

Demystifying artificial intelligence for end-users: findings from a participatory machine learning show

Katharina Weitz, Ruben Schlagowski, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Weitz, Katharina, Ruben Schlagowski, and Elisabeth André. 2021. "Demystifying artificial intelligence for end-users: findings from a participatory machine learning show." *Lecture Notes in Computer Science* 12873: 257–70. https://doi.org/10.1007/978-3-030-87626-5_19.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:

Deutsches Urheberrecht

Weitere Informationen finden Sie unter: / For more information see:

<https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/>



Demystifying Artificial Intelligence for End-Users: Findings from a Participatory Machine Learning Show

Katharina Weitz^(✉), Ruben Schlagowski, and Elisabeth André

Lab for Human-Centered AI, University of Augsburg, Augsburg, Germany
{katharina.weitz,ruben.schlagowski,elisabeth.andre}@uni-a.de
<https://hcai.eu>

Abstract. Interactive and collaborative approaches have been successfully used in educational scenarios. For machine learning and AI, however, such approaches typically require a fair amount of technical expertise. In order to reach everyday users of AI technologies, we propose and evaluate a new interactive approach to help end-users gain a better understanding of AI: A participatory machine learning show. During the show, participants were able to collectively gather corpus data for a neural network for keyword recognition, and interactively train and test its accuracy. Furthermore, the network’s decisions were explained by using both an established XAI framework (LIME) and a virtual agent. In cooperation with a museum, we ran several prototype shows and interviewed participating and non-participating visitors to gain insights about their attitude towards (X)AI. We could deduce that the virtual agent and the inclusion of XAI visualisations in our edutainment show were generally rated positively by participants, even though the frameworks we used were originally designed for experts. When comparing both groups, we found that participants felt significantly more competent and positive towards technology compared to non-participating visitors. Our findings suggests that the consideration of specific user needs, personal background, and mental models about (X)AI systems should be included in the XAI design for end-users.

Keywords: Explainable AI · Virtual agents · Artificial intelligence · Neural networks · Edutainment

1 Introduction

The results of the survey of the European Commission [2] in 2017 revealed that the attitude people have towards artificial intelligence strongly depends on their level of knowledge and information about these systems. Accordingly, one goal of the research community should be to provide means that help to better educate end-users of AI-based technologies and research methods by which public understanding of AI-systems can be improved. Explainable AI (XAI) refers to

methods that illustrate and make the decisions of AI-systems more understandable. However, authors like Miller et al. [25] point out that XAI approaches focus too much on the needs for AI-experts and do not consider the needs of end-users. Although there are first approaches that introduce X(AI) to end-users in a playful way, these are targeting individuals or small groups [4]. To enable a large group of end-users to understand the abilities and limitations of AI systems, we presented a public interactive machine learning show (ML-show) in the German museum in Munich to over 2200 visitors¹. Within this show, participants were able to collectively train an artificial neural network for audio keyword recognition after collecting a corpus of audio samples. After about 20 min of training, the audience was able to test how well the keyword classifier performed to gain a better understanding of the system’s limited accuracy. After testing the network’s performance, participants were shown information, detailing why the system made right or wrong predictions. This information was presented with the help of a the visual XAI framework LIME [28] and a virtual agent, which was previously found to have a potential positive impact on user trust [37]. After the show, participants were asked to fill out a questionnaire about their personal impressions of AI and XAI as well as the virtual agent.

By comparing the questionnaire results of ML-show attendees with baseline data, which was gathered in a separate questionnaire with non-participating museum visitors, we investigated the following research questions:

1. How do end-users perceive a participatory ML-show?
2. How do ML-show attendees differ from non-participants in terms of their attitude towards AI and self-estimated competence regarding AI?
3. How do ML-show attendees differ from non-participants in terms of attitude towards technical systems in general?

In addition to these experiment-related results, our paper presents a novel, participatory approach combining virtual agents with XAI methods to introduce machine learning topics to large user groups.

