Journal on Multimodal User Interfaces (2021) 15:87–98
https://doi.org/10.1007/s12193-020-00332-0

ORIGINAL PAPER

"Let me explain!": exploring the potential of virtual agents in explainable AI interaction design

Katharina Weitz[1] · Dominik Schiller[1] · Ruben Schlagowski[1] · Tobias Huber[1] · Elisabeth André[1]

Received: 2 November 2019 / Accepted: 16 June 2020 / Published online: 9 July 2020

Abstract
While the research area of artificial intelligence benefited from increasingly sophisticated machine learning techniques in recent years, the resulting systems suffer from a loss of transparency and comprehensibility, especially for end-users. In this paper, we explore the effects of incorporating virtual agents into explainable artificial intelligence (XAI) designs on the perceived trust of end-users. For this purpose, we conducted a user study based on a simple speech recognition system for keyword classification. As a result of this experiment, we found that the integration of virtual agents leads to increased user trust in the XAI system. Furthermore, we found that the user's trust significantly depends on the modalities that are used within the user-agent interface design. The results of our study show a linear trend where the visual presence of an agent combined with a voice output resulted in greater trust than the output of text or the voice output alone. Additionally, we analysed the participants' feedback regarding the presented XAI visualisations. We found that increased human-likeness of and interaction with the virtual agent are the two most common mention points on how to improve the proposed XAI interaction design. Based on these results, we discuss current limitations and interesting topics for further research in the field of XAI. Moreover, we present design recommendations for virtual agents in XAI systems for future projects.

Keywords Explainable artificial intelligence · Interpretable artificial intelligence · Virtual agents · Human-agent interaction · Deep learning · Trust

1 Introduction

The research area of artificial intelligence benefited from increasingly sophisticated machine learning techniques in recent years. As an effect, a variety of use cases for these new technologies found their way into the everyday lives of a multitude of users. As an example, automatic speech recognition is already powering a new generation of voice assistants like Amazon's Alexa, Google's Assistant or Apple's Siri.

While those advancements are leading to improved and more intuitive ways of interacting with AI systems, the underlying algorithms are growing in complexity and therefore decreasing the system's comprehensibility. This is not only complicating matters for machine learning engineers and practitioners, who are working on improving the performance of their models, but also proposes some new challenges when it comes to end-user related human-computer interaction. Evidence suggests that a lack of transparency, with respect to the decisions of an AI system, might have a negative impact on the trustworthiness of a system. This lack of trustworthiness can also decrease the overall user-experience [13,35].

The reemerging research field of XAI [11] investigates approaches to address this problem. One goal of XAI is the development of innovative explanation algorithms which are promising to grant new insights into state of the art machine learning black box models, and thereby helping the user better understand and trust an AI system [4,16,19]. Many approaches are relying on visualisation techniques like saliency maps to highlight the parts of the input that were most relevant for the decision of a model. Although those efforts achieved remarkable progress in recent years, concerns have been expressed that the development of expla-

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12193-020-00332-0) contains supplementary material, which is available to authorized users.

✉ Katharina Weitz
weitz@hcm-lab.de

[1] Department of Computer Science, Human-Centered Multimedia, Augsburg University, Universitätsstraße 6a, Augsburg, Germany

nation methods has been focused too much on building solutions for AI-experts while neglecting end-users [21]. Weitz et al. [44] concluded that those approaches are not yet at a point where they can be utilized to benefit the user directly.

A potential next step that steers explainable artificial intelligence research into a more user-centered direction has been proposed by De Graaf and Malle [6]. They suggested that humans are approaching the explanations of an AI system with the same attitude of expectation, they are employing towards another human. Therefore the generation of explanations within the bounds of the conceptual and linguistic framework of human behaviour could greatly improve the transparency and explainability of AI systems towards end-users. Van Mulken et al. [38] additionally found out that personified agents can have positive effects on the perceived difficulty of processing technical information in human–computer interaction while not decreasing the overall objective performance of the user.

Similar results were found in Weitz et al. [45], where we showed that end-users trust an AI system for speech recognition more when a virtual agent is presenting XAI visualisations.

Based on the results of this study, our paper evaluates in detail which aspects of a virtual agent are relevant to support XAI visualisations. To this end, we focus on assessing the effect of different levels of anthropomorphism/human-likeness of an agent (voice, visualisation, and the content of what is said).

For this evaluation we conducted a user study in which a virtual agent presented XAI visualisations to users of a simple Artificial Neural Network (ANN)-based speech recognition model, which classifies audio keywords based on visual representations of the audio signal (spectrograms). To this end, we split the participants into three groups, which interacted with different versions of the same virtual agent (text, voice, and visual presence), and a baseline group without a virtual agent. We are aiming to examine the following research questions:

1. Does the usage of a virtual agent positively impact the perceived trustworthiness of AI systems like deep neural networks?
2. Which of the three modalities of a virtual agent that we tested (pure information in form of text, voice, and visual presence) are important for an impact on the perceived trustworthiness of an AI system?
3. How are the presented XAI visualisations perceived and rated by users?
4. How does the use of virtual agents affect the perception of the presented XAI visualisations?

In order to answer the first and second research question, we formulated a directional hypothesis that is evaluated within the scope of a contrast analysis. To calculate the effect size, we used the recommendations for contrast analyses from Perugini et al. [24].

For our hypothesis we assume a linear trend, which means that the general trust increases depending on the virtual agent group where the baseline group without agent has the lowest general trust score, followed by the text agent group, the voice agent group, and the virtual agent group with the highest scores in general trust.

The third and fourth research question will be evaluated qualitatively as well as quantitatively by performing an ANOVA to determine the impact of the different virtual agent modalities on the rating of XAI visualisations of the participants.

Overall, our paper contains three contributions:

1. We present a novel XAI interaction design where we employ a virtual agent to present XAI visualisations for a simple ANN-based speech recognition model which classifies audio keywords.
2. We conducted a user-study to empirically verify the impact of the human-likeness of a virtual agent on the helpfulness of XAI visualisations and perceived trust in the system.
3. Based on the results of this study we are presenting suggestions to improve the integration of virtual agents in XAI interaction designs.

