It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI

ANDREA PAPENMEIER and DAGMAR KERN, GESIS – Leibniz Institute for the Social Sciences, Germany
GWENN ENGLEBIENNE, University of Twente, Netherlands
CHRISTIN SEIFERT, University of Duisburg-Essen, Germany and University of Twente, Netherlands

Automated decision-making systems become increasingly powerful due to higher model complexity. While powerful in prediction accuracy, Deep Learning models are black boxes by nature, preventing users from making informed judgments about the correctness and fairness of such an automated system. Explanations have been proposed as a general remedy to the black box problem. However, it remains unclear if effects of explanations on user trust generalise over varying accuracy levels. In an online user study with 959 participants, we examined the practical consequences of adding explanations for user trust: We evaluated trust for three explanation types on three classifiers of varying accuracy. We find that the influence of our explanations on trust differs depending on the classifier's accuracy. Thus, the interplay between trust and explanations is more complex than previously reported. Our findings also reveal discrepancies between self-reported and behavioural trust, showing that the choice of trust measure impacts the results.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI);**

Additional Key Words and Phrases: Explainable AI, machine learning, minimum explanations, user trust, explanation fidelity

ACM Reference format:
Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 35 (March 2022), 33 pages.
https://doi.org/10.1145/3495013

In recent years, the need for a right to explanation has grown, both for legal and ethical reasons. With the European **General Data Protection Regulation** (**GDPR**) coming into effect in 2018, a debate on a "right to explanation" has arisen. It has been argued that a "good" and ethical explanation should be sound, i.e., faithful to the underlying reasoning [20, 26]. However, state-of-the-art

Partly funded by the Cancer Research Center Cologne Essen (CCCE).

Authors' addresses: A. Papenmeier and D. Kern, GESIS—Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6–8, Cologne 50667, Germany; emails: {andrea.papenmeier, dagmar.kern}@gesis.org; G. Englebienne, University of Twente, Drienerlolaan 5, Enschede 7522 NB, Netherlands; email: g.englebienne@utwente.nl; C. Seifert, University of Duisburg-Essen, Girardetstr. 2, Essen 45131, Germany and University of Twente, Drienerlolaan 5, Enschede 7522 NB, Netherlands; email: christin.seifert@uni-due.de.

1073-0516/2022/03-ART35
https://doi.org/10.1145/3495013

machine learning algorithms such as deep neural networks are highly complex—providing full transparency can be overwhelming to the user [50]. Gilpin et al. [15] therefore suggest to account for the limited human attention span and cognitive abilities when designing explanations, e.g., by showing what brought about a single decision, rather than explaining the machine learning model as a whole. In general, state-of-the-art explanations for current machine learning systems select only a fraction of the available information [43]. To comply with the GDPR, the selected information needs to ensure fidelity: Article 22 imposes a right to "information about the logic involved". Pervasive yet unfaithful explanations that do not represent the inner logic of a system, therefore, do not suffice for the GDPR. However, an experiment by Langer, Blank, and Chanowitz [27] from the field of psychology shows that the informativeness of explanations may play a minor role in human-human interaction and Eiband et al. [12] provided an indication that this holds also true for their experiment in human-machine interaction. Literature also claims that explanations have a positive effect on user trust [4, 16, 40, 51]. However, recent research provides conflicting evidence concerning this claim: Lim et al. find a positive effect of specific explanation types on trust [31], whereas other research finds no impact of transparency on trust [8, 9]. Studies on user trust are mostly focusing on a single classifier, evaluating transparency effects on a static level of classification accuracy, e.g., [8, 31, 39]. It remains therefore unclear to what extent the observed effects generalise beyond the evaluated accuracy level. Moreover, a variety of metrics for user trust are currently used, ranging from behavioural measures [11, 53, 54] to questionnaires [12, 38, 54], reducing the comparability of results. In consequence, there are more open questions than answers.

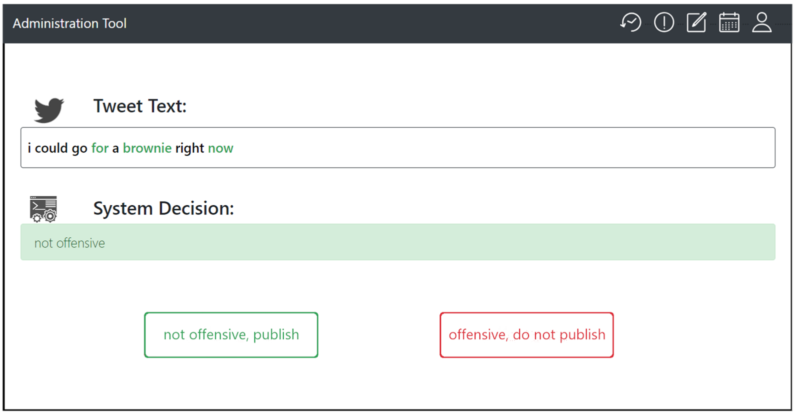
To holistically assess how explanation fidelity impacts the level of user trust, we take a broader view of different classifiers at varying levels of accuracy. We set up an online user study with 959 participants and evaluate user trust with a 3 (high accuracy, medium accuracy, and antagonistic accuracy) x 3 (faithful explanation, random explanation, and no explanation) experimental design. Inspired by the experiment by Langer, Blank, and Chanowitz [27], we measure the practical implications of implementing explanations, i.e., how user behaviour and the user's attitude towards the system change with the presence and informational content of an explanation. We, therefore, implement an automated support system that helps users in a classification task. We choose the use case of offensive language detection in Tweets (see Figure 1). Furthermore, we set out to compare behavioural measures of trust to self-reported trust.

Our work demonstrates the complexity of the interplay between trust, explanation fidelity, and model accuracy: In our experiment, the impact of adding explanations to a decision-support system varied for different classifier accuracy levels. Our findings further show that self-reported measures of trust cannot be used interchangeably with behavioural measures of trust. With the insights from our large-scale quantitative evaluation, we challenge the common claim of explanations improving user trust. In particular, our contributions are (1) empirical insights into the impact of explanation faithfulness on the user-system relationship at varying levels of accuracy, and (2) a comparison of two common metrics (self-reported and behavioural measures) that have been previously used to assess trust.

In this article, we discuss related concepts and research work that form the basis of our study design. We further describe the technical implementation of algorithms and explanations used for the experiment and report results and discuss our findings and their impact on machine learning practitioners.

1 RELATED WORK

Previous work provides insights into the generation of explanations, as well as how user trust is influenced by explanations and how trust can be evaluated in the context of **artificial intelligence** (**AI**) systems.



Tweet Text:

i could go for a brownie right now

System Decision:

not offensive

Fig. 1. Screenshot of the decision-support system used in our user study: We evaluated explanations for automated decision-making with regard to three explanation types. The user study is based on a scenario in which participants judge Tweets (top) with the help of a support system (middle) and indicate the offensiveness of the Tweet (buttons, bottom).

1.1 Trust in AI

In human-human relationships, trust is understood as the willingness to put oneself at risk while believing that a second party will be benevolent [42]. It is part of a relationship and represents an individual, subjective experience rather than a universal, objective characteristic of a user or a system [35]. Although a general definition for trust in human-machine relationships does not exist, most definitions agree that trust means that the system behaves as the user expects [35]. It is often connected with a sense of risk connected to a decision [23, 28], although the consequences of faulty behaviour or wrong decisions do not have to be severe [23]. In case of great risk, users of a system might even choose to ignore the system and rely on themselves instead [23]. For the remainder of the article, we adopt the definition of trust given by [35]:

Definition 1. In this article, we understand **user trust** as the extent to which the trustee believes that an automated system will behave as expected.

Moreover, trust builds up over time [42] and shows phases of development and refinement [54]. Yu et al. [54] report that the trust-building process between a user and a decision support system develops in three phases: a learning phase in the beginning with strong adaption of trust, an adjustment phase with minor changes to the trust level, and a more critical fine-tuning phase, showing that trust-building is not a linear process. During repeated interaction, trust develops according to the experienced accuracy of a system. For a system with varying accuracies (100%, 90%, 80%, and 70% accuracy), Yu et al. [54] showed that the self-reported trust levels strongly correlate with system accuracy during the last interactions. Yin, Wortman Vaughan, and Wallach [53] investigated whether this dynamic can be disrupted by displaying a different accuracy level. In their experiment, users adapted their trust level to an explicitly stated accuracy level but changed the trust level to the experienced accuracy during repeated interaction [53]. Misclassifications (i.e., system's predictions not corresponding to user's predictions) play a special role for user trust in automated decision making. If the system's behaviour does not correspond to how the user expects the system to behave, the "expectation mismatch" leads to a decrease in user trust [16]. But not only the user's perception of a system determines the level of trust. The user's ability to trust and prior experiences with comparable systems influence trust as well [24].

