How model accuracy and explanation fidelity influence user trust in AI

Andrea Papenmeier[1,2] , **Gwenn Englebienne**[2] , **Christin Seifert**[2]
[1]GESIS - Leibniz Institute of the Social Sciences, Germany
[2]University of Twente, Netherlands
andrea.papenmeier@gesis.org, g.englebienne@utwente.nl, c.seifert@utwente.nl

Abstract

Machine learning systems have become popular in fields such as marketing, financing, or data mining. While they are highly accurate, complex machine learning systems pose challenges for engineers and users. Their inherent complexity makes it impossible to easily judge their fairness and the correctness of statistically learned relations between variables and classes. Explainable AI aims to solve this challenge by modelling explanations alongside with the classifiers, potentially improving user trust and acceptance. However, users should not be fooled by persuasive, yet untruthful explanations. We therefore conduct a user study in which we investigate the effects of model accuracy and explanation fidelity, i.e. how truthfully the explanation represents the underlying model, on user trust. Our findings show that accuracy is more important for user trust than explainability. Adding an explanation for a classification result can potentially harm trust, e.g. when adding nonsensical explanations. We also found that users cannot be tricked by high-fidelity explanations into having trust for a bad classifier. Furthermore, we found a mismatch between observed (implicit) and self-reported (explicit) trust.

1 Introduction

The need for explanations of machine learning algorithms has been identified in the past [Richardson and Rosenfeld, 2018; Goodman and Flaxman, 2016] and led to the emergence of the research field of explainable artificial intelligence (xAI). Several researchers argue that explanations have a positive effect on user trust [Biran and Cotton, 2017; Glass *et al.*, 2008; Preece, 2018; Vorm, 2018] and that inappropriate trust impairs the human machine interaction [Preece, 2018; Ribeiro *et al.*, 2016]. However, explanations do not necessarily have to deliver accurate information about the machine learning algorithm. Yet, untruthful explanations with low fidelity to the machine learning model can appear plausible to the user [Lipton, 2016]. It has not yet been established how characteristics such as fidelity of an explanation impact user trust.

We therefore investigate how varying explanation fidelity influences the user's trust into an automatic decision system. Using the scenario of a "social media administrator" with the task to detect offensive language in Tweets, we develop three machine learning classifiers able to process textual input and classify the texts into "offensive" and "not offensive" classes at varying levels of accuracy. Furthermore, we implement and validate the automatic generation of explanations at high fidelity and low fidelity levels. We measure the trust and perceived understanding in a user study with 327 participants in order to compare different classifier-explanation combinations. Our research was driven by the following questions:

RQ1: What influence does the accuracy of an automatic decision system have on user trust?
RQ2: How do the presence and the level of fidelity of explanations influence user trust?

Our key findings show that explanations affect user trust in a variety of ways, depending on the overall accuracy of the system, the fidelity level of the explanation, and the user's level of consciousness. In general, the systems' accuracy levels were most decisive for user trust: the higher the accuracy, the higher the user's trust. The influence of explanation fidelity differs depending on the model accuracy: We see that for systems with medium accuracy, a high-fidelity explanation does not harm user trust, while a low-fidelity explanation does. Yet, for a system with high accuracy, any explanation (high-fidelity as well as low-fidelity) leads to a decrease in trust. We conclude that the interplay between explanation fidelity and user trust is more complex than pictured in literature to date. Furthermore, our findings show a discrepancy between how users act and what users report, which should be taken into account when evaluating user trust.

With our research, we contribute empirical evidence of the relation between accuracy, fidelity, and user trust to the xAI community. Other than related research, we focus on the practical implications of explainability and their effect on the relationship with the user. Furthermore, we test an observational measure of trust as an objective method complementing traditional self-reported trust questionnaires.

In this paper, we first review the existing literature on explanations and user trust in AI. We then derive the structure for a user study (section 3 and 4) aiming to test the influence of explanation fidelity and classifier accuracy on user trust. Finally, we present and discuss the results in section 5.

