Explaining Machine Learned Relational Concepts in Visual Domains – Effects of Perceived Accuracy on Joint Performance and Trust

Anna Magdalena Thaler (anna-magdalena.thaler@stud.uni-bamberg.de)
Cognitive Systems Group, University of Bamberg,
An der Weberei 5, 96049 Bamberg, Germany

Ute Schmid (ute.schmid@uni-bamberg.de)
Cognitive Systems Group, University of Bamberg
An der Weberei 5, 96049 Bamberg, Germany

Abstract

Most machine learning based decision support systems are black box models that are not interpretable for humans. However, the demand for explainable models to create comprehensible and trustworthy systems is growing, particularly in complex domains involving risky decisions. In many domains, decision making is based on visual information. We argue that nevertheless, explanations need to be verbal to communicate the relevance of specific feature values and critical relations for a classification decision. To address that claim, we introduce a fictitious visual domain from archeology where aerial views of ancient grave sites must be classified. Trustworthiness among other factors relies on the perceived or assumed correctness of a system's decisions. Models learned by induction of data, in general, cannot have perfect predictive accuracy and one can assume that unexplained erroneous system decisions might reduce trust. In a 2×2 factorial online experiment with 190 participants, we investigated the effect of verbal explanations and information about system errors. Our results show that explanations increase comprehension of the factors on which classification of grave sites is based and that explanations increase the joint performance of human and system for new decision tasks. Furthermore, explanations result in more confidence in decision making and higher trust in the system.

Keywords: Explainability; verbal explanations; prediction errors; trust; relational learning.

Introduction

With the success of machine learning in many real-world application domains, it has been recognized that it is important for machine learned models to be transparent and comprehensible for humans (Miller, 2019). The focus of Explainable Artificial Intelligence (XAI) research is mostly on post hoc explanations for back box models such as deep neural network models for image classification (Guidotti et al., 2019; Rudin, 2019). That is, it is shown which information in the input image contributes mostly to the systems classification decision. In other domains – especially such where data is available in form of symbolic representations such as feature vectors – intrinsically interpretable machine learning approaches can be used (Rudin, 2019) such as different variants of decision rules (Lakkaraju et al., 2016) or logical rules (Muggleton et al., 2018). However, such white box models, although they are in principle understandable by humans, can be too complex to be grasped by direct inspection and need to be augmented with an explanation mechanism.

Over the last years, different taxonomies for explanation mechanisms have been proposed (e.g., Adadi & Berrada, 2018; Miller, 2019). Explanations can be characterized by their modality – being either verbal (symbolic) or visual where visual explanations typically highlight pixels or areas in the input image. Furthermore, explanations can be local – explaining the classification decision for the current instance – or global – explaining the learned model as a whole. Which type of explanation is helpful depends on the recipient and the application domain (Miller, 2019). Explanations can help model developers to identify overfitting to irrelevant information or unfair biases. Explanations can be given to end-users, for instance in the context of personal recommenders (Tintarev & Masthoff, 2012). For specialized domains such as medical diagnosis or quality control in industrial production, explanations need to be addressed to domain experts (Holzinger et al., 2019).

Many expert domains rely on interpreting and classifying visual information, such as X-rays or microscopic images. While visual explanations in the form of highlighting are helpful to detect overfitting, they are mostly not expressive enough to communicate decision relevant information (Rabold et al., 2018). For instance, an expert might recognize the presence of tumor cells in a tissue sample. But to understand why the system returns a specific tumor class, the size, the form, or the spatial relations between the tumor and other tissue might be important (Bruckert et al., 2020). Consequently, even for visual domains, there is a need for verbal explanations to inform about classification relevant information concerning feature values or relations between different parts.

Besides understanding the reason for a system decision, human decision makers can profit from transparent communication of the predictive accuracy of a machine learned model to assess the reliability of system outputs. This can be realized by explicit communication of the degree of uncertainty of a specific decision (Bykov et al., 2020) or about differences in precision and recall for specific classes (Katsikopoulos et al., 2020). This kind of information has been shown to be helpful to decrease the cry wolf effect in human-machine interaction (Breznitz, 2013).

