating a system artifact behind a curtain was in John F. Kelley's dissertation [12]. In his dissertation, he worked on natural language queries that could understand user input with an accuracy of up to 97%. This was achieved by experimenting in another room, where his team helped the computer understand the user's input in the software's early design. This continued until the computer could understand the user queries without help. Since then, the WoZ has become one of the standard methods used in Experimental Psychology, Human Factors, Ergonomics, and Usability Engineering, in which researchers or experimenters settle down in different rooms observing the behavior of system artifacts[13].

[1] Https://qiita.com/Atelier-Akihabara/items/716e88b306db12751c9b

In the field of Human-Robot Interaction, Bartneck et al. argue that robots can be either fully autonomous or controlled by the researcher (whether in the background or in front of users) [2]. These days, WoZ is mainly used to test and improve artificial intelligence or natural language programs. Numerous studies have benefited from the WoZ method, with each study employing it in a distinct manner. In some studies WoZ method used to relocate the robot to particular position in the room (eg., [6]), in some studied it has been used when the robot failed to maintain the conversation with the user (eg., [5, 6, 11], changing the state of emotion expression of the robot (eg., [5, 11])). Manavi et al. used WoZ to simulate the functionality of a remote app for the Pepper robot, allowing care professionals to control and program it [16]. They showed the app to the care professionals, asking them to interact with the app. When they pressed a button or used a feature on the robot, a researcher used WoZ to issue the command to the robot, which a care professional then executed on the app.

WoZ is a low-cost factor for testing a system artifact that is either a prototype or not fully functional. This method helps researchers and developers identify system interaction weaknesses and challenges during the interaction.

3 DESIGN

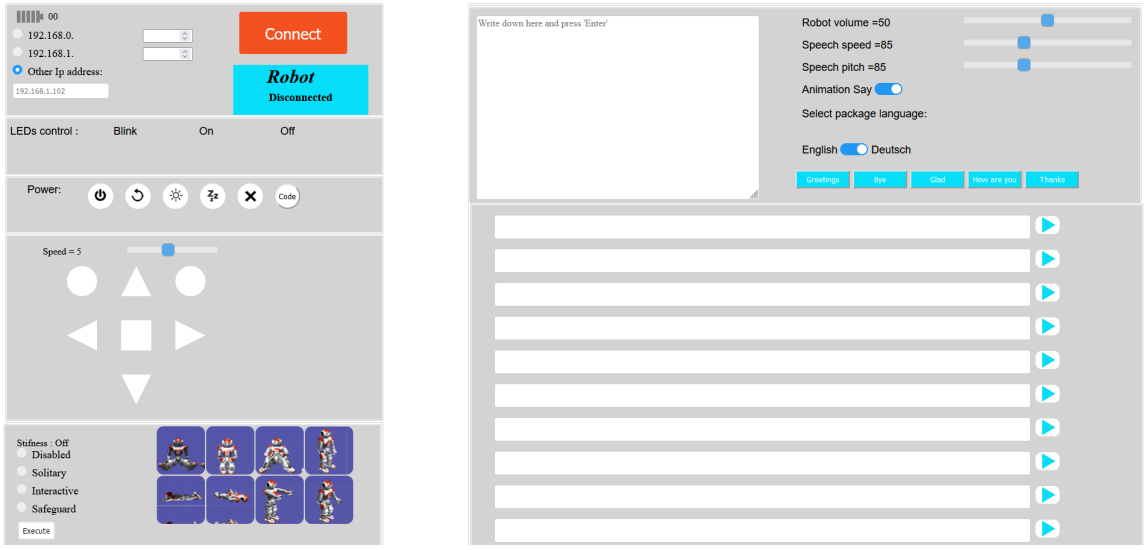

Fig. 1. The web dashboard

This web-based dashboard serves as a remote control for both Pepper and NAO robots. It utilizes JavaScript and the Qi messaging (v1) library by Aldebaran to communicate with the robot through NAOqi (robot operating system) [1]. We made several versions of the web dashboard over the years. Some extra features were added (changing the level of stiffness, power section, navigation, and different types of Speech-to-Text), and some features were removed (changing the colors of the eyes LEDs). Some features of the robots, such as the camera and microphone, were not implemented in the dashboard's design. Hence, there was no intention to collect personal data through the dashboard or monitor the user; instead, we used it as a mechanism to facilitate a smooth dialog flow. This dashboard is designed to examine the simulated autonomy of the robot and test the user interfaces of early applications for the robot. Therefore, it has never been used to spy on users. It does not record any activity that the researcher performs via the dashboard or on

the robot, and it does not connect to the robot's camera or microphone. The reason was its nature as an assistant for the researcher to examine the robot or remote application under development. The web dashboard has the following features (see Figure 1):

