Has Wizard of Oz Testing Passed Its Use By Date?

ANONYMOUS AUTHOR(S)*

ACM Reference Format:
Anonymous Author(s). 2023. Has Wizard of Oz Testing Passed Its Use By Date?. 1, 1 (September 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Artificial intelligence (AI) is creeping into our lives from the smart-home speaker to the personalization and recommendation systems infused in the music services we use everyday. Indeed it is difficult to think of a single user facing product whose experience may be enhanced through the incorporation of AI. At the same time, advancements in the underlying technology have enabled many to easily build and train sophisticated models once reserved to experts in the field[1]. For instance, some AI frameworks allow designers to simply provide training data with the models being generated with a couple of clicks[2].

Despite these advancements in AI technology a key question surrounding the use of AI in consumer facing applications remains: will the envisioned AI create an experience individuals or even teams would actually want to use? Answering such a question late in the development process of an AI model can be costly both in time and development resources despite the foregoing advancements. For instance, it could be quite costly to collect, clean and annotate training data to build a model designers believe will create a favourable experience. What is clearly necessary is some method which allows designers to evaluate the user experience of their envisioned AI application that is both low cost and allows them to rapidly iterate on the design.

One popular method for evaluating such envisioned AI experiences for several decades has been the Wizard of Oz (WoZ) methodology first discussed by Kelley[2]. The WoZ methodology calls for a user test, whereby participants are told they will be interacting with some type of AI agent or product that uses AI in some way. However, in reality the behaviour of the AI is directly controlled by an experimenter who can observe the user's interaction. This methodology has been used to evaluate a number of products including conversational interfaces[6], robotics[3], and even autonomous vehicles[1, 9]. So popular is this methodology that there are even literature reviews of WoZ studies solely in Human-Robot Interaction[5].

Despite its prevalence, one question to the knowledge of the author that has not been discussed by the community is whether WoZ as an evaluation method remains relevant today given the numerous changes in both technology, applications and the people using such technology since WoZ was first proposed in 1984. To explore this question further, I will examine some of the benefits and challenges associated with using the WoZ methodology as presented by Kelley[2]. I will then

[1]One popular example being Tensorflow: https://www.tensorflow.org
[2]Examples of which include Google Vertex, Amazon SageMaker and Microsoft's Azure Machine Learning platforms

XXXX-XXXX/2023/9-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

conclude that WoZ is still a valid evaluation method, however key changes need to be made for it to be relevant today.

2 STRENGTHS OF WOZ EVALUATION

The Wizard of Oz methodology as a tool for evaluating envisioned AI powered experiences provides a number of key benefits over other evaluation methodologies. First, the WoZ methodology does not require any sophisticated AI model to be trained or complete end-to-end system to be built. Instead designers can simply "mock up" the experience using a low cost design tool (e.g., Figma, Powerpoint) and some mechanism for the Wizard to control the experience in a manner similar to how an AI agent would control it. Indeed, there have been a number of toolkits discussed in the literature that allow individuals with limited knowledge about the WoZ methodology to put together a WoZ experiment with minimal effort [4, 7]. In addition, since experiences can be created so rapidly, it allows designers and engineers to experiment across the design space of potential AI powered experiences.

Second, the WoZ evaluation methodology allows individual AI experiences to be completely isolated from each other, which would otherwise be difficult to perform in conventional software development where typically components must be built together for any meaningful experience to be evaluated. Consequently, designers can focus on specific scenarios they want to test. Relatedly, since technology that is known to cause latency and technical issues (e.g., speech recognition in the case of a conversational agent) are not present in the WoZ study, designers can focus on evaluating whether the envisioned scenarios result in a positive user experience without the underlying technology potentially confounding their results.

Lastly, an added benefit of a well designed WoZ experiment is the capability to extensively log all interaction between the user and the agent. Such data could prove invaluable in training early models of a given AI experience, particularly when it would be otherwise difficult to collect. For instance, an organization developing a new product would likely have no user collected data to train their models. In addition, since the AI is directly controlled by a human experimenter, data scientist get access to human decisions which may be seen as a gold standard when training many AI models.

3 DRAWBACKS OF WOZ EVALUATION

In spite of its popularity and the benefits highlighted in the preceding section, there are several key drawbacks to the WoZ methodology as it is employed today. Specifically, there are concerns about how data is collected using the WoZ methodology, how the confederate tasked with the Wizard behaves during the study and lastly concerns about the participants who play such a pivotal role in these studies.

A key benefit of running a WoZ study is the collection of rich interaction data between the user and Wizard which could form the basis of training data used to train an AI model. However, what is unclear is whether the data collected during a WoZ study is sufficient to train a model with a desirable level of accuracy. Indeed, it is not uncommon for modern deep reinforcement learning models to require potentially gigabytes or even terabytes of data to train a model with a reasonable degree of accuracy. Relatedly, some models may require specific types of data that may not be easily collectable during a WoZ study. For instance, it seems unlikely that using a driving simulator for a WoZ study of an envisioned autonomous vehicle would provide enough varied types of driving patterns nor provide the rich sensor data a modern computer vision based driving system would need. Another key concern when developing some AI powered experiences is the collection of sufficient longitudinal data to train such models to adapt to changing user behaviour and preferences. In particular, many recommendation systems used in a variety of applications

requires such data to effectively predict how such trends change over time to deliver the most relevant recommendations.

