Wizard of Errors: Introducing and Evaluating Machine Learning Errors in Wizard of Oz Studies

Anniek Jansen
a.jansen1@student.tue.nl
Department of Industrial Design, Eindhoven University of Technology
Eindhoven, The Netherlands

Sara Colombo
s.colombo@tue.nl
Department of Industrial Design, Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

When designing Machine Learning (ML) enabled solutions, designers often need to simulate ML behavior through the Wizard of Oz (WoZ) approach to test the user experience before the ML model is available. Although reproducing ML errors is essential for having a good representation, they are rarely considered. We introduce Wizard of Errors (WoE), a tool for conducting WoZ studies on ML-enabled solutions that allows simulating ML errors during user experience assessment. We explored how this system can be used to simulate the behavior of a computer vision model. We tested WoE with design students to determine the importance of considering ML errors in design, the relevance of using descriptive error types instead of confusion matrix, and the suitability of manual error control in WoZ studies. Our work identifies several challenges, which prevent realistic error representation by designers in such studies. We discuss the implications of these findings for design.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Interactive systems and tools*; *Systems and tools for interaction design.*

KEYWORDS

Wizard of Oz, Machine Learning, Machine Learning Errors, User Experience Design, User Experience Analysis, Interaction Design, Computer Vision, Prototyping Methods

ACM Reference Format:
Anniek Jansen and Sara Colombo. 2022. Wizard of Errors: Introducing and Evaluating Machine Learning Errors in Wizard of Oz Studies. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '22 Extended Abstracts), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3491101.3519684

1 INTRODUCTION

Machine Learning (ML) is becoming an increasingly important asset for designers as it provides, among others, new interaction possibilities (e.g. [12, 13]) and new personalization techniques (e.g. [8, 25, 29]). ML provides the functional backbone of applications such as voice assistants, recommender systems, and face recognition. It can be embedded into consumer products in various ways, from narrow backend functionalities such as email spam filtering, to complex autonomous systems like robots or self-driving cars.

Despite its potential, research has highlighted that designers find it challenging to use ML as a design material and to prototype with ML [6, 27, 28], especially in the early phases of a design process. Training a custom ML model is a resource-intensive process and does not fit with the 'fail fast, fail often' approach typical of the design process early stages [6]. To overcome the difficulty of embedding working ML models in early prototypes to assess ML-enabled solutions, the Wizard of Oz (WoZ) method can be used [15]. In this method, the wizard - a designer behind the screen who mimics the technology, can act like an ML model. The goal is to provide the user with the impression of interacting with a working system, and to test their reactions to the experience it generates [5]. This method is often used in the fields of design and HCI to prototype non-existing (or not yet realized) technologies and therefore represents a valid alternative to simulate the behavior of ML even when a model has not been trained yet. WoZ has already been applied to evaluate ML models in previous studies (e.g. [7, 16, 24]). However, both these studies and additional research [5, 23] show that simulating ML models behaviors in WoZ is no easy task. Among the challenges identified is the difficulty to realistically reproduce ML errors, because they are unlike human errors [3, 26]. For instance, Riek [19] shows that only 3.7% of the WoZ studies in the Human-Robot Interaction domain include deliberate errors. ML errors are an intrinsic feature of ML models, and omitting them in a WoZ can lead to findings that are not representative of the user experience that is being simulated [3].

Advantages of considering ML errors early on. Being able to successfully include ML errors in the early evaluation phases of a design process would allow designers to assess users' reactions to different performance levels and errors of an ML model. This would have two main advantages. Firstly, it would allow designers to test the overall acceptability of a solution well in advance. Although the accuracy of an ML model is unknown before a model is trained, it can greatly influence the user experience (UX). For instance, in some applications like movie recommender systems, an 80% accuracy score may be acceptable, while in others, such as in diagnostics systems based on image recognition (e.g. to detect skin conditions [14]) the same accuracy might be deemed unacceptable. Testing the impact of different accuracy levels on the UX early on would prevent the development of systems that are unsuccessful from a UX viewpoint.