IJCAI 2019 Workshop on Explainable Artificial Intelligence (xAI), August 11th, 2019. Macau, China

2 Related Work

Artificial intelligence and machine learning algorithms are nowadays employed in a variety of areas. In safety-critical applications such as terrorism detection [Ribeiro *et al.*, 2016] or autonomous robotics [Richardson and Rosenfeld, 2018], faulty behaviour needs to be avoided at all costs. Furthermore, machine learning systems treating sensitive data such as credit ratings [Domingos, 2012] or health applications [Goodman and Flaxman, 2016] need to communicate what brought about a single decision. [Goodman and Flaxman, 2016; Wachter *et al.*, 2017; Selbst and Powles, 2017] discuss a "right to explanation" or "right to information" as a consequence of the General Data Protection Regulation (GDPR) introduced in the EU in 2018. Overall, those systems not only need to be right in a high number of cases, but right for the right reasons [Preece, 2018].

2.1 Explanations in AI

When being confronted with new information, humans incorporate them in mental models. Explanations are a tool to build and refine inner knowledge models [Miller, 2018]. For an engineer working on a machine learning system, understanding underlying principles and consequences of the system's behaviour is a necessary step in designing a system that is "right for the right reasons" [Preece, 2018]. On the user side, explanations have a positive effect for the ability to predict the system's performance correctly [Biran and Cotton, 2017]. [Ribeiro *et al.*, 2018] found that explanations increase the user's ability to predict the classifier decision, while decreasing the time needed to reach a judgement. Their within-subject study design, however, could have led to familiarisation and hence an overrating of explanations.

In recent years, machine learning algorithms show a trend towards increasing accuracy, but also increasing complexity. In general, the higher the accuracy and complexity, the lower the explainability [Chen *et al.*, 2018; Richardson and Rosenfeld, 2018]. An interpretable machine learning system is either inherently interpretable (e.g. decision trees, linear models [Biran and Cotton, 2017]), or is capable of generating descriptions understandable to humans [Lipton, 2016]. [Lipton, 2016] points out that a retrospectively added explanation does not guarantee fidelity, "however plausible they appear".

To achieve explainability, [Chen *et al.*, 2018] developed an add-on explanation system for texts based on mutual information analysis and measure the explanations fidelity to the underlying model with good results. [Feng *et al.*, 2018] went a step further with an image classification system and add-on textual explanatory mechanism. However, they also show that their high-fidelity explanations are nonsensical for humans. In human-human explanations, people tend to question underlying principles of events by comparing it to known concepts. "Why A, why not B?" is a common question during this thought process [Miller, 2018]. [Chen *et al.*, 2018] suggests showing reference cases in automatic decision systems: similar cases with a different predicted class, or dissimilar cases (counterfactuals) [Hendricks *et al.*, 2018]. Approximating elements of an opaque system is another method of achieving interpretability. [Domingos, 2012] argues that most high-dimensional real-world application data is "concentrated on or near a lower-dimensional manifold" and suggests dimension reduction techniques to reduce the complexity of a system to a human-comprehensible level. [Chen *et al.*, 2018] suggests salience map masks on input features to point the attention towards features that are decisive in a sample, e.g. single words in texts. [Goodman and Flaxman, 2016] suggests a "minimum explanation", showing at least how input features relate to the prediction of a classifier.

In summary, [Chen *et al.*, 2018]'s model-agnostic explanations combined with [Goodman and Flaxman, 2016]'s definition of minimum explanations provide a basis for examining the influence of explanation fidelity and model accuracy. However, following [Lipton, 2016], explanation fidelity needs to be validated computationally.

2.2 Trust in AI

Literature suggests that insights into the system functioning and decision process increase trust [Biran and Cotton, 2017; Glass *et al.*, 2008; Preece, 2018; Vorm, 2018]. In the field of computer science, most definitions agree in that trust relates to the assurance that a system performs as expected [Mohammadi *et al.*, 2013]. Since trust is placed in an agent by another agent, it is not an objective measure but a subjective experience of an individual [Mohammadi *et al.*, 2013]. [Körber, 2018] developed a trust metric for automated systems based on a model of human-human trust. It consists of 19 self-report items measuring the trust factors reliability, predictability, the user's propensity to trust, as well as the attitude towards the system's engineers and the user's familiarity with automated systems.