The remainder of this paper is divided into six sections. Section 2 introduces the related work with respect to current state of the art XAI systems and explanations from the standpoint of human-like interactions. In Sect. 3 we describe the implemented speech recognition system and the XAI algorithm LIME [26] that we used for our study. Then in Sect. 4 we present our experimental setup in detail, followed by the results of our study in Sect. 5. Finally, we are closing the paper with a discussion of results in Sect. 6 as well as a final conclusion and an outlook for future work in Sect. 7.

2 Related work

In this section we are presenting an overview of related work in the area of explainable AI. We split this overview in two subsections regarding the technical aspects of XAI as well as the human interactive aspects of explanations in general.

2.1 Explainable AI

Current state of the art approaches for artificial intelligence are increasingly relying on the deployment of Machine

Learning (ML) models across a wide range of application areas. Specifically, Deep Learning methods, which are able to automatically learn abstract high level features from low-level or even raw input, are continuously growing in popularity. On the one hand such models are often able to push results further beyond the state of the art in human-computer-interaction related prediction problems like automatic speech recognition [47], activity recognition [42], or sensing a users affective state [41]. On the other hand, they are confronted with a trade-off between accuracy and comprehensibility, where the inner workings of the applied models are becoming increasingly more complex and therefore opaque, which makes them impossible to comprehend for humans [29].

This problem is being addressed by recent research within the realm of XAI and interpretable AI. Since both terms are often used synonymously [10], we will subsume work from both fields under the term XAI. [18] pointed out that there are varying definitions for the exact goal of XAI in the literature. For this paper, we adopt the view of Gilpin et al. [10], who state that the goal of XAI is the description of a system, its internal states, and its decisions in such a way that this description can be understood by human beings.

We focus on XAI approaches that attempt to shed light on specific decisions of incomprehensible ML models by visually highlighting parts of the input data according to their relevance for that decision. XAI visualisations are suitable for classifications on the basis of visual input, which is easier to interpret for humans than raw data [22]. One of the first approaches for creating XAI visualisations was to use gradients with respect to the input [34] to measure how much a small change in each input part would affect the prediction. Later, the Grad-CAM algorithm [31] used gradients with respect to intermediate results of the prediction model to achieve the same goal. Another method to determine relevance is Layerwise-Relevance Propagation [2], which uses the intermediate results of the model during the prediction to calculate the contribution of each part of the input to the overall prediction. While these concepts can theoretically be used on various machine learning models, the deployed algorithms are often optimized for neural networks and require adjustments for other models. A different take on the explanation of opaque machine learning models is proposed by Ribeiro et al. [26] in the form of the LIME framework. In contrast to the aforementioned methods, model-agnostic approaches like LIME have the advantage of being universally applicable, independent of the input data or the utilized machine learning algorithm. LIME addresses the task of generating XAI visualisations by training a simple machine learning model to approximate the underlying model locally. To accomplish this, the image to be explained is divided into superpixels by a segmentation algorithm. These superpixels are then randomly grayed out to create a new dataset with perturbed sample instances. This dataset is used to train a more explainable model, which uses the presence of superpixels as input, to predict the decisions of the original model. By analyzing this explainable model, the effect that each superpixel has on the overall prediction of the original model can be assessed.

2.2 Explanations and human-like interactions

Recent research indicates that modern XAI approaches, which are aiming at unraveling the non-linear maze of state of the art machine learning models, are not quite meeting the requirements to impart enough insightful information to the end-user.

While Ribeiro et al. [26] used three different tasks to demonstrate the usefulness of LIME for human end-users, each of these tasks was tailored to show a specific problem. It is unclear whether these results generalize to more complex problems. Layerwise-Relevance Propagation and Grad-CAM were evaluated in more general settings. Alqaraawi et al. [1] conducted a user study investigating the usefulness of Layerwise-Relevance Propagation visualisations on the 2012 version of the PASCAL Visual Object Classes dataset and Selvaraju et al. [32] used the 2007 version of the same dataset to evaluate Grad-CAM. In both studies, the participants were able to correctly guess the model's prediction based on the XAI visualisations about 60 % of the time (60.07% for LRP, 61.23% for Grad-CAM). While this was better than the respective comparison groups, it is not a good result by itself. This indicates that there is potential for XAI visualisations but that they are not yet on a level where they can be used on their own.

A promising way to increase the accessibility of XAI visualisations is to incorporate them into a human-like interaction. De Graaf and Malle [6] hypothesized that people are applying human traits to AI systems and will, therefore, expect explanations within the conceptual and linguistic framework used to explain human behaviours. They argue that people are more likely to form a correct mental model of an AI system and recalibrate their trust in the system if it communicates explanations in a human-like way. The challenging nature of this task is reflected by the recent growth in research addressing that topic. Richardson and Rosenfeld [27] are proposing a taxonomy of interpretability to answer the questions of why-, what-, how-, and when a machine learning driven human-agent system should generate an explanation for the user. However, they focused on self-explaining agents instead of investigating how agents can improve existing XAI methods.Broekens et al. [3] investigated the effect of different types of explanations for agents that are driven by a belief-desire-intention model (BDI-Agents). Similar to the work presented in this paper, they conducted a user study to investigate explainability in the context of BDI-Agents. As a result of their study, they came up with a set of guidelines for the future development of

explainable BDI-Agents. For example, they point out that the overall goal an agent aims to achieve should be presented transparently and that certain additional information is required, depending on the type of action the agent takes.

Miller [20] conducted an interdisciplinary survey with the goal of exploring ways to define, generate, select, present, and evaluate explanations with a focus on XAI. He came up with four major findings regarding the properties of explanations in human-like interactions:

- `Explanations are contrastive`: People tend to not ask why something happened, but rather why something happened instead of something else. They are therefore implicitly creating a reference between the actual occurrence of an event and their own expectations.
- `Explanations are selected`|: People are rarely expecting an explanation to be covering all potential causes. An explanation is rather communicated by highlighting one specific reason.
- `Probabilities probably don't matter`|: Causal explanations are more important than pure correlations. Therefore, explanations with the highest probabilities are not necessarily the best explanations for a user.
- `Explanations are social`|: Explanations are a transfer of knowledge as part of a conversation or interaction. This also involves queries by the interlocutor that receives the explanation as well as adaption to this his or her preferences (e.g., style of communication or available background knowledge).

