# “Do you trust me?” Increasing user-trust by integrating virtual agents in explainable AI interaction design

**Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, Elisabeth André**

**Angaben zur Veröffentlichung / Publication details:**

Weitz, Katharina, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. “‘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 - IVA ’19, Paris, France, July 02 - 05, 2019*, edited by Catherine Pelachaud, Jean-Claude Martin, Hendrik Buschmeier, Gale Lucas, and Stefan Kopp, 7–9. New York, NY: ACM Press. https://doi.org/10.1145/3308532.3329441.



# "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design

Katharina Weitz
Human-Centered Multimedia,
University of Augsburg
Augsburg, Bavaria
weitz@hcm-lab.de

Dominik Schiller
Human-Centered Multimedia,
University of Augsburg
Augsburg, Bavaria
schiller@hcm-lab.de

Ruben Schlagowski
Human-Centered Multimedia,
University of Augsburg
Augsburg, Bavaria
schlagowski@hcm-lab.de

Tobias Huber
Human-Centered Multimedia,
University of Augsburg
Augsburg, Bavaria
huber@hcm-lab.de

Elisabeth André
Human-Centered Multimedia,
University of Augsburg
Augsburg, Bavaria
andre@hcm-lab.de

## 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. This development led to an on-going resurgence of the research area of explainable artificial intelligence (XAI) which aims to reduce the opaqueness of those black-box-models. However, much of the current XAI-Research is focused on machine learning practitioners and engineers while omitting the specific needs of end-users. In this paper, we examine the impact of virtual agents within the field of XAI on the perceived trustworthiness of autonomous intelligent systems. To assess the practicality of this concept, we conducted a user study based on a simple speech recognition task. As a result of this experiment, we found significant evidence suggesting that the integration of virtual agents into XAI interaction design leads to an increase of trust in the autonomous intelligent system.

## CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; **Empirical studies in interaction design**.

## KEYWORDS

explainable artificial intelligence, interpretability, virtual agents, human-agent interaction, deep learning, trust

**ACM Reference Format:**
Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *ACM International Conference on Intelligent Virtual Agents (IVA '19), July 2–5, 2019, PARIS, France.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3308532.3329441


*IVA '19, July 2–5, 2019, PARIS, France*

ACM ISBN 978-1-4503-6672-4/19/07.
https://doi.org/10.1145/3308532.3329441

## 1 INTRODUCTION

Recent advancements in the field of automatic speech recognition and natural language processing are already powering a new generation of speech 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 intelligent systems, the underlying algorithms are growing in complexity and therefore decreasing the system's comprehensibility. Evidence suggests that a lack of transparency, with respect to the decisions of an autonomous agent, might have a negative impact on the trustworthiness of a system, which hurts the overall user-experience in return [5, 13].

The reemerging research field of explainable artificial intelligence (XAI) [3] 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 to better understand and trust a system [1, 8, 9]. Although those efforts achieved remarkable progress in recent years, concerns have been expressed that the development of explanation methods has been focused too much on building solutions for AI-Experts while neglecting end-users [10]. These reservations have been backed by recent studies on state of the art XAI-Methods, which concluded that those approaches are not yet at a point where they can be utilized to benefit the user directly [16]. De Graaf and Malle [2] hypothesized that people are applying human traits to autonomous intelligent systems (AIS) 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 AIS and recalibrate their trust in the system if it communicates explanations in a human-like way. In this paper, we aim to find out if a personified virtual agent can be applied within the field of XAI to make an AIS more trustworthy for end-users. We are specifically examining the following research question: Does the incorporation of a virtual agent into XAI approaches positively impact the perceived trustworthiness of complex intelligent systems like Deep Neural Networks (DNN)? To investigate the potential added value of employing virtual agents for XAI-tasks, we conducted a study in which an agent presented

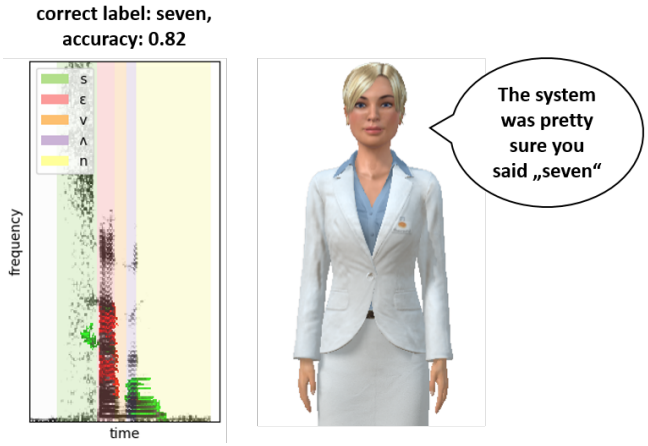


**Figure 1: The visual explanation output for a correct prediction of the keyword "seven" (left) and the virtual agent communicating redundant information (right).**

explanations of a simple speech recognition system to 30 participants.