When focusing on effective explanations for users rather than the model developers themselves (as strongly advocated by Miller et al., (2017)), it becomes relevant to understand what characteristics of an explanation make it helpful to the user. The notion of helpfulness has been mostly defined as effective transfer of knowledge (Lombrozo, 2009). In the context of XAI, this means that a machine learning model has uncovered information which has not previously been available to the human decision maker and communicates this information in a way which gives new insights to the human and in consequence allows for more effective and efficient decision making. Machine learning approaches which support this type of joint human and machine learning have been classified as ultra-strong (Muggleton et al., 2018). A study by Muggleton et al. (2018) demonstrated that, in a relational learning classification task, the participants were not able to learn the concepts by themselves, but did so easily when provided with an explanation. Comprehensibility of the explanations have been assessed as the average accuracy a participant achieves when classifying new samples in the same domain.

The relational domains investigated in the study of Muggleton et al. (2018) has been the family domain involving relations between pairs of persons and an isomorph fictitious chemistry domain with relations between substances. While these domains are purely symbolic, we are interested in relational learning in visual domains. Therefore, we adapted a classification task from Rabold et al. (2018), in which aerial images of fictitious ancient grave sites need to be assigned to their age of origin as Viking Age or Iron Age. This his artificial domain as a cover story ensures that no prior knowledge about archeology is required or helpful to comprehend the classification of the stimuli.

Besides the helpfulness of explanations for comprehension, the effect of explanations on trust is often discussed (Miller, 2019; Ribeiro et al., 2016). In general, trust – be it in humans or machines – is the result of experience of interactions (Miller, 2019). To assume a system to be more trustworthy just because it can provide an explanation is an oversimplification. Explanations might be right for the wrong reasons or even wrong for the wrong reasons depending on the machine learned models predictive accuracy and on the fidelity of the explanation (Schramowski et al., 2020). Trust might be appropriate or not and explanations should help humans to develop appropriate trust and understand better when to trust and when not to trust system decisions.

To investigate the effect of erroneous system decisions, in our study, the presented machine learned model system has been designed such that it is not perfectly accurate. Depending on the specific class, a system decision is more or less reliable. To our knowledge, there exists no prior empirical study that systematically compares the combined effect of explanations and system error information on comprehensibility and trust.

Table 1: Stimuli following the 5-4 category structure (see Medin & Schaffer, 1978) with the respective class assigned.

Example	O	S	A	N	Class
e1	**1**	**1**	**1**	*0*	Iron
e2	**1**	**1**	0	*1*	Iron
e3	**1**	0	**1**	*0*	Iron
e4	0	**1**	**1**	*1*	Iron
e5	0	**1**	**1**	*0*	Iron
e6	**0**	**0**	**0**	*0*	Viking
e7	**0**	**0**	1	*1*	Viking
e8	**0**	1	**0**	*0*	Viking
e9	1	**0**	**0**	*1*	Viking

Note. O: Orientation (1: North, 0: East); S: Shape (1: narrow, 0: wide); A: Ascending order of inner stones (1: yes, 0: no); N: Number of outer stones (1: many, 0: few). The values that are relevant for the category assignment are written in bold, the irrelevant factor in italics.

In the following, we will first introduce the visual relational domain of ancient grave sites. Afterward, we will present an experiment to explore the effect of verbal explanations and system error information on comprehensibility and trust. Furthermore, joint performance, as well as confidence in joint decision making, is assessed where joint performance means that a human receives the systems classification decision and can decide whether to follow it or not.