Based on these works we argue that one should take inspiration from the human explanation process when developing XAI interaction designs.

3 Keyword classifier and explainable AI implementation

Out of the XAI approaches previously discussed in Sect. 2, we chose the LIME framework by Ribeiro et al. [26] to explain the automatic recognition of spoken keywords within our user study. The underlying algorithm creates XAI visualisations for any classification or regression based system by approximating predictions locally with an an explainable model.

As we stated before, the LIME algorithm only creates interpretable visualisations. However, it is often difficult to extract meaningful information from the visual representation of raw audio data (i.e., a wave form). We therefore chose to present our XAI visualisations in the form of highlighted spectrograms (see Fig. 1). Spectrograms are visual representations (images) of audio samples and display sound pressure

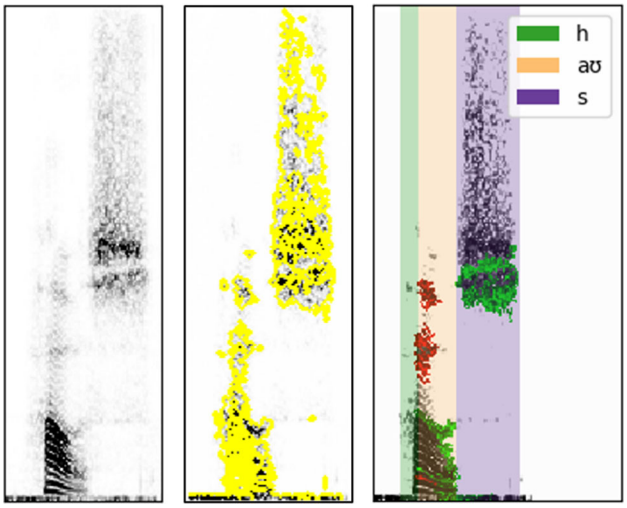

Fig. 1 A spectrogram of an audio sample (left), its segmentation into superpixels (center) and the output for the user containing LIME visualisations and additional phoneme information (right)

levels as pixel values over the dimensions time (x-axis) and frequency (y-axis). Figure 1 illustrates the spectrogram for the spoken input word 'house' on the left as an example. These spectrograms are calculated from the respective audio signal and used as input for our classification model.

As prediction model we used the neural network architecture proposed by Sainath and Parada [28]. This ANN-based classification model uses a convolutional neural network to generate abstract features based on mel-frequency cepstrum coefficients (MFCCs) which are derived from the spectrograms of the raw audio wave forms. These features are then fed into a fully-connected layer which finally predicts the target class, which is one of the keywords (labels) of the training dataset (see Fig. 2).

We trained our model on the speech command dataset provided by Warden [43]. This dataset consists of instances from 35 different spoken words and was specifically designed to train and evaluate audio keyword classification systems. The comparably high ratio of samples per class to the overall number of classes and the high variance with respect to speakers and sound-quality, enabled us to train a fairly robust model for our specific use case.

To generate a visual explanation for a specific prediction (keyword) of our classification model, the input-spectrogram is first segmented by the Felzenszwalb's algorithm for image segmentation [7] (see Fig. 1, centered image) into so-called superpixels. Subsequently, the generated superpixels are randomly greyed out to conceal the visual information they contain. Afterwards, a more explainable model is trained to predict the decisions of the original model on those perturbed images based on a binary vector that encodes which superpixels were grayed out in the input image. Analyzing this new model enables us to assess the effect that each superpixel has

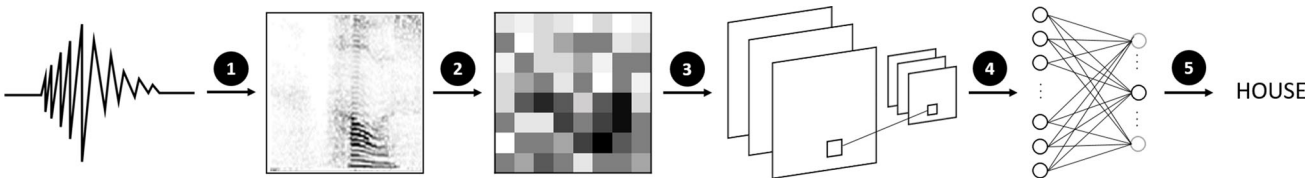

Fig. 2 XAI visualisation of the spoken keyword "seven". With every XAI visualisation the predicted label and the prediction accuracy of the speech recognition system were displayed

on the overall prediction of the model. Finally, superpixels that are found to have a significant impact in favor of a specific label are highlighted green for the user, whereas red highlighted segments speak against the predicted label (see Fig. 1, right image).

To further enhance the explainability of the LIME visualisations, we are also presenting a phoneme based segmentation of the input-word to the user. Phonemes are small units of sound that can be used to distinguish one word from another. Therefore they are particularly well suited to assist with the establishment of a relation between the way humans understand spoken language and the visualisations provided by our system. The phoneme segmentation of the spectrogram is generated through the WebMAUS tool developed by Kisler et al. [15]. An example of this segmentation for the spoken word 'house' can be seen in Fig. 1 in the right image.