2 Related Work

2.1 Virtual Agents in Education and Edutainment

The use of animated agents in education scenarios was found to have positive effects on students learning performance in various studies since the 90s. Lester et al. [21] called this phenomenon the **persona effect**, which links the presence of life-like agents with positive effects on student’s perception of their learning experience and motivation. This was confirmed by Mulken et al. [34], who found out that certain tests were perceived as less demanding when presented by an agent. They argued that the persona effect can also help to reduce fear that would

¹ The presented study in this paper as well as the collected data have been approved by the data protection officer of University of Augsburg.

normally arise when using common educational material. Hammer et al. [8] found similar effects for social robots that were used to foster well-being of elderly people. In their study, elderly people felt more confident when interacting with the robot than with tablet computers and perceived the social robot as less complex. In the study of Jin et al. [16], similar phenomena were observed while using a virtual agent in a computer-aided educational test. During the test, users felt entertained by the virtual agent which led to increased attention and test performance. The observed positive effects during the presence of virtual agents scale with the quality of the agent: The more life-like or realistic the virtual agent appears, the better. As such, the quality of agent features such as human-like voice, gestures, facial expression, eye gaze, and body movement plays an important role [23]. In the edutainment sector, technologies like virtual reality and virtual agents are often used to make communication of knowledge playful and entertaining (e.g., virtual museums [20]). Carolis and Rossano [1] used agents to teach children about healthy nutrition in a enjoyable way. Ming et al. [26] used a virtual reality setting to help participants to learn Mandarin. Here, users interacted with a virtual agent to practice the pronunciation of words.

These examples of the successful integration of virtual agents in education and edutainment settings represent a promising approach that we used in our field study to help end-users to understand the abilities and limitations of AI systems.

2.2 Explainable AI

With the rise of advanced machine learning models such as deep neural networks in a wide application field, the resulting lack of transparency and comprehensibility of an AI's decisions can not only be challenging for engineers and scientists, but also have negative impact on the perceived trustworthiness and user-experience of end-users [12, 33]. For this paper we adapt the view of Gilpin et al. [5], who stated that the goal of XAI is the description of a system, its internal states and its decisions, so that this description can be understood by humans. A common approach to shed light on to the decisions of deep neural networks is the highlighting of regions of the input data (e.g., images), that are important for specific decisions, resulting in visual explanations (for the interested reader, we recommend the papers of [14, 36]).

Within our user study, we used the LIME framework proposed by Ribeiro et al. [28] to highlight relevant areas within the spectrograms of audio samples that are classified as keywords by a neural network.

2.3 Trust in Technical Systems

One common definition of trust in human-agent interaction sees trust as “the attitude that an agent will help achieve an individual’s goals in a situation characterized by uncertainty and vulnerability.” [19, p. 51] The approach of Hoff and Bashir [11] distinguishes dispositional, situational, and learned trust. Dispositional trust refers to long-term tendencies based on biological and environmental

influences. Situational trust describes external aspects (e.g., the environment) as well as internal aspects (e.g., the mood of the user) in a specific situation. Learned trust refers to the experience a user has already gained, for example with virtual agents. A distinction has also been made between different types of trust, such as distrust and overtrust. Marsh and Dibben [22] present a continuum of trust, where distrust is placed on the negative end, and trust on the positive end of the trust scale. Overtrust refers to unrealistic expectations in the system, which can result in the system being ascribed skills that it does not possess and can end in its misuse [19].

In our participatory ML-show, we focus on situational and perceived trust (not distrust or overtrust), i.e., the trust in an AI system that incorporates a virtual agent for XAI purposes in a public participatory ML-show.

3 Field Study

For six months, visitors of the museum were able to take part in a participatory ML-show. Here, a neural network for audio keyword speech recognition [30] was trained to learn a new keyword live in the show. Simultaneously, a virtual agent named Gloria was displayed on a screen and communicated with the audience via speech output (see right subfigure in Fig. 1).



Fig. 1. *Left:* Stand at the museum, which was used to ask museum visitors about their attitude towards AI and XAI. *Right:* Begin of a public participatory machine learning show visited by non-experts in the museum.

Additionally, during three days, we gathered baseline data using a paper-based questionnaire oriented on the questions used in the Eurobarometer report [2]. We also recorded the affinity of museum visitors for technology using the TA-EG questionnaire [17]. We made sure that we only questioned people who did not visit our participatory ML-show (see left subfigure in Fig. 1).

3.1 Demonstrator Setup

The demonstrator (see Fig. 2) mainly consisted of a **demonstration PC** including a high performance GPU (Nvidia GTX 1060) for improved training performance and a smartphone which was used to record and transmit audio samples

for training and prediction of the neural network over WLAN. The demonstration PC was connected to a **beamer** which displayed the virtual agent, the generated spectrograms, and the XAI visualisations generated by the XAI framework. In parallel, the demonstration PC hosted a website providing audio recording and transmission functionalities on a server in the local network. An android app containing a browser window was used on the **smartphone** to access the site when the audience recorded the audio samples. For the recognition of audio keywords we used the neural network architecture proposed by Sainath & Parada [30]. This prediction model uses mel frequency cepstrum coefficients (MFCCs) that are calculated from audio spectrograms as input data for multiple convolutional layers for the calculation of abstract features. These features are subsequently passed on to a fully connected layer for the final target class prediction.