Category Learning in the Relational Ancient Grave Domain

Abstracting rules and forming concepts are basic constituents of human cognition. To investigate context effects on classification learning, Medin and Schaffer (1978) created a simple two-class problem with a set of stimuli based on four binary visual features, such as color, form, size, and position. Depending on the feature values, instances belong to one of two classes. We introduce the structure of the Medin and Schaffer stimuli on the ancient grave domain we use in our experiment. The four features are orientation, shape, order of inner stones, and number of outer stones (see Table 1) For instance, example e1 from Table 1 has the values <North, narrow, ascending, few>. The last feature – number of outer stones – is irrelevant for the class decision. An instance is assigned to the category Iron Age when two or more of the three relevant features have value 1 and to Viking Age when two or all three of them are 0. None of the features alone offers a sufficient cue for a clear categorization (Medin & Schaffer, 1978). The 2-of-3 rules which characterize the true classes are given in Figure 1. Note that there are no rules for the class Viking as this automatically applies if none of the Iron rules are true. The rules can be easily described verbally as "If a grave has value North on the dimension orientation and value narrow on the dimension shape, then it belongs to Iron Age".

Following Medin and Schaffer's (1978) 5-4 category structure, to learn the specific feature combinations, participants are presented nine labeled training examples –

(1) **North ∧ Narrow → *Iron***

(2) **North ∧ Ascending → *Iron***

(3) **Narrow ∧ Ascending → *Iron***

Figure 1: Classification rules for Iron Age graves.

five examples for Iron Age ("positive") and four examples for Viking Age ("negative") without any further information about the relevant dimensions or possible values. An example of an aerial view is given in Figure 2.

In contrast to the original Medin and Schaffer stimuli, the third feature (order of inner stones) is not a simple visual pattern, but a more general concept based on evaluating the relationship between the sizes of the entire row of an arbitrary number of at least three stones. This is a recursive relation which is true when – starting with the leftmost stone – each stone in the sequence of inner stones is larger than its left neighbor.

The correct rule to classify all possible stimuli of this ancient grave domain can be learned from the examples given in Table 1 using the inductive logic (ILP) programming system Metagol (Muggleton et al., 2018). This fact gives evidence that the nine examples constructed in accordance with the Medin and Schaffer category structure are sufficient to infer the target concept. Recent work in explainable artificial intelligence demonstrates (1) how verbal explanations can be generated from rules learned with ILP (Siebers & Schmid, 2019), and (2) how ILP can be applied to generate local explanations in a model agnostic way from a deep learning model such as a CNN for image classification (Rabold et al., 2019).

Experiment

For the introduced visual relational domain, we hypothesize that (1) verbal explanations increase the trust in an imperfect system, the joint performance of human and machine, the participant's confidence in their decisions, and their comprehension of the rules. We predict that (2) information about system errors increases the understanding of such errors. Furthermore, we assume (3) additive effects of verbal explanations and error information on trust, performance, and confidence.

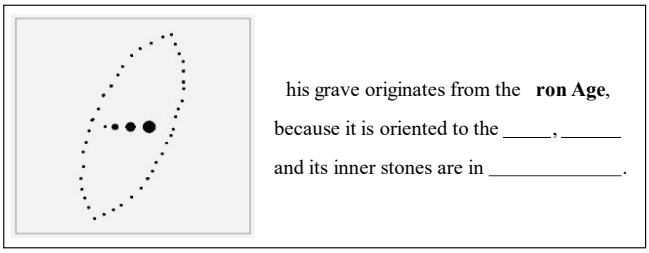

Figure 2: Stimulus example <1111> with the system's recommendation (line 1) and the verbal explanation (line 2-3) of the relevant features for the classification.

Table 2: Experimental design with the three experimental (EG) and the control group (CG).

		Verbal explanations	
		available	not available
System error information	available	*EG 1*	*EG 3*
	not available	*EG 2*	*CG*

Method

Design and Procedure The experiment is based on a complete 2×2 between-subject factorial design with the factors *verbal explanations* and *error information*, which is summarized in Table 2. The main parts of the experiment are: (1) initial unguided concept learning, (2) being shown new instances together with machine learned class decisions by the system with (EG 1, EG 2) or without (EG 3, CG) explanation, (3) joint decision making where participants receive system decisions (again with or without explanation) which they can follow or not, and (4) unguided classification of further graves by the participants. After part (2), the system error information has been given.

