Color for Characters - Effects of Visual Explanations of AI on Trust and Observability

Tim Schrills(✉) and Thomas Franke

Universität zu Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany
{schrills, franke}@imis.uni-luebeck.de

Abstract. The present study investigates the effects of prototypical visualization approaches aimed at increasing the explainability of machine learning systems in regard to perceived trustworthiness and observability. As the amount of processes automated by artificial intelligence (AI) increases, so does the need to investigate users' perception. Previous research on explainable AI (XAI) tends to focus on technological optimization. The limited amount of empirical user research leaves key questions unanswered, such as which XAI designs actually improve perceived trustworthiness and observability. We assessed three different visual explanation approaches, consisting of either only a table with classification scores used for classification, or, additionally, one of two different backtraced visual explanations. In a within-subjects design with N = 83 we examined the effects on trust and observability in an online experiment. While observability benefitted from visual explanations, information-rich explanations also led to decreased trust. Explanations can support human-AI interaction, but differentiated effects on trust and observability have to be expected. The suitability of different explanatory approaches for individual AI applications should be further examined to ensure a high level of trust and observability in e.g. automated image processing.

Keywords: Human-AI interaction · Explainable AI · Machine learning · Trust in Automation · Human-automation interaction

1 Introduction

A central problem with many artificial intelligence (AI) systems is that they typically do not explain to users how they calculate their results [1], which makes it difficult for users to trace back, understand and, finally, rely on the outcome. Hence, a rising demand regarding AI research is to focus on technological approaches to achieve explainable AI (XAI) systems [2]. XAI addresses this challenge by adding information on how AI generates classifications, e.g. by analyzing how changing input affects result [3]. Technology connected to XAI is particularly important in the field of machine learning (ML), because of deep neural networks (DNNs), where sub-symbolic procedures are performed. Accordingly, the present study refers to the use of DNNs as one key approach in ML.

While users and application contexts can differ with regard to what information is needed to explain the behavior of ML systems, a key question from the perspective of

H. Degen and L. Reinerman-Jones (Eds.): HCII 2020, LNCS 12217, pp. 121–135, 2020.
https://doi.org/10.1007/978-3-030-50334-5_8

human factors is: what is a good explanation? Moreover, how can the perceived quality of XAI approaches be measured? Here it is important to note that many AI systems can be viewed as automated inferences that were previously carried out by humans [4]. Research on automation has long demonstrated perceived trustworthiness as a key variable [4–6]. A good AI explanation should thus aim to establish trust in the system used. In addition, a key requirement in the field of AI is that errors can be detected with as little effort as possible [7]. This is because ML systems can change their rules at runtime and without communicating this explicitly, thus resulting in a changing or, at very least, opaque rule systems, thereby limiting transparency for users. Good AI explanations should therefore also target increased perceived observability of AI systems [8].

Currently, XAI research mostly focuses on developing explanations in order to optimize the classification algorithm (i.e., technical optimization) and does not examine how users perceive different forms of explanations, e.g. [9]. So far, the usefulness of an explanation is measured by its mere impact on optimizing the algorithm. In the case of an AI classification, for example, this could be done by calculating the extent to which a particular data element (e.g., image pixel, word) affects the classification [10]. Due to their high complexity and sheer amount of information, technically optimized explanations may differ from those that are optimized to users' needs [11]. Nevertheless, it is clear that in many contexts the human user will be responsible for evaluating and accepting the proposed decision alternatives or actions of AI systems [12]. However, there is currently a lack of research regarding how good existing XAI approaches perform to increase traceability, and therefore trust and observability.

Explainability plays a role in many AI application contexts, e.g. image processing or natural language processing [13]. Consequently, the type of artifact an explanation can consist of is also strongly context-dependent, e.g., textual, graphical; see also [14]. However, given the early stage of development of empirical human factors XAI research in this area, it is expedient to reduce the complexity of relevant application contexts and choose experimental tasks with a low level of complexity and high experimental control. In order to investigate how explanation-approaches differ with regard to their usefulness for users (in terms of perceived trustworthiness and observability), we therefore focus on a very basic character recognition AI system, providing explanations by presenting visualizations based on the input.

The objective of the present research was to investigate the effect of visual explanations in ML systems on perceived trustworthiness and observability. To this end, we focused on two prototypical XAI visualization approaches and studied the user-related effects.