4 Study

To investigate the effect of agents in combination with XAI visualisations, we conducted a user study with 60 participants. Each participant was given the same ten prescribed English keywords (i.e., dog, four, happy, core, on, right, eleven, two, seven, cat) to speak into our speech recognition system. Only eight of those keywords were part of the training data, whereas the remaining two words (i.e., core and eleven) were unknown to the classification system and would therefore be wrongly classified for sure. The intention behind this was to verify that the generated explanations help the user understand both correct and incorrect predictions. In order to reduce statistical deviations of the prediction model and the explanation framework we chose keywords which we found to reliably produce comprehensible explanations in advance. Prior to the test, the supervisor introduced the simple graphical user interface (GUI) to the participants and a textual cover story provided detailed instructions about how to read the systems' explanations and spectrograms. Then, every participant interacted with the GUI and spoke a predefined and fixed sequence of the ten chosen keywords into a microphone. After each recording, the audio data was classified by the model and a XAI visualisation for this classification was displayed together with the predicted label and

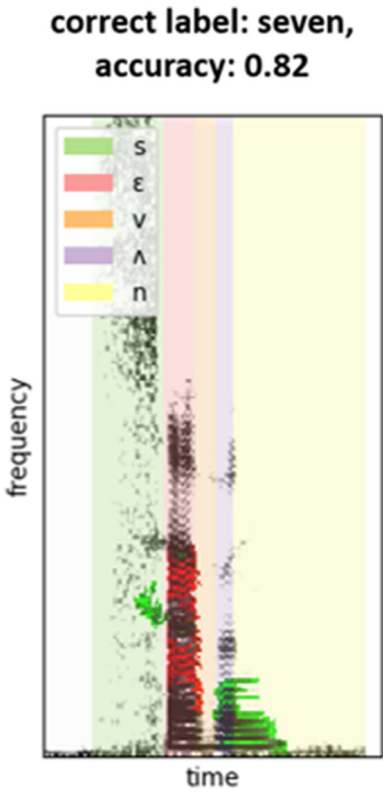

Fig. 3 Modalities of the conducted user study. Four different groups received information to understand the prediction of a speech recognition system

the prediction accuracy of the speech recognition system (see Fig. 3). For wrong classifications, the XAI visualisations for the three predictions with the highest probability were presented. Before continuing to the next keyword, the participants rated the helpfulness ('not helpful', 'helpful', and 'don't know') of the XAI visualisation in a questionnaire. To examine the influence of the human-likeness of a virtual agent on the XAI interaction design, some participants received information by a virtual agent in addition to the XAI visualisations. To this end, we split the 60 participants evenly into four test groups of 15: text agent group (only textual information), voice agent group (only information via voice), virtual embodied agent group (visual presence and voice), and a no agent group (see Table 1). The no agent group received only the XAI visualisations without further commentary. The

Table 1 Demographic information of the participants

Characteristic	Agent Text	Voice	Virtual	No Agent
n	15	15	15	15
Age				
M	25.7	25.0	28.2	27.27
SD	3.99	5.6	8.6	5.19
Gender				
Male	12	11	10	12
Female	3	4	5	3
Experience				
Voice assistants	11	13	10	8
Audio processing	5	4	7	5
Virtual agents	5	4	6	6

other three groups received additional information in varying modalities from a virtual agent named Gloria (see Fig. 4). The information given by the agent was selected dynamically from a set of phrases that were designed by our team in advance. These phrases were designed to communicate the following information:

- Acknowledgement of user inputs, e.g., "Ok the system got that!"
- Comments on the prediction accuracy of the neural network, e.g., "The system was pretty sure you said seven!"
- Comments on important phonemes within the output of the XAI framework, e.g., "Phoneme number two was found to have a particularly positive effect towards the prediction."

The text agent group received only the textual output of Gloria's comments in a separate GUI. The voice agent group, in contrast, received the same information via text-to-speech provided by Amazon Polly.[1] The third group saw, in addition to the speech output, the virtual presence of a 3D-character designed by the Charamel GmbH,[2] which lip-synced the phrases and performed body gestures while communicating (see Fig. 5).

After the experiment, all participants rated their impression of and their trust in the system and answered the Trust in Automation (TiA) questionnaire [14]. Additionally, we used a combination of 7-point Likert scales and open form questionnaires to collect qualitative and quantitative user feedback.[3] Furthermore, the user's individual impressions

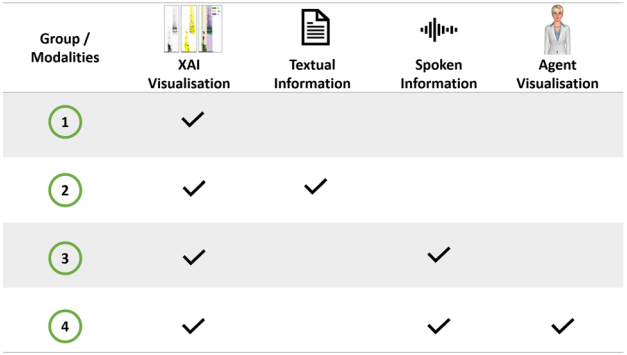

Fig. 4 Schematics of the used speech recognition system. (1) A spectrogram is generated from the raw audio wave form. (2) The spectrogram is used to calculate 20 MFCCs. (3) The MFCCs are fed into a convolutional neural network. (4) The learned features are then forwarded to the fully connected layers of the network. (5) Finally the output of the network is mapped to the corresponding target class

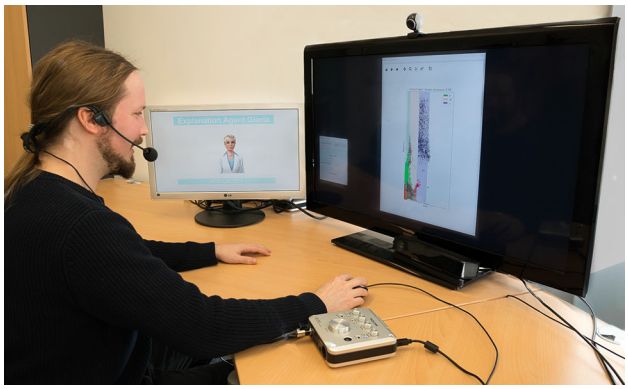

Fig. 5 Setup of the experiment for participants in the virtual embodied agent group

of Gloria were queried, if the participant was part of one of the virtual agent groups. The participants rated how they perceived Gloria in terms of her helpfulness (i.e., "The information Gloria gave me helped me to understand the decisions of the system"), comprehensibility (i.e., "Gloria's answers are understandable"), trustworthiness (i.e., "Gloria is trustworthy"), interaction (i.e., "I would interact with Gloria again), and likability (i.e., "I liked Gloria"). Participants of the text agent group also were asked to asses how often they had read the text information of Gloria on a 7-point Likert scale (1 = never, 7 = always). The results of our study will be presented in the next section of this paper.