- **Connection section** to connect to the robot via its IP address
- **Led section** to turn on, off, and blink the eyes of the robot
- **Power section** to Power off, restart, sleep, and wake up the robot, cancel all events, and launch the special program.
- **Navigation section** to move the robot around at a certain speed
- **Stiffness section**: Pepper and NAO have four stage: 1) Disable which put the robot on sleep mode; 2) Solitary which is somewhere between interactive and safeguard; 3) Interactive which has active voice interaction and robot random movement; 4) Safeguard which is the robot is alive, but its on freeze mode.
- **Text-to-Speech (TTS) section** for quick voice interaction between user and robot. It can also change the robot's language, activate and deactivate the robot's body movement when it's talking, adjust the speed and tone of TTS, and utilize shortcuts to say common phrases.
- **Pre-written TTS** for the robot to say about 20 pre-written phrases.

The stiffness of the robot, which became the most essential part of the dashboard, was due to its ability to make the robot alive in conditions where Choregraphe and the smartphone app could not be connected, or the robot became unresponsive. This unresponsiveness often occurred after the robot was transported to new places. Then the robot started reporting errors and requested a reboot. It usually did not come alive, and we could not connect via smartphone or the Choregraphe app. Then we used the dashboard and experimented with the stiffness section, and the robot came alive. This made the web dashboard an essential part of our research over the years. Additionally, we provided students with the opportunity to expand and enhance their design and functionality. Still, it remained the same, as we only focused on its functionality rather than design, and used it as a primary tool. It is always used as a backbone to ensure the robot's workability and is left aside when everything is running smoothly.

The primary difference between the TTS and pre-written sections lies in the ease of use of the TTS. In the TTS section, the user can adjust the volume, speech pitch, and voice characteristics of the robot, as well as the language installed on it. In some cases, when our target group was so close to the robot, the animation on the robot could potentially harm the users. Additionally, object detection on the robot is quite accurate; however, it did not work as expected on some occasions, causing it to hit the human even slightly. Therefore, an animation switch was added to prevent the robot's movement during the conversation. Some buttons were designed to quickly say phrases such as greeting, goodbye, thanks, etc. The Text field in this section was designed so that the researcher could write the text, and then, after pressing the button, the robot would read the text aloud, clearing the text field for subsequent use. The pre-written section is designed for scenario-based conversation. For example, the user can write 20 examples and click the play button before each text field to have the robot read them aloud. This especially helped us during the children's robot interaction, where the robot did not understand the children, we could type that they did a great job, and convince them to continue their practice.

The programming button had one purpose to launch a high-level command in the script. Its purpose was to test a specific action for developing the Android application. Nevertheless, the dashboard's main purpose was to provide low-level control of NAO and Pepper robots. This simplicity causes the dashboard to execute the command almost in real-time with close to zero latency.

4 ETHICAL CONSIDERATIONS

Is it right to make users believe robots are autonomous when they are not? In the movie "Wizard of Oz", when Dorothy realized there is a man behind the curtain who controls the machine, but not a real wizard, she felt betrayed and said to the Wizard *"Oh, you are very bad man!"* In response, the wizard (Professor Marvel) says *"oh no my dear, I am a very very good man, but I am just a very bad wizard!"* A lack of transparency and honesty can negatively impact the project and users dealing with system artifacts.

Researchers using the WoZ method should be transparent about when and how it is used. Once the truth is revealed, trust can be severely damaged in the system or the person operating it. In social robotics, the primary goal is to create meaningful interactions between the user and the robot. When WoZ is used, this interaction can be perceived as fake, potentially breaching user trust.

It is therefore important to clearly explain to users and, in the case of care homes to other stakeholders such as nurses, social services, and care home management, which parts of the interaction are driven by WoZ and which are already embedded in the robot's algorithms (i.e., "robot-native"). Equally important is transparency regarding the data recorded during these interactions—what is captured, how it is stored, and who has access to it. Social robots are intended for social interaction, which means users may experience private moments with them. These moments should not be shared unknowingly or livestreamed. Users must be informed whether the wizard sees a live video, a transcript, or nothing at all—for instance, if the wizard is simply in the same room and only observes directly.