2 Background

Currently, many technical XAI approaches are developed with the aim of increasing the reliability and observability of ML systems – especially with regard to the detection of classification errors in trained systems, e.g. [15]. Often, those are local explanations, focusing to explain a specific outcome, not the general, cf. [16]. Most DNNs also display a value about how e.g. secure the classification is: a classification score.

Without additional explanations, however, it is impossible for users to understand exactly how these values were generated. To address this, within the area of visual classification analyses, pixel-wise backtracing is used to generate visual explanations based on the original input. One aim of an explanation should be to have the highest possible fidelity, i.e. address the connection between input and output in the best possible way [17]. Thus, for visual explanations, it makes sense to rely on methods that use backtracing and generating e.g. pixel-based heat maps. For example, a frequently examined approach in this area is Sensitvity Analysis (SA) [18]. Here, a systematic change of the input stimulus is used to identify which image components are particularly important for the classification. While this approach can reveal insights into the machines' information processing and possible errors, it is one-dimensional, as well as unipolar, and therefore not recommended to enhance understanding of the processes [19]. One further, recently proposed, approach is Layer-wise Relevance Propagation (LRP), in which the results are traced back inside the network in order to more precisely identify which pixels contribute for and which against the calculated classification [20]. This results in two-dimensional data, allowing better representations of networkactivity.

Still, those XAI visualization approaches highlight the relevance of the stimulus components for the classification of every pixel and thus produces results very rich in data. This data contains information about the relevance of every pixel for the chosen classification and about every other classification not chosen. Hence, these can be labelled omni-explanations. Since the presented information can be very complex, this could result in information overload for users [21], may hindering the development of trust, cf. [22].

A proposed procedure to increase both trustworthiness and observability, without inflating the information complexity, is based on how humans understand explanations in general. [14] argues that explanations could substantially benefit from relying on counterfactual thinking. This refers to a concept where hypothetical pasts are constructed and their effects assessed, in order to assess the present situation more easily [23]. Counterfactual explanations [24] or contrast cases 23 work in a structurally similar way as counterfactual thinking. Counterfactual explanation means that a reason (e.g. for a classification) does not only answer why the classification is correct, but also why other (almost equally) probable classifications are not [24]. For example, why the handwritten letter i was classified as i and not as the next-most probable option, j. The advantage of counterfactual explanations is that the amount of irrelevant information is reduced, because all information that also speaks for other, similar explanations is filtered out. At the same time, the extent to which the information shown describes the functioning of the system remains high. Therefore, counterfactual explanations could particularly meet users' demand for information and increase trustworthiness and observability of ML systems. Still, counterfactual explanations may result in incomplete explanations, since only one alternative is compared to the given classification. [25] found incompleteness to negatively affect how users evaluate an explanation. However, in this case, completeness refers to the proportion of available information on different process variables, and not to the complexity of the comparison with other possible outcomes. First studies using counterfactual explanations in XAI have shown to be effective at the technical level [26].

While first studies found positive results when presenting explanations in AI systems, e.g. [25, 27], there are no studies specifically comparing the effects of structurally different prototypical visual explanation-approaches and, to the best of the authors' knowledge, no quantitative user studies have yet been carried out to evaluate this way of generating explanations empirically.

2.1 Trust in AI Systems

In order for additional explanations in AI systems to have a positive effect on the user, they must address crucial interaction variables. As introduced above, the requirements for a successful user-AI interaction are closely related to the creation of trust (i.e., users' perceived trustworthiness of a system), see also [28].

So far, empirical research on trust in ML systems is limited. First research indicates that trust in AI systems can have a decisive effect on the behavior of users and can influence the extent to which e.g. provided classifications are perceived as useful [29]. Further, recent studies found that the level of trust to be significantly influenced by information related to the processing of the input data, e.g. [27]. As explicated above, it remains important to present only relevant influencing variables in order to make it as easy as possible for users to gain knowledge about the processing path and not e.g. conflict with previous mental models [30]. This specific perspective on trust in processing is – besides the purpose of the system and aspects related to the concrete performance – an important focus of the field of applied ML [31].

2.2 Observability of AI Systems

Since systematic biases can occur in the training of neural networks due to various factors, e.g. data set or sequence of training, cf. [32], a decisive success factor in the cooperation between humans and AIs is whether detected failures can be correctly attributed to the system [33]. However, for this to be possible, users must be able to understand the state of the system when the failure occurs. In this sense, observability refers to the system property of how correct conclusions about a system state can be drawn from the given output [34]. Here, the state of a ML system also includes the current, exact way of processing inputs to outputs, e.g. classification. Ensuring observability is a prerequisite for the design of highly functional systems in the context of human-centered development [35, 36]. [37] showed that in systems that do not guarantee complete reliability, additional information is important for the inspection of the data; an explanation, for example, would be additional information. Considering that a major part of ML systems is still characterized by limited reliability, it is particularly important to investigate observability as a target variable of XAI. Finally, systems that have higher observability may also obtain higher understandability of classifications results.