CHI '22 Extended Abstracts, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9156-6/22/04.
https://doi.org/10.1145/3491101.3519684

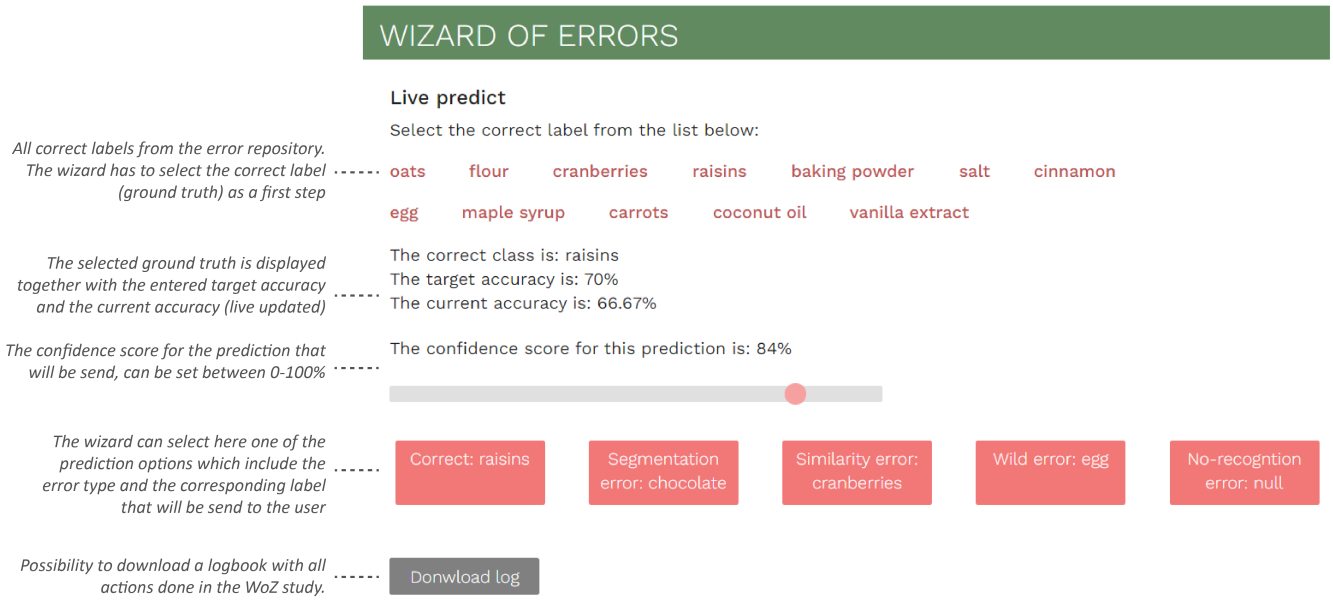


Figure 1: Prediction interface with explanations at the left-hand side

Secondly, testing different ML errors in advance would allow designers to create user interfaces that respond or adapt to ML errors, making the overall UX more pleasant or understandable. Previous studies have shown that ML errors can greatly impact the UX [1, 10, 11], the mental models users form to interact with the system, and their trust in the ML model [17]. Different error types could differently influence the UX, e.g. categorizing a Golden Retriever as a Labrador might be perceived as an understandable ML error by users, while mistaking a dog for a bird might cause more frustration and distrust in the ML-based system. Considering different error types and testing their effects on the UX early on would inform design decisions to ensure that the system can 'fail gracefully' [18], for instance giving users the possibility to report an error, or providing information on the types of errors that might occur, and their causes.

In this paper, we introduce Wizard of Errors (WoE), a web interface that facilitates the inclusion of ML errors in WoZ studies. This interface can be used in WoZ studies where the wizard can recognize the ground truth, i.e. the correct outcome of the ML prediction, from the input data. This happens, for instance, in object recognition, where the designer is able to recognize an object correctly, the same way an ML model is expected to do. We tested the WoE prototype with 10 design students to investigate if designers can successfully mirror ML errors during a WoZ study, in order to have a more representative simulation of the UX in the early phases of a design process. This study aims to contribute to the current discussion on designing with ML, by showing how ML errors can be included and tested in WoZ studies, and by uncovering designers' challenges and misconceptions in mirroring an ML model behavior.