## 2 EXPERIMENT

To investigate the effect of virtual agents in explainable AI approaches with respect to trustworthiness, we conducted a user study with 30 participants. Every participant interacted with a simple graphical user interface and spoke a predefined and fixed sequence of ten chosen keywords into a microphone. After each phrase, the recorder audio data was classified by a neural network and a visual explanation of this classification was displayed (see Figure 1).

The neural network model we used for keyword classification is based on [12] and has been trained on a dataset for limited-vocabulary speech recognition [15]. The model uses visual representations (spectrograms) of the audio data as input. Only eight of the given keywords were part of the training data, whereas the remaining two words were unknown to the classification system and would, therefore, be wrongly predicted for sure. This verifies that the generated explanations help the user understand both correct and incorrect classifications and therefore increase the transparency of the network's decisions in both situations.

For generating visual explanations we chose the *LIME* framework by Ribeiro et al. [11]. The underlying algorithm highlights areas within the spectograms that provide evidence in favor of the prediction in green and areas that speak against it in red (Fig 1). Since spectrograms are difficult to interpret for people without prior experiences in audio processing, we also presented a phoneme based segmentation of the input-word to the user. This phoneme segmentation of the spectrogram is generated through the WebMAUS tool developed by Kisler et al. [7] (see Figure 1, left side).

The participants were split into two test groups of equal size: Group A (Agent) and group B (no Agent). Group A consisted of 15 participants ($M$ = 28.2, $SD$ = 8.6, 10 male, 5 female). Group B also consisted of 15 participants ($M$ = 27.27, $SD$ = 5.19, 12 male, 3 female). The persons who already had experience in the use of language agents, audio processing, and virtual agents were distributed almost identically in both groups.

Group B received only the visual explanations without further commentary from a virtual agent as well as textual information about the predicted label and accuracy of each classification. Prior to the test they received written instructions about the XAI output and about spectrograms. Group A received the same information and additionally interacted with the virtual agent Gloria, designed by the Charamel GmbH [1], which commented on the user's interactions and the system's predictions and explanations (see Figure 1). The interaction with the agent was designed as such, that group A did not receive any additional information in comparison to group B.

After the experiment, the participants answered the *Trust in Automation questionnaire* [6]. Additionally, the individual impressions of Gloria were queried if the participant was part of group A.

## 3 RESULTS

### 3.1 Agent Evaluation

The evaluation of the agent covered the following areas: sympathy, repeated interaction, trustworthiness, comprehensibility of it's statements and help in understanding the system decision. Participants evaluated each area on a 7-point Likert scale (1=disagree, 7=fully agree). As a result of an additional open question section, we found two criteria that were positively evaluated 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 verbal comments, in particular those referring to highlighted phonemes, supportive.

### 3.2 Trust Evaluation

First, we evaluated the general trust value by examining the data from the *Trust in Automation Questionnaire*. Across all items, there was a statistically significant difference between group A and group B, where group A ($M$ = 5.12, $SD$ = 0.69) rated the system more trustworthy than group B ($M$ = 4.48, $SD$ = 0.86), $t(28) = 3.29$, $p$ = .001, $g$ = 1.17 (large effect).

Based on the general trust value, we exploratively investigated which items of the *Trust in Automation Questionnaire* differed the most between the two groups. For this, a one-way MANOVA was performed. The result was statistically significant, Pillai's Trace = 0.67, $F(12, 17) = 2.87$, $p$ = .023. Therefore, follow-up tests for the selected items were conducted. Holm correction for multiple testing was applied. Between groups A and B, statistically significant differences were found for three items (see Table 1). Group A rated the system as less deceptive and was less wary of the system than group B. Additionally, group A trusted the system more than group B.

## 4 DISCUSSION OF RESULTS

The aim of our user study was to answer the research question whether a virtual agent in combination with XAI has a positive effect on the trustworthiness of a AIS. Examining the results, we

[1] https://vuppetmaster.de/

**Table 1: Results of the follow-up tests of the one-way MANOVA. Degrees of freedom with decimal places are the result of applying the Welch-test**

| Variable | Agent (A) *M(SD)* | No Agent (B) *M(SD)* | *t*(df) | *p* |
|---|---|---|---|---|
| is deceptive | 1.67(1.05) | 3.53(1.88) | *t*(21.88) = -3.35 | .014* |
| behaves underhanded | 1.27(0.59) | 2.20(1.37) | *t*(19.06) = -2.42 | .104 |
| suspicious actions | 1.80(1.37) | 2.80(1.90) | *t*(28) = -1.65 | .290 |
| wary of the system | 1.53(0.83) | 3.80(1.86) | *t*(19.41) = -4.31 | .002* |
| has harmful outcome | 1.40(0.91) | 2.00(1.00) | *t*(28) = -1.72 | .290 |
| confident in the system | 4.53(0.99) | 4.13(0.99) | *t*(28) = 1.11 | .417 |
| provides security | 3.73(1.67) | 3.47(1.25) | *t*(28) = 0.50 | .624 |
| has integrity | 4.80(1.66) | 5.40(1.30) | *t*(28) = -1.10 | .861 |
| is dependable | 4.60(1.35) | 3.67(1.48) | *t*(28) = 1.82 | .275 |
| reliable system | 4.73(0.96) | 3.60(1.40) | *t*(28) = 2.58 | .069 |
| trust the system | 5.00(1.13) | 3.47(0.99) | *t*(28) = 3.94 | .003* |
| familiar with the system | 5.27(1.87) | 4.33(1.54) | *t*(28) = 1.49 | .294 |