3 Present Research

The objective of the present research was to examine how different prototypical visualizations that aim to explain AI results affect the perceived trustworthiness and observability of an image classification system. Specifically, we compare the effects of the presentation of classification-scores with the additional presentation of two different backtraced visual explanations, whereby one is based on counterfactual reasoning. First, we assume that information about the classification process increases perceived trustworthiness and observability. This leads to the following two hypotheses:

H1.1) Visual explanations of classifications based on the input stimulus lead to higher trust in the system.

H1.2) Visual explanations of classifications based on the input stimulus lead to higher observability of the system.

Second, we assume that counterfactual explanations lead to higher trust and observability compared to omni-explanations by reducing the risk of information overload, which is why the following hypotheses can be formulated:

H2.1) Counterfactual explanations lead to higher trust than omni-explanations.

H2.2) Counterfactual explanations lead to higher observability than omni-explanations.

Finally, higher observability can also be expected to have an effect on the general comprehensibility of AI systems, hence:

H2.3) Counterfactual explanations are rated as more understandable than omni-explanations.

To examine these hypotheses, we choose basic character classification task. Here, no user-evaluated machine-learning systems, generating a backtraced visual explanation, exist. The low complexity of this task contributes to the fact that the resulting effects of the visualization can be clearly evaluated on basis of the input and are not significantly influenced by other variables, resulting in an ambiguous estimation. Thus, in tasks where the user has to make a greater effort to evaluate the results of the AI, the willingness to do so could adversely affect the experiment.

4 Method

4.1 Experimental Procedure

The invitation to participate in the study was distributed over e-mail and social networks. All participants were offered to enter a prize draw for a gift coupon. Psychology students of the local university were also offered course credit for taking part in the study. This research complied with the American Psychological Association Code of Ethics. Informed consent was obtained from each participant. The N = 83 users who completed the study had an average age of 25.4 years (SD = 8.1), 71% were female, 27% male and 1 person did not indicate gender. Participants generally had a rather low level of previous knowledge on AI systems (Mdn = 1.75, IQR = 1.25 – 2.75, possible score values 1-6

with a label of both poles as "true" and "false", score based on 4 items depicted in Table 1, Cronbach's alpha = .76) and an affinity for technology interaction score (M = 3.43, SD = 1.17) close to average value in the general population (3.5, see [38]).

Table 1. Translated texts of newly constructed items to assess previous AI knowledge; presented at the beginning of the study.

Please indicate to what extent you agree with the following statements	
01	I have already dealt with machine learning in the AI field
02	I understand how data is processed in neural networks.
03	I know technical approaches to increase the explainability in the field of machine learning
04	I have already dealt with the topic of AI

4.2 XAI Visualization Approaches

The study material was created using a public platform provided on the basis of [19]. The platform consists, among other things, of a neural network for number recognition, which was trained based on MNIST data records [39]. In addition, LRP [19] for the given input stimuli was calculated and provided. LRP Formula LRP Epsilon was selected and the beta value was set to 1.

In addition to the result of the classification, a classification-score overview was provided, in which a classification value is displayed for each potential digit. These represent the calculated probability for each possible category based on output-layer values of the DNN. Furthermore, heat map visualizations were provided for each input, as shown on the left in Fig. 1. The middle depiction – the LRP based visualization – is a manipulation of the actual input stimulus depicted on the left. In the present experiment, LRP-based images are used for the omni-explanation, since in LRP the value of each pixel for or against the classification is visualized. Stronger red hues indicate a higher relevance for the classification and stronger blue hues against the classification. If the respective pixel has no or very little influence, it is displayed in white.