A variety of prior research has examined how model accuracy influences user trust. Data from several studies on model accuracy and user trust show that users adapt their trust level and perception of reliability to the experienced accuracy of a system [11, 38, 54]. The experienced accuracy is, therefore, an important factor for user trust in AI.

Another factor for user trust is the transparency of an AI system. Transparency answers the question "How does the system work?" [32], irrespective of whether the information used to answer the question make particular sense to a human. The level at which transparency makes sense to a human is the level of interpretability [3]. We agree with [32] and [3] and adopt the following definitions of algorithmic transparency and interpretability for this article:

Definition 2. In this article, we understand **transparency** as the extent to which information is exposed to a system's inner workings.

Definition 3. In this article, we understand **interpretability** as the extent to which transparency is meaningful to a human.

In order to provide interpretability to the user, the systems need to be equipped with mechanisms that communicate information about the system's workings to the users [3]. Rader, Cotter, and Cho [41] further add that transparency concerns "non-obvious" information about the system that the user would not be able to acquire without additional explanatory mechanisms. Following this definition, we use the term "explanation" in this article as follows:

Definition 4. In this article, we understand **explanations** as the mechanisms by which a system communicates information about its inner workings (transparency) to the user.

When looking at the interplay between transparency and user trust, studies provide conflicting results. Cramer et al. [9] tested the effects of transparency on users by comparing an art recommender without explanation to a system with an informative explanation (answering "why" some art was recommended) and a system with random information (showing how confident the classifier was about the recommendation). The results of their user study (N = 82) showed a correlation between self-reported 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. Furthermore, they found that showing the classifier's confidence level for each decision is not perceived as information that improves transparency [9]. Cheng et al. [8] use two levels of transparency (no transparency vs. minimum explanation) and two levels of interactivity (no interactivity vs. user interactivity) to investigate explanations in a user study with 199 participants. While their systems have a positive influence on understanding, they do not affect user trust. Schmidt, Biessmann, and Teubner [49] even find a negative effect of transparency on user trust when using the individual decision's confidence levels as explanations. This indicates that transparency could be harmful in some situations, while probably being helpful in others. Investigating in more detail in what cases an explanation can be helpful for users, Dzindolet et al. [11] find that explaining the reasons for misclassification improves user trust, while not providing an explanation for misclassifications decreases trust.

They also find a correlation between trust and reliance on a system (acceptance of its prediction), implying that behavioural measures can be used to approximate trust.

1.2 Evaluation of Trust

User trust in computer systems can be measured with subjective measures (e.g., questionnaires or interviews) or with behavioural measures (e.g., observation of behaviour) [36]. Different questionnaires have been developed to measure user trust in computer systems via self-reporting [22, 24, 33]. For example, based on models of interpersonal trust, Körber [24] develops a model of trust in automated systems describing trust along six dimensions: reliability, predictability, the user's propensity to trust, as well as the attitude towards the developers, the user's familiarity with automated systems and general trust in automation. Other works in the area of user trust in AI have used single questions that asked participants to rate their trust on a scale, e.g., on a 5-point scale in [5], on a 7-point scale in [54], on a 9-point scale in [55].

> *Definition 5.* In this article, we understand **self-reported user trust scores** as metrics that are based on participant's self-reflective answers to questionnaires.

However, using a questionnaire requires the ability to reflect on the relationship with the system, which can be difficult for some respondents. An alternative for measuring trust is the observation of behaviour. Inspired by [19], Poursabzi-Sangdeh et al. [39] suggest the "weight of advice" as a behavioural measure for trust in regression tasks. They determine the influence of a system's recommendation (advice) on the user's decision by asking participants to make a prediction about an apartment price before and after seeing the estimation given by a machine learning system. They calculate the difference between both prices given by the user (before and after) and divide it by the difference between the system's prediction and the user's initial prediction. The stronger the user adapts the price to the system's prediction, the higher the calculated "weight of advice". Vorm [51] reports "willingness to accept a computer-generated recommendation" as an observable sign of trust. Similarly, Yin, Wortman Vaughan, and Wallach [53] use the "switching rate" of participants as a measure of trust in binary classification tasks. They count how often users adapt their prediction to match the system's prediction and normalise over the amount of opportunities to see such behaviour (i.e., count how often the participant initially disagreed with the system). All three measures ("weight of advice", "willingness to accept a computer-generated recommendation", and the "switching rate") assume that users are affected by the system's output and change their behaviour. They rely on observing the user's behaviour in absence of the system in question, compared to the behaviour with the system present. With a study design that did not allow for before-after comparisons, Dzindolet et al. [11] and Schmidt, Biessmann, and Teubner [49] only use count of how often participants "relied" on the system, i.e., how often they agreed with the classifier, as a measure of trust. Although not a direct measure of trust, reliance is often used in this field to evaluate user's attitude towards the usage of a system [36]. The aforementioned measures focus solely on the user's behaviour with respect to the system, not on the ground truth. Whether the users achieve a better result when being supported by a system needs to be evaluated independently.

> *Definition 6.* In this article, we understand **behavioural user trust measures** as metrics based on observable behaviour of participants that functions as a practical indicator or proxy for user trust.

In recent years, research has started to compare and evaluate different types of trust measures. Buçinca et al. [5] found discrepancies between participants self-evaluation and their behaviour when evaluating inductive (similar data points from the training set) and deductive (individual features) explanation types: Although participants self-reported more trust in a system with deductive explanations, they showed a higher classification performance (participants' classification with respect to the ground truth) with inductive explanations. Yu et al. [54] compare behavioural and self-reported trust measures by calculating correlation. They measure self-reported trust by asking participants to indicate their level of trust on a 7-point Likert scale (from 1 = "distrust" to 7 = "trust") and behavioural trust by counting how often participants change their own prediction to match the classifier's prediction. In a small user study (N = 21), they found strong correlation between both the switching rate and the trust users report. Schaffer et al. [48] likewise set out to compare a behavioural measure with self-reports of trust, yet arrive at the opposite conclusion. They compare the "adherence" to a food recommender's advice to self-reported trust and find only a weak relationship between both measures. In their setup, however, the real adherence to the recommendation is difficult to assess, as it is not known how participants would have reacted without the recommender. To capture both aspects (behavioural and subjective), Williams et al. [52] suggest combining self-reports and behavioural measures in HCI research.

1.3 Explanations for AI

To investigate which influence the informational richness of an explanation plays in human-human interaction, Langer, Blank, and Chanowitz [27] conducted a study for the field of psychology. Participants were confronted with a request (letting someone jump the queue at a copy machine) in combination with either some type of explanation for the request (meaningful or nonsense), or without an explanation. For settings without severe consequences, compliance rates were similar for both explanation types, but lower for cases without explanation. However, when consequences arose from compliance (the person asking to jump the queue had a big pile of paper to copy), compliance was equally low for all three conditions. Eiband et al. [12] found a first indication that the results from [27] also hold for human-machine interaction. In a small user study (N = 30 over 3 conditions), they evaluated informative sentences and "placebic" sentences (i.e., sentences without relevant information) to explain parts of a nutritional adviser system and compared both conditions to a condition without explanations. In their experiment, both explanation types led to equal trust scores, while the absence of explanations resulted in lower user trust.

In recent years, machine learning models show a trend towards increasing accuracy, but also increasing complexity. The higher the complexity and accuracy of a system, the lower its transparency [7, 46]. Especially for systems that are too complex to understand fully, trust decides whether a user relies on the system or not [28]. **Explainable AI** (**XAI**) aims at helping users understand the workings of an AI system while maintaining high performance [3]. It is often mentioned in the literature that explanations have a positive effect on user trust [4, 16, 40, 51]. However, the more complex a model is, the more it needs to be simplified to match the human attention span and cognitive abilities [26]. Kulesza et al. [26] define a good explanation as truthful and complete, but not overwhelming. While completeness describes how much information is given about the system, truthfulness describes how well the information explains the underlying mechanism. Explanations can be truthful, i.e., faithful, to the underlying machine learning model, yet are not necessarily meaningful to a human observer [14, 32]. Conversely, explanations that are meaningful and persuasive for humans do not necessarily reflect the underlying model [32]. The latter could lead to a false feeling of understanding the workings of complex systems [47]. In fact, Bussone, Stumpf, and O'Sullivan [6] found an indication that extensive explanations can lead to over-reliance and inappropriate trust in a system.

Definition 7. In this article, we understand **explanation faithfulness** as the level to which an explanation is faithful to the underlying machine learning model, irrespective of its level of completeness.