The moderator of the show, who was instructed in advance, operated the main application by using a step-by-step structured GUI that enabled him or her to (1) start and stop the training process of the neural network, (2) start prediction for a recorded audio sample, (3) review transmitted audio files, and (4) calculate XAI visualisations after prediction.

The virtual agent Gloria developed by Charamel² was integrated into a separate website that was hosted locally and displayed with a browser on the demonstration PC. Communication between the virtual agent and the main application was implemented with WebSockets.

To generate a XAI visualisation for a specific prediction of our classification model, we first used the Felzenszwalb’s algorithm for image segmentation [3] to generate so-called superpixels. The LIME framework subsequently determined the most relevant superpixels for the three top predictions and colored them in green (if they contributed positively to a prediction) or red (if they contributed negatively to a prediction). In order to make the resulting XAI visualisations better readable, we used the webMAUS API [18] to highlight areas within the spectrograms that contain the phonemes of the actually spoken word (ground truth).

3.2 Study Procedure

The new keyword that was trained to the neural network was freely selected by the audience during a discussion at the start of the show (see Fig. 3 for the study procedure). Afterwards, the visitors recorded a training dataset for the selected word by passing around the smartphone with a connected high quality microphone.

As soon as about 80 audio samples were recorded and transmitted to the demonstration PC, the moderator used pre-programmed software functionalities to label the samples and merge them with a subset of the the speech command dataset provided by Warden [35] (we used data for 11 classes/keywords, 80 samples each) to create the training corpus. Then, the moderator started the training process of the prediction model. To give the participants a feeling of

² <https://vuppetmaster.de/>.

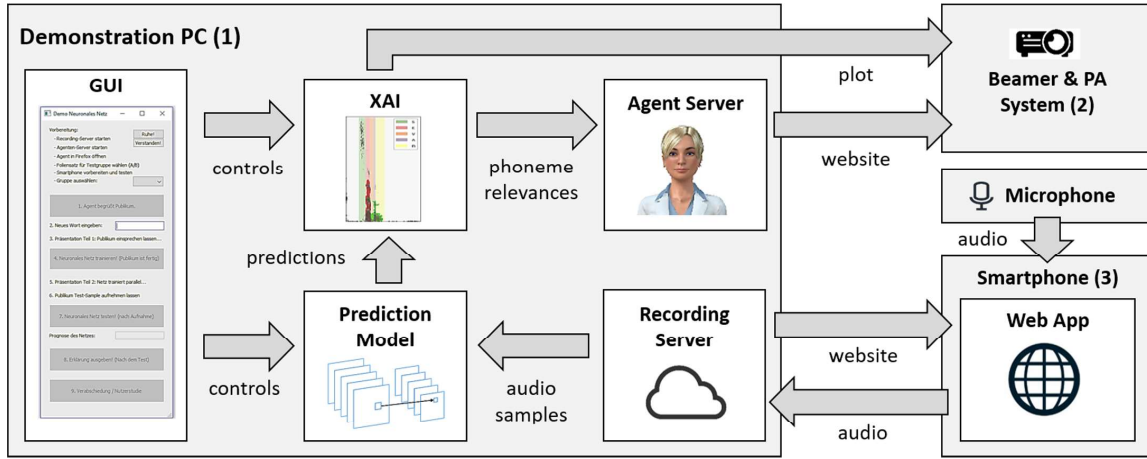


Fig. 2. Demonstrator overview as used in the field study. *Left: (1) Demonstration PC*, running the neural network for keyword recognition and calculating the XAI visualisations. *Right: (2) Beamer & PA system* were used for agent displaying & sound, the XAI visualisations, and the presentation slides for the show. *(3) Smartphone* and microphone for recording the audio samples.

how good the classifier was after this relatively short amount of time (the typical validation accuracy was about 80%), we decided to not use any pre-trained networks and instead train the network from scratch for each show. While the model was trained, visitors were given a 20-minute lecture on how neural networks for speech recognition work and how the LIME framework can be used to understand the classifiers decisions in this context. As soon as the lecture was finished, the moderator stopped the training. Afterwards, the network could be tested by volunteering participants multiple times by speaking both known and unknown keywords into the microphone. The resulting audio samples were transmitted to the demonstration PC and passed on to the classifier. Together with prediction results, the XAI visualisations generated by the XAI framework LIME were displayed for the audience.