The calculation of the stimuli for the counterfactual representation was based on the omni-visualization. The aim here is to mark the areas of the image that speak for the classification, but not for the next likely classification. To achieve this, in addition to calculating the omni-visualization of the assumed classification, the omni-visualization of the second most probable classification (e.g. not 3 but 0) was calculated. In the following step, the color values were modified so that: 1) one pixel, which had the same color in both images, was displayed as white and 2) a pixel, which in the selected classification was maximally red and – in the alternative classification – was maximally blue, received the darkest color. Smaller color distances result in a less intense color tone. Pixels within which the selected classification was blue (negative) and the alternative classification was red (positive) were also displayed in white. The results of the calculation of the counterfactual visualizations are generally colorless, i.e. in gray scales. A yellow overlay was used to increase similarity between omni- and

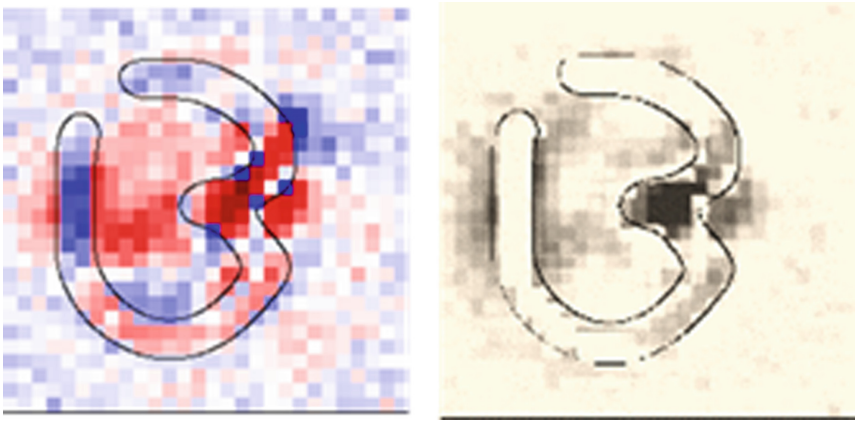

Fig. 1. Example of a Visualization for the omni-condition (left) and for the counter-condition (right).

counter-condition resulting in higher comparability with the colored images of the omni-condition while safeguarding clear discriminability of the two visualization approaches for the participants.

For this experiment, the stimuli set consisted of twenty different characters, which were manipulated for each condition, resulting in 60 different images. These were created before the experiment and represent the characters 0–9, as well as other partly ambiguous symbols. For each digit, a scoring overview (for all conditions), an omni-visualization as well as a counterfactual visualization were created.

4.3 Design and Procedure

The present study was set up as an online experiment in the German language. Three different conditions were defined that presented different information visualizations: in the first, the participants were only shown the input image, the AI classification and the classification scoring, but no additional visual or backtraced explanations (baseline-condition). In the second condition, the omni-visualization was additionally shown (omni-condition). In the third condition, the counterfactual visualization was shown, and the bar of the alternative number was marked in the scoring overview (counter-condition). In the course of the experiment, demographic data was first collected; then the three different conditions were explained in randomized order. Users were asked to perform the study only on sufficiently large screens. Examples for the different conditions are depicted in Fig. 1. Users were shown 20 randomly selected stimuli.

4.4 Scales and Measures

Trust was measured after each stimulus using the 5-item facets of system trustworthiness (FOST) scale [40]; see translated items 1–5 in Table 2. This allowed the assessment of key subfacets of trust for each trial (i.e., maximum scale length deemed acceptable for 20 presentations). The averages were calculated within each condition as

one single value per condition for each participant, independent of the actual presented stimuli. Finally, the comprehensive Trust in Automation scale [41] was assessed at the end of the survey for each of the 3 experimental conditions, together with another FOST assessment in order to check the validity of the FOST scale in the context of the present study. As in previous research [40], both scales converged (with $r = .59$ for baseline-condition, $r = .57$ for omni-condition and $r = .69$ for counter-condition, all $p < .001$).

Table 2. Translated item texts of the Facets of System Trustworthiness (FOST) scale and all newly constructed scales. Reversed Items have been marked with "r" behind the number.

Please indicate to what extent you agree with the following statements	
01	The system's classification is reliable
02	The system's classification is precise
03	The system's classification is traceable
04	I can trust the system's classification
05r	I cannot depend on the system's classification
06	With the help of the visualization I am able to identify wrong mechanisms of the AI
07	I agree with the classification
08	The visualization provides a good explanation for the classification

Unfortunately, there is no generally accepted procedure to assess the observability of AI systems. Hence, we developed a single-item measure focusing on the key aspects of observability of enabling users to deduce systems states and detect failures. Item 6 in Table 2 shows the translated item text. To also operationalize the understandability of given explanations, we also added another item (see Item 8 in Table 2 for translated item text). Furthermore, for each stimulus, the level of agreement with the classification was measured in order to be able to identify effects of agreement related to trust and observability or vice versa; the translated text can be found in Item 7 of Table 2. For all additional items, we used the same Likert scale as for the FOST scale (see Table 2) coded as 1–6. The response scale had gradual formulations from "completely disagree" to "completely agree". The Trust in Automation scale was measured using a 7-point Likert scale. Reliability was good for all multi-item scales (see Table 3).