2 MACHINE LEARNING FAILURES AND ERRORS

ML errors can occur in many forms and can have different origins. In this paper, we focus on errors that may occur during the user's interaction with ML-enabled solutions due to an incorrect ML prediction. During the remainder of this paper, the term *ML error* will be used to describe such types of errors.

A standard method to report ML errors is a confusion matrix. Here, errors are described as False Positives (FPs) and False Negatives (FNs). However, the confusion matrix, especially for a multi-class classification model (where the prediction is not a binary yes/no, but is based on three or more labels) [22] is hard to interpret for ML non-experts since the terminology is confusing and the reading direction and structure are unintuitive [2, 20]. To overcome these issues and make errors more understandable to designers, we adopted the terminology introduced by Bott and Laviola [4] where they differentiate between four types of errors. The definition of the errors is tailored to their use case - classifying handwritten mathematical equations, but it can be generalized into the following definitions:

- *Segmentation error*: incorrect segmentation of the data that results in an incorrect prediction (e.g. recognizing a face in a photo in the background of a photo instead of the person in the foreground)
- *Similarity error*: incorrect prediction that is somewhat related to the correct answer (e.g. recognizing sugar in a photo as salt)
- *Wild errors*: incorrect prediction that appears to have no relationship with the correct answer (e.g. recognizing sugar in a photo as a carrot)

- *No-recognition error*: failing to give a prediction (e.g. a face recognition system not responding when a face appears in front of the camera)

These descriptive error types can be used to create an error repository for an ML model, which contains all possible error options for each correct label. Such a repository can then be used in a WoZ study. Compared to the confusion matrix, these four error types are expected to be easier to understand and to better support designers in testing the users' response to different types of errors.

3 THE WIZARD OF ERRORS INTERFACE

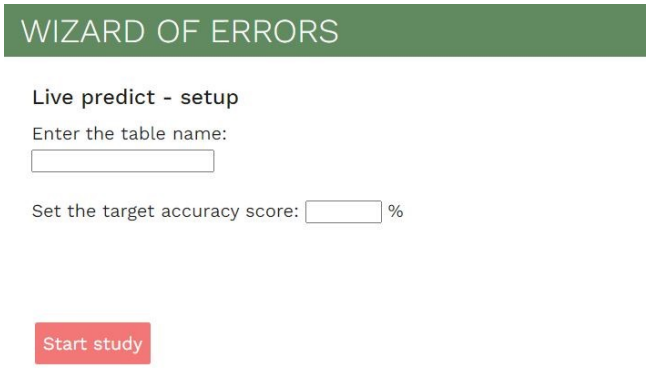
To enable the inclusion of the four ML error types in WoZ studies, we developed *Wizard of Errors* (WoE). WoE is a web interface, which aims to help designers test the effects of different ML errors on users' experience while interacting with an ML-enabled system. The WoE application can be connected to different prototyping platforms (e.g. Processing, JavaScript, Python, Arduino) to perform WoZ studies. During the study, the WoE enables designers to simulate the predictions of supervised ML models by selecting either correct or incorrect labels (i.e. ML errors). For each correct label, four incorrect labels can be selected, based on the four different types of errors. Designers can also set an ML model target accuracy, so they can adjust the number of errors dynamically throughout the test, in order to reach the desired accuracy they intend to assess.

The WoE application consists of two elements: (i) *the error repository*; (ii) *The WoE interface*. The error repository needs to be created by the designer in advance, as part of the WoZ setup, and it consists of a table that includes all the possible correct labels, and the corresponding incorrect labels - one for each of the four error types (see Table 1). Once the error repository is uploaded to the WoE interface, the wizard-designer can decide to select a target accuracy, which will guide them in selecting the right amount of ML errors (see Figure 2a).