For trust in automatic classification systems, misclassifications (i.e. the system's prediction does not correspond to the user's prediction) play a special role, as they can lead to a decrease in user trust [Glass *et al.*, 2008]. [Vorm, 2018] reports "willingness to accept a computer-generated recommendation" as an observable sign for trust. [Yu *et al.*, 2017] found that users are able to detect the accuracy of a classifier without being told explicitly, and adjust their trust accordingly. [Cramer *et al.*, 2008] tested the effects of transparency on user perception, finding a correlation between perceived understanding and trust, but no evidence for a direct influence of transparency on trust. They hypothesise that transparency also discloses system boundaries and unfulfilled preferences, ultimately cancelling out any positive effects. [Langer *et al.*, 1978] found in a user study that the pure presence of an explanation, regardless of the content, can make a difference in how people react to requests. Without explanation, humans complied significantly less with a request than in cases where an explanation was given. They compared nonsensical and meaningful explanations, but found only little difference in their power of persuasion [Langer *et al.*, 1978]. They explain this behaviour with the state of "mindlessness", triggering an automatic script "comply if reason is given", no matter the given reason. The mindless state, however, is revoked if complying leads to stronger consequences. In an attentive state, the explanation does make a difference: People were more likely to comply when an informative explanation was given, as compared to a nonsensical one [Langer *et al.*, 1978].

Overall, evaluating trust implies measuring the subjective experience of users. Since [Langer *et al.*, 1978] observed a difference in user trusting behaviour between a "mindful" and a "mindless" state, trust should be evaluated both subjectively and objectively, e.g. using a questionnaire and observation.

3 Study Design

As trust is a subjective experience, it must be evaluated in a user study. We use the following scenario for a user study: the social media presence of a company that targets teenagers and young adults (15-20 years old). The use case task is to identify offensive texts with the support of a machine learning system. To measure the influence of accuracy and explanation fidelity on user trust, we establish 9 conditions: three classifiers (high, medium, low accuracy), each with three explanation types (high-fidelity, low-fidelity, no explanation), see table 1. To avoid learning and familiarisation effects, we use a between subject design, with each participant being assigned to one condition at the beginning of the survey. As trust builds during repeated interaction [Rempel *et al.*, 1985], we show participants a subset of 15 Tweets. We construct 10 disjunct subsets (cf. sec 4), to reduce the impact of specific wording or topics. At the start of the survey, each participant is randomly assigned to one subset.

		Classifier Accuracy		
		high	medium	low
Explan.	high-fidelity	$C_{0.97}^{high}$	$C_{0.76}^{high}$	$C_{0.03}^{high}$
	low-fidelity	$C_{0.97}^{low}$	$C_{0.76}^{low}$	$C_{0.03}^{low}$
	no	$C_{0.97}^{no}$	$C_{0.76}^{no}$	$C_{0.03}^{no}$

Table 1: Classifier-explanation conditions

Apparatus & Procedure

The user study is set up as an online study on the soSci platform[1]. Participants are asked to access the survey via an online link on their private device, with small screens (e.g. smartphones) being excluded to ensure proper image scaling. Consistent with the use case scenario, screenshots of a fictive social media management platform show the input texts, decisions and explanations. The screenshots have a ratio of 900px (width) to 253px (height).

The study consists of three blocks. In the first block, the participant is asked to manually classify 15 Tweets as offensive or not offensive. The second block introduces the automatic decision system, asking to classify 15 "very similar" Tweets, which are in fact identical to those in the first block. The Tweets are pre-classified and displayed according to one of the 9 conditions. Finally, the last block contains questions to measure perceived understanding, trust (including an attention check), and the demographic background. The survey was tested in a pilot with 11 participants.

Measures & Analysis

We measure perceived understanding and trust quantitatively. For perceived understanding, we ask three statements to be rated on a 5-points Likert scale and take the average as a single score per participant. To measure trust, we observe how the system influences the participant's judgement by comparing the manual classifications of the first (without system) and second block (with system). We define changing a classification in favour of the system's prediction but away from the truth as a sign for being convinced and trusting the system. The opposite behaviour (changing towards the truth but away from the system's prediction) is interpreted as a sign for mistrust. We normalise the number of changes by the number possibilities to see the behaviour in question (e.g. a highly accurate classifier offers only once the possibility to contradict the truth in favour of its prediction). As a subjective, self-reported trust measure, we use the questionnaire of [Körber, 2018], taking the mean score over all 19 items for a single trust score per participant. We use the two-sided Mann-Whitney U test with Bonferroni correction to compare two score samples.

Participants

Participants were recruited via the science crowdsourcing platforms Prolific[2] and SurveyCircle[3]. In total, 327 participants took part in the main user study with an average age of 29.4 years (SD=8.8), with 56% females and 43% males. Two participants reporting the third gender. 57% self-assessed their English as equivalent to a native speaker, but all participants claimed to be fluent in English. 41 data points were invalidated due to failed attention check and survey completion level, resulting in 286 valid cases.