In accordance with the ultra-strong machine learning proposition (Muggleton et al., 2018) and the current research on human-AI partnership (Nguyen et al., 2018), we assume that explanations support joint decision making of humans and AI systems. That is, the AI system provides information and insights which the human alone would not have been able to come up with and at the same time allows the human to evaluate this information based on his or her experience. Consequently, the proposition of the AI system can be either accepted or rejected by the human based on the provided explanation. Combining the strength of both human and AI should result in a higher performance – reflected in the number of correct decisions – than of the AI system or the human alone. Furthermore, there is empirical evidence (Ai et al., 2021; Muggleton et al., 2018) that explicit verbal explanations support human understanding and learning. Consequently, giving explanations can result in better human performance in new tasks which have to be solved without support of the AI system.

The general procedure has been the following: First, the participants received a short introduction to machine learned classifiers as decision support systems and that such AI systems can be helpful for complex classification tasks. Afterward, they were made familiar with the ancient grave domain and were informed that a machine learned model is available to propose the age of a grave. Subsequently, the participants were shown nine aerial images of ship settings following the 5-4 structure from Table 1 divided into a group of "Iron Age" graves and of "Viking Age" graves. They were instructed to acquaint them with the two types of graves, and it can be assumed that participants in consequence, generalized at least a partial representation for the two types of graves. Following this concept learning task, a set of nine new sample stimuli was presented

together with the system's recommendations (see Figure 2, line 1). Part of the participants (EG 1 and EG 2) in addition was given verbal explanations (see Figure 2, line 2-3). These verbal explanations were systematically generated by the authors and not an actual XAI system to eliminate possible confounding factors (for explanation generation see Siebers & Schmid, 2019). Afterward, all participants received information about systems predictive accuracy, stating that the system had a success rate of 83.3% and part of the participants (EG 1 and EG 3) were additionally provided with information about the different error rates for Iron Age and Viking Age graves. In the following joint decision task, participants were given 12 new grave sites with the system's recommendation – again with verbal explanations for EG 1 and EG 2. For the grave site stimuli presented on positions 6 and 11 faulty system recommendations were given. For each of the presented grave images, participants were asked to decide which Age the grave originated from. hey could follow the system's recommendation or not. Additionally, they had to rate how confident they were with their decision from "just guessing" to "absolutely sure" on a 7-point scale. Finally, participants had to classify six new samples without the system's help to assess transfer (unguided classification). The experiment concluded with a trust questionnaire, an assessment of demographic information, and a debriefing.

Materials The grave stimuli were manually created using the software GIMP according to the binary codes in Table 1. Combining the four dimensions with the two possible values creates $2^4 = 16$ distinct stimuli for the category membership. However, as the experimental design required a higher number of grave stimuli, they were all slightly modified so that no ship setting resembles another one. Therefore, stimulus characteristics for the non-relational features are provided by the decision boundaries adopted from Rabold et al. (2018) and illustrated in (see Table 3). The row of inner stones consists of three to five stones in up to five different sizes. To be classified as ascending, all inner stones from left to right need to grow in size, but not necessarily linearly.

An example of a verbal explanation that has been used in the experiment is shown in Figure 2. Although mentioning two of the three relevant features would be enough, the system points out all three (if applicable) to prevent misunderstandings. The information about system errors was given as a statement before the actual classification task with the system's aid. Participants were told that the system is generally more susceptive to errors with Viking graves (19%) than Iron Age graves (3%) and were asked to be more attentive for "Iron Age" recommendations. his error allegedly occurred because the system confuses inner and outer stones for precisely horizontally aligned stimuli (0° rotation from the axis), and mistakenly classifies them as originating from the Iron Age. To assess trust, the shortened version of the questionnaire for human-computer trust by Madsen and Gregor (2000) has been used with the subscales *perceived understandability*, *perceived reliability,* and *faith*.

Table 3: Decision boundaries for binary value assignment to the non-relational features.