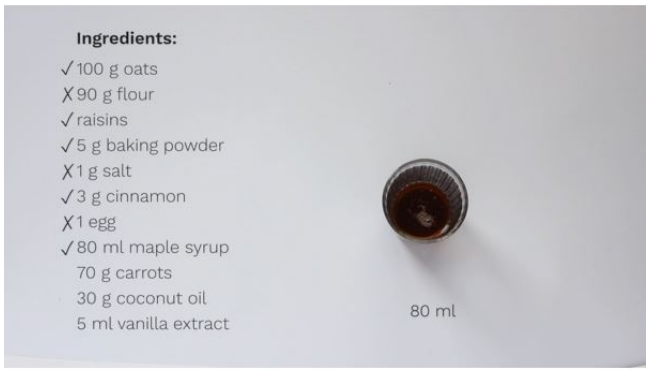
The WoE interface is used during the WoZ study to simulate different errors during the user's interaction with the ML-enabled system (see Figure 1 for an overview of the system). The WoE interface is controlled by the wizard-designer, who can insert ML errors during the WoZ test, in order to determine how users react to potential ML errors. The interface allows designers to test three elements that potentially affect the overall UX: (i) the accuracy of the ML model; (ii) the types of ML errors that can occur; (iii) the confidence score for the predictions (a number between 0-1 or 0-100% that indicates the probability associated to that label. A high confidence score does not mean that the label is correct, just that the machine associates that label to the input with a high probability).

4 METHOD

To assess if designers could mirror the behavior of an ML model in a WoZ study, we tested the WoE interface with 10 design students. We aimed to investigate if designers would be able to simulate the ML behavior, including its errors, through the WoE interface, and what issues they would encounter. As a design case for our test, we used a smart kitchen countertop inspired by the IKEA concept kitchen [21], which adopts ML image recognition to identify ingredients placed on the countertop. We asked participants to simulate the



(a) Setup page



(b) A screenshot of the video with ingredient list that was added to the prediction interface in Figure 1

Figure 2: Elements of the WoE interface

image recognition (ML) system performance through WoE, as if they were testing the user experience of the smart countertop in a WoZ study. Because in-person interactions were not possible, due to COVID-19 restrictions, the tests were performed online. In order to create the setting of a WoZ test, a video was generated, which simulated a sequence of user's actions, i.e. a series of ingredients being added to different bowls on the countertop. The video also showed an ingredient list on the screen. This video was shown next to the WoE interface (see Figure 2b) and participants were asked to use it as an input for simulating the ML system behavior in correctly or incorrectly classifying the ingredient being added to the bowl. We chose to simulate a user's interactions through a video because it allowed for consistency between the different tests - by providing all participants with the same inputs, and because in-person studies were not allowed. For this study, an error repository was created by the researchers and uploaded to WoE. An extract of the error repository can be found in Table 1.

4.1 Participants

Participants were selected through snowball sampling [9]. Students who were known to have worked with ML in a design project were approached, and they referred other students who also had experience with using ML in a design project. 10 participants were recruited (n=5 female; age: 22-27, M:23.5, SD: 1.71). All participants

Table 1: A selection from the error repository

ID	correctAnswer	segmentationError	similarityError	wildError	noRecognitionError
0	oats	cinnamon	flour	carrots	null
1	flour	salt	oats	maple syrup	null

were Bachelor (n=1) or Master (n=9) Industrial Design students from a technical university. All of them had experience with using ML in at least one design project, up to six projects.

4.2 Procedure

The study consisted of three parts and was conducted entirely online through a video-conference platform. The study protocol complied with the university Ethical Review Board procedures and all data were managed in accordance with GDPR regulations.

Part I - pre-experiment interview. For each subject, a pre-experiment semi-structured interview and a short survey were used to get insights into the participant's experience with and knowledge of ML errors.