4 Experiment

Dataset

We use a dataset of offensive language and hatespeech[4] provided by [Davidson *et al.*, 2017]. It contains Tweets labelled by at least 3 annotators, of which we use only those data points with an inter-annotator agreement of 100%. The final dataset contains 4324 Tweets with a class balance of 1:1. We randomly split the data set into training (80%) test (20%). The Tweets are preprocessed with a conversion to lower cases, common contraction solving (e.g. "we're"), deletion of retrospectively added signifiers (e.g. "RT" indicating a Re-Tweet), deletion of non-alphabetic characters (all besides hashtags), and replacement of URLs and user names by dummy handles. The texts are tokenized on whitespaces.

Classifiers

For the system with **high accuracy**, we adopt the setup used by [Chen *et al.*, 2018]. They use a convolutional neural network (CNN) for sentiment analysis. We implement the CNN using the *Keras*[5] Python library. $C_{0.97}$achieves an accuracy of

[1] https://www.soscisurvey.de/en/index accessed on 24.02.2019

[2] https://prolific.ac accessed on 24.02.2019

[3] https://www.surveycircle.com accessed on 24.02.2019

[4] https://github.com/t-davidson/hate-speech-and-offensive-language accessed on 24.02.2019

[5] https://keras.io accessed on 24.02.2019

0.97 on the test set. For the classifier with **medium accuracy** ($C_{0.76}$) we adapt the approach of [Davidson *et al.*, 2017], which uses logistic regression to identify offensive language and hate speech, achieving an F1-score of 0.9 on their test set. The logistic regression classifier is implemented with the *scikit-learn*[6] Python library, with an L-BFGS optimiser. We adjust all positive coefficients of $C_{0.76}$to a value of 1.0 and all negative to -1.0 to reach the final (medium) accuracy of 0.76 on the test set. The **low accuracy** classifier is essentially equal to $C_{0.97}$, but trained on a training set with inversed labels. $C_{0.03}$'s accuracy on the test set with non-inversed labels is 0.03.

Explanations

For generating explanations, we focus on input features (single words) and influence on the prediction, as suggested in the minimum explanation setup by [Goodman and Flaxman, 2016]. Following [Feng *et al.*, 2018], we highlight the most decisive words in the texts by colour. To convey just enough explanation, we highlight between $\frac{1}{3}$ and $\frac{1}{4}$ of the texts. The Tweets contain on average between 14 and 15 words, which results in $k = 4$ highlighted words per Tweet. To generate **high-fidelity explanations** for $C_{0.97}$and its inverse-label counterpart $C_{0.03}$, we use the L2X algorithm suggested by [Chen *et al.*, 2018] on top of the CNN to select the most decisive features. For $C_{0.76}$, we use the learned model coefficients. The **low-fidelity explanations** should not provide useful information about the underlying model, but should only be visually similar to the high-fidelity explanations. To generate such nonsensical explanations, we draw words uniformly at random from the texts.

As the explanation is not tied to the classification result, the system can also show **no explanations** by not highlighting any word ($k = 0$) but still show the classifier's prediction.

Subset Sampling

It is not feasible to show the complete dataset to the participants during the user study, we therefore display a subset of 15 Tweets. To avoid affects from specific wording or topics in the subset, we generate subsets by drawing 15 Tweets at random from the test set. The subset is only kept if it is non-overlapping with previously drawn subsets, has a class balance similar to the test set, and if the classifiers' accuracies are equal to those on the test set. We then select the 10 subsets with the closest feature distribution compared to the training set, using the Kullback-Leibler Divergence (KLD) with Laplace smoothing (k=1).

Explanation Evaluation

For computer-generated explanations, it is possible that (1) the explanations constructed to have a high fidelity are meaningful to humans but are not faithful to the model, that (2) low-fidelity explanations nonetheless convey information about the classifier, and that (3) the explanations in the subsets show a different fidelity as those in the whole test set.