While some machine learning models are transparent and interpretable on low complexity levels (e.g., decision trees, linear models [4]), deep learning models are inherently intransparent "black box" models that can retrospectively be equipped with an explanation mechanism [32]. An example for such model-agnostic add-on systems are LIME and the Anchors, both developed by Ribeiro, Singh, and Guestrin [43, 44]. Likewise aiming for add-on explanations for text input, Arras et al. [2] use heat map masks to point the attention towards features that are decisive in a sample, e.g., single words in texts. Feng et al. [14] used an image classification system and added post-hoc question-answering. They eliminated words from the post-hoc questions that are irrelevant for the answer result, reducing the texts to a bare minimum that shows the most important words. Although the reduced texts still lead to high accuracy answers, they find in a user study that those texts, although highly faithful to the underlying system, are nonsensical for humans: A picture showing a sunflower with the question "What color is the flower?" is answered correctly with "yellow", even when reducing the question to its – for the classifier – most decisive feature "flower?". Their participants gave significantly more incorrect answers on the reduced questions than on the non-reduced ones, while the accuracy of the machine learning system even increased its accuracy on the reduced texts.

Besides a potential benefit for user trust, explanations can help to foster the understanding of a system. Kulesza et al. [25] compared users' mental models of a decision-support system they interacted with, with and without explanations. In their experiment, explanations helped the user to form a better mental model of the system. Additionally, they evaluated whether the user's classification after each decision would increase a classifier's performance. Although users with explanations were slower in the classification task (which is, considering they have to read the explanations, not surprising), their classifiers performed better than those of users without explanations. These two analyses show that the mental model improves with explanations. Similarly, Oduor and Wiebe [38] found that the modality of an explanation influences the self-reported understanding of a system. They used the scenario of a city manager (N = 90) overseeing and taking decisions for a whole city using a support system. The system had a decision tree that generated suggestions for management opportunities. Self-reported understanding of the system was highest with textual explanations for each decision, while giving a graphical representation of the model or giving no explanation yielded equal understanding.

Other research works have explored different explanation types. For a knowledgeable audience (e.g., developers), Liao, Gruen, and Miller [29] collected information needed for questioning an AI system. They provide a comprehensible list of question types that users might have about an AI system, such as knowing the system's performance, why (or why not) a certain prediction was made, and what data the system was trained on. Other research focuses on evaluating the effect of specific explanations. Rader, Cotter, and Cho [41], for example, conducted a user study (N = 68) with four explanation types ("what", "why", "how", and "objective" explanations). All four types surprised the users, which suggests that they did not match what the users expected to see. Lim, Dey, and Avrahami [31] likewise compare explanation contents ("why", "why not", "what if", and "how to"). In their setup, "why" and "why not" explanations are the only ones to improve user trust and understanding of the system. Ribera and Lapedriza [45] suggest to adapt the explanation style to the type of user (AI developers, domain experts, end-users), arguing that "why" and "why not" explanations are more suitable for end-users. Similarly, Nourani, King, and Ragan [37] see differences between experts and lay-users when detecting and reacting to classification errors of an AI support system. In their study (N = 116), experts adapt their trust in a system quickly when encountering an error, whereas a misclassification does not have an equally strong effect on the trust of lay-users. Poursabzi-Sangdeh et al. [39] examined the impact of "explanatory depth" by comparing a short explanation (explaining the impact of two input features) to a more complete explanation (eight features). While they did not find any impact on user trust between conditions, they observed that users were better at predicting the system's behaviour with shorter explanations. Lim and Dey [30] similarly focused on how much information they revealed about a system, but for varying levels of accuracy. They use two types of faithful explanations, but with varying amounts of completeness (extensive explanation, short explanation, no explanation). Spanning the experimental conditions along the dimension of explanations and accuracy (50%, 60%, 70%, 80%, 90%, 100% accuracy), they find that extensive explanations increase perceived accuracy for high-accuracy systems, but decreased perceived accuracy for low-accuracy systems. Their results show that explanations might have a different influence on the user's perception, depending on how accurate the system performs. Findings of the impact of explanations can therefore not be generalised for systems with a different accuracy level.

Explanations in machine learning can either aim at explaining a single decision of a classifier (local explanation, e.g., the words that have the strongest influence on the classification of a text) or aim at explaining the model (global explanation, e.g., by visualising a decision tree) [18]. The graphical representation of the decision trees in Odour and Wiebe's experiment [38] (explained above) is an example of a global explanation. However, global understanding can also result from repeated interaction with local explanations [43]. Goodman and Flaxman [17] argue that a local explanation for machine learning systems must at least show how input features impact the classifier's output. An example of such a "minimum explanation" is the "**Learning to Explain**" (**L2X**) algorithm by Chen et al. [7] that selects the *k* words of an input text that have been most decisive for the classifier's decision. They evaluate the faithfulness of their explanations by reducing the text input to the selected words, re-classifying the reduced texts, and subsequently counting how often the classifier gives the same output. If the selected words are not a good representation of the underlying model, the classifier would not be able to produce the same labels for the reduced text. Another approach to explain parts of the machine learning system was suggested by Zhou et al. [55], who point out that connecting a data point with similar training data points could increase transparency. Their initial user study (N=22) indicates that such explanations might increase user trust for high-accuracy systems.

1.4 Summary

From the literature research, we see that user trust relates strongly to the system's accuracy [53, 54]. Furthermore, multiple aspects of explanations and their effect on user trust have been researched. Eiband et al. [12], for example, have made a first step towards examining the effect of explanation faithfulness on user behaviour, yet with a single classifier, resulting in a constant accuracy across conditions. Likewise, prior research has focused on a single independent variable (e.g., explanation type with a fixed accuracy [39], or accuracy with a fixed explanation type [53]). It remains unclear whether the insights into user trust with respect to explanation types generalise over classifiers with different accuracy levels. Lim and Dey [30] provided an indication that explanations have a different effect on user perception for different classifier accuracies. However, their work focuses on the completeness of an explanation rather than on explanation faithfulness. We are therefore interested in how the effect of model accuracy on trust and the effect of explanation faithfulness on trust work together. In this article, we investigate the interplay between model accuracy and explanation faithfulness with respect to user trust in automated decision-making. As global understanding can arise from a series of local explanations [43], we focus on local explanations in this research work. Some information that can be given in an explanation have been shown not to have an effect on users: Showing the classifier's confidence for a classification result was not perceived as containing information about the system [9], and stated accuracy was overwritten over repeated interaction [53]. We, therefore, choose to apply minimum explanations for text input [17] by highlighting words that are decisive for the classification result [7].

To date, user trust in automated decision-making systems has been measured in multiple ways (c.f. Mohseni, Zarei, and Ragan [36] for a comprehensive overview). So far, only a few works have researched the comparability of subjective and behavioural measures [48, 54]. None of them uses a validated trust questionnaire as a subjective measure. It, therefore, remains an open question to what extent the findings of studies with behavioural measures can be compared with insights from self-reports. We, therefore, use a validated questionnaire and a behavioural measure in our experiment and assess whether one can serve as a proxy for the other and whether prior studies using different measures can be compared to each other.

2 STUDY DESIGN

We set up a user study to measure user trust in a decision-support system. We first evaluate the participants' classification performance to find out whether a decision-support system has practical consequences for how users accomplish a decision task. Subsequently, we evaluate the users' trust in the decision-support system. On the one hand, using questionnaires to evaluate trust requires the participants to actively reflect on their relationship with the system, which can be time-consuming and difficult for participants unfamiliar with this activity. On the other hand, there is no standard method that is widely adopted for measuring trust by observing behaviour during a user study. We, therefore, chose to employ both a behavioural measure (switching rate towards the classifier in binary classification [53]) as well as a validated questionnaire (19-items questionnaire to evaluate user trust in automation [24]) to measure user trust. More specifically, we address the following research questions:

RQ1 To what extent is *behavioural trust* influenced by the accuracy and explanation presence and the faithfulness of a decision-support system?

H1.1 Following the observations reported in [12] and [27], we expect users to put equal trust in systems showing an explanation (faithful explanation and random explanation),

H1.2 but lower trust for systems without explanation.

RQ2 How do system accuracy and explanation presence and faithfulness impact the user's perception of trust, i.e., the *self-reported trust*?

H2.1 Following the observations reported in [12] and [27], we expect users to put equal trust in systems showing an explanation (faithful explanation and random explanation),

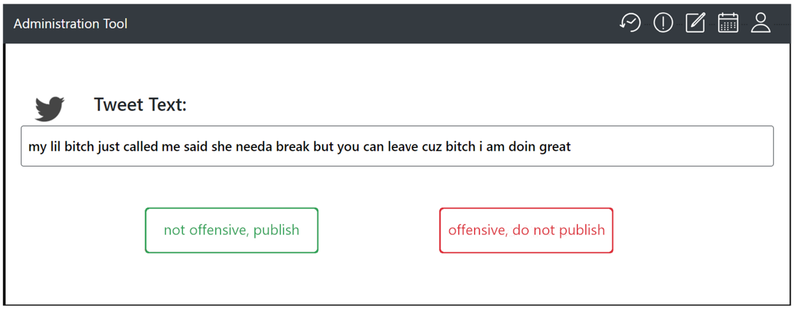
H2.2 but lower trust for systems without explanation.