Part II - the experiment. The design case of the smart countertop was introduced to the participant, together with the aim of the WoZ study, i.e. simulating the behavior of an ML system for image recognition, including its errors. After describing the WoE interface, participants were asked to set the target accuracy score to 50% and to start the video. The video showed an ingredient being poured into a bowl, and the task of the participant was to classify that ingredient, to simulate the ML prediction. After the ingredient was added, the video paused. The participant was then asked to (i) select the ground truth from the upper list (see Figure 1), (ii) choose the confidence score for their prediction, and (iii) simulate the ML prediction by selecting either the correct label or one of the ML errors. If the correct label was chosen, a check mark would appear next to the ingredient in the video. In case of an error, a cross mark would be shown. This was meant to help the participant keep track of the number of correct predictions and errors they had simulated during the study. After recording the prediction, the video would continue by showing the next ingredient. The same procedure was repeated for 12 ingredients in total.

The WoE interface also showed the real-time accuracy score, based on the number of (in)correct predictions made up to that point. This allowed the participant to adjust their behavior over time, in order to reach the pre-defined target accuracy, i.e. make more correct predictions if the accuracy score was below the target accuracy score, or more incorrect predictions if the accuracy score was above the target. Once the video ended, the participant was asked to follow the same procedure, watching the video again, this time with a target accuracy score set to 70%. Participants were asked to think aloud during the whole study, to explain the reasoning behind their choices and actions. All participants' interactions with WoE (i.e. logs) were recorded for subsequent analysis.

Part III - post-experiment interview. At the end of the experiment, a semi-structured interview was conducted and participants filled out a post-experiment survey. The interview and the survey were aimed to assess the participant's understanding of the error types, their view on how different error types could influence the UX, and what error they found most difficult to make.

5 PRELIMINARY RESULTS

Results stem from a thematic analysis of qualitative data (i.e. audio recordings and interviews), as well as from the analysis of quantitative data (i.e. Likert scales from the survey and logs of the WoE system). Since the sample was small and homogeneous, we see these results as interesting initial findings that need to be validated through a larger study.

5.1 Machine Learning errors

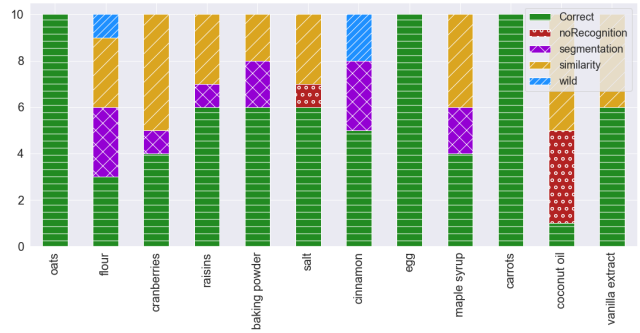
Participants indicated that before working with ML, they were unaware of possible ML errors, but they became familiar with them during their first project. They assessed errors as increasingly important throughout the design process (Ideation phase: M=2.9, SD= 1.5; Realization phase: M=6.4, SD= 0.5). This assessment barely changed after the experiment (Ideation phase: M=3.0, SD= 1.7; Realization phase: M=6.5, SD= 0.5).

The descriptive types of ML errors used in this experiment were new to all participants. While most ML errors were clear, mainly due to their self-explanatory names, the segmentation error was more difficult to understand and only two participants could correctly define it afterwards. During the study, the participants could read the explanation by hovering over the buttons, but this option was hardly used.

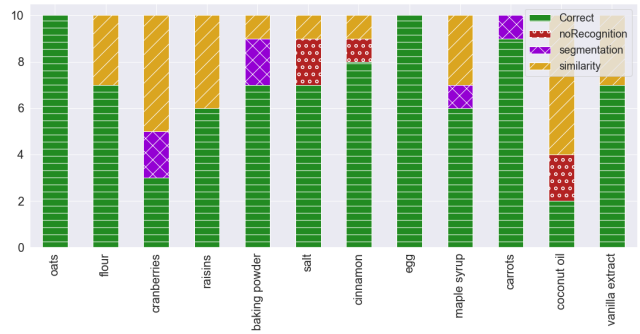
Moreover, participants expected the different error types to have different impacts on the UX. While the similarity error was seen by most as less harmful because it was expected to be understandable by humans, the other errors were considered to have a more negative impact on the UX. Participants expected these errors to decrease the trust in the model or elicit frustration. *"But if I put a carrot on the table and it is like "Oh strawberry" then I would be like, this is a stupid system it cannot even recognize such a simple thing then I must just work horribly"* (P2).