To validate that the selected features are an actual representation of the classifier's reasoning, we reduce the texts of the test set to the k selected features and subsequently let the respective classifier predict the label. If the explanations have a high fidelity to the underlying model, the reduced texts should lead to the same predictions as the original texts. We use the prediction for the original texts as ground truth for the reduced texts to calculate the label agreement. We repeat the evaluation for each subset to confirm that the fidelities of the explanations in the subsets do not differ from those of the complete test set. Table 2 shows that the high-fidelity explanations are enough to reproduce the original prediction of all three classifiers, even when reducing the texts to a single word. The low-fidelity explanations, on the contrary, cannot reliably reproduce the original predictions. We conclude that (1) the high-fidelity explanations indeed faithfully represent the underlying classifier, and that (2) a random selection of words is not explanatory for the classifiers.

	$C_{0.97}^{high}$	$C_{0.97}^{low}$	$C_{0.76}^{high}$	$C_{0.76}^{low}$	$C_{0.03}^{high}$	$C_{0.03}^{low}$
$k = 1$	0.97	0.58	1.00	0.64	0.97	0.59
$k = 4$	**0.98**	**0.74**	**1.00**	**0.77**	**0.97**	**0.72**
$k = all$	1.00	1.00	1.00	1.00	1.00	1.00
$\bar{x}_{subs}$	0.97	0.74	1.00	0.64	0.97	0.74
s_{subs}	0.03	0.13	0.00	0.12	0.03	0.10

Table 2: Label agreements evaluating the fidelity of explanations. Showing the accuracy of reduced texts when prediction of complete text is set as ground-truth for test set (top) and subsets (bottom, $k = 4$). Class balance of 50:50 for both the test set and each subset.

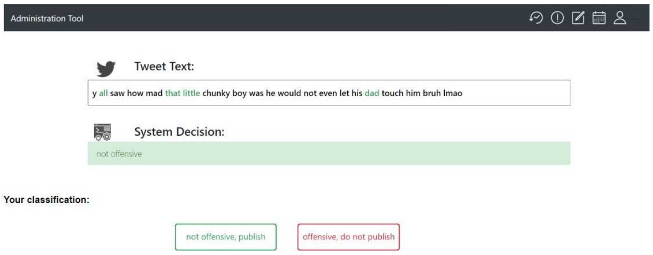
On average, the mean fidelities ($\bar{x}_{subs}$) of the subsets cor2respond to those on the complete test set. Only few subsets show differing fidelities, e.g. for $C_{0.97}^{low}$, two sets have a fidelity lower than the average (0.53), while one subset shows a higher fidelity level (0.93). The standard deviation of the subset fidelities (s_{subs}) are higher for randomly selected explanations than for high fidelity explanations, which was to be expected.

Graphical User Interface

For testing the effect of explanations of an automatic decision tool on users, we create an authentic and modern web interface with a minimalistic design, as to not distract the user from the main task. Figure 1 shows the "Administration Tool", a software tool to support a social media administrator in detecting offensive content.

5 Results and Discussion

Table 3 reports the mean scores and their standard deviations for *self-reported trust, perceived understanding*, and


Administration Tool

Tweet Text:

y all saw how mad that little chunky boy was he would not even let his dad touch him bruh lmao

System Decision:

not offensive

Your classification:

Figure 1: Screenshot of the graphical user interface in the survey

[6] https://scikit-learn.org accessed on 24.02.2019

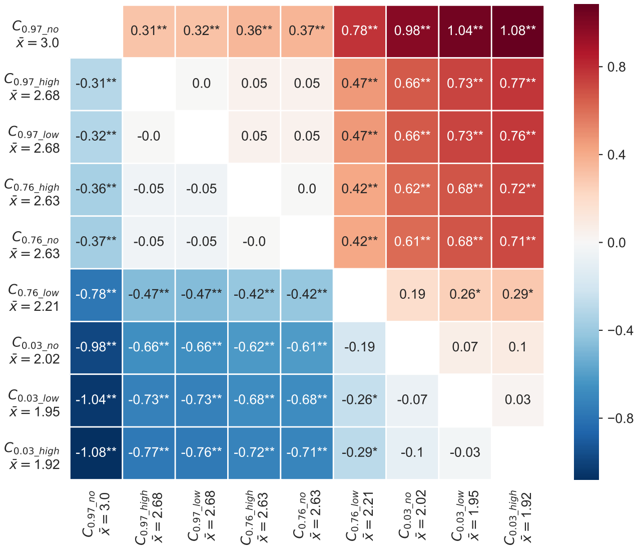


Figure 2: Comparison of self-reported trust scores ordered by mean ($\bar{x}$), value reporting difference of means ($\bar{x}_{row} - \bar{x}_{column}$), asterisk reporting significance (* significant at $\alpha = \frac{0.05}{9}$, ** significant at $\alpha = \frac{0.01}{9}$)