Dimension	Characteristics
Orientation	North: ± 21° from vertical East: ± 21° from horizontal
Shape	Narrow: axis ratio < 0.5 (*Mdn* = 0.36) Wide: axis ratio > 0.5 (*Mdn* = 0.66)
Number of outer stones	Many: > 35 (*Mdn* = 39) Few: < 25 (*Mdn* = 22)

Participants An international online questionnaire in both English and German ensured the large number of 243 voluntary participants, that had started the study. Out of the 213 completed questionnaires, 23 participants had to be excluded from the analysis post hoc because of low scores (< 3) on the 7-point scales ranging from "strongly disagree" to "strongly agree" for language comprehension (n = 4), distraction (n = 9) and efforts taken during the study (n = 3) as well as participants with implausibly fast reaction times (relative speed index > 2.0 *SD*) or response biases (n = 1). The final sample size comprised of 190 participants (63% women, 1% diverse), which were randomly distributed under the restriction of equal group sizes. Participants were mostly students from various study programs (63%) and employees from various working backgrounds (25%), ranging from age 18 to 67 (M = 27.2 years, SD = 10.09 years). The three experimental groups EG 1, EG 3, and CG did not differ significantly in age or gender distribution. EG 2 demonstrated a significantly higher proportion (p = .009) of male participants (43%) and a higher average age with greater variance (M = 31.7 years, SD =15.5 years) compared to EG 3 (M = 24.7 years, SD =6.9 years).

Results

A two-way analysis of variance (ANOVA)[1] was performed for the factors *verbal explanation* and *error information*. An a priori power analysis for medium effect size (f = .25, α = .05, 1-β = .90) gave a minimum required sample size of 171 participants, which has been met in the present study.

Joint Decision Making The descriptive findings on the joint human-machine performance are illustrated in Figure 3. The analysis of the performance scores showed a significant main effect for the factor verbal explanations, $F(1, 186) = 22.38$, $p < .001$, $\eta_{p2} = .11$. The mean confidence scores during this joint decision task of EG 2, EG 1, CG and EG 3 were $M_{E2} = 5.21$ ($SD_{E2} = 0.97$), $M_{E1} = 4.91$ ($SD_{E1} = 1.03$), $M_C = 4.73$ ($SD_C = 1.23$), $M_{E3} = 4.35$ ($SD_{E3} = 1.24$). We found a significant main effect for verbal explanations on the confidence in the decisions, $F(1, 186) = 10.03$, $p = .002$, $\eta_{p2} = .05$. The ANOVA on the confidence ratings in

[1] We decided for multiple ANOVAs and not a single MANOVA because of low positive or no significant correlation between the dependent variables.

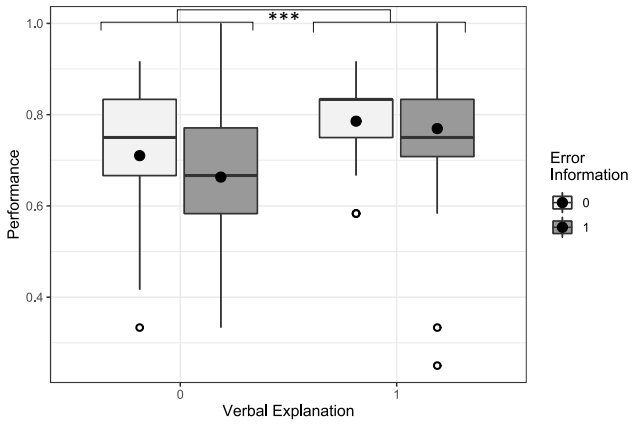

Figure 3: Classification accuracy in the joint decision task dependent on the verbal explanation and error information: 1 = available, 0 = not available (performance sores: min = 0, max = 1).

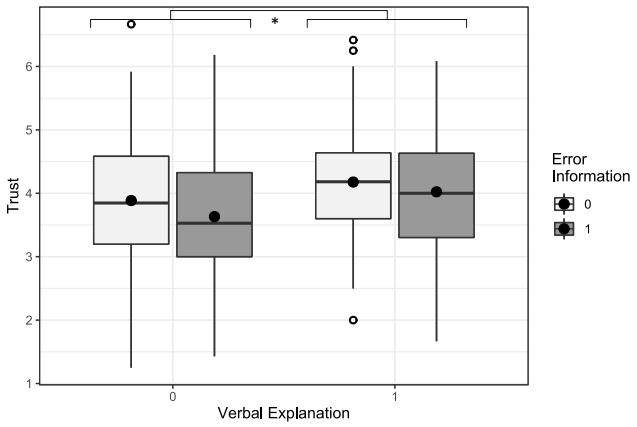

Figure 4: Trust scores dependent on verbal explanation and error information: 1 = available, 0 = not available (trust scores: min = 1, max = 7).

the joint decision task was the only one that showed a significant main effect for the factor error information, $F(1, 186) = 4.40$, $p = .037$, $\eta_{p2} = .02$.