RQ3 Do the measures of behavioural trust and self-reported trust *correlate*?

H3 We expect to see comparable results for both behavioural and self-reported trust measures, as they should measure the same variable, only in different ways.

2.1 Scenario

We created a web interface that should help social media administrators in detecting offensive text or hate speech with the support of a machine learning system. We chose "automatic detection of offensive text in Tweets for a youth platform" as our use case, because it showcases an issue with understandable consequences that a broad audience can relate to. As the youth is a user group that



Tweet Text:

my lil bitch just called me said she needa break but you can leave cuz bitch i am doin great

Fig. 2. Screenshot of the graphical user interface for classification without the support system.

needs protection, the consequences of faulty classifications are easy to understand for participants. To ensure that all participants have the same understanding of offensiveness, we established the following rules for offensive Tweets (based on [21]):

- Containing hateful language: any comment that disparages a person or a group on the basis of some characteristic such as race, colour, ethnicity, gender, sexual orientation, nationality, and religion
- Containing pornographic language: explicit sexual subject matter for the purposes of sexual arousal and erotic satisfaction
- Containing vulgar language: coarse and rude expressions, which include explicit and offensive reference to sex or bodily functions
- Not only single words can be offensive, but also the meaning of a text. A text can be offensive without explicitly mentioning offensive words.

Figure 1 shows a screenshot of the graphical user interface. The “social media administrators”, i.e., our participants, see the text of a Tweet, followed by the system’s prediction if the text shown is detected as offensive or not. With one of the two buttons “not offensive, publish” or “offensive, do not publish”, the social media administrator submits her own decision. To provide the user an explanation of the system’s decision, the most decisive words in the text are highlighted with colour, as suggested in the literature [2, 7]: red if the system predicts offensiveness, and green if it predicts a non-offensive Tweet. Figure 2 shows the web interface without system support, while the interface with system support is shown in Figure 1.

2.2 Dataset

We use a dataset curated by Davidson et al. [10] that consists of Tweets with and without offensive language and hate speech. However, the dataset contains inconclusive annotations with an inter-annotator agreement of less than 100% (at least three annotators per Tweet) and shows class imbalance towards the non-offensive class. Although ambiguous cases with low inter-annotator agreement are cases where a support system would be most helpful, it is impossible to draw conclusions based on an unclear ground truth. We, therefore, included only data points with a clear consensus among the annotators (inter-annotator agreement of 100%). To assess the impact of this selection on the classification performance of participants and the resulting need for support, we provide the participants’ classification accuracies in Section 4.2. We sample an equal amount of data points from each class (offensive/not offensive) to make the classifier evaluation fair and provide enough possibilities to observe switching between classes. The final dataset contains 2,162 Tweets labelled as containing offensive language and 2,162 Tweets labelled as not containing offensive language.

Table 1. Classifier-explanation Conditions

		Classifier Accuracy		
		high	medium	antagonistic
Expl.	faithful	C_HF	C_MF	C_AF
	random	C_HR	C_MR	C_AR
	no	C_HN	C_MN	C_AN

2.3 Experimental Conditions

So far, investigating explanation types on user trust is usually done at a fixed accuracy level. Explanation types (e.g., explanatory depth in [39], explanation type in [31], explanation content in [41]) serve as the independent variable, while classification accuracy is a constant. The accuracy level, however, affects user trust as well [11, 53]. We, therefore, examine the effect of explanations at varying levels of model accuracy. To derive insights for realistic settings, we choose two realistic accuracy levels

(1) high accuracy between 90% and 100% accuracy;
(2) medium accuracy around 75%.

We introduce a third accuracy level with a classifier being obviously wrong

(3) "antagonistic" accuracy between 0% and 10%.

With the third accuracy, we span the whole dimension of algorithmic support from high support to almost no helpful support. We refrained from using a random classifier at 50% accuracy, as such a system would have still provided support in 50% of the cases and hence might result in unclear trust levels. With the antagonistic accuracy, however, we include a system that does not provide helpful support, as it is wrong in most cases. We do not expect participants to perceive the antagonistic classifier as "deceptive", i.e., intentionally misleading, since we neither introduce techniques of persuasion nor deliver evidence for why the obviously incorrect predictions could be correct after all. For the explanation types, we follow the experiments in [12, 27] and set up

(1) faithful explanations (i.e., explanations that convey information about the inner workings of the system);
(2) random explanations (i.e., explanations that do not convey information about the inner workings on the system);
(3) no explanations.

The experiment design is a 3 (high accuracy, medium accuracy, and antagonistic) x 3 (true explanation, random explanation, and no explanation) design, resulting in nine conditions in total. The notation used in this article is shown in Table 1. In addition to the nine conditions, a baseline is conducted. Participants in the baseline group are not confronted with the automated decision-making system, but instead only see the Tweets (see Figure 2).

2.4 Method

We use a between-subjects design. At the beginning of the experiment, each participant is randomly assigned to exactly one condition, with an equal distribution of participants over the conditions. As the participants need to have enough interaction with the system to build an intuition for the system's accuracy, they are presented with a subset of 15 Tweets. We chose this subset size as a balance between having sufficient experience with the system and not being overwhelmed by lengthy interactions. Selecting 15 Tweets out of the dataset bears the risk of not representing the dataset sufficiently. Side effects of specific wording used in the subset or topics treated in the Tweets could influence the perception and opinion of participants. We, therefore, construct 10 non-overlapping subsets of 15 Tweets (see Section 3), to which participants are assigned uniformly at random at the start of the experiment. The order of the Tweets is randomised for all participants. The experiment has received clearance from the institution's ethics committee.

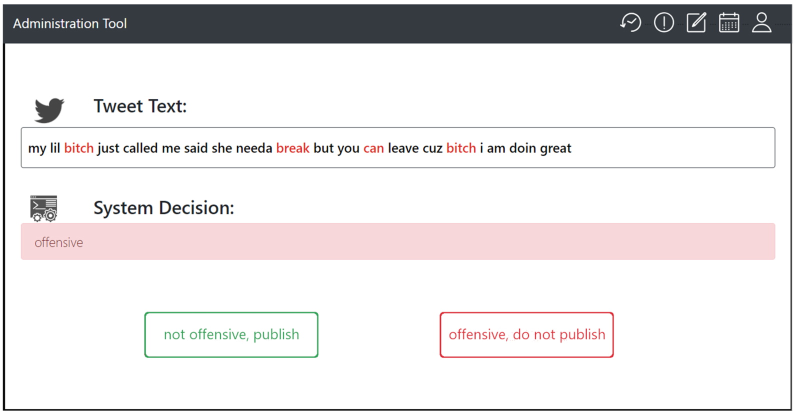
2.5 Apparatus & Procedure

The study is set up as an online study for laptops and tablets on the SoSci platform.[1] Overall, we structured the experiment task in two blocks. In the first block, the scenario and web interface are introduced, without an example Tweet to prevent priming effects. Participants are then confronted with the 15 Tweets of their subset. The Tweets appear in random order for each participant. The Tweets are presented within screenshots of the web interface at a ratio of 900 px (width) to 253 px (height), which is the reason why devices with small displays such as smartphones were excluded from the study. Participants are asked to classify the Tweets into "offensive" and "not offensive" without the support system (see Figure 2). Subsequently, in the second block, the automated decision-making system is introduced in a short paragraph without an explanation of how predictions and explanations are generated. In the baseline condition, participants are not confronted with the automated decision-making system and instead see a reminder of the definition of offensiveness. Participants then classify the 15 Tweets from block one again, again in random order. This setup of using the same Tweets in blocks 1 and 2 was needed to measure the "switching rate" of participants, i.e., the influence of a support system on the participants' behaviour. When re-classifying the same Tweets, participants could feel pressured to repeat their initial classifications in block 1. They might force themselves to remember their classifications and stick to their initial classification, even if they changed their opinion. We wanted to liberate the participants from this pressure and allow for changes in opinion after seeing the system's prediction in block 2. We, therefore, told participants at the start of block 2 that they will now judge 15 "very similar" Tweets. Except in the baseline, the Tweets are pre-classified using the algorithm described in Section 3 and displayed according to one of the nine conditions (Figure 3 depicts the interface with C_HF). Participants were neither told their own classification performance nor the one of the system they interacted with, as [53] found that displayed accuracy information will be overwritten by experienced accuracy in the long run. After the task, participants filled in Körber's validated questionnaire [24] about trust in automated systems with an added attention check question. Additionally, participants answered questions about their demographic background and had the opportunity to leave a comment about the system and the survey at the end. The survey was successfully tested in a pilot with 11 participants. Figure 4 shows an overview of the experiment structure.