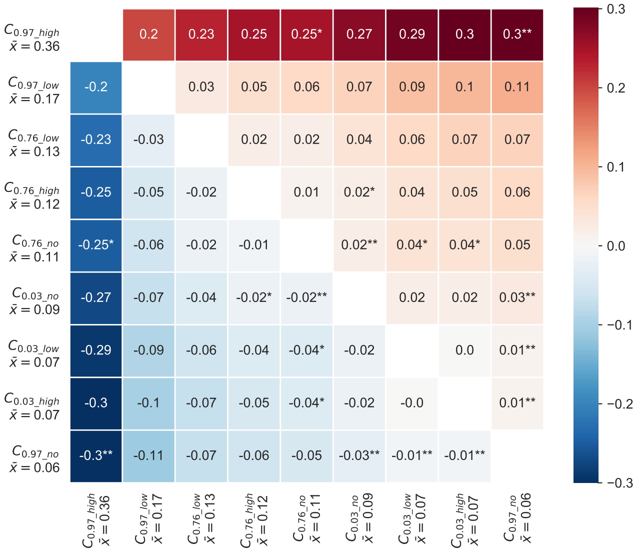


Figure 3: Comparison of observed trust scores (relative changes towards classifier away from truth) ordered by mean, value reporting difference of means ($\bar{x}_{row} - \bar{x}_{column}$), asterisk reporting significance (* significant at $\alpha = \frac{0.05}{9}$, ** significant at $\alpha = \frac{0.01}{9}$)

perceived predictability (an individual item from the self-reported trust questionnaire). Figures 2-3 show the differences in means between all conditions and the significance of the sample comparisons (denoted with an asterisk) for self-reported trust and *observed trust* trust. All significance scores use Bonferroni correction to account for the multiple comparisons bias. Table 4 presents the results of observed trust in changes towards or away from the truth and the classifier's prediction, respectively.

5.1 Model Accuracy

Our results suggest that model accuracy has a stronger influence on user trust than explanation fidelity. Figure 2 shows that when ordered by trust score, no system of $C_{0.03}$ is ranked higher than any system of $C_{0.76}$ or $C_{0.97}$, and no system of $C_{0.76}$ is ranked higher than any system of $C_{0.97}$. These findings are in line with the research of [Yu *et al.*, 2017]. It also aligns with the "expectation mismatch" described in [Glass *et al.*, 2008]: A classifier with high accuracy leads to fewer mismatches with the user's expectations, which in turn does not decrease the trust. We observe the same trend for user ratings of predictability: low accuracy systems are rated to be less predictable than high accuracy systems. Both classifiers, however, are objectively equally predictable, since they behave exactly the same ($C_{0.03}$always returns the opposite label from $C_{0.97}$; their results hence have equal entropy). This suggests that user's perception is heavily influenced by accuracy levels.

Condition	self-rep. trust	perceived understanding	
		understanding	predict.
$C_{0.97\_high}$	2.7 ± 0.4	3.9 ± 0.9	3.0 ± 0.8
$C_{0.97\_low}$	2.7 ± 0.5	3.7 ± 0.9	2.9 ± 0.8
$C_{0.97\_no}$	3.0 ± 0.5	4.1 ± 0.7	3.2 ± 0.7
$C_{0.76\_high}$	2.6 ± 0.5	3.8 ± 0.8	2.9 ± 0.6
$C_{0.76\_low}$	2.2 ± 0.5	2.9 ± 1.0	2.3 ± 0.8
$C_{0.76\_no}$	2.6 ± 0.5	3.7 ± 0.7	2.7 ± 0.8
$C_{0.03\_high}$	1.9 ± 0.4	2.5 ± 1.2	2.0 ± 0.6
$C_{0.03\_low}$	2.0 ± 0.4	2.5 ± 1.1	1.8 ± 0.7
$C_{0.03\_no}$	2.0 ± 0.5	2.9 ± 1.2	2.1 ± 0.8

Table 3: Means and standard deviations for self-reported trust, perceived understanding, and predictability scores.