Overall Trust The overall trust scores are descriptively illustrated in Figure 4. We found a significant main effect for the verbal explanations on the reported overall trust in the system, $F(1, 186) = 5.46$, $p = .021$, $\eta_{p2} = .03$.

Unguided Classification The correct application of the classification rules without the help of the system, shows a similar pattern to the previous findings. The mean accuracy in the transfer task was $M_{E2} = 0.78$ ($SD_{E2} = 0.19$), $M_{E1} = 0.77$ (SDE1 = 0.20), MC = 0.67 (SDC = 0.19), ME3 = 0.65 (SDE3 = 0.21). For the verbal explanation factor, the ANOVA showed a significant main effect on the individual performance in the transfer task, $F(1, 186) = 12.11$, $p < .001$, performance in the transfer task, $F(1, 186) = 12.11$, $p < .001$, $\eta_{p2} = .06$.

Exploratory Analysis

The exploratory ANOVAs on socio-demographics showed a significant main effect for the factor gender on the decision' confidence, $F(2,184) = 4.89$, $p = .009$. A background in computer science was found to interact with the factor error information on the confidence scores, $F(1,182) = 5.08$, $p = .025$ and with the factor verbal explanation on the joint performance scores, $F(1,182) = 4.04$, $p = .046$. Overall, the variables trust, joint performance, and confidence during the joint decision task demonstrate the same descriptive pattern: Groups with verbal explanations reported and scored higher than groups without the system's recommendations, and groups with error information reported and scored lower than groups without the display of error proneness. This results in the following order: EG 2 > EG 1 > CG > EG 3.

The only exception of this structure lays in the reports for error understanding. EG 1 ($M_{E1} = 4.09$, $SD_{E2} = 1.25$) with both verbal explanations and error information reported the highest understanding before EG 2 ($M_{E2} = 3.60$, $SD_{E2} =$

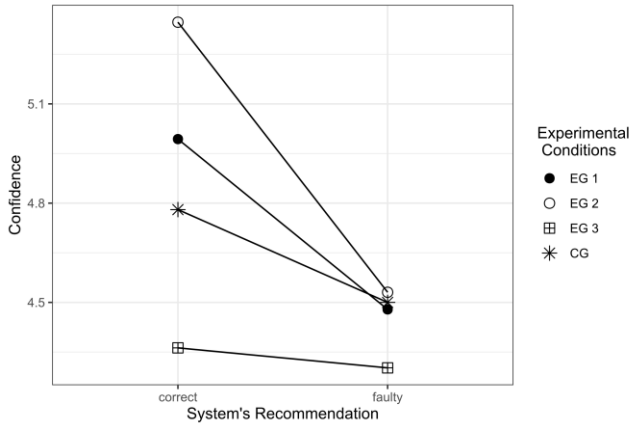

Figure 5: Mean confidence in decisions dependent on the validity of the recommendation and experimental condition.

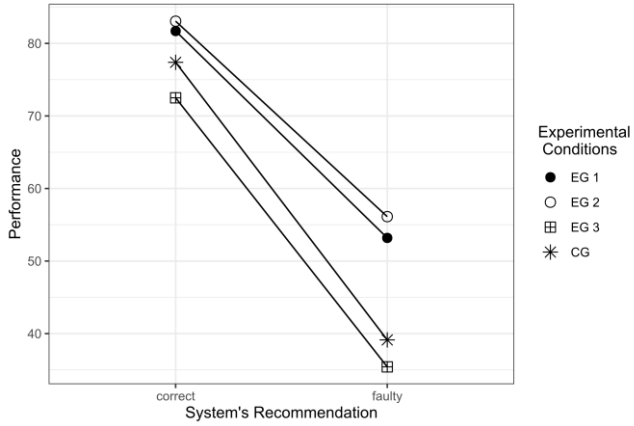

Figure 6: Mean decision accuracy in the joint decision task dependent on the validity of the recommendation and experimental condition.