5 Results

In this section, we describe the results of our study starting with a comparison of trust values between the different test-groups. To calculate the required sample size for the test-group comparison, we performed an a-priori power analysis.

[1] https://aws.amazon.com/de/polly/.

[2] https://vuppetmaster.de/.

[3] The translated version of the german questionnaire can be found in supplementary material.

With a desired power of 0.80, an alpha value of 0.05 and an effect size of 0.45 (based on the large effect size resulted in Weitz et al. [45]), we calculated a required sample size of 60, which would result in a expected power of 0.82. After evaluating the results, the actual effect size of 0.42 showed that an actual power of 0.75 was achieved. In addition to the group comparison, we report the evaluation of our virtual agent Gloria, followed by the ratings and the feedback for the XAI visualisations.

5.1 Test-group comparison on trust

To answer our first and second research question, we evaluated the general trust value by examining the data from the TiA questionnaire using a contrast analysis, depending on the hypothesis stated in Sect. 1. Contrast analysis is a specific way of analysis for testing directional hypotheses (planned contrasts), that uses linear contrast coefficients to weight the means of the groups that are compared. This method offers insights into group differences as it gives the possibility to define specific and more precise comparisons between groups. We specifically chose this methodology, since it leads to a higher power, makes post-hoc testing obsolete and the effect-sizes are easier to interpret. The results of our contrast analysis showed a linear trend $R^2 = .16$, $F\ (3,56) = 3.45$, $p = .02$, indicating that as the human-likeness of the agent increases, general trust increased proportionately.

The planned contrast revealed that the human-likeness significantly increased in the text agent ($M = 4.89$, $SD = 0.95$), voice agent ($M = 5.12$, $SD = 0.79$), and virtual embodied agent group ($M = 5.42$, $SD = 0.69$), compared to the no agent group ($M = 4.48$, $SD = 0.86$), $b = .68$, $(t) = 3.19$, $p = .001$, $f = 0.42$ (medium effect).[4]

These findings support our hypothesis about a linear trend of the observed user trust regarding the chosen modalities, rising from no agent group over text and speech groups up unto virtual embodied agent group.

5.2 Agent evaluation

Second, we analysed how the agent Gloria was perceived by the participants in the three groups with agent (text agent, voice agent, and virtual embodied agent). The evaluation of the agent Gloria covered the following areas: sympathy, repeated interaction, trustworthiness, comprehensibility of her statements and helpfulness in understanding the system's decision (see Fig. 6). Participants evaluated each area on a 7-point Likert scale (1 = disagree, 7 = fully agree). For each item Gloria received the lowest average rating by the participants of the text agent group. For being comprehensible, trustworthy, and likable Gloria received the highest average ratings from the voice agent group. Participants in the voice agent group also most often wanted to interact with Gloria again. The highest rating for Gloria being helpful was given by the virtual embodied agent group.

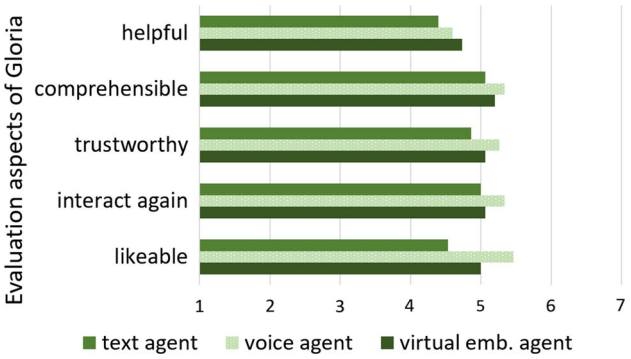

Fig. 6 Evaluation of five different aspects of the virtual agent Gloria. The rating was scaled between 1=disagree to 7=fully agree

As a result of the evaluation of the open questions, two areas were found to be assessed positively by the participants:

- Appearance of the virtual agent: Facial expressions, voice and gestures were emphasized as appealing.
- Interactions with the virtual agent: The participants indicated that they found verbalization of the visualisation (e.g., the reference to relevant phonemes) supportive.

Participants within the embodied agent group mentioned that the body gestures of Gloria (e.g., pointing on the spectrogram) were perceived as helpful to draw attention to the XAI visualisation.

5.3 Evaluation of explanations

To answer the third and fourth of our research questions, the participants gave feedback at the end of the study as to whether the given XAI visualizations were sufficient and which aspects or further explanations they would find helpful. The ANOVA reveals that the difference between the four groups were not significant, $F\ (1, 58) = 0.47$, $p = .495$, which means the ratings of the LIME visualisations do not differ between the four groups. Figure 7 displays the ratings of the participants on whether the given XAI visualisations were sufficient. Additionally, it can be seen that the average ratings of each group did not reach values above 5. This shows that there is still room for improvement within the XAI methods used in our study.

Many participants stated that they would have found detailed information in linguistic form (see some examples in Table 2) and comparative information helpful. Here, visual as well as linguistic comparisons were mentioned by the par-

[4] For calculating the effect size, we used the recommendations for contrast analyses from Perugini et al. [24]

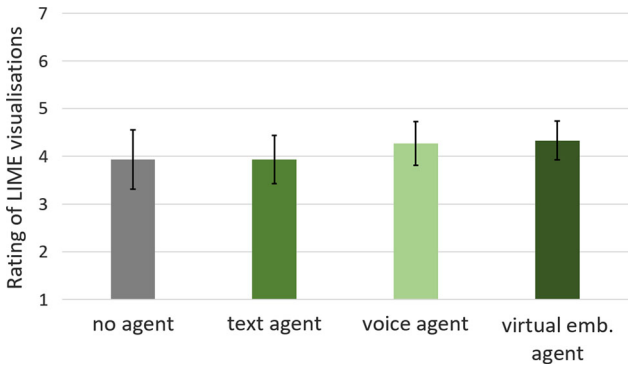

Fig. 7 Rating of the participants whether the displayed XAI visualisations were sufficient. The rating was scaled between 1 = disagree to 7 = fully agree. Error bars represent the standard error

ticipants. Also the analysis of the feedback suggests that participants would have liked to see more interaction with the virtual agent as well as with the XAI visualisations (e.g., clicking on superpixels or a label to get more detailed information).