5.2 Explanations

In our experiment, the presence of an explanation did not have a positive effect on self-reported trust in any of the conditions (figure 2). Adding an explanation to the system decreased the trust in the case of $C_{0.97}$ and did not influence trust in $C_{0.03}$. For $C_{0.76}$, the type of explanation was crucial for its influence – a high-fidelity explanation did not decrease trust levels significantly, while a low-fidelity explanation did.

For $C_{0.97}$, $C^{no}_{0.97}$ shows better results than $C^{high}_{0.97}$ and $C^{low}_{0.97}$. With $C^{no}_{0.97}$, there is no "expectation mismatch" as no explanation is given and accuracy is high. The explanations of $C^{high}_{0.97}$, however, are built on statistical rather than causal relations, while $C^{low}_{0.97}$'s explanations are random. As humans make sense of new observation by using previously learned knowledge, i.e. assuming human-like reasoning strategies even for an algorithm, seeing any of those two explanations leads to a deceptive experience. Contrarily, the observed trust measure (figure 3) shows a significantly lower score for $C^{no}_{0.97}$ than for $C^{high}_{0.97}$, meaning that participants show a higher willingness to accept the predictions of $C^{high}_{0.97}$ than for $C^{no}_{0.97}$.

			Truth	
			Towards	Away
Classifier Prediction	$C^{high}_{0.97}$	Towards	0.32	0.36
		Away	0.00	0.02
	$C^{low}_{0.97}$	Towards	0.19	0.17
		Away	0.00	0.03
	$C^{no}_{0.97}$	Towards	0.26	0.06
		Away	0.00	0.02
	$C^{high}_{0.76}$	Towards	0.27	0.12
		Away	0.09	0.01
	$C^{low}_{0.76}$	Towards	0.19	0.13
		Away	0.19	0.05
	$C^{no}_{0.76}$	Towards	0.33	0.11
		Away	0.05	0.01
	$C^{high}_{0.03}$	Towards	0.17	0.07
		Away	0.11	0.07
	$C^{low}_{0.03}$	Towards	0.25	0.07
		Away	0.10	0.00
	$C^{no}_{0.03}$	Towards	0.33	0.09
		Away	0.04	0.00

Table 4: Influence of systems on user labelling behaviour: relative changing frequencies when confronted with system prediction, per classifier, normalised over opportunities.

[Langer *et al.*, 1978] noticed a difference between a "mindless" (non-attentive) and a "mindful" (attentive) state, which could be the explanation for the difference between a self-reported (attentively) and an observed (non-attentively) measure. Users do not report different trust levels for $C^{high}_{0.97}$ and $C^{low}_{0.97}$, but they more often follow $C^{high}_{0.97}$'s recommendation than $C^{low}_{0.97}$'s (table 4) – their behaviour is hence influenced by the level of truthfulness.

Unlike $C_{0.97}$, $C_{0.76}$ shows an equal self-reported trust score for $C^{high}_{0.76}$ and $C^{no}_{0.76}$ and a significantly lower score for $C^{low}_{0.76}$. Making three to four mistakes on each subset, it is imaginable that users are more conscious about the classifier's behaviour than they are with $C_{0.97}$ due to the higher error rate. Having at least an indication of the reasons for misclassifications ($C^{high}_{0.76}$) could in turn increase the trust. For $C^{low}_{0.76}$, the "expectation mismatch" is twofold, bringing together misclassifications and nonsensical explanations. Looking at the observed trust, $C^{high}_{0.76}$ has the highest trust rate, while $C^{low}_{0.76}$ has the highest rate of mistrust.

$C_{0.03}$ did not show evidence of diverting self-reported trust scores for any of the three explanation types. The same homogeneity is found in the observed trust scores, for both trust and mistrust. This suggests that users are not fooled by a bad classifier and do not trust it, no matter the explanation given.

5.3 Objective Trust Measure

Self-reporting requires users to have the ability to reflect on and process their relationship with the system. Using an objective measure for trust avoids the necessity of this ability. In our results, we see that the observed (hence potentially unconscious) trust scores do not always align with the self-reported trust scores. Although users of $C^{no}_{0.97}$ have the highest self-reported trust score, they are not as easily "lured" towards a wrong classification as users of $C^{high}_{0.97}$. If this is due to an actual gap between actions and reflections of users, the observation measure could be interesting for xAI practitioners as it shows how users actually interact with a system. However, as our results of the observed trust measure are ambiguous and have high variance, the measure should be validated in future research.