2.6 Participants

In total, 1,092 participants were recruited via the crowdsourcing platform "Prolific",[2] of which 133 responses were excluded due to a failed attention check question, resulting in 959 valid responses. 74 participants (30 m, 44 f, 0 d) took part in the baseline, while 885 participants (378 m, 500 f, 7 d)

[1] https://www.soscisurvey.de/en/index, accessed on 04.06.2021.
[2] https://prolific.ac, accessed on 04.06.2021.


Administration Tool

Tweet Text:

my lil bitch just called me said she needa break but you can leave cuz bitch i am doin great

System Decision:

offensive

Fig. 3. Screenshot of the graphical user interface with the support system in condition *C_HF*.

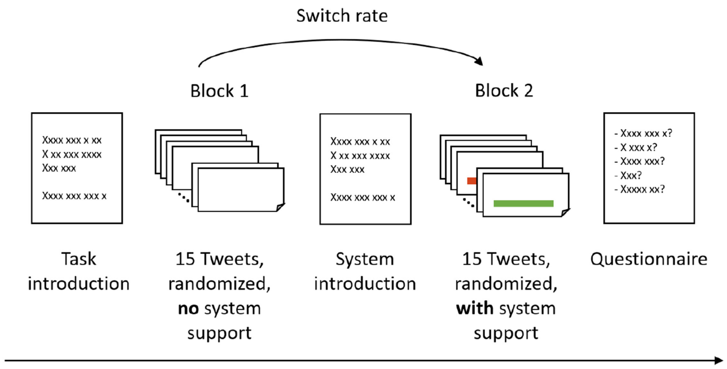

Switch rate
Block 1
Block 2
Task introduction
15 Tweets, randomized, **no** system support
System introduction
15 Tweets, randomized, **with** system support
Questionnaire

Fig. 4. Overview of the experiment design.

were confronted with the AI support system. Participants were distributed uniformly at random over the nine conditions named in Table 1: 91 in *C_HF*, 102 in *C_HR*, 92 in *C_HN*, 96 in *C_MF*, 103 in *C_MR*, 100 in *C_MN*, 100 in *C_AF*, 98 in *C_AR*, and 103 in *C_AN*. The sample size per condition was chosen to match the required sample size according to Slovin's formula [1] with a population of all internet users (≈4.5 billions) and an error tolerance of 0.10. All 1,092 participants received a financial allowance of 1.10 EUR for an average completion time of 12 minutes. Most participants, 48%, were between 18 and 30 years old. 16% reported an age between 31 and 40, while 36% were older than 40. All participants described themselves as being "fluent" in English, with 64% classifying their English as a native language or equal to native speakers.

2.7 Measures & Analysis

We measure trust in two different ways. First, we use a behavioural measure and observe how the system influences the participant's opinion by comparing their classifications of the first block (without system) and second block (with system). We adopt the definition of behavioural trust through observation used by Yin, Wortman Vaughan, and Wallach [53], measuring how often participants alter their prediction to match the system's prediction. We normalise over the possibilities to see such behaviour. We measure behavioral trust t_b as

$$t_b = \frac{\text{number of changes towards the system's prediction between blocks 1 and 2}}{\text{number of disagreements with the system's prediction in block 1}}$$

$$= \frac{\sum_{n=1}^{15} |P1_n - S_n| \cdot \begin{cases} 1 & \text{if } P2_n = S_n, \\ 0 & \text{otherwise.} \end{cases}}{\sum_{n=1}^{15} |P1_n - S_n|} \in [0, 1],$$

with $P1_n$ being the participant's classification in block 1 for Tweet n, respectively, $P2_n$ in block 2, and S_n being the system's prediction shown in block 2. Note that participants do not see the system's prediction in block 1 and therefore do not consciously "disagree" with the system in block 1.

Secondly, we measure subjective, self-reported trust using the questionnaire of Körber [24] that contains 19 items drawing on the six aspects of trust: reliability/competence, predictability/understanding, familiarity, the intention of developers, propensity to trust, trust in automation [24]. The complete questionnaire is publicly available in [24]. We take the mean score over all 19 items to compute a single trust score per participant, accounting for negatively formulated items (reverse items) by calculating the inverted score (i.e., 5 - rating). We further zoom into the six sub-concepts of trust that are present in Körber's questionnaire by calculating the average scores for items related to the sub-concepts.

For significance tests concerning the trust measures, we use a two-way ANOVA as it is robust against non-normally distributed data. The independent variables are the explanation type (faithful, random, and none) and the classifier accuracy (high, medium, and antagonistic), while the dependent variable is either the self-reported trust score or the behavioural trust score. To compare differences between conditions with respect to a single variable, we use a one-way ANOVA. For all post-hoc analyses, we use the two-sided Mann–Whitney U-test, applying Bonferroni correction to the p-values to counteract the repeated testing problem. For testing correlations, we use Pearson's correlation coefficient and additionally report the r-squared measure.

3 MACHINE LEARNING MODELS

The following section describes the implementation and evaluation of three machine learning classifiers and post-hoc explanations (faithful and random). Our focus lies on achieving three systems with sufficiently different classification behaviour.

We split the dataset described in the study design section into 80% training data and 20% testing data. All Tweets are preprocessed by resolving contractions (e.g., "we're" to "we are"), lowercasing, deletion of special characters, replacing URLs and user names by placeholder versions (e.g., "http://website.com/website" and "@username"), and tokenizing on white spaces. For instance the following Tweet:

```
@WBUR: A smuggler explains how he helped fighters along the "Ji-
hadi Highway": http://t.co/UX4anxeAwd
```

is processed into:

```
@username a smuggler explains how he helped fighters along the
jihadi highway http://website.com/website
```

We implement three machine learning classifiers to predict the label of a Tweet ("offensive", "not offensive") at three different accuracy levels. As an explanatory mechanism, we generate faithful and random explanations by highlighting words in the Tweets. On average, the Tweets contain

Table 2. Confusion Matrix for the High-accuracy Classifier

		True Class	
		offensive	not offensive
Pred. Class	offensive	813	32
	not offensive	35	785

Table 3. Confusion Matrix for the Medium-accuracy Classifier

		True Class	
		offensive	not offensive
Pred. Class	offensive	550	267
	not offensive	128	720

between 14 and 15 words. The system gives local explanations that show the most important words for the decisions, highlighting not more than ⅓ (i.e., 4) words of the Tweets.

3.1 High-Accuracy Classifier

As a high-accuracy classifier, we implement the **convolutional neural network** (CNN)[3] proposed by Chen et al. [7], who identified sentiment in short movie reviews. We achieve an accuracy of 0.96 on our test set (see Table 2 for details).

Besides the classification system, Chen et al. present an add-on explanatory system "L2X" that selects the k most decisive features of an input vector. The selected words are based on statistical evidence (mutual information) from the training set that lead to the classifier's decision. This explanatory mechanism satisfies our definition of faithfulness, as it is faithful to the underlying model [7]. We, therefore, use Chen et al.'s L2X algorithm to select the $k = 4$ words in the Tweets that have the most impact on the decision. For example, the high-accuracy classifier labels the following Tweet as "offensive" and highlights four tokens for the faithful explanation (marked in bold):

> `my lil` **`bitch`** `just called me said she needa` **`break`** `but you` **`can`**
> `leave cuz` **`bitch`** `i am doin great`

3.2 Medium-Accuracy Classifier

For a medium-accuracy classifier, we decided against random classification (yielding an accuracy of 0.50 on our class-balanced dataset), because we aimed at providing high-fidelity explanations for each classifier. A random classifier does not have an underlying decision structure, which makes it impossible to provide faithful explanations. Davidson et al. [10], who curated the dataset we use, have implemented a logistic regression classifier and achieved an F1-score (the weighted average of precision and recall) of 0.90. We use a logistic regression model[4] that yields an accuracy of 0.95 on our test set. To reduce the accuracy level, we round the model coefficients to −1 for all negative coefficients and 1 for all positive coefficients. The final accuracy on our test set is 0.76 (see Table 3 for details).

For faithful explanations of this classifier, we use the model coefficients of the words in the Tweet that lean towards the prediction (i.e., positive coefficients if prediction is "offensive", negative if

[3] implemented using https://keras.io, accessed on 04.06.2021.

[4] implemented with https://scikit-learn.org, accessed on 04.06.2021.