5.2 Mirroring Machine Learning

The main focus of the study was to evaluate if participants could mirror the behavior of an ML model by simulating ML errors. While all participants had worked with ML before and trained models themselves, their behavior did not match with that of an ML model. Several misconceptions surfaced during the experiment, as explained below. Nevertheless, the WoE interface was successful in encouraging wizards to make ML errors and to reach the target accuracy score, although they were hesitant to make ML errors in the beginning (for $accuracy = 50(\%) : M = 59.10, SD = 10.47$; for $70(\%) : M = 68.72, SD = 9.67$). Participants claimed that WoE was a useful tool to test the interactions in an early stage: *"I think it would be helpful to explore the opportunities and limitations of ML*



(a) First cycle with target accuracy score 50%



(b) Second cycle with target accuracy score 70%

Figure 3: Overview of the predictions for each ingredient

in combination with a specific project. Normally this is difficult to do because in an early stage of a project a full system is often not working already." (P6). However, they also indicated that it was hard to assess a realistic number of errors to include (P9) and that they would like to have something that could be applicable even earlier in the design process, to evaluate the potential impact of errors (P3, P4 and P5).

Logical errors are easier to make. The ML errors made during the experiment showed that participants mainly selected similarity errors and hardly any wild errors (see Figure 3). From the think-aloud transcripts, it became clear that participants based their choices mainly on the label associated to each error type and not the error type itself (e.g. choosing the 'sugar' label for salt, instead of choosing a similarity error). By using labels, participants often selected an understandable or logical error from a human perspective. They also expressed it felt unnatural to send an unrelated error. As participant 9 stated: *"For me as a wizard, a no-recognition error would be weird to make because I know what it is and rationalizing why it would not recognize it at all would be a bit weird, in my opinion."*

Projecting personal knowledge. Similarly to the tendency to select logical errors, participants based the behavior of the ML model on their own knowledge. This was mainly visible when coconut oil was shown in the video, and 85% of the predictions were an ML error - while the mean for all other ingredients, excluding coconut oil, was 31.8%. As participant 7 put it: *"I myself have never used coconut oil before, so I am not really sure if it is an oil or if it looks like this."*

The tendency to rely on their own knowledge could also be seen in participants' assumption that common ingredients (in their food culture) are easier to recognize: *"If there are ingredients that are exotic, then it might be easier to make [a no-recogntion error]"* (P1). Moreover, they did not consider that ingredients can exist in different shapes and forms. Eggs and carrots were seen as very easy to classify for the machine because they had a distinctive shape, but participants did not take into account that they could exist in many forms and shapes, which would make it harder for the ML model to recognize them correctly.

Higher confidence score for correct answers. To check if participants gave higher confidence scores to correct predictions, a linear regression was calculated. A significant regression was found ($F(1, 238) = 35.618, p < .001$), with an R^2 of .130 and an $R^2_{Adjusted}$ of .127, indicating that the error type (correct vs incorrect) is a significant predictor for the confidence score ($\beta = .361, t(238) = 5.968, p < .001$) and that the confidence score increased when a correct prediction was selected. This differs from a real ML model, where a confidence score can be equally high for incorrect predictions.

ML model self-awareness. Finally, participants expected the model to be aware that it was making a mistake and mentioned that it would be useful if it would indicate to the final user what type of error it was making: *"I think for the user it is very helpful to know that there is an error, and they know what kind of error it is."* (P3). A possible explanation for this misconception might be that the participants had only trained and tested an ML model and not deployed it live, since in training and testing supervised classification models the correct label is available and makes it possible to detect errors - although not the error types.

6 DISCUSSION

The preliminary results presented in the previous section provide insights into what aspects of mirroring an ML model are difficult for wizard-designers. These findings have implications both for designing WoZ studies for ML applications, and for designing with ML in general.