1.40) and EG 3 (M_{E3} = 3.04, SD_{E3} = 1.65) with only one type each, and CG (M_C = 2.97, SD_{E3} = 1.47) without any explanation/information ranked last. Concerning the error understanding values, the factor verbal explanations reached significance, $F(1, 186) = 15.91$, $p < .001$, $\eta_p^2 = .08$, but no main effect for the factor error information.

The confidence during the joint decision task and the overall reported trust correlated significantly, $t(188) = 3.36$, $p < .001$, resulting in a sample estimate correlation of $r = .24$. Additionally, we post-hoc explored the performance and confidence for the two faulty recommendations in the joint decision task demonstrated in Figure 5 and Figure 6. The ANOVA showed a significant main effect for explanations on the accuracy scores for faulty recommendations, $F(1,186) = 9.06$, $p = .003$, $\eta p^2 = .05$. Similarly, we found a significant main effect on the confidence ratings for decisions with faulty predictions $F(1,116) = 4.85$, $p = .030$, $\eta_p^2 = .04$. All ANOVAs above demonstrated no significant interaction of the two explanation/information types.

Discussion

The findings confirm that verbal explanations increased the performance in the joint decision task and enhanced the confidence within the decisions. Groups with verbal explanations stated higher trust in the system and – measured in the transfer task accuracy – had a better comprehension of the classification rules. Error information affected the understanding of incorrect system answers, however, failed to reach significance. We also found no additive effect of verbal explanations and error information as later, contrary to our expectations, decreased all main dependent variables.

From our exploratory analysis, we conclude that verbal explanations not only encourage in (correctly) deciding contrary to the system's recommendations but also enhance the confidence for not relying on the system. The significant main effect for gender might be influenced by the higher proportion of male participants in the group EG 2 that received solely verbal explanations. The moderating effect of background in AI suggests that those participants are more aware that in machine learned models, errors can occur.

A potential limitation of the present study lays in the way the system error information was presented. This additional information about when a system is more prone to errors might have increased the task complexity. This higher demand for cognitive capacity could have affected the performance and confidence scores and be a possible explanation for the unexpected lower scores. However, we do not assume that this potentially higher complexity affected the reported trust.

Conclusion

We presented an experiment exploring the effect of explanations in human-AI partnerships on joint performance, human learning, and trust. Generating explanations to make machine learned black box classifiers comprehensible to humans, is a very active area of current AI research. Explainable AI (XAI) has been mainly concerned with explanations for image classifications. Many XAI approaches propose visual highlighting of relevant parts in the input image as an explanation. We presented the ancient grave domain as a relational visual domain where highlighting alone is not enough to convey the information relevant to classify an object. Explanations were presented in explicit verbal form. Generating verbal explanations from reasoning traces has already been proposed in the context of expert system research (Clancey, 1983). However, classic symbolic AI is well suited to deal with explicit, symbolic inputs, but not well suited for images. In our work, we assume an underlying hybrid system which combines black box deep learning for image classification with white box symbolic explanations (Rabold et al., 2019).

In the presented experiment, we addressed verbal explanations for image classification together with information about system errors. This second aspect, to our knowledge, has not been addressed in empirical work on XAI before. However, since machine learned models cannot be 100 percent correct by design, communication of system errors and system uncertainty is important for justified trust. The empirical findings show that verbal explanations of the classification decisions of a machine learned model improve the overall trust in the system. Explanations not only help to perform better and feel more confident in a classification task with faulty predictions but also enhance the comprehension of the general ground truth rules underlying the classification.

If the understanding of errors in an XAI application is crucial for domain experts or end-user, then it might be useful to offer both (local) verbal explanations and general (global) information for contexts in which the system is more prone to errors. However, when the focus lays in developing appropriate trust in a system, higher confidence in the interaction and the general performance, the results of our study suggest that verbal explanations for system decisions on specific instances are sufficient. Nevertheless, further experiments are necessary to give more insights into the interaction between explanations and perceived error such that future XAI systems support adequate trust resulting in human-AI partnerships which exceed the performance of a human or machine alone.