6 Discussion

The primary goal of our user study was to examine whether a user interface featuring a virtual agent has a positive effect on the perceived trustworthiness of an ANN-based classification model for an end-user. Here, we investigated whether the modalities (pure information in form of text, voice, or visual presence), that were chosen for the communication of the classifier's prediction results and their XAI visualisations, had a significant impact on the perceived trustworthiness. Furthermore, we examined the overall perceived quality of the generated XAI visualisations. For this purpose, we analysed additional feedback from the four different participant groups.

Within the first subsection, we discuss our findings regarding perceived user trust. In the second subsection, we discuss the free-form feedback of the participants as well as our findings regarding the effects of our virtual agents on user ratings of the XAI visualisations.

6.1 Agent-user interface design and perceived trust

Examining the results of our study, we were able to empirically verify our hypotheses that

- The users' trust in an ANN-based classification model benefited from additional text output given by a virtual agent.
- The users' trust in an ANN-based classification model benefited from speech output provided by the virtual agent compared to text output.
- The users' trust in an ANN-based classification model benefited from the provided visual presence of a virtual agent performing additional lip-synchronisation and body gestures compared to raw speech output.

Our results are contrasting the study by Van Mulken et al. [39], in which no significant increase in trustworthiness through the personification of user interfaces could be determined. They argued that this might have been caused by an insufficient quality of virtual agents at that time. This suggestion provides a possible explanation for our deviating result, since the advancements in technology enabled us to employ a more lifelike and realistic virtual agent in our study. This is reflected in the ratings of our agent, which are all well above average in the voice agent and virtual embodied agent group (see Fig. 6).

Our study examined the relationship between the "human-likeness" of a virtual agent and how this influences perceived user trust. The overall impression from our results is that the more human-like XAI interactions appear, the more the users tend to trust the classification model whose predictions are explained. As virtual embodied agents offer simulated human-like behaviour, such as lip synchronization and body language along with speech output, their potential for trust-oriented XAI interaction-design seems intuitive, but it was not yet verified prior to this study. Our study gives first indications that design choices of a virtual agent influence humans' trust in an AI system. This information is a crucial step in establishing appropriate trust [17] in AI systems in the future. Knowing the means by which a user's trust can be influenced might help to increase awareness towards such methods.

However, when analysing trust one has to be careful, since trust is a complex concept that can be influenced by various aspects. Hoff and Bashir [12] presented a three-layered framework, consisting of dispositional trust, situational trust, and learned trust. In our study we focused primarily on the situational trust which is strongly dependent on the situational context. This context is further divided into external and internal factors. External factors include task difficulty (i.e., spectrograms), the type of system (i.e., text-, voice-, virtual embodied agent vs. no agent), and system complexity (i.e., ANN). Among others, internal factors include subject matter (e.g., background in signal processing) and self-confidence. While influences attributable to dispositional and learned trust were not explicitly addressed in our study, these could be used in further work to make more precise statements about perceived trust.

Table 2 Evaluation of the LIME explanations

Kind of information	Example feedback of the users
Linguistic information	"Detailed answers for wrong words"
	"A verbal explanation of why some sounds were not understood"
	"Explanations for the individual case, if something is not recognized and what exactly the problem was."
	"To tell me which phoneme had a very beneficial effect on the prediction, this could be used more."
	"How does the system work in the background?"
Comparative information	"Comparisons of similar sounding words"
(visual & linguistic)	"In case of incorrect predictions, additional windows with analysis of the correct label."
	"More detailed explanation of what should be heard and what was actually heard (in the diagram)."
	"In case of wrong classification also visualisation of the actual class would be helpful."
	"It is not clear what the word would look like if it were spoken perfectly."

Answers from participants to the question which further explanations they would have found helpful

6.2 XAI visualisation feedback

Besides the impact of virtual agents on the perceived trustworthiness of the ANN-based classification model, we wanted to investigate (1) how the presented XAI visualisations are perceived and rated by participants and (2) how virtual agents affect this perception of XAI visualisations. We found that

- Participants wanted more information in linguistic form.
- Participants asked for comparative information in visual as well as linguistic form.
- Participants would have preferred further interaction with the system (e.g., to ask questions).

As the ratings of the visual explanations were not particularly high (average around 4 with a maximum rating of 7), there is still a high potential for improvement regarding the visual explanations we used in our experiment. A cause for this may be the complexity of the visual explanations, as they require some basic understanding of spectrograms and how to read them.

From the results of our study, a tendency can be observed regarding the participants' rating of the quality of the XAI visualisations (see Fig. 7), where the no agent and text agent group rated the XAI visualisations as less sufficient than participants in the voice agent and virtual embodied agent group. This result reflects the findings on user's trust towards the system discussed in the previous subsection. A possible cause for this might be a cognitive bias such as the halo effect [37]. The halo effect states that a positive impression of a person about an object in one area positively influences their opinion in other areas. In our study, the perceived trustworthiness of the ANN-based classification model could have positively influenced the ratings of the participants towards the XAI visualisations. The aforementioned observation provides first indications that cognitive biases may occur during interaction with XAI systems. Whether and to what extent cognitive biases influence the perception of XAI should therefore be the focus of further studies.