5 Results

In order to test the hypotheses, the data was analyzed with repeated measures ANOVA containing the three conditions. The violation of the sphericity assumption was controlled according to Mauchly's method [42], and in the case of a significant result, the correction supposed by [43] was performed. If the results were significant, additional post-hoc comparisons between the individual conditions were carried out; these were

each carried out on the basis of a familywise performed Bonferroni correction [44]. The evaluation of the statistical power was based on the recommendation given by Cohen [45]. The Alpha level for all test to be rated as significant was $p < .05$.

Table 3. Reliability analysis and descriptive statistics for used multi-item scales and descriptive statistics for used single-item scales.

Scale	Cronbach's alpha	*Mean*	*SD*
Trust in automation	.82	4.37	0.85
FOST baseline-condition	.93	4.43	0.79
FOST omni-condition	.86	4.18	0.73
FOST counter-condition	.90	4.86	0.71
Observability baseline-condition	–	3.47	1.21
Observability omni-condition	–	3.83	1.08
Observability counter-condition	–	3.95	1.04
Understandability baseline-condition	–	3.84	1.22
Understandability omni-condition	–	4.17	1.00
Understandability counter-condition	–	3.95	1.20
Agreement baseline-condition	–	4.84	0.71
Agreement omni-condition	–	4.67	0.79
Agreement counter-condition	–	4.86	0.63

5.1 Hypotheses Testing

Descriptive statistics for all tested variables can be found in Table 3. The FOST scale was used to examine hypotheses H1.1 and H2.1, i.e. to determine trust in each condition. Although the ANOVA performed revealed a significant result, as expected, with $F(2, 164) = 10.06$, $p < .001$, $\eta^2 = .109$), post-hoc tests revealed (see Table 4) that the difference between the baseline-condition and the two experimental conditions was not as expected, as the baseline- and counter-conditions was not found to be significantly different. On this basis, H1.1 needs to be rejected. Only the omni-condition was significantly below both the baseline- and the counter-condition in the assessment of trust. Accordingly, this supports H2.1.

Table 4. Post-hoc performed pairwise t-test for Hypothesis 1.1 and 2.1. N = 83 for all tests. P-values are two-tailed. Familywise Bonferroni-Correction has been applied. Cohens'd has been calculated following [46].

Post-hoc tested groups FOST	*t*	*p*	*Cohen's d*
FOST baseline – FOST omni	3.24	.006	0.32
FOST baseline – FOST counter	0.76	>.999	0.07
FOST omni – FOST counter	−4.35	<.001	0.42

To test hypothesis H1.2 and H2.2, the values of the item to assess observability were examined (see Table 5). Significant results were also found here with $F(2, 164) = 7.25$ $p < .001$, $\eta^2 = .109$). Post-hoc tests showed that, as expected, the omni-condition and the counter-condition received better evaluations than the baseline-condition (i.e. without a visual explanation of the procedure). While H1.2 was confirmed by this, there was no difference between the omni-condition and the counter-condition, which is why H2.2. was rejected. The examination of hypothesis H2.3 did not reveal a significant result with $t(82) = 1.64$, $p = .106$, $d = 0.20$, both conditions showed almost equal ratings (see Table 3).

Table 5. Post-hoc performed pairwise t-test for Hypothesis 1.2 and 2.2. $N = 83$ for all tests. *P*-values are two-tailed. Familywise Bonferroni-Correction has been applied. Cohens'd has been calculated following [46].

Post-hoc tested groups	*t*	*p*	*Cohen's d*
Observability baseline - Observability omni	−2.95	.012	0.31
Observability baseline - Observability counter	−3.36	.003	0.42
Observability omni - Observability counter	−0.94	>.999	0.11

5.2 Exploratory Analysis

We further evaluated the correlations of level of agreement, and trust and observability for each condition. For trust, all correlations were significant, while this was only true for observability in the counter-condition (see Table 6).

Table 6. Descriptive statistics for agreement and correlation with trust and observability N = 83 for all tests. P-values are two-tailed. Familywise Bonferroni-Correction has been applied. Cohens'd has been calculated following [46].