Table 4. Confusion Matrix for Antagonistic Classifier

		True Class	
		offensive	not offensive
Pred. Class	offensive	43	786
	not offensive	805	31

"not offensive"). The coefficients satisfy our definition of being faithful to the model as they are part of the model itself. If more than $k = 4$ words are eligible, we randomly draw four words from the set of eligible words. The model coefficients are available per unique word. If a coefficient is selected as an explanation, we highlight all occurrences of that word in the Tweet. All occurrences of a word share the same coefficient and therefore contribute equally to the decision. The medium-accuracy classifier processes the above Tweet as follows for faithful explanations (marked in bold):

my lil **bitch** just called me said she needa break but you can
leave cuz **bitch** i **am** doin great

3.3 Antagonistic Classifier

To reach a low accuracy, we use the same implementation as for the high-accuracy classifier, but train the model on the training data with inversed labels. The accuracy of the classifier on the test set is 0.04 (see Table 4 for details). We again use the L2X algorithm to highlight the most decisive words for faithful explanations.

3.4 Random Explanations

For all three systems, we use the same mechanism to generate random explanations. "Random" here means that the explanations do not communicate information about the classifier's working. Therefore, we draw $k = 4$ words uniformly at random from the Tweet to be highlighted. As for the faithful explanations of the medium-accuracy classifier, all occurrences of the selected words are highlighted. Selecting words at random is completely independent of the classification algorithm, and can therefore not disclose information about the classifier. To give an overview over all conditions, we display an example Tweet for all accuracy levels and explanation types in Table 5, where the background colour depicts the classifier's prediction and the font colour signifies the words selected as an explanation. To evaluate whether the random explanation mechanism produces explanations that are by chance faithful to the underlying decision mechanism, we conduct additional experiments presented in Section 3.6.

3.5 Subset Selection

As described in Section 2.4, we generate 10 subsets of Tweets to reduce effects of wording or topics in single Tweets and have an overall diverse representation of the test set. Each subset contains 15 Tweets selected randomly from the test set. Additionally, we require the subset to be non-overlapping with previously drawn subsets. The subset is furthermore only accepted if it has a class balance similar to the test set (1:1), and if the classifiers perform equally on the subset as on the whole test set (0.96, 0.76, and 0.04). We then calculate how representative the words in the selected Tweets are for the whole test set by comparing the feature distribution of the subset with that of the test set using **Kullback–Leibler Divergence** (**KLD**) with Laplace smoothing (k = 1).

Table 5. Example Tweet in All Conditions

		Classifier Accuracy		
		high	medium	antagonistic
Explanation	faithful	you **bitches** you aint **shit** i swear you think you **poppin** cause you instagram **famous**	**you** bitches **you** aint shit i swear **you** think **you** **poppin** **cause** **you** instagram famous	you **bitches** you aint **shit** i swear you think you **poppin** cause you instagram **famous**
	random	**you** bitches **you** aint **shit** i swear **you** think **you** poppin cause **you** **instagram** famous	you bitches you **aint** **shit** i swear you **think** you poppin **cause** cause you instagram famous	**you** bitches **you** aint **shit** i swear **you** think **you** poppin cause **you** **instagram** famous
	no	you bitches you aint shit i swear you think you poppin cause you instagram famous	you bitches you aint shit i swear you think you poppin cause you instagram famous	you bitches you aint shit i swear you think you poppin cause you instagram famous

Bold text represent the explanations. Background color indicates the classifier's decision: red for an offensive Tweet, green for not offensive.

We generate a quantity of 100 such subsets. The 10 subsets with the smallest score (and therefore highest similarity to the training dataset) are chosen as the final set of subsets.

3.6 Explanation Evaluation

As suggested by Chen et al. [7], we validate the fidelity of our explanations by reducing the Tweets to the highlighted words, classifying the reduced texts, and subsequently comparing the original prediction to the prediction on the reduced texts (label agreement). The validation procedure is necessary to avoid three problems:

(1) Since the L2X algorithm provides post-hoc explanations, there is no guarantee that it in fact represents the behaviour of the underlying CNN, and the selection of four words could be too small to represent the decision basis of the classifiers. Table 6 shows that the faithful explanations reach an accuracy close to 100% when selecting four words as an explanation for all three classifiers. We conclude that they faithfully represent the classifier's behaviour as only those four words are enough to lead to the same result.

(2) Since Tweets in the dataset have a maximum of 140 characters, the information density is presumably high. Selecting words at random for the random explanations yields the risk of selecting informative words by chance and hence delivering faithful explanations even in the random explanation conditions. Even with a random selection, we expect the selected words to contain some informative content, yet it should not be enough for the classifier to always come to the same prediction when classifying the reduced texts. Table 6 shows that using a random selection of words does not always lead to the same predictions. The classifiers only reach a label agreement between 0.72 and 0.77 when selecting four random words. A classifier that randomly guesses the label would achieve an agreement of 0.50, showing that in our case, about every second word is, by chance, a meaningful word—leading to a label agreement of 0.74 on average.

Table 6. Label Agreements Evaluating the Faithfulness of Explanations

	C_H		C_M		C_A	
	F	R	F	R	F	R
$a_{testset}$, $k = 4$	0.98	0.74	1.00	0.77	0.97	0.72
$a_{testset}$, $k = all$	1.00	1.00	1.00	1.00	1.00	1.00
$\bar{a}_{subsets}$, $k = 4$	0.97	0.74	1.00	0.64	0.97	0.74

Showing the label accuracy a on Tweets reduced to k words when the prediction on the non-reduced Tweets is set as ground-truth, for the test set and mean accuracy $\bar{a}$ across subsets.

(3) The explanation of faithfulness in the subsets could by chance deviate from the faithfulness in the whole test set. However, from Table 6, we see that the averaged agreement over all subsets is indeed similar to those on the whole test data.

4 RESULTS

The collected and anonymised data from the user study is publicly available online.[5] We first analyse whether demographic variables are *confounding factors* for user trust. To investigate the effect on user trust in detail, we present the results of the *behavioural trust* measure (**RQ1**) and the *self-reported trust* measure (**RQ2**). We use the Mann–Whitney U-test to test for significant differences between conditions and adjust the p-values with the Bonferroni correction to account for the repeated tests bias. We report the p-values after the Bonferroni correction. Finally, we analyse the correlation between the two measures of trust with Pearson's correlation coefficient rho (**RQ3**).

4.1 Are Demographic Characteristic Confounding Factors for User Trust?

We tested whether *demographic factors (age and gender)* impact the participant's perception concerning the AI support system to account for possible confounding variables. For each condition, we compared the self-reported trust scores and behavioural trust scores of participants younger than 30 years (N = 569) with those of 30 years and older (N = 316), and those of male participants (N = 378) with those of female participants (N = 500). The threshold age of 30 was chosen to distinguish between "digital natives" who grew up with technology as an inherent part of their daily lives, and "digital immigrants". We did not compare the third gender due to the insufficient sample size of N = 7. Overall, there was no significant impact of age or gender on either of the two measures in any condition. To furthermore evaluate whether the language level impacts the participants' understanding of the Tweet texts, we calculate the participants' classification performance (in terms of accuracy with respect to the ground truth) over the first block of Tweets (15 Tweets without the support system). Overall, there is a significant difference ($p < 0.001$) in classification performance between participants with a language level of native speakers or equal to native speakers (M = 80%, SD = 12%) and those who are not native speakers (M = 77%, SD = 12%) in block 1. However, concerning self-reported trust and behavioural trust, the effects of language are marginal: There is no significant effect in any condition of either measure. Moreover, the ratio of native speakers and non-native speakers is comparable in each condition. We, therefore, conclude that the level of language does not significantly impact the findings on trust. The average performance furthermore shows that constraining the dataset to datapoints with 100% inter-annotator agreement in the ground truth (see Section 2.2) did not entirely eliminate the need for a support system

[5] https://git.gesis.org/papenmaa/tochi_trustinai.

Table 7. Means and Standard Deviations of Participants' Classification Performance on Block 1 (No Support) and Block 2 (System Support)

	Performance block 1 M (SD), significance	**Performance block 2** M (SD), significance	**Difference between blocks** difference, significance
C_HF	0.78 (0.12)	0.83 (0.12)**	+0.5**
C_HR	0.80 (0.11)	0.82 (0.12)*	+0.2
C_HN	0.80 (0.12)	0.84 (0.11)**	+0.4*
C_MF	0.79 (0.12)	0.81 (0.11)	+0.2
C_MR	0.80 (0.12)	0.79 (0.14)	−0.1
C_MN	0.80 (0.12)	0.81 (0.13)	+0.1
C_AF	0.78 (0.12)	0.74 (0.18)	−0.4**
C_AR	0.79 (0.13)	0.77 (0.14)	−0.2**
C_AN	0.79 (0.11)	0.74 (0.14)*	−0.5**
baseline	0.78 (0.12)	0.79 (0.11)	+0.1

Significance tests with respect to baseline performance, where **: p < 0.01, *: p < 0.05. Pairwise comparison with Mann–Whitney U-Test with a two-sided alternative. Difference in performance between blocks is shown in right column, with tests for significance with respect to baseline difference.

by selecting only "easy-to-classify" cases. Misclassifications in block 1 leave room for reconsidering the classification in block 2, possibly with the help of the support system.