In parallel to the show, the virtual agent Gloria commented on the training, communicated the classifier’s prediction results, and commented on the XAI visualisations (e.g., “The most relevant phoneme for the prediction of *< keyword >* was...”).

3.3 Evaluation Method

After the show, participants were asked to complete a questionnaire, either online or on paper. In addition to the collection of demographic information, the following questions were included:

Agent & (X)AI system Evaluation. To evaluate the virtual agent Gloria, we used 5 items on a 7-point Likert scale (e.g., “I liked Gloria”) and free-form feedback. We collected participants’ feedback about the AI system by using 3 items a 7-point Likert scale (e.g., “I would use the AI system”). To gain insights of the

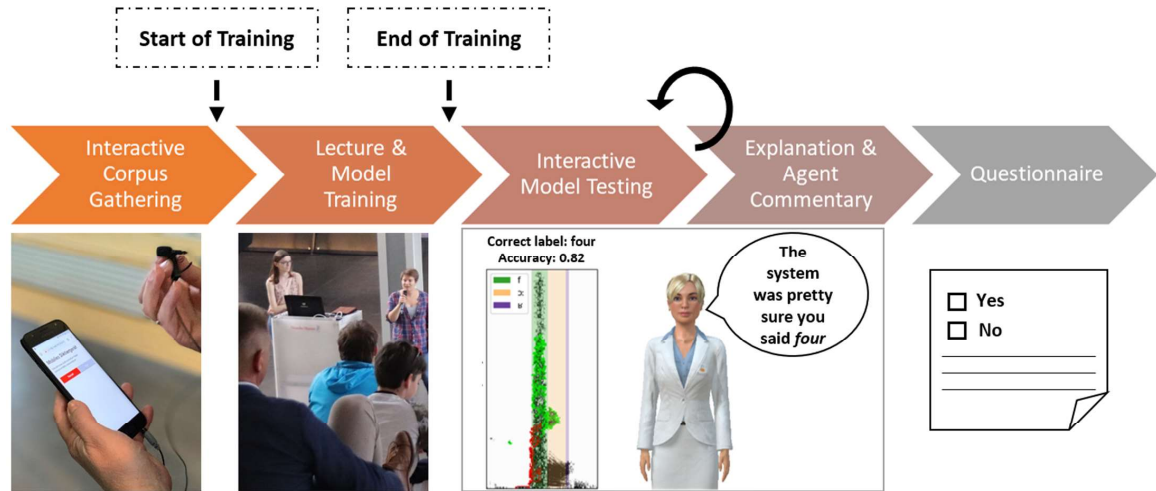


Fig. 3. Procedure during the ML-show: An interactive audio-corpus was collected and used to train a Neural Network during a lecture about Machine Learning. Afterward, the model was tested by the participants. The virtual agent Gloria presented the results and XAI visualisations. At the end of the ML-show, participants had to answer a questionnaire.

perceived helpfulness of the XAI visualisations, we asked 1 item on a 7-point Likert scale (i.e., “Were the explanations sufficient?”) and a free-form question about which additional information would be helpful for them to understand the AI system.

Technical Affinity. To measure the technical affinity of participants using the TA-EG questionnaire [17] was queried.

Trust. Subjective trust was assessed with the Trust in Automation (TiA) questionnaire [15].

Attitude Towards AI. At the end of the questionnaire, additional questions about the participant’s general knowledge attitude towards AI and XAI were posed (e.g., “How would you rate your knowledge of AI?” and “In general, what is your attitude towards artificial intelligence?”).

4 Results

4.1 Information About Participants

ML-show Participants A total of 65 public participatory machine learning shows with an average of 35 participants each were held. A total of 2275 museum visitors took part in the study, of which 51 completed the subsequent questionnaire. Due to missing data in some questionnaires, 47 participants (24 male, 22 female, 1 non-binary) between 13 and 80 years ($M = 42.07$, $SD = 22.6$) were included in the final analyses presented in this paper. The educational

background of the participants was mixed and ranged from “no degree” to “university degree”. Most participants had no previous knowledge or experience in the use of virtual agents, voice assistants, or audio processing. 88% of the participants stated that they have already heard of the term AI, but only 11% of them rated their AI knowledge as extensive. Most of the participants either had a balanced or a positive view about the impact on AI in the future (see left side of Fig. 4). A majority of the participants saw XAI as an important topic, especially for researchers, companies, and end-users (see right side of Fig. 4). For politicians, participants rated the importance of XAI less compared to the other stakeholders.