How the robot is guided also matters. Whether the wizard is visible or not can affect users' understanding, although Nasir et al. [18] and Bradley et al. [4] saw no significant change in user behavior if WoZ was used or not. In our work, we preferred to be in the same room as the user when acting as wizards. This allowed users and nearby staff to see what we were doing on the computer and how that influenced the robot's behavior. However, it is also technically possible for the wizard to control the robot from another room, remaining invisible to users. While this hidden approach might enable more seamless interactions, it lacks transparency and prevents users from understanding what the wizard does and what the robot's true capabilities are. It may also create false expectations and simulate functionality that does not actually exist.

All of this raises the broader question: when is it acceptable to simulate robot behavior using the WoZ method? The answer is complex, involving many dimensions such as culture, context, personality, and trust. We argue that it depends on the specific context, the relationship between users and wizards, and the functions that the wizard performs.

Regarding context, there is a significant difference between using WoZ in a research project and in a market-ready product. The setting also matters; a living lab or research facility sets expectations that may justify the use of WoZ. Another key factor is the relationship with users. There should be mutual trust and awareness. Users should know who the wizard is and what their role is. We have long-standing relationships with the care homes in which we work and have always been transparent about our methods. Using WoZ in a familiar, trusted environment feels far less deceptive than in a new setting.

Finally, the acceptability of WoZ also depends on the scope of the wizard's role. If the wizard controls the entire interaction, the implications are different than if they are only stepping in occasionally. For example, to answer an unexpected question or to initiate a new action after the previous one ends. Partial wizarding to bridge technical limitations is likely to be more acceptable, especially when transparently communicated.

5 REFLECTION AND FUTURE WORK

We developed the web-based dashboard to navigate and control the basic functions of the Pepper robot. Over time, we added more functions to this platform to test the usability of the remote app we were developing for the care professionals. The ease of maintenance and troubleshooting has led to the researcher's long-term use of this platform since then. On some occasions, when we used the platform to prolong the conversation between the user and the robot, it often caused people who interacted with the robot to continue talking to it and even like it. This, however, shows the drawback of overstimulating the power of robots. Although understanding what must be designed for a dialog system is helpful, ethical considerations for implementing such a system are crucial.

The web dashboard is quite handy in h Human-Robot Interaction, although some ethical concerns are related to the WoZ method. In most cases, we declare that we used the WoZ method; in some cases, we needed to use it spontaneously to make the conversation smooth, especially for small girls who interacted with the robot. The Pepper robot sometimes failed to react to the voices of small children (mostly the voices of small girls) due to the low frequency of their voices. This always causes frustration in small children, and regarding gender, it can also be felt as discrimination. Even though it is not intentional, in a group setting, it can cause a reduction in the confidence of children or even older adults when robots do not understand them. Here, we often used WoZ to interpret what the user had said.

6 CONCLUSION

The Web Dashboard, which was intended to be a supporting platform for a larger system to control and program a humanoid robot, evolved into a crucial platform for testing various projects, both with and without a dialog system. It occurred using the WoZ method, in which we tested different dialog systems and applications powered by this dashboard. Surprisingly, the WoZ method became a permanent element in our research. Especially when it comes to a vulnerable target group with either hearing impairments or low voice frequency, the dashboard, combined with WoZ, served as a bridge between the project's smoothness and success in the early design stage. One important key is the ethical consideration. We highly recommend that the other researcher be transparent with the users, inform them about the use of such a method, and obtain their agreement to participate in it during the experiment.

REFERENCES

[1] R. Aldebaran. 2023. Aldebaran. *n.d. Aldebaran* 6 (2023). https://www.aldebaran.com/en

[2] Christoph Bartneck, Tony Belpaeme, Friederike Eyssel, Takayuki Kanda, Merel Keijsers, and Selma Šabanović. 2024. *Human-robot interaction: An introduction.* Cambridge University Press.

[3] Frank Baum. 1900. The Wonderful Wizard of Oz. *New York: Sterling Publishing* (1900).

[4] Jay Bradley, David Benyon, Oli Mival, and Nick Webb. 2010. Wizard of Oz experiments and companion dialogues. In *Proceedings of HCI 2010.* BCS Learning & Development.

[5] Felix Carros, Berenike Bürvenich, Ryan Browne, Yoshio Matsumoto, Gabriele Trovato, Mehrbod Manavi, Keiko Homma, Toshimi Ogawa, Rainer Wieching, and Volker Wulf. 2022. Not that Uncanny After All? An Ethnographic Study on Android Robots Perception of Older Adults in Germany and Japan. In *Social Robotics: 14th International Conference, ICSR 2022, Florence, Italy, December 13–16, 2022, Proceedings, Part II* (Florence, Italy). Springer-Verlag, Berlin, Heidelberg, 574–586. https://doi.org/10.1007/978-3-031-24670-8_51

[6] Felix Carros, Johanna Meurer, Diana Löffler, David Unbehaun, Sarah Matthies, Inga Koch, Rainer Wieching, Dave Randall, Marc Hassenzahl, and Volker Wulf. 2020. Exploring human-robot interaction with the elderly: results from a ten-week case study in a care home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* 1–12.