Acknowledgments

This work has been partially funded by the German Research Foundation (DFG), project Dare2Del (318286042) within the priority program within the priority program Intentional Forgetting. The authors cordially thank Johannes Rabold for training the ILP system Metagol. The paper is based on research done in the bachelor thesis of the first author as part of the psychology degree (B.Sc.) at the University of Bamberg.

The study complied with the ethical standards of the APA (see Standards 3.10 and 8.01—8.09 in the "Ethical

Principles of Psychologist and Code of Conduct," APA, 2002).

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Ai, L., Muggleton, S. H., Hocquette, C., Gromowski, M., & Schmid, U. (2021). Beneficial and harmful explanatory machine learning. *Machine Learning*.

Breznitz, S. (2013). *Cry Wolf: The Psychology of False Alarms*. Taylor and Francis.

Bruckert, S., Finzel, B., & Schmid, U. (2020). The Next Generation of Medical Decision Support: A Roadmap Toward Transparent Expert Companions. *Frontiers in Artificial Intelligence*, *3*, Article 507973, 75.

Bykov, K., Höhne, M. M.-C., Müller, K.-R., Nakajima, S., & Kloft, M. (2020, June 16). *How Much Can I Trust You? -- Quantifying Uncertainties in Explaining Neural Networks*. https://arxiv.org/pdf/2006.09000

Clancey, W. J. (1983). The epistemology of a rule-based expert system —a framework for explanation. *Artificial Intelligence*, *20*(3), 215–251.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Data Mining and Knowledge Discovery*, *9*(4), e1312.

Katsikopoulos, K. V., Şimşek, Ö., Buckmann, M., & Gigerenzer, G. (2020). *Classification in the wild: The science and art of transparent decision making*. The MIT Press.

Lakkaraju, H., Bach, S. H., & Jure, L. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. *Proc. Int. Conf. on Knowledge Discovery & Data Mining*, *2016*, 1675–1684.

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?". *Cognition*, *110*(2), 248–253.

Madsen, M., & Gregor, S. (2000). Measuring human–computer trust. *Proc. of 11th Australasian Conf. on Information Systems*, 6–8.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38.

Miller, T., Howe, P., & Sonenberg, L. (2017). *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*. http://arxiv.org/pdf/1712.00547v2

Muggleton, S. H., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A., & Besold, T. (2018). Ultra-Strong Machine Learning: Comprehensibility of programs learned with ILP. *Machine Learning*, *107*(7), 1119–1140.

Nguyen, A. T., Kharosekar, A., Krishnan, S [Saumyaa], Krishnan, S [Siddhesh], Tate, E., Wallace, B. C., & Lease, M. (2018). Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking. In *Proc. of the 31st Annual ACM Symposium on User Interface Software and Technology* (pp. 189–199).

Rabold, J., Deininger, H., Siebers, M., & Schmid, U. (2019). Enriching visual with verbal explanations for relational concepts–combining LIME with Aleph. In Workshops of *Joint Europ.Conf. on Machine Learning and Knowledge Discovery in Databases* (pp. 180-192). Springer.

Rabold, J., Siebers, M., & Schmid, U. (2018). Explaining Black-Box Classifiers with ILP – Empowering LIME with Aleph to Approximate Non-linear Decisions with Relational Rules. In F. Riguzzi et al. (Ed.), *Lecture Notes in Artificial Intelligence, Inductive Logic Programming: Proc. 28th Int. Conf.* (pp. 105–117). Springer.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "Why Should I rust You?": Explaining the Predictions of Any Classifier. In *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Demonstrations.* Association for Computational Linguistics.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., & Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, *2*(8), 476–486.

Siebers, M., & Schmid, U. (2019). Please delete that! Why should I? Explaining learned irrelevance classifications of digital objects. *KI - Künstliche Intelligenz*, *33*(1), 35–44.

Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 399–439.