Another key concern of contemporary WoZ studies is the role of the Wizard, who tries to mimic the behaviour of the envisioned AI model/agent. Hence, the Wizard is the cornerstone in most WoZ studies and consequently can be a potential source for error. In particular, for some AI experiences being evaluated is that the agent's interaction with the user is indirect. Consider a recommendation system that suggest products to a user. The input to such a system would likely be what products the user has purchased in the past or searched for but may also consist of a number of user behavioural factors such as the amount of time they spent on a given product page. The output from the recommendation system would be a list of products the system thinks is relevant for the user. Hence, it is difficult for the Wizard to measure and determine what to look for with the user interacting with the system to ultimately generate its product recommendations. Contrast that with a conversational agent where the user is directly interacting with the agent through spoken words and the agent has a finite set of reasonable responses. Another challenge surrounding the Wizard's behaviour is their potential lack of consistency in how they perform, particularly when multiple Wizards are employed during a study or between experiment sessions which may impact the reliability of the data collected. For instance, different Wizards may handle the same interaction from a given user differently. Relatedly, it is unclear whether Wizard taking on the role of an AI agent can effectively mimic the behaviour of the system particularly given it is difficult to predict how such a system would behave without first training it. As an exercise for the reader, consider the potential failure modes for a generative AI model such as Chat-GPT. A final challenge surrounding the behaviour of Wizards in WoZ studies relates to their use in collaborative tasks. Specifically, as AI technology improves, it is becoming increasingly likely that AI will not just interact with one user but potentially several users. Despite this, the WoZ methodology makes a critical assumption that the interaction is solely between a single user and the agent.

One critical assumption made by the WoZ methodology is that participants believe they are interacting with an AI agent even though in reality that agent is controlled by an experiment. This suspension of disbelief is critical, since participants could behave differently if they think the agent is actually an experimenter. However, how can we be certain as practitioners of the WoZ technique that this is actually occurring? Without any certainty regarding whether the participant truly believes the agent they were interacting with was AI, the results of such studies can be questionable at best.

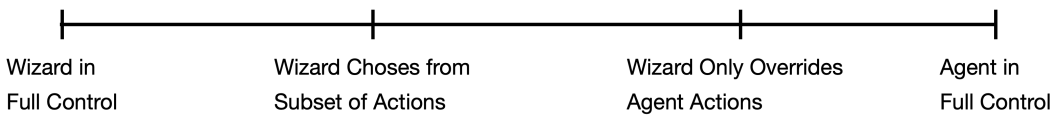

Fig. 1. Spectrum of Agency the Human Wizard has on AI

4 A RETHINK OF WOZ?

Given the preceding discussion, the reader might reasonably believe that the WoZ methodology is well-past its use-by date given the contemporary challenges raised. However, all hope is not lost. In particular, there are a number of solutions to these challenges worthy of consideration.

In terms of concerns surrounding the quantity and quality of data being collected during WoZ studies, considering where, when, and how WoZ is applied to a given problem can go a long way towards addressing these very issues. First and foremost, data scientists and other such relevant stakeholders need to consulted to determine what data needs to be collected and how such data

collection can take place during the study. Second, WoZ studies are usually performed at beginning of AI product lifecycle with Wizard in control of every action the agent takes. However, adopting an iterative approach where WoZ is applied throughout the product lifecycle from before an envisioned product is developed through to release would allow the data collected to adapt to the needs of the organization. For instance at the start of a product lifecycle the key question may be which AI features are relevant for the user. However, as a preliminary AI model is trained a key question might be whether the predictions made by the agent concur with how a human might perform the task. Consequently, the WoZ methodology may need to be adopted so the amount of the agency the Wizard has over the agent changes as the model gets developed (Figure 1) [8]. Whereas a Wizard might have full control of the agent at the beginning of the product lifecycle, their agency might be limited to overriding the occasional errant action from the trained model just prior to the product being shipped. Third, the apparatus used to conduct the WoZ study needs to be instrumented to collect interaction data during the study. Testing such apparatus through a pilot study to ensure that experiment data is indeed being correctly collected is also crucial. Fourth, the collection of sufficient amounts of data, especially longitudinal data could be achieved by scheduling follow up WoZ studies with the same participants (perhaps further down the product lifecycle), using crowdsourcing platforms such as Amazon Mechanical Turk or Prolific to recruit many participants at once. Relatedly, data collected in another product, an external data source or in an earlier study could potentially be used to effectively bootstrap a WoZ project to a desired level of simulation. Irrespective of which approaches are employed to deal with the issue of data quality and quantity, one key consideration are potential legal and ethical concerns which is why relevant stakeholders (e.g., legal department, institutional review board) should be considered. There are a number of key ways a Wizard's behaviour can be adapted to address the concerns facing the evaluation of modern AI applications. In particular, to address the lack of consistency of performance for different Wizards performing the role as well as between experiment sessions, training can play a key role. Specifically, Wizards should be provided with a guide along with accompanying resources (e.g., training videos, audio of conversational agents in action) that explains their role in the experiment and how they should behave. In the case of WoZ studies where the Wizard and participant aren't directly interacting with each other, it should be made clear to the Wizard what they are allowed to observe (inputs) and what set of actions they are allowed to take (outputs). Developers, data scientists, designers and other relevant stakeholders need to be involved in determining which inputs and outputs are reasonable for the Wizard. In addition, Wizards need to given adequate opportunity to practice their role by way of several pilot sessions before any experiment session begins. Extending the WoZ methodology to support environments where the agent must collaborate with multiple users in a given experiment session is a topic that has received some attention recently in the literature. For instance Simpson et al.[6] proposes a collaborative WoZ environment with the Wizard selecting responses from a user interface through keyboard shortcuts. However, such an approach may not be appropriate in all scenarios. Hence, we will likely need guidelines on how to design these interfaces to be effective for Wizard who now must observe and interact with multiple participants. Whether participants in a WoZ genuinely believe they are interacting with an autonomous AI agent (as opposed to the human controlling it) can be measured in a couple of ways. Firstly, participants could be provided a survey directly after interacting with the Wizard where they are simply asked to score whether the agent was entirely autonomous versus controlled by a human. Another potential solution is simply telling participants that in one trial block of the study they will be interacting with a human, which is played by the confederate and for the other trial block they will be interacting with an autonomous AI agent which in reality is controlled by the confederate (WoZ) without telling them in what order each agent appears. After each trial block participants are asked to score the degree of autonomy of the agent they were