A result from our free-form feedback showed that participants wished for more linguistic explanations. This aligns with the social characteristic of explanations found by Miller [20], since it underlines the participants need for selective information and causality within the explanation. Our simple implementation of linguistic explanations in the text, voice, and virtual embodied agent groups, which highlights the most relevant phoneme to the user, already illustrates the usefulness of this concept. This corresponds to the findings of Siebers and Schmid [33], who suggested that adding textual explanations can redirect the focus of the user towards important areas, and of Schmid [30], who pointed out that additional textual explanations enable the inclusion of causal relations among other information. Park et al. [23] introduced a concept to generate such explanations for a visual question answering system by using recurrent neural networks to generate textual explanations based on an input image, a question, and visual explanations of the predicted answer. In the same way one could use the visual explanations we implemented in this paper to generate additional linguistic explanations for the agent which correspond to the specific input.

In addition to linguistic explanations, the supplementary use of advanced body gestures could help the agent to point

at certain regions of the visualisation more precisely and thus simulate a more natural behaviour. To achieve this, one could build up on the already existing body of work that addresses the topic of automatic gesture generation [5,9,25].

Another aspect that emerged from the evaluation of the free form questionnaire was that the participants wanted information that was prepared in such a way that particularly intuitive comparisons could be made. A possible cause for this might be the specific way of integrating XAI visualisations in our system, which does not show the visualisation of the correct keyword in the case of misclassification. Instead, we displayed only three visualisations corresponding to the top predictions of our classifier. In some cases those visualisations did not contain the word that was actually spoken by the participant. Here, participants missed additional information which would have enabled them to interpret the explanation in the correct context. This insight supports the thesis of Miller [20], according to which people prefer to ask why one prediction was made instead of another. To enable such a comparison, an explanation design could benefit from additionally displaying example explanations of inputs that have been classified correctly.

The participants feedback suggest that they would have preferred to interact more with the system, for example, to ask questions when they do not understand something. This insight corresponds to Miller's findings stating that explanations have social characteristics since they represent a transfer of knowledge in the context of conversations [20]. Conversations are one of the most important ways for people to exchange and share knowledge [8] and therefore are one of the main characteristics of human-to-human explanations. This characteristic has also been investigated in human-computer interaction by Susan Robinson and Henderer [36]. They found that users most often reacted to utterances of a conversational agent with queries. It would be interesting to experiment with the application of more mature conversational agent architectures in this area, since those should be able to to respond adequately to questions and also to deal with queries from the user. Modern neural network based architectures like the ones proposed by Wu et al. [46] and Vinyals and Le [40] are already enabling natural user adaptive conversations with a virtual agent. Combining such conversational capabilities with the textual explanation approaches, like the one by Park et al. [23] we mentioned before, could lead to a more natural interaction and improved transfer of knowledge.

7 Conclusion

Within this paper we explored the potential of virtual agents to explain the decisions of an ANN-based classification model to end-users. To this end, we conducted a user-study in which we presented XAI visualisations of the decisions from a speech recognition system to the user. While one test group only received the XAI visualisations, three test groups were presented additionally different modalities of a virtual agent (text, voice, or virtual presence). The results of our study show a linear trend of the user's perceived trust in the used ANN-based classification model regarding the chosen modalities, rising from no-agent group over text and speech groups up unto virtual embodied agent group. By analyzing the participants' free-form feedback, we additionally found that:

- End-users want additional linguistic explanations.
- End-users want explanations to be suitable for intuitive comparisons.
- End-users want to interact with the agent, e.g., by asking questions.

The results of our study are inline with our initial assumption that the end-users' experience could benefit from a more human-like XAI interaction design. Based on our findings, we argue that there lies vast potential in the use of virtual agents to achieve this design goal.

Acknowledgements Open Access funding provided by Projekt DEAL. This work has received funding from the DFG under project number 392401413 (DEEP) and from the BMBF within the project "VIVA", Grant Number 16SV7960.

Compliance with ethical standards

Conflicts of interest The authors declare that they have no conflict of interest.

References

1. Alqaraawi A, Schuessler M, Weiß P, Costanza E, Berthouze N (2020) Evaluating saliency map explanations for convolutional neural networks: a user study. arXiv:2002.00772
2. Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10(7):e0130140. https://doi.org/10.1371/journal.pone.0130140