5.4 Limitations

In this study, we make use of minimum explanations which show the relation between the input and output but do not deliver information about the inner structure of a classifier. The influence of the task (difficulty) and explanation visualisation (detailedness) should be further investigated. It should also be tested in future research whether users accept only high-fidelity explanations, or likewise accept explanations that look meaningful to a human but are not faithful to the underlying machine learning algorithm. The study results are furthermore limited by the cultural background of the participants (mainly Caucasians). The results therefore cannot be generalised across cultural backgrounds and the connected general attitude towards technology.

6 Conclusion

This paper presents empirical evidence for the impact of model accuracy and explanation fidelity on user trust. We generated minimal explanations with high and low fidelity for three systems with different performance levels. We then validated the explanations' fidelity level and tested differences in nine conditions (3 model accuracy levels x 3 explanation fidelity levels) in a user study.

Our findings show that explanations affect user trust in a variety of ways, depending on the overall accuracy of the system, the fidelity of the explanation, and the user's level of consciousness. Participants showed the most trust in systems without explanations, i.e. minimum explanations can potentially harm, but not improve user trust. We argue that the act of reconciling conflicting information of the mental model and the given explanations counts as a deceptive experience and therefore affects the user's trust negatively. If an explanation is added to a system (e.g. for increasing user's understanding of the system), its fidelity is crucial for user trust. We saw that for systems with a medium accuracy ($C_{0.76}$), a high-fidelity explanation does not harm user trust, while a low-fidelity explanation decreases trust. Overall, the model's accuracy levels showed the most impact on trust levels. We furthermore found that users' awareness level influences their perception of trust. The results found from self-reported trust measures show a different picture than when objectively observing trust via the participant's actions.

Further research with more rich explanations and a detailed investigation of trust factors is needed to examine potential positive effects of explanations on user trust. The development of trust over time should also be researched in the future, to give practical directions to xAI practitioners implementing explanations in productive systems.

References

[Biran and Cotton, 2017] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 8, 2017.

[Chen *et al.*, 2018] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 883–892. PMLR, 10–15 Jul 2018.

[Cramer *et al.*, 2008] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, 2008.

[Davidson *et al.*, 2017] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515, 2017.

[Domingos, 2012] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[Feng *et al.*, 2018] Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, 2018.

[Glass *et al.*, 2008] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236. ACM, 2008.

[Goodman and Flaxman, 2016] Bryce Goodman and Seth Flaxman. Eu regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38, June 2016.

[Hendricks *et al.*, 2018] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Generating counterfactual explanations with natural language. In *ICML Workshop on Human Interpretability in Machine Learning*, 2018.

[Körber, 2018] Moritz Körber. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer, 2018.

[Langer *et al.*, 1978] Ellen J Langer, Arthur Blank, and Benzion Chanowitz. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of personality and social psychology*, 36(6):635, 1978.

[Lipton, 2016] Zachary Lipton. The mythos of model interpretability. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. ICML, 2016.

[Miller, 2018] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2018.

[Mohammadi *et al.*, 2013] Nazila Gol Mohammadi, Sachar Paulus, Mohamed Bishr, Andreas Metzger, Holger Könnecke, Sandro Hartenstein, Thorsten Weyer, and Klaus Pohl. Trustworthiness attributes and metrics for engineering trusted internet-based software systems. In *International Conference on Cloud Computing and Services Science*, pages 19–35. Springer, 2013.

[Preece, 2018] Alun Preece. Asking 'why' in ai: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018.

[Rempel *et al.*, 1985] John K Rempel, John G Holmes, and Mark P Zanna. Trust in close relationships. *Journal of personality and social psychology*, 49(1):95, 1985.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[Ribeiro *et al.*, 2018] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

[Richardson and Rosenfeld, 2018] Ariella Richardson and Avi Rosenfeld. A survey of interpretability and explainability in human-agent systems. In *XAI Workshop on Explainable Artificial Intelligence*, pages 137–143, 2018.

[Selbst and Powles, 2017] Andrew D Selbst and Julia Powles. Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4):233–242, 2017.

[Vorm, 2018] Eric S Vorm. Assessing demand for transparency in intelligent systems using machine learning. In *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pages 1–7. IEEE, 2018.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[Yu *et al.*, 2017] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317. ACM, 2017.