4.2 RQ1: To What Extent is Behavioural Trust Influenced by the Accuracy and Explanation Presence and Faithfulness of a Decision-support System?

The participants classified Tweets as "offensive" or "not offensive" during the experiment. First, they classified 15 Tweets without the support of the AI support system (block 1). After a short introduction to the support system, they classified those 15 Tweets again (block 2), this time (except in the baseline condition) with the help of the AI support system. We measured behavioural trust by observing how often the participants switch their decision towards the classifier's recommendation between block 1 (without system) and block 2 (with system). A prerequisite for analysing trust via switching behaviour is that switching occurs because of the system introduced in block 2, not merely because of a training effect. We therefore first analysed how accurate participants classified the Tweets in blocks 1 and 2.

During block 1, participants of all conditions have an equal performance (one-way ANOVA, p = 0.955). However, in block 2, the performances differ (one-way ANOVA, p < 0.001). Participants could have a learning effect and have higher performance on the second passing than on the initial try. If this was true, the performances should increase equally in all conditions. In the baseline condition (without a system in block 2), the performance increases on average by 1% (see Table 7). However, participants interacting with *C_AF*, *C_AR*, and *C_AN* make significantly more misclassifications in block 2 than participants of the baseline – their performance drops by 2%–5%. Contrarily, in *C_HF* and *C_HN*, the increase in performance are significantly stronger than in the baseline. One participant of *C_HF* summarises his opinion on the system in the optional comment box at the end of the study as follows:

> (P47) "Difficult to distinguish between what is offensive and what you personally think is offensive which is where the system decision can be helpful"

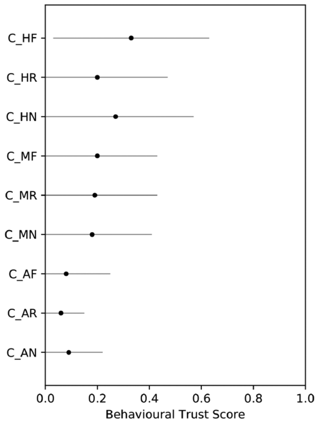


Fig. 5. Means (dots) and standard deviations (lines) of behavioural trust scores on a scale of 0 to 1 ("0" indicating low trust, "1" indicating high trust), measured as switching towards classifier's prediction normalised over amount of opportunities.

Table 8. Mean Values and Comparison of Behavioural Trust Scores Across Conditions, Grouped by Model Accuracy

		M = 0.33 *C_HF*	M = 0.20 *C_HR*	M = 0.27 *C_HN*	M = 0.20 *C_MF*	M = 0.19 *C_MR*	M = 0.18 *C_MN*	M = 0.08 *C_AF*	M = 0.06 *C_AR*	M = 0.09 *C_AN*
C_HF	M = 0.33		**		**	**	**	**	**	**
C_HR	M = 0.20	**								
C_HN	M = 0.27							**	**	**
C_MF	M = 0.20	**						**	**	*
C_MR	M = 0.19	**						**	**	
C_MN	M = 0.18	**						**	**	
C_AF	M = 0.08	**		**	**	**	**			
C_AR	M = 0.06	**		**	**	**	**			
C_AN	M = 0.09	**		**	*					

Where *: $p < 0.05$ and **: $p < 0.01$. Pairwise comparison with Mann–Whitney U-Test with two-sided alternative and p-values adjusted with Bonferroni correction.

We derive two insights from the analysis of the participants' classification performance: (1) The classification performance barely changes in the baseline condition when classifying the same Tweets a second time. Larger differences in classification behaviour between blocks 1 2 can therefore not be explained by a learning effect. (2) Some conditions with a support system show a significantly higher increase (*C_HF*, *C_HN*) or decrease (*C_AF*, *C_AR*, *C_AN*) in performance as compared to the baseline without a system. Since this observation cannot be explained by a training effect, other factors seem to influence the decisions of the participants.

Considering the results from analysing the participants' performances in blocks 1 and 2, we now investigate the classification and switching behaviour in more detail. The two-way ANOVA of the behavioural trust scores shows a significant difference between conditions ($p < .001$). The highest behavioural trust score appears in *C_HF* (M = .33, SD = .30), while the lowest score is found for *C_AR* (M = .06, SD = .09), see Figure 5. Comparing the systems with high accuracy (see Table 8), giving no explanation or giving a faithful explanation elicits equal trust levels, while *C_HR* is rated significantly lower than *C_HF*. *C_HF* is not rated significantly worse than *C_HN*, showing that adding a faithful explanation does not harm behavioural user trust. For both the medium-accuracy systems and antagonistic systems, adding an explanation does not have a significant impact: In both cases, a random explanation does not harm trust, nor does a faithful explanation harm or increase trust levels, as compared to the condition without any explanation.

Comparing the conditions with different classifier accuracies (see Table 8), our findings show that all systems with a random explanation had equally low behavioural trust scores—no significant difference in means could be proven. The same holds true for *C_HN* which is not significantly different from any condition with a medium classifier. *C_HR* also does not show a significant difference from any other condition besides *C_HF*. It needs to be noted that the variances of behavioural trust scores are high for some conditions, especially for the high-accuracy classifier. As the high-accuracy classifier only makes very few mistakes (0.5 mistakes over 15 Tweets on average), the number of opportunities to observe trust (i.e., switching towards the classifier) is lower than in

Table 9. Means and Standard Deviations of Switching Towards the Truth and Away from the Truth Per Condition

	towards truth M (SD)	**away from truth M (SD)**
C_HF	0.33 (0.31)*	0.03 (0.05)*
C_HR	0.20 (0.28)	0.03 (0.06)**
C_HN	0.27 (0.31)	0.03 (0.05)**
C_MF	0.24 (0.27)	0.04 (0.06)
C_MR	0.20 (0.28)	0.07 (0.11)
C_MN	0.19 (0.27)	0.04 (0.07)
C_AF	0.10 (0.23)**	0.08 (0.16)
C_AR	0.09 (0.20)**	0.06 (0.08)
C_AN	0.09 (0.16)**	0.09 (0.12)
baseline	0.23 (0.27)	0.06 (0.09)

Significant differences to baseline reported with asterisks, where * = $p < 0.05$, ** = $p < 0.01$. Pairwise comparison with Mann–Whitney U-Test with two-sided alternative.

other conditions. Users can only show this behaviour if they misclassified the Tweet in block 1, as the classifier mostly agrees with the ground truth.

Table 9 shows the normalised switching behaviour towards and away from the truth for all nine conditions and the baseline condition, in which participants classified the Tweets in both experiment blocks without the AI support system. It serves to validate that differences in switching behaviour (behavioural trust measure) are indeed originating from the influence of the system they interacted with, and not merely an artifact of interacting with the texts a second time (blocks 1 2). Similar to the analysis of participants' classification accuracy, we compare all nine conditions with the baseline and test for significant differences. Participants' switching behaviour diverts significantly from the baseline in several conditions, generally following the classifier's prediction: users of an antagonistic classifier (*C_A*) switch significantly less *towards* the truth, while users of the high-accuracy classifier (*C_H*) switch significantly less *away* from the truth. However, this holds true for all explanations (faithful, random and none) in both cases. Overall, this supports the validity of the behavioural trust results, as changes divert significantly from the baseline condition.

4.3 RQ2: How do System Accuracy and Explanation Presence and Faithfulness Impact Self-reported Trust?

We measured *self-reported trust* and illustrate the scores in Figure 6. The mean values are reported in Table 10. Participants show the highest self-reported trust in condition *C_HN* (M = 3.0, SD = 0.48), while the lowest trust score is found in *C_AF* (M = 1.9, SD = 0.45). Overall, the two-way ANOVA indicates a significant difference in means ($p < 0.001$). Table 10 also reports the pairwise comparison of conditions using the two-sided Mann–Whitney U-test with Bonferroni correction. For antagonistic systems (*C_A*), no significant difference in self-reported trust scores can be found between the three explanation types (faithful explanations, random explanations, and no explanations). However, in our experiment, for both "realistic" accuracy levels (i.e., high and medium), giving an explanation has the potential to reduce user trust significantly: The scores of *C_HR* and *C_HF* are both significantly lower than *C_HN*, and the trust in *C_MR* is significantly lower than in *C_MN*. Only *C_MF* does not receive significantly lower (but also not higher) self-reported trust

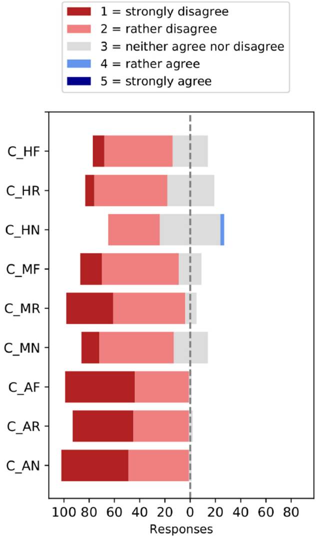


Fig. 6. Likert plots of self-reported trust scores per condition on a 5-point scale ("1" indicating low trust, "5" indicating high trust). Aggregation of 19 items, values in the plot rounded to the next integer value.