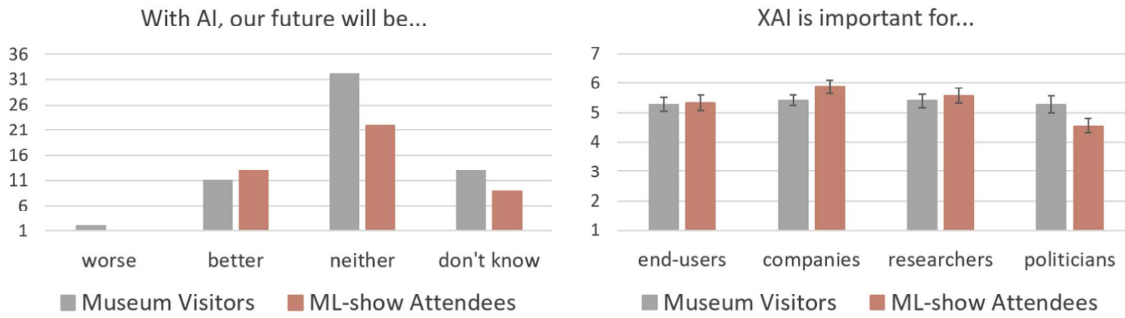


Fig. 4. *Left:* ML-show attendees and non-participating museum visitors answers to the question “What future do you think we will have with AI?”. *Right:* Rating of the participants, whether XAI is important for different stakeholders (1=disagree; 7=fully agree). Error bars represent the standard error.

Baseline Participants. 59 museums visitors took part in our field survey which we used as baseline for the comparison with the ML-show participants. We had to remove the answers of one visitor due to too much unanswered questions. Therefore, for the following analyses, answers from 58 museum visitors (29 female, 29 male) between 8 and 66 years ($M = 30.3$, $SD = 16.5$) are considered. The educational background of the participants was mixed and similar distributed as in the ML-shows.

Most of the visitors had heard about AI in the last 12 month (97%), but only 24% of them had heard about XAI. Non-participants saw XAI and the future of AI similar than ML-show attendees (see Fig. 4).

4.2 Results of the ML-show

Agent & (X)AI Rating. Participants gave the virtual agent Gloria a rating of $M = 3.9$ which was slightly above average (7-point Likert scale). The LIME visualisations were rated with $M = 4.15$ slightly higher than the virtual agent. Investigating whether the participants would use such an AI system, the rating was with $M = 3.06$ beyond average (7-point Likert scale). In response to the

free-form question about what additional information they would have liked to see in Gloria's explanations, users indicated that they would have liked more details (e.g., "What does Gloria calculate in the training phase?").

Correlations for ML-show Participants. To examine potential connections of the educational background, gender, technical affinity, and age of the participants on questionnaire items like trust in the AI system, virtual agent impression, and the helpfulness of the XAI visualisations, we calculated Pearson's product-moment correlations.

We found a significant weak positive linear relationship between perceived trust in the presented AI system and educational background ($r = .47$, $p < .05$), where participants with an higher educational background tend to trust the AI system more. Neither age nor gender had a significant impact on subjective trust in the AI system, as we did not find any significant correlations for these variables. For the impression of the agent as well as the helpfulness of the XAI visualisations, we did not find correlations for age, gender, and educational background of the participants.

4.3 Comparison Between Participating and Non-participating Museum Visitors

We used two one-way MANOVAs to examine if there were any significant differences compared to the non-participating museums visitors (baseline). Holm correction for multiple testing was applied.

Attitudes Towards AI. We conducted a MANOVA to evaluate whether there was a difference between baseline museum visitors and ML-show participants in (1) the perceived knowledge about AI as well as (2) their attitude about the impact of AI on our lives in the future and (3) in their attitude towards AI. We found no significant differences for these three variables, $F(3, 100) = 1.76$, $p = .16$, Pillai's Trace = 0.51.

Technical Affinity. To evaluate the TA-EG questionnaire, we looked at the four subscales (excitement, competence, negativity, and positivity) using a one-way MANOVA. The MANOVA showed significant differences between the groups for the TA-EG variables, $F(4, 100) = 28.58$, $p < .001$, Pillai's Trace = 0.53. To find out on which subscales of the TA-EG significant differences exist, we then performed an ANOVA that revealed significant differences for the subscales competence $F(1, 103) = 23.15$, $p < .001$, excitement $F(1, 103) = 5.03$, $p < .03$, and positivity $F(1, 103) = 96.15$, $p < .001$.