Condition	*Mean*	*SD*	*Pearson's r for [Condition] Trust*	*Pearson's r for [Condition] Observability*
Agreement baseline-condition	4.84	0.71	.70*	.08
Agreement omni-condition	4.67	0.79	.70**	.04
Agreement counter-condition	4.86	0.63	.67**	.30*

* indicates $p < .05$ and ** indicates $p < .01$

6 Discussion

6.1 Summary of Results

The objective of the present research was to examine how different prototypical visualizations that aim to explain AI results affect the perceived trustworthiness and

observability of a ML system. Overall, we found a mixed pattern of results regarding our hypotheses, indicating that relationships between variables may – in part – be more complex than expected.

The assumption, that additional information about the classification process based on the input material generally contributes to an increase of trust (H1.1), could not be confirmed; the representation of the omni-condition led to lower values than the baseline-condition. However, the assumption that an explanation based on the approach of counterfactual explanations achieves better values than the omni-condition was confirmed (H2.1, small significant effects).

The hypotheses on observability showed the opposite pattern. Here, results showed that the support by visual information increased the perceived observability (H1.2), but contrary to hypothesis H2.2, the two examined XAI visualization approaches did not differ with regard to the observability rating (only very weak and insignificant advantage of counter-condition compared to omni-condition). The examination of the hypothesis for the understandability of explanations (H2.3) did not show significant results and the effect size was too small to be further considered. Finally, we found strong and significant correlations between agreements to classification and perceived trust (see Table 6).

6.2 Implications

Results of the present study show that the visualization of the process of automated AI classifications can have an influence on how the system is evaluated in terms of trust and observability. However, this influence must be evaluated in a non-linear way – trust in AI can also decrease by adding additional information [30]. In line with previous research, we assume the poor performance of the omni-condition to be because information overload occurred and users were not able to build up additional trust [47]. As expected, this effect did not occur within the counter-condition (H2.1).

We further suspect that the result regarding the explanations' understandability to possibly be based on information overload too – the users may have relied on a heuristic process to judge the explanation due to the high amount of information the system generated, similar to the effort heuristic [11]. Yet, this remains speculative and needs further research before allowing to draw firm conclusions. Furthermore, the strong correlation between agreement and perceived trust needs to be further examined. Previous research has already shown that trust is depended on the predictability of a system [48]. Accordingly, we assume that trust in the context of AI also depends on whether expectations regarding the classification (i.e. based on the input) are violated. This especially applies to systems, where users need only little effort to build and verify expectations.

With regard to the observability of the system, the data indicated that the examined XAI visualizations were helpful for users. This lends support to previous notions [49], that a system's observability is therefore crucial for cooperation between humans and AI. This is also in line with our expectations, since without additional information about the process (as in the baseline-condition), conclusions about the actual functioning of the system can only be drawn with considerable additional effort and only to

a lesser extent. Hence, it seems inevitable to address this issue more strongly in further research on the cooperation between AI and humans.

Concerning both primary variables examined, trust and observability, it can be seen that overall the counterfactual explanations proposed by [14] showed a positive tendency to be effective.

6.3 Limitations and Further Research

Some limitations and reveals many interesting open questions for future investigations. First, the dependent constructs (perceived trust and observability) need to be addressed more specifically in isolated tasks. Additionally, a more selective definition of measurable and distinguishable facets of explanations, as suggested by [17] could be helpful.

The task carried out in this experiment was kept as simple as possible. The aim was not to add any further possibly disturbing variables induced by a higher complexity of the task, and to present an AI system that functions as reliably and correctly as possible. Future studies should further control and specifically investigate the agreement of users with the AI's classification. On top of that, more complex tasks, e.g. the extraction of sentiment from texts, need to be examined, because the cooperation between human and intelligent system is necessary. However, future research has to consider tasks humans cannot evaluate without additional effort.

Ultimately, we focused on quantitative data in the present investigation to examine our hypotheses. Yet given the potential complexity of users' cognitive processing of different XAI visualization approaches, further research should also more closely examine aspects like visualization comprehension and the development of mental models of AI systems with qualitative methods. Detailed empirical quantitative investigations of cognitive processes involved in the processing of counterfactual explanations and other explanatory approaches should follow these qualitative studies to further advance the effectiveness of XAI in human-machine interaction [50, 51].