When designing WoZ studies to test user experiences and interactions based on ML, one needs to take into account the fact that wizards will not be able to ignore their human rationale, therefore their behavior will not be a good representation of the ML behavior. Our study showed that the WoE interface and the ML errors can already contribute to a better representation of ML models in WoZ studies and can trigger designers to reflect on the importance of testing different types of ML errors. However, it also is clear that this is not sufficient, as designers' misconceptions prevent them from properly mirroring ML models. One way to overcome this limitation is to recommend wizards certain error types during the use of the WoE interface, for instance in case certain error types are more rarely selected, compared to others.

Another option is to allow designers to pre-define how many errors for each type will be sent by the WoE interface, and let the interface randomly assign those errors, while asking designers only

to select the ground truth. While limiting designers' ability to adjust the number and type of errors in real-time, this would make the study more realistic.

Next to that, designers not only need to test different error types, but also vary on the number and the moment of occurrence to explore the effect of all these factors on the user's experience and their trust in the ML predictions. While this study only focused on using WoE in the early phase of a design process, ML errors can also be introduced during other phases. For instance, in the ideation and conceptualization phase, cards with the error types, potential consequences and options for adjusting the design can help to consider errors from the beginning. When designing a UX wireframe, the designer can go over each step and consider what would happen if each type of error occurred, to check if their interface is able to fail gracefully [18] or if it needs to be adjusted so that it can fail gracefully.

7 LIMITATIONS AND FUTURE WORK

A limitation of the WoE interface, and WoZ testing with ML in general, is that it is only applicable in cases where the humans are able to make the correct predictions - i.e. recognize the ground truth, during the study. However, this still leaves out a considerable number of ML models within supervised learning. This study only covered object recognition, but we expect the error types to be also transferable to other subsections of computer vision and potentially also to other types of supervised classification. Further research should explore and validate the use of ML in these applications.

Another limitation of this study is that it was conducted with students from one Industrial Design faculty, which could give a biased view on how design students, and designers in general, approach and use ML. Therefore, we consider these findings as preliminary results and future studies should be conducted with a larger and more diverse sample to confirm their validity. Furthermore, the WoE interface should be tested in a WoZ study with actual users, to gather more insights on designers' behaviors in a real context.

8 CONCLUSION

Using ML in prototypes during the early phases of a design process is challenging. In this study, we introduced Wizard of Errors (WoE), a prototyping tool for conducting WoZ studies on ML-enabled interactions. The WoE interface facilitates the inclusion of ML errors into WoZ studies, since these are essential for UX assessment but are currently rarely included. In WoE we used four descriptive error types instead of the confusion matrix and evaluated with design students if ML errors are important to consider and how relevant the four error types are in comparison to the confusion matrix. Moreover, during a WoZ simulation, we evaluated if it is possible for a designer to simulate an ML model in terms of ML error behavior. The error types showed to have potential to be used during WoZ studies, but we also identified several challenges that still prevent the designer from realistic error representation in WoZ. In this study, we designed and tested the WoE interface to embed ML errors in WoZ studies, and provide preliminary knowledge that can help both design researchers and practitioners in this field to consider ML errors as a regular component of the design process for ML-enabled solutions.

REFERENCES

[1] Abhay Agarwal and Marcy Regalado. n.d.. A Design Language for Human-Centered AI. Retrieved on April 6, 2021 from https://linguafranca.polytopal.ai.

[2] Emma Beauxis-Aussalet, Joost van Doorn, and Lynda Hardman. 2018. Supporting End-User Understanding of Classification Errors. In *Proceedings of the 36th European Conference on Cognitive Ergonomics* (Utrecht, Netherlands) *(ECCE'18)*. Association for Computing Machinery, New York, NY, USA, Article 10, 8 pages. https://doi.org/10.1145/3232078.3232096