We then used post-hoc tests to investigate the direction of these differences. For this purpose, we use t-tests or Wilcoxon tests if the requirements for the

t-test were not met³ to evaluate whether there was a difference between baseline museum visitors and ML-show participants in (1) the perceived technical competence as well as (2) their excitement towards technology and (3) their positivity towards technology. Our results (Fig. 5) show:

- **Competence:** Participants of the ML-show ($M = 3.00$, $SD = 0.63$) feel more competent about technology compared to the baseline museum visitors ($M = 2.14$, $SD = 1.08$), $Z = 692$, $p < .001$.
- **Excitement:** Participants of the ML-show ($M = 3.15$, $SD = 0.94$) do not feel more excited compared to the baseline museum visitors ($M = 2.71$, $SD = 1.03$), $t = -2.24$, $p = .05$ ⁴.
- **Positivity:** Participants of the ML-show ($M = 3.62$, $SD = 0.83$) feel more positive towards technology compared to the baseline museum visitors ($M = 2.32$, $SD = 0.52$), $Z = 216$, $p < .001$.

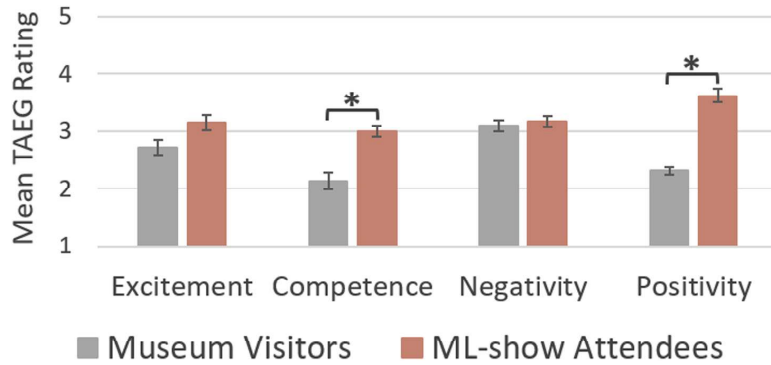


Fig. 5. Mean TA-EG ratings by category for the ML-show participants and for non-participating museum visitors. Subcategories *competence* and *positivity* indicate significant differences between the two groups ($*p < .001$). Error bars represent the standard error.

5 Discussion

We can overall conclude that end-users were receptive towards XAI-visualisations in our ML-show, even though visualisation methods in use were not specifically designed for end-users without any background knowledge in the research fields of AI and XAI. Furthermore, our field study helped us to gain initial insights concerning end-users' views about (X)AI and virtual agents in a participatory ML-show, which we would discuss in the following.

³ The Mann-Whitney U-test is the non-parametric equivalent of the t-test for independent samples and is used when the conditions for a parametric procedure are not met (in our case: homogeneity of variances and a non-normal distribution of the data).

⁴ This result was no longer significant due to the alpha error correction.

5.1 Take Users' Attitudes and Experiences into Account

The correlation analysis of our data revealed a connection between the educational background and the perceived trust in our AI system. This result encourages XAI design that fits the user's educational background. As part of our study was a presentation on the basic functioning of neural networks, speech recognition and XAI, better educated participants might have been more receptive to knowledge transfer. Thus, they might have understood the XAI visualisations better which might have resulted in increased trust. Miller [24] argued that explanations for AI systems have to be based on the expectations and needs of humans. Heimerl et al. [10] found out that more XAI information about an emotion recognition system leads not automatically to higher trust in the AI. They concluded that users tend to transfer their own mental models about emotions to the AI. Therefore, having the mental model of users in mind [29] when personalizing XAI for different stakeholders and different AI scenarios is an important step to adjust XAI to the "right amount" for individual users [32].

Here, trust models such as those of Sanders et al. [31] and Hancock et al. [9], which indicate that different components (e.g., agent characteristics, user attributes as well as situation characteristics) have an impact on trust, can be used to examine possible variables that might influence user trust in XAI scenarios.

5.2 Think About Who You Want to Reach with XAI Edutainment

The results of our study show that users who participate in an ML-show differ in aspects of technical affinity from non-participating museum visitors. Due to our study design, which did not contain a pre-study questionnaire, we cannot tell whether the differences occurred due to more technically affine museum visitors being more likely to participate in the ML-show, or whether the observed differences were a result of the ML-show itself. However, there are indications that the interaction with the AI system and the virtual agent in the ML-show could have influenced participants technical affinity. Reich-Stiebert et al. [27] reported in their study similar findings. They stated that positive attitudes towards robots increased among people who had the opportunity to be part of the prototyping process. Even though evidence suggests that virtual agents can have positive effects on user trust in XAI applications [37], it is not quite clear which factors play a role and need to be considered when designing user interfaces. According to Gulz and Haake [7], gender stereotypes is one factor that has a slight influence on the perception of virtual pedagogical agents. Whether the external appearance of a virtual agent (e.g., female or male virtual agent) plays a role in subjective trust for an AI system, or whether they can increase perceived helpfulness of an XAI setting, is still unclear.