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052
2. Weld, D.S., Bansal, G.: The Challenge of Crafting Intelligible Intelligence. ArXiv180304263 Cs. (2018)
3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for Deep Neural Networks. ArXiv171106104 Cs Stat. (2017)
4. Lee, J.D., See, K.A.: Trust in automation: designing for appropriate reliance. Hum. Factors **46**, 50–80 (2004)
5. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**, 1243–1270 (1992). https://doi.org/10.1080/00140139208967392
6. Muir, B.M., Moray, N.: Trust in automation. Part II experimental studies of trust and human intervention in a process control simulation. Ergonomics. **39**, 429–460 (1996). https://doi.org/10.1080/00140139608964474

7. Nushi, B., Kamar, E., Horvitz, E.: Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. ArXiv180907424 Cs Stat. (2018)
8. Lim, B.Y., Dey, A.K.: Assessing demand for intelligibility in context-aware applications. In: Proceedings of the 11th international conference on Ubiquitous computing (Ubicomp 2009). p. 195. ACM Press, Orlando (2009). https://doi.org/10.1145/1620545.1620576
9. Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. **73**, 1–15 (2018). https://doi.org/10.1016/j.dsp.2017.10.011
10. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. ArXiv160204938 Cs Stat. (2016)
11. Kruger, J., Wirtz, D., Van Boven, L., Altermatt, T.W.: The effort heuristic. J. Exp. Soc. Psychol. **40**, 91–98 (2004). https://doi.org/10.1016/S0022-1031(03)00065-9
12. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI 2018), pp. 1–18. ACM Press, Montreal QC (2018). https://doi.org/10.1145/3173574.3174156
13. Amershi, S., et al.: Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI 2019), pp. 1–13. ACM Press, Glasgow (2019). https://doi.org/10.1145/3290605.3300233
14. Miller, T.: Explanation in Artificial Intelligence: Insights from the Social Sciences. ArXiv170607269 Cs. (2017)
15. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.-R., Samek, W.: The LRP toolbox for artificial neural networks. J. Mach. Learn. Res. **17**(1), 3938–3942 (2016)
16. Hoffman, R.R., Mueller, S.T., Klein, G., Litman, J.: Metrics for explainable AI: Challenges and prospects. ArXiv Prepr. ArXiv181204608. (2018)
17. Ras, G., van Gerven, M., Haselager, P.: Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. ArXiv180307517 Cs Stat. (2018)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626. IEEE, Venice (2017). https://doi.org/10.1109/ICCV.2017.74
19. Samek, W., Wiegand, T., Müller, K.-R.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. ArXiv170808296 Cs Stat. (2017)
20. Binder, A., Bach, S., Montavon, G., Müller, K.-R., Samek, W.: Layer-Wise Relevance Propagation for Deep Neural Network Architectures. Information Science and Applications (ICISA) 2016. LNEE, vol. 376, pp. 913–922. Springer, Singapore (2016). https://doi.org/10.1007/978-981-10-0557-2_87
21. Timmermans, D.: The impact of task complexity on information use in multi-attribute decision making. J. Behav. Decis. Mak. **6**, 95–111 (1993). https://doi.org/10.1002/bdm.3960060203
22. Furner, C.P., Zinko, R.A.: The influence of information overload on the development of trust and purchase intention based on online product reviews in a mobile vs. web environment: an empirical investigation. Electron. Mark. **27**, 211–224 (2017). https://doi.org/10.1007/s12525-016-0233-2
23. Roese, N.J., Morrison, M.: The psychology of counterfactual thinking. Hist. Soc. Res. Sozialforschung 16–26 (2009)
24. Sokol, K., Flach, P.: Glass-Box: explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Organization, Stockholm, Sweden, pp. 5868–5870 (2018). https://doi.org/10.24963/ijcai.2018/865