[3] Andrew Begel, John Tang, Sean Andrist, Michael Barnett, Tony Carbary, Piali Choudhury, Edward Cutrell, Alberto Fung, Sasa Junuzovic, Daniel McDuff, Kael Rowan, Shibashankar Sahoo, Jennifer Frances Waldern, Jessica Wolk, Hui Zheng, and Annuska Zolyomi. 2020. Lessons Learned in Designing AI for Autistic Adults. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 46, 6 pages. https://doi.org/10.1145/3373625.3418305

[4] Jared N. Bott and Joseph J. Laviola Jr. 2015. The WOZ Recognizer: A Wizard of Oz Sketch Recognition System. *ACM Trans. Interact. Intell. Syst.* 5, 3, Article 15 (Oct. 2015), 38 pages. https://doi.org/10.1145/2743029

[5] Jacob T. Browne. 2019. Wizard of Oz Prototyping for Machine Learning Experiences. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, Article LBW2621, 6 pages. https://doi.org/10.1145/3290607.3312877

[6] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739

[7] Andrew Finke. 2019. Lake: A Digital Wizard of Oz Prototyping Tool. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3308455

[8] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. 2016. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 3558–3565. https://doi.org/10.1145/2851581.2856492

[9] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* 32, 1 (1961), 148–170.

[10] Matthew K. Hong, Adam Fourney, Derek DeBellis, and Saleema Amershi. 2021. Planning for Natural Language Failures with the AI Playbook. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 386, 11 pages. https://doi.org/10.1145/3411764.3445735

[11] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[12] Gierad Laput, Robert Xiao, and Chris Harrison. 2016. ViBand: High-Fidelity Bio-Acoustic Sensing Using Commodity Smartwatch Accelerometers. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. Association for Computing Machinery, New York, NY, USA, 321–333. https://doi.org/10.1145/2984511.2984582

[13] Gierad Laput, Yang Zhang, and Chris Harrison. 2017. Synthetic Sensors: Towards General-Purpose Sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3986–3999. https://doi.org/10.1145/3025453.3025773

[14] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. 2020. A deep learning system for differential diagnosis of skin diseases. *Nature medicine* 26, 6 (2020), 900–908.

[15] David Maulsby, Saul Greenberg, and Richard Mander. 1993. Prototyping an Intelligent Agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 277–284. https://doi.org/10.1145/169059.169215

[16] Andrea Isabell Müller, Veronika Weinbeer, and Klaus Bengler. 2019. Using the Wizard of Oz Paradigm to Prototype Automated Vehicles: Methodological Challenges. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings* (Utrecht, Netherlands) *(AutomotiveUI '19)*. Association for Computing Machinery, New York, NY, USA, 181–186. https://doi.org/10.1145/3349263.3351526

[17] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[18] Google PAIR. 2019. People + AI Guidebook. pair.withgoogle.com/guidebook.

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

[20] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 153 (Oct. 2020), 22 pages. https://doi.org/10.1145/3415224

[21] Sly Golovanov. 2015. "IKEA Concept Kitchen 2025" April 21, 2015. [YouTube video]. https://www.youtube.com/watch?v=qD60cBQOABY

[22] Alaa Tharwat. 2020. Classification assessment methods. *Applied Computing and Informatics* 17, 1 (2020), 168–192.

[23] Philip van Allen. 2018. Prototyping Ways of Prototyping AI. *Interactions* 25, 6 (Oct. 2018), 46–51. https://doi.org/10.1145/3274566

[24] Sruthi Viswanathan, Behrooz Omidvar-Tehrani, Adrien Bruyat, Frédéric Roulland, and Antonietta Maria Grasso. 2020. Hybrid Wizard of Oz: Concept Testing a Recommender System. 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, 1–7. https://doi.org/10.1145/3334480.3383097

[25] Qian Yang, Nikola Banovic, and John Zimmerman. 2018. Mapping Machine Learning Advances from HCI Research to Reveal Starting Places for Design Innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3173574.3173704

[26] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T. Iqbal, and Jaime Teevan. 2019. Sketching NLP: A Case Study of Exploring the Right Things To Design with Language Intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300415

[27] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730

[28] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[29] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning Adaptive Mobile Experiences When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 565–576. https://doi.org/10.1145/2901790.2901858