[7] Felix Carros, Tobias Störzinger, Anne Wierling, Adrian Preussner, and Peter Tolmie. 2022. Ethical, Legal & Participatory Concerns in the Development of Human-Robot Interaction. *i-com* 21, 2 (2022), 299–309. https://doi.org/10.1515/icom-2022-0025

[8] Xiaohui Chen, Katherine Luo, Trevor Gee, and Mahla Nejati. 2024. Does chatgpt and whisper make humanoid robots more relatable? *arXiv preprint arXiv:2402.07095* (2024).

[9] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing* 4, 4 (2005), 18–26. https://doi.org/10.1109/MPRV.2005.93

[10] Paul Green and Lisa Wei-Haas. 1985. The rapid development of user interfaces: Experience with the Wizard of Oz method. In *Proceedings of the human factors society annual meeting*, Vol. 29. SAGE Publications Sage CA: Los Angeles, CA, 470–474. https://doi.org/10.1177/154193128502900515

[11] Nora Hille, Berenike Bürvenich, Felix Carros, Mehrbod Manavi, Rainer Wieching, Yoshio Matsumoto, and Volker Wulf. 2023. The Effects of Android Robots Displaying Emotion on Humans: Interactions between Older Adults and Android Robots. arXiv:2303.13624 [cs.HC] https://arxiv.org/abs/2303.13624

[12] Jeff Kelley. 2025. *Jeff Kelley's Personal Web Page.* https://www.musicman.net/kelley.html Accessed: 08. May 2025.

[13] John F. ("Jeff") Kelley. 2018. Wizard of Oz (WoZ): a yellow brick journey. *J. Usability Studies* 13, 3 (May 2018), 119–124.

[14] Andre W Kushniruk, Hrishikesh S Nair, and Elizabeth M Borycki. 2025. The Pepper Robot in Healthcare: A Scoping Review. *Envisioning the Future of Health Informatics and Digital Health* (2025), 101–105.

[15] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 10 - Usability testing. In *Research Methods in Human Computer Interaction (Second Edition)* (second edition ed.). Morgan Kaufmann, Boston, 263–298. https://doi.org/10.1016/B978-0-12-805390-4.00010-8

[16] Mehrbod Manavi, Felix Carros, Isabel Schwaninger, Richard Paluch, Dave Randall, Maximilian Krueger, Ilayda Tufan, Rainer Wieching, Peter Tolmie, Claudia Mueller, et al. 2025. "Making Robots Usable": Empowering Care Professionals to Customize Social Robots in Care. *International Journal of Social Robotics* (2025), 1–22. https://doi.org/10.1007/s12369-025-01238-1

[17] Mehrbod Manavi, Felix Carros, David Unbehaun, Clemens Eisenmann, Lena Müller, Rainer Wieching, and Volker Wulf. 2025. From idle to interaction – assessing social dynamics and unanticipated conversations between social robots and residents with mild cognitive impairment in a nursing home. *i-com* 24, 1 (2025), 227–242. https://doi.org/10.1515/icom-2024-0046

[18] Jauwairia Nasir, Pierre Oppliger, Barbara Bruno, and Pierre Dillenbourg. 2022. Questioning wizard of oz: effects of revealing the wizard behind the robot. In *2022 31st IEEE international conference on robot and human interactive communication (RO-MAN)*. IEEE, 1385–1392.

[19] Sam Nussey. 2021. EXCLUSIVE SoftBank shrinks robotics business, stops Pepper production- sources. Retrieved 2024.10.07 from https://www.reuters.com/technology/exclusive-softbank-shrinks-robotics-business-stops-pepper-production-sources-2021-06-28/

[20] Sara Paul and Maria Rosala. 2024. *The Wizard of Oz Method in UX.* https://www.nngroup.com/articles/wizard-of-oz Accessed: 18. May 2025.

[21] Siddhartha Shibi and Sohail Zaidi. 2024. STEM Approach to Enhance Robot-Human Interaction Through AI Large Language Models and Reinforcement Learning. In *2024 IEEE Integrated STEM Education Conference (ISEC)*. IEEE, 1–2.