5.3 Trust and Distrust Are Important Components in XAI Interaction Design

The ML-show participants had an average trust rating of about 4, which is slightly above average. This positive tendency towards trusting AI systems incorporating XAI and virtual agents has been previously reported by Weitz et al. [37] and demands an ethical perspective on systems that have the potential to increase user trust. In this manner, Gilpin et al. [6] stated that XAI cannot be equated with reliability and responsibility of an AI system. Hoffmann [13] makes similar statements, demanding that distrust and mistrust must also be included in the evaluation of XAI systems. We argue that ethical XAI systems should therefore be able to (1) Encourage user trust if a system performs well, (2) prevent distrust if a system performs badly, and (3) prevent overtrust if a system cannot live up to expectations.

As an average prediction accuracy of about 80% after 20 min of training was far from being perfect, a variety of wrong classifications occurred during the show and resulted in a demystification of AI systems. In fact, it might also have encouraged more distrust into XAI systems for users that were originally trusting AI systems, as they were most likely used to much better prediction models in their everyday lives.

6 Conclusion

We presented a novel public participatory machine learning show where we let visitors of a museum train a neural network together in order to clarify and demystify opportunities and limits of AI systems. During the show, we used a virtual agent and a XAI framework to provide participants with additional information about the decision-making processes of the neural network during a speech recognition task. By examining the results of a post-study questionnaire, we could deduce that the virtual agent and the inclusion of XAI visualisations in our edutainment show were generally rated positively by participants, even though the frameworks we used were originally designed for experts. We also found a correlation between trust in our AI system and the educational background of the participants. Compared to non-participating museum visitors, ML-show participants felt more competent and positive about technology. During the discussion of our results, we pointed out possible causes and limitations of our findings and concluded that consideration of specific user needs, personal background (e.g., education), and mental models is a promising approach for XAI design for end-users.

Acknowledgements. This work was partially funded by the Volkswagen Stiftung in the project AI-FORA (Az. 98 563) and by the German Federal Ministry of Education and Research (BMBF) in the project DIGISTA (grant number 01U01820A). We thank Deutsches Museum Munich, who made it possible for us to conduct the study.

References

1. De Carolis, B., Rossano, V.: A team of presentation agents for edutainment. In: Proceedings of the 8th International Conference on Interaction Design and Children, pp. 150–153. IDC 2009, ACM, New York, NY, USA (2009)
2. European Commission: Special Eurobarometer 460 (2017)
3. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. *Int. J. Comput. Vis.* **59**(2), 167–181 (2004)
4. Fulton, L.B., Lee, J.Y., Wang, Q., Yuan, Z., Hammer, J., Perer, A.: Getting playful with explainable AI: games with a purpose to improve human understanding of AI. In: CHI Conference on Human Factors in Computing Systems, pp. 1–8. CHI EA 2020, Association for Computing Machinery, Honolulu, HI, USA (2020)
5. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an approach to evaluating interpretability of machine learning (2018)
6. Gilpin, L.H., Testart, C., Fruchter, N., Adebayo, J.: Explaining explanations to society (2019)
7. Haake, M.: Virtual pedagogical agents-beyond the constraints of the computational approach (2006)
8. Hammer, S., Kirchner, K., André, E., Lugrin, B.: Touch or talk? Comparing social robots and tablet pcs for an elderly assistant recommender system. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 129–130. HRI 2017, ACM, New York, NY, USA (2017)
9. Hancock, P.A., Billings, D.R., Schaefer, K.E., Chen, J.Y., De Visser, E.J., Parasuraman, R.: A meta-analysis of factors affecting trust in human-robot interaction. *Hum. Factors* **53**(5), 517–527 (2011)
10. Heimerl, A., Weitz, K., Baur, T., Andre, E.: Unraveling ml models of emotion with nova: multi-level explainable AI for non-experts. *IEEE Transactions on Affective Computing*, p. 1 (2020)
11. Hoff, K.A., Bashir, M.: Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Factors* **57**(3), 407–434 (2015)
12. Hoffman, J.D., et al.: Human-automation collaboration in dynamic mission planning: a challenge requiring an ecological approach. *Proc. Hum. Factors Ergon. Soc. Ann. Meet.* **50**(23), 2482–2486 (2006)
13. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: challenges and prospects (2018)
14. Huber, T., Weitz, K., André, E., Amir, O.: Local and global explanations of agent behavior: Integrating strategy summaries with saliency maps. *CoRR abs/2005.08874* (2020)
15. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* **4**(1), 53–71 (2000)
16. Jin, S.A.A.: The effects of incorporating a virtual agent in a computer-aided test designed for stress management education: the mediating role of enjoyment. *Comput. Hum. Behav.* **26**(3), 443–451 (2010)
17. Karrer, K., Glaser, C., Clemens, C., Bruder, C.: Technikaffinität erfassen-der fragebogen ta-eg. *Der Mensch im Mittelpunkt technischer Systeme* **8**, 196–201 (2009)
18. Kisler, T., Reichel, U., Schiel, F.: Multilingual processing of speech via web services. *Comput. Speech Lang.* **45**, 326–347 (2017)
19. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. *Hum. Factors* **46**(1), 50–80 (2004)