Table 10. Mean Values and Comparison of Self-reported Trust Scores Across Conditions, Grouped by Model Accuracy

		M = 2.68 *C_HF*	M = 2.75 *C_HR*	M = 3.03 *C_HN*	M = 2.47 *C_MF*	M = 2.17 *C_MR*	M = 2.61 *C_MN*	M = 1.93 *C_AF*	M = 1.95 *C_AR*	M = 1.97 *C_AN*
C_HF	M = 2.68			**	*	**		**	**	**
C_HR	M = 2.75			**	**	**		**	**	**
C_HN	M = 3.03	**	**		**	**	**	**	**	**
C_MF	M = 2.47	*	**	**		**		**	**	**
C_MR	M = 2.17	**	**	**	**		**	**	*	*
C_MN	M = 2.61			**		**		**	**	**
C_AF	M = 1.93	**	**	**	**	**	**			
C_AR	M = 1.95	**	**	**	**	*	**			
C_AN	M = 1.97	**	**	**	**	*	**			

where *: $p < 0.05$ and **: $p < 0.01$ (after Bonferroni correction). Pairwise comparison with Mann–Whitney U-Test using a two-sided alternative.

scores than the same system without explanations (*C_MN*). We conclude that the explanation type of an explanation is irrelevant for user trust if the classifier is sufficiently accurate (*C_H*), but plays a role for realistic but imperfect systems (*C_M*). Adding an explanation does not improve self-reported user trust in our experiment.

The self-reported ratings correspond to the results of behavioural trust in the antagonistic conditions. Yet, for a system with medium accuracy, there is a discrepancy between self-reported trust and behavioural trust. The self-reports show significant differences between *C_MR* and both *C_MF* and *C_MN*, while the behaviour shows no differences between the medium-accuracy conditions.

Furthermore, Table 10 allows the comparison across different explanation types and accuracies, showing that *C_MN* even reaches a trust score equal to that of two conditions with higher accuracies (*C_HF* and *C_HR*). In conditions without explanations, user trust adapts to the accuracy level: *C_HN* has a higher score than *C_MN*, which has a higher score than *C_AN*.

Körber's trust in automation questionnaire [24] also allows to investigate six sub-concepts of trust: perceived reliability/competence, understanding/predictability, familiarity, intention of developers, propensity to trust, and trust in general. The results of all six sub-concepts are presented in Figures 7 to 12. In the following paragraphs, we spotlight the four most relevant sub-concepts. For the pairwise post-hoc comparison of conditions, we provide the results of all six sub-concepts online.[6]

Understanding/Predictability. The questions used to measure self-reported understanding of the system revolve around assessing the transparency and predictability of a system (positive items: "The system state was always clear to me.", "I was able to understand why things happened.";

[6] https://git.gesis.org/papenmaa/tochi_trustinai.

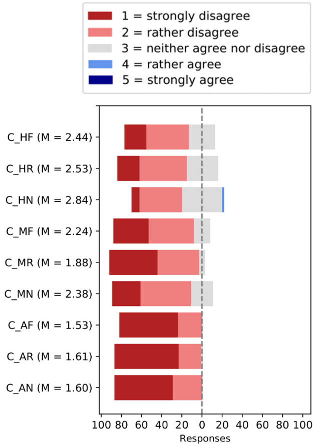


Fig. 7. Reliability/Competence.

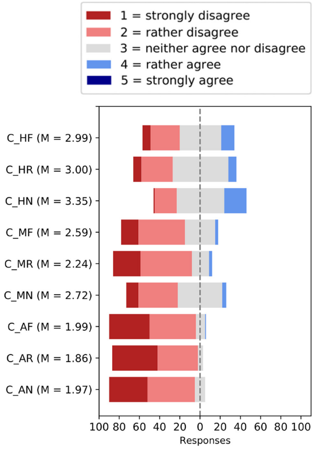


Fig. 8. Understanding/Predictability.

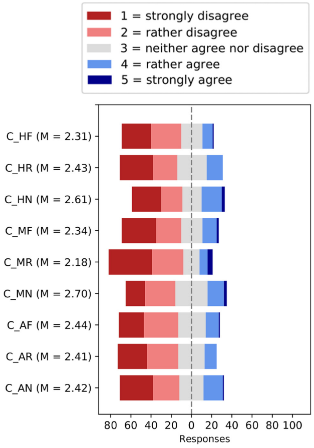


Fig. 9. Familiarity.

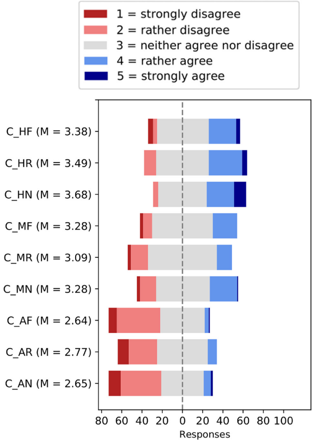


Fig. 10. Intention of Developers.

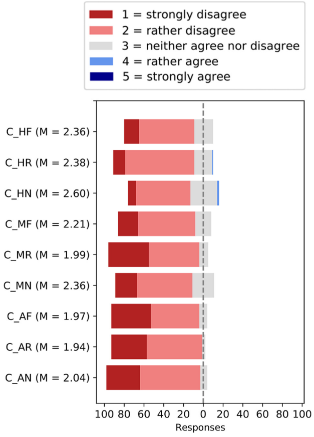


Fig. 11. Propensity to Trust.

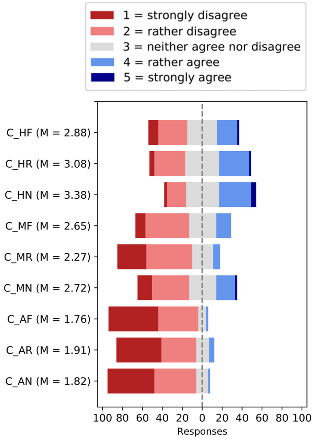


Fig. 12. Trust in Automation.

negative items: “The system reacts unpredictably.”, “It’s difficult to identify what the system will do next.”). The results (depicted in Figure 8) show that the user’s perception diverts from objective measures of predictability: Objectively, both *C_HN* and *C_AN* are equally predictable, yet the systems yield significantly different scores (p < 0.001). The highest understanding and predictability was ascribed to *C_HN*, while *C_AR* receives the lowest score.

For high-accuracy conditions, faithful and random explanations are both rated significantly lower than *C_HN*. For a medium-accuracy classifier, the faithful explanation (*C_MF*) leads to an equal self-reported understanding as *C_MN*. Overall, these findings are in line with the results of the general trust score.

Across accuracies, *C_MR* leads to the same feeling of understanding as systems with low accuracy, showing that random explanations have great potential for confusing the user. Contrarily, *C_MN* does not deliver explanations to the user and is rated to bring about the same understanding as *C_HF* and *C_HR*.

Intention of Developers. The participant’s perception concerning the developers of the automated decision-support system is measured with the items “The developers are trustworthy” and

“The developers take my well-being seriously”. The results show two effects (see Figure 10). First, all systems with *C_A* have comparable scores (no significant differences), no matter the explanation type. Secondly, none of the faithful explanation conditions showed to a significant rise in perceived benevolence of the developers (compared to the same system without any explanations).

Additionally, the results for this sub-concept also suggest that participants did not perceive *C_A* as particularly deceptive but rather ill-performing. Although the mean scores in *C_A* are lower than in *C_M* and *C_H*, there is room for lower scores (all three conditions lead to scores around 2.69 on a 5-point Likert scale). Looking at other sub-concepts shows that participants do not hesitate to give lower scores, e.g., scores around 1.6 for the reliability of *C_A* or around 1.94 for understanding of *C_A*. This conforms with comments of the participants:

> (P266) “Looks like a neural net or similar just starting out. Word choice seemed arbitrary.”

Some participants even tried to justify bad decisions of *C