interacting with. Using such an approach provides researchers with a human baseline they can use to compare whether the Wizard condition did indeed suspend the disbelief of the user. However, whilst such an approach may be suitable for certain AI experiences (i.e. conversational agent) it may not be suitable for all AI experiences being evaluated. For instance, it might be difficult for the participant to observe whether an agent is playing the role of a human confederate in a product recommendation system.

5 CONCLUSION

It seems increasingly likely that artificial intelligence will become ubiquitous in our everyday lives as it makes it weaves its way into the products and services we use. However, for such technology to be a success it needs to behave in a manner that users will want. At the same time, incorporating AI into products represents a significant risk for such organizations given its black-box nature making it difficult to effectively determine what resources are required to effectively train the models. Hence, the evaluation of such models through methodologies like Wizard of Oz evaluation provide a potential way for organizations to mitigate this risk and deliver experiences users will want to have. However, rethinking key aspects of the WoZ methodology such as actively considering what data needs to be collected, how Wizards should behave in such experiments, and whether participants have indeed suspended disbelief during the study are crucial for this method to deliver on its promise of providing inexpensive, rapid and reliable testing of envisioned AI experiences.

REFERENCES

[1] Isabel, Beggiato Matthias, Halama Josephine, Krems Josef F Hensch Ann-Christin, and Neumann. 2020. How Should Automated Vehicles Communicate? – Effects of a Light-Based Communication Approach in a Wizard-of-Oz Study, Neville Stanton (Ed.). *Advances in Human Factors of Transportation*, 79–91.

[2] J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2 (1 1984), 26–41. Issue 1. https://doi.org/10.1145/357417.357420

[3] Edith Law, Vicky Cai, Qi Feng Liu, Sajin Sasy, Joslin Goh, Alex Blidaru, and Dana Kulić. 2017. A Wizard-of-Oz study of curiosity in human-robot interaction. *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 607–614. https://doi.org/10.1109/ROMAN.2017.8172365

[4] Martin Porcheron, Joel E. Fischer, and Michel Valstar. 2020. NottReal: A Tool for Voice-based Wizard of Oz studies. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3405755.3406168

[5] Laurel D Riek. 2012. Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *J. Hum.-Robot Interact.* 1 (7 2012), 119–136. Issue 1. https://doi.org/10.5898/JHRI.1.1.Riek

[6] James Simpson, Hamish Stening, Patrick Nalepka, Mark Dras, Erik D Reichle, Simon Hosking, Christopher J Best, Deborah Richards, and Michael J Richardson. 2022. DesertWoZ: A Wizard of Oz Environment to Support the Design of Collaborative Conversational Agents. *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, 188–192. https://doi.org/10.1145/3500868.3559711

[7] Anoop K. Sinha, Scott R. Klemmer, Jack Chen, James A. Landay, and Cindy Chen. 2001. SUEDE: Iterative, informal prototyping for speech interfaces. *Conference on Human Factors in Computing Systems - Proceedings* (2001), 203–204. https://doi.org/10.1145/634067.634189

[8] Aaron Steinfeld, Odest Chadwicke Jenkins, and Brian Scassellati. 2009. The Oz of Wizard: Simulating the Human for Interaction Research. *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, 101–108. https://doi.org/10.1145/1514095.1514115

[9] Peter Wang, Srinath Sibi, Brian Mok, and Wendy Ju. 2017. Marionette: Enabling On-Road Wizard-of-Oz Autonomous Driving Studies. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 234–243. https://doi.org/10.1145/2909824.3020256