20. Lepouras, G., Vassilakis, C.: Virtual museums for all: employing game technology for edutainment. *Virtual Real.* **8**(2), 96–106 (2004)
21. Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., Bhogal, R.S.: The persona effect: affective impact of animated pedagogical agents. In: *Proceedings of the conference on Human factors in computing systems CHI 1997*, pp. 359–366. ACM Press, Atlanta, Georgia, United States (1997)
22. Marsh, S., Dibben, M.R.: Trust, untrust, distrust and mistrust – an exploration of the dark(er) side. In: Herrmann, P., Issarny, V., Shiu, S. (eds.) *iTrust 2005*. LNCS, vol. 3477, pp. 17–33. Springer, Heidelberg (2005). https://doi.org/10.1007/11429760_2
23. Mayer, R.E., DaPra, C.S.: An embodiment effect in computer-based learning with animated pedagogical agents. *J. Exp. Psychol. Appl.* **18**(3), 239–252 (2012)
24. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2018)
25. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: beware of inmates running the asylum (2017)
26. Ming, Y., Ruan, Q., Gao, G.: A mandarin edutainment system integrated virtual learning environments. *Speech Commun.* **55**(1), 71–83 (2013)
27. Reich-Stiebert, N., Eyssel, F., Hohnemann, C.: Involve the user! changing attitudes toward robots by user participation in a robot prototyping process. *Comput. Hum. Behav.* **91**, 290–296 (2019)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: *Proceedings of the 22Nd ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, New York, NY, USA (2016)
29. Rutjes, H., Willemsen, M., IJsselsteijn, W.: Considerations on explainable ai and users’ mental models. In: *Where is the Human? Bridging the Gap Between AI and HCI*. Association for Computing Machinery Inc, United States (May 2019)
30. Sainath, T.N., Parada, C.: Convolutional neural networks for small-footprint keyword spotting. In: *Proceedings of Interspeech, 2015*, pp. 1478–1482. ISCA Archive, Dresden, Germany (2015)
31. Sanders, T., Oleson, K.E., Billings, D.R., Chen, J.Y.C., Hancock, P.A.: A model of human-robot trust: theoretical model development. *Proc. Hum. Factors Ergon. Soc. Ann. Meet.* **55**(1), 1432–1436 (2011)
32. Schneider, J., Handali, J.: Personalized explanation in machine learning: a conceptualization. *arXiv preprint [arXiv:1901.00770](https://arxiv.org/abs/1901.00770)* (2019)
33. Stubbs, K., Hinds, P.J., Wettergreen, D.: Autonomy and common ground in human-robot interaction: a field study. *IEEE Intell. Syst.* **22**(2), 42–50 (2007)
34. Van Mulken, S., André, E., Müller, J.: The persona effect: how substantial is it? In: Johnson, H., Nigay, L., Roast, C. (eds.) *People and computers XIII*, pp. 53–66. Springer, London (1998). https://doi.org/10.1007/978-1-4471-3605-7_4
35. Warden, P.: Speech commands: a dataset for limited-vocabulary speech recognition (2018)
36. Weitz, K., Hassan, T., Schmid, U., Garbas, J.U.: Deep-learned faces of pain and emotions: elucidating the differences of facial expressions with the help of explainable AI methods. *tm-Technisches Messen* **86**(7–8), 404–412 (2019)
37. Weitz, K., Schiller, D., Schlagowski, R., Huber, T., André, E.: “Let me explain!”: exploring the potential of virtual agents in explainable AI interaction design. *J. Multimodal User Interfaces* **15**(2), 87–98 (2021)