3. Broekens J, Harbers M, Hindriks K, Van Den Bosch K, Jonker C, Meyer JJ (2010) Do you get it? user-evaluated explainable BDI agents. In: German conference on multiagent system technologies. Springer, pp 28–39
4. Chen JYC, Procci K, Boyce M, Wright J, Garcia A, Barnes MJ (2014) Situation awareness-based agent transparency. US Army Research Laboratory
5. Chiu CC, Marsella S (2011) How to train your avatar: a data driven approach to gesture generation. In: International workshop on intelligent virtual agents. Springer, pp 127–140, https://doi.org/10.1007/978-3-642-23974-8_14
6. De Graaf MMA, Malle BF (2017) How people explain action (and autonomous intelligent systems should too). In: AAAI 2017 fall symposium on AI-HRI, pp 19–26
7. Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vis 59(2):167–181. https://doi.org/10.1023/B:VISI.0000022288.19776.77
8. Garrod S, Pickering MJ (2004) Why is conversation so easy? Trends Cogn Sci 8(1):8–11. https://doi.org/10.1016/j.tics.2003.10.016
9. Gatt A, Paggio P (2014) Learning when to point: a data-driven approach. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers, pp 2007–2017
10. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an approach to evaluating interpretability of machine learning. arXiv:1806.00069
11. Gunning D (2017) Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA)
12. Hoff KA, Bashir M (2015) Trust in automation: integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434. https://doi.org/10.1177/0018720814547570
13. Hoffman JD, Patterson MJ, Lee JD, Crittendon ZB, Stoner HA, Seppelt BD, Linegang MP (2006) Human-automation collaboration in dynamic mission planning: a challenge requiring an ecological approach. In: Proceedings of the human factors and ergonomics society annual meeting, vol 50(23), pp 2482–2486. https://doi.org/10.1177/154193120605002304
14. Jian JY, Bisantz AM, Drury CG (2000) Foundations for an empirically determined scale of trust in automated systems. Int J Cognit Ergon. https://doi.org/10.1207/S15327566IJCE0401_04
15. Kisler T, Reichel U, Schiel F (2017) Multilingual processing of speech via web services. Comput Speech Lang 45:326–347. https://doi.org/10.1016/j.csl.2017.01.005
16. Lane HC, Core MG, Van Lent M, Solomon S, Gomboc D (2005) Explainable artificial intelligence for training and tutoring. University of Southern California/Institute for Creative Technologies, Tech. rep
17. Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80
18. Lipton ZC (2018) The mythos of model interpretability. Commun ACM 61(10):36–43. https://doi.org/10.1145/3233231
19. Mercado JE, Rupp MA, Chen JY, Barnes MJ, Barber D, Procci K (2016) Intelligent agent transparency in human-agent teaming for multi-UxV management. Hum Factors 58(3):401–415. https://doi.org/10.1177/0018720815621206
20. Miller T (2018) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007
21. Miller T, Howe P, Sonenberg L (2017) Explainable AI: beware of inmates running the asylum. In: IJCAI International joint conference on artificial intelligence, arXiv:1712.00547
22. Montavon G, Samek W, Müller KR (2017) Methods for interpreting and understanding deep neural networks. Digit Signal Process 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011
23. Park DH, Hendricks LA, Akata Z, Rohrbach A, Schiele B, Darrell T, Rohrbach M (2018) Multimodal explanations: Justifying decisions and pointing to the evidence. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pp 8779–8788, https://doi.org/10.1109/CVPR.2018.00915
24. Perugini M, Gallucci M, Costantini G (2018) A practical primer to power analysis for simple experimental designs. Int Rev Soc Psychol 31(1):20. https://doi.org/10.5334/irsp.181
25. Ravenet B, Clavel C, Pelachaud C (2018) Automatic nonverbal behavior generation from image schemas. In: Proceedings of the 17th international conference on autonomous agents and multiagent systems, pp 1667–1674
26. Ribeiro MT, Singh S, Guestrin C (2016) Why should i trust you? Explaining the predictions of any classifier. In: Proceedings of the 22Nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1135–1144, https://doi.org/10.1145/2939672.2939778
27. Richardson A, Rosenfeld A (2018) A survey of interpretability and explainability in human-agent systems. In: Proceedings of the 2nd workshop of explainable artificial intelligence, pp 137–143
28. Sainath TN, Parada C (2015) Convolutional neural networks for small-footprint keyword spotting. Proc Interspeech 2015:1478–1482
29. Samek W, Wiegand T, Müller KR (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 pp 1–8
30. Schmid U (2018) Inductive programming as approach to comprehensible machine learning. In: Proceedings of the 7th workshop on dynamics of knowledge and belief (DKB-2018) and the 6th workshop KI & Kognition (KIK-2018), co-located with 41st German conference on artificial intelligence, vol 2194
31. Selvaraju RR, Das A, Vedantam R, Cogswell M, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: The IEEE international conference on computer vision (ICCV) 2017, pp 618–626
32. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359
33. Siebers M, Schmid U (2018) Please delete that! why should I? Explaining learned irrelevance classifications of digital objects. KI - Künstliche Intelligenz. https://doi.org/10.1007/s13218-018-0565-5
34. Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. http://arxiv.org/abs/1312.6034, arXiv:1312.6034
35. Stubbs K, Hinds PJ, Wettergreen D (2007) Autonomy and common ground in human-robot interaction: a field study. IEEE Intell Syst 22(2):42–50. https://doi.org/10.1109/MIS.2007.21
36. Susan Robinson MI David Traum, Henderer J (2008) What would you ask a conversational agent? observations of human-agent dialogues in a museum setting. In: Proceedings of the 6th international conference on language resources and evaluation (LREC'08), European Language Resources Association
37. Thorndike EL (1920) A constant error in psychological ratings. J Appl Psychol 4(1):25–29
38. Van Mulken S, André E, Müller J (1998) The persona effect: how substantial is it? In: People and computers XIII. Springer, pp 53–66, https://doi.org/10.1007/978-1-4471-36057_4
39. Van Mulken S, André E, Müller J (1999) An empirical study on the trustworthiness of life-like interface agents. In: Human–Computer interaction: communication, cooperation, and application design, proceedings of 8th international conference on human–computer interaction, 1999, pp 152–156
40. Vinyals O, Le QV (2015) A neural conversational model. arXiv preprint arXiv:1506.05869

41. Wagner J, Schiller D, Seiderer A, André E (2018) Deep learning in paralinguistic recognition tasks: are hand-crafted features still relevant? Proc Interspeech 2018:147–151
42. Wang J, Chen Y, Hao S, Peng X, Hu L (2018) Deep learning for sensor-based activity recognition: a survey. Pattern Recognit Lett 119:3–11. https://doi.org/10.1016/j.patrec.2018.02.010
43. Warden P (2018) Speech commands: a dataset for limited-vocabulary speech recognition. arXiv:1804.03209v1
44. Weitz K, Hassan T, Schmid U, Garbas JU (2019a) 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, https://doi.org/10.1515/teme-2019-0024
45. Weitz K, Schiller D, Schlagowski R, Huber T, André E (2019b) "Do you trust me?": Increasing user-trust by integrating virtual agents in explainable ai interaction design. In: Proceedings of the 19th ACM international conference on intelligent virtual agents. ACM, New York, NY, USA, IVA '19, pp 7–9, https://doi.org/10.1145/3308532.3329441
46. Wu J, Ghosh S, Chollet M, Ly S, Mozgai S, Scherer S (2018) Nadia: Neural network driven virtual human conversation agents. In: Proceedings of the 18th international conference on intelligent virtual agents. ACM, pp 173–178, https://doi.org/10.1145/3267851.3267860
47. Zhang Z, Geiger J, Pohjalainen J, Mousa AED, Jin W, Schuller B (2018) Deep learning for environmentally robust speech recognition: an overview of recent developments. ACM Trans Intell Syst Technol (TIST) 9(5):49:1–49:28, https://doi.org/10.1145/3178115