*$p < .05$, one-tailed.

found that users had significantly more trust in the explanations that were presented by the agent. Furthermore, significant differences between the groups were found both in positively asked questions as well as in negatively asked questions within the *Trust in Automation Questionnaire*:

- The users found the system to be less deceptive when the explanation results were presented by the agent.
- The users were less wary towards the system.
- The users trusted the system more when the explanations were presented by the agent.

However, trust is a complex concept that can be influenced by various aspects. Hoff and Bashir [4] 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 (spectrograms), the type of system (agent vs. no agent), and system complexity (DNN). 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.

Our results are contrasting a study by Mulken et. al [14] in 1999, in which no significant increase in trustworthiness through the personification of user interfaces could be determined. In their paper the authors 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 was additionally reflected in the overall positive evaluation results for the virtual agent used in our study.

## 5 CONCLUSION

Within this paper we examined the impact of virtual agents within the field of XAI on the trustworthiness perceived by human end-users. To this end, we conducted a user-study in which we presented visual explanations of predictions made by an automatic speech recognition system to users. Based on the results from our study we found that users had significantly more trust in the intelligent system that were presented by a virtual agent as compared to users that solely received the visual output of an AIS. Our results show that the combination of XAI methods with linguistic information presented by an agent can be beneficial for bringing trustworthy AI systems to end-users and thus contribute towards a responsible AI.

## ACKNOWLEDGMENTS

This work has been funded by the Bundesministerium für Bildung und Forschung (BMBF) within the project "VIVA", Grant Number 16SV7960.

## REFERENCES

[1] Jessie Y. C. Chen, Katelyn Procci, Michael Boyce, Julia Wright, Andre Garcia, and Michael J. Barnes. 2014. Situation Awareness-Based Agent Transparency. *US Army Research Laboratory* (2014).

[2] Maartje M A De Graaf and Bertram F Malle. 2017. How People Explain Action (and Autonomous Intelligent Systems Should Too). In *AAAI 2017 Fall Symposium on AI-HRI*. 19–26.

[3] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA)* (2017).

[4] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors* 57, 3 (2015), 407–434. https://doi.org/10.1177/0018720814547570

[5] Joshua D. Hoffman, Michael J. Patterson, John D. Lee, Zachariah B. Crittendon, Heather A. Stoner, Bobbie D. Seppelt, and Michael P. Linegang. 2006. Human-Automation Collaboration in Dynamic Mission Planning: A Challenge Requiring an Ecological Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 23 (2006), 2482–2486. https://doi.org/10.1177/154193120605002304

[6] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

[7] Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45 (2017), 326 – 347. https://doi.org/10.1016/j.csl.2017.01.005

[8] H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. 2005. *Explainable artificial intelligence for training and tutoring.* Technical Report. University of Southern California/Institute for Creative Technologies. https://apps.dtic.mil/dtic/tr/fulltext/u2/a459148.pdf

[9] Joseph E. Mercado, Michael A. Rupp, Jessie Y.C. Chen, Michael J. Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors* 58, 3 (2016), 401–415. https://doi.org/10.1177/0018720815621206

[10] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. *IJCAI International Joint Conference on Artificial Intelligence* (2017). arXiv:1712.00547

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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, 1135–1144. https://doi.org/10.1145/2939672.2939778

[12] Tara N Sainath and Carolina Parada. 2015. Convolutional Neural Networks for Small-footprint Keyword Spotting. In *Proceedings of Interspeech, 2015*. 1478–1482. https://www.isca-speech.org/archive/interspeech{_}2015/papers/i15{_}1478.pdf

[13] K. Stubbs, P. J. Hinds, and D. Wettergreen. 2007. Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *IEEE Intelligent Systems* 22, 2 (2007), 42–50. https://doi.org/10.1109/MIS.2007.21

[14] Susanne Van Mulken, Elisabeth André, and Jochen Müller. 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*. 152–156.

[15] Pete Warden. 2018. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition. (2018). arXiv:1804.03209v1 https://arxiv.org/pdf/1804.03209.pdf

[16] Katharina Weitz, Teena Hassan, Ute Schmid, and Jens Garbas. 2018. Towards Explaining Deep Learning Networks to Distinguish Facial Expressions of Pain and Emotions. In *Forum Bildverarbeitung 2018*. KIT Scientific Publishing, 197–208.