25. Kulesza, T., Stumpf, S., Burnett, M., Kwan, I.: Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI 2012), p. 1. ACM Press, Austin (2012). https://doi.org/10.1145/2207676.2207678
26. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual Visual Explanations. ArXiv190407451 Cs Stat. (2019)
27. Bigras, E., et al.: In AI we trust: characteristics influencing assortment planners' perceptions of AI based recommendation agents. In: Nah, F.F.-H., Xiao, B.S. (eds.) HCIBGO 2018. LNCS, vol. 10923, pp. 3–16. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91716-0_1
28. Breuer, C., Hüffmeier, J., Hibben, F., Hertel, G.: Trust in teams: a taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. Hum. Relat. (2019). https://doi.org/10.1177/0018726718818721
29. Zanker, M.: The influence of knowledgeable explanations on users' perception of a recommender system. In: Proceedings of the sixth ACM conference on Recommender systems (RecSys 2012), p. 269. ACM Press, Dublin (2012). https://doi.org/10.1145/2365952.2366011
30. Springer, A., Whittaker, S.: "I had a solid theory before but it's falling apart": polarizing effects of algorithmic transparency. arXiv preprint arXiv:1811.02163 (2018)
31. Hengstler, M., Enkel, E., Duelli, S.: Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. Technol. Forecast. Soc. Change. **105**, 105–120 (2016). https://doi.org/10.1016/j.techfore.2015.12.014
32. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2015), pp. 1721–1730. ACM Press, Sydney (2015). https://doi.org/10.1145/2783258.2788613
33. Krause, J., Perer, A., Bertini, E.: A user study on the effect of aggregating explanations for interpreting machine learning models. In: ACM KDD Workshop on Interactive Data Exploration and Analytics (2018)
34. Kalman, R.E.: A new approach to linear filtering and prediction problems. J. Basic Eng. **82**, 35–45 (1960)
35. Billings, C.E.: Human-centered aviation automation: principles and guidelines (1996)
36. Johnson, M., Bradshaw, J.M., Feltovich, P.J.: Tomorrow's human–machine design tools: from levels of automation to interdependencies. J. Cogn. Eng. Decis. Mak. **12**, 77–82 (2018). https://doi.org/10.1177/1555343417736462
37. Rovira, E., McGarry, K., Parasuraman, R.: Effects of imperfect automation on decision making in a simulated command and control task. Hum. Factors J. Hum. Factors Ergon. Soc. **49**, 76–87 (2007). https://doi.org/10.1518/001872007779598082
38. Franke, T., Attig, C., Wessel, D.: A personal resource for technology interaction development and validation of the affinity for technology interaction (ATI) Scale. Int. J. Hum.-Comput. Inter. **35**, 456–467 (2019). https://doi.org/10.1080/10447318.2018.1456150
39. Deng, L.: The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process. Mag. **29**, 141–142 (2012). https://doi.org/10.1109/MSP.2012.2211477
40. Franke, T., Trantow, M., Günther, M., Krems, J.F., Zott, V., Keinath, A.: Advancing electric vehicle range displays for enhanced user experience: the relevance of trust and adaptability. In: Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2015). pp. 249–256. ACM Press, Nottingham (2015). https://doi.org/10.1145/2799250.2799283

41. 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**, 53–71 (2000). https://doi.org/10.1207/S15327566IJCE0401_04
42. Mauchly, J.W.: Significance test for sphericity of a normal n-Variate distribution. Ann. Math. Stat. **11**, 204–209 (1940). https://doi.org/10.1214/aoms/1177731915
43. Greenhouse, S.W., Geisser, S.: On methods in the analysis of profile data. Psychometrika **24**, 95–112 (1959). https://doi.org/10.1007/BF02289823
44. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**, 65–70 (1979)
45. Cohen, J.: A power primer. Psychol. Bull. **112**, 155–159 (1992). https://doi.org/10.1037/0033-2909.112.1.155
46. Dunlap, W.P., Cortina, J.M., Vaslow, J.B., Burke, M.J.: Meta-analysis of experiments with matched groups or repeated measures designs. Psychol. Methods **1**(2), 170 (1996)
47. Kizilcec, R.F.: How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI 2016). pp. 2390–2395. ACM Press, Santa Clara (2016). https://doi.org/10.1145/2858036.2858402
48. Biros, D.P., Fields, G., Gunsch, G.: The effect of external safeguards on human-information system trust in an information warfare environment. In: Proceedings of the 36th Annual Hawaii International Conference on System Sciences 2003, p. 10. IEEE, Big Island (2003). https://doi.org/10.1109/HICSS.2003.1173894
49. Christoffersen, K., Woods, D.: How to make automated systems team players. Adv. Hum. Perform. Cogn. Eng. Res. pp. 1–12 (2002). https://doi.org/10.1016/S1479-3601(02)02003-9
50. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G.: Explanation in human-AI systems: a literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint arXiv:1902.01876 (2019)
51. Hoffman, R.R., Klein, G., Mueller, S.T.: Explaining explanation for "Explainable Ai". Proc. Hum. Factors Ergon. Soc. Annu. Meet. **62**, 197–201 (2018). https://doi.org/10.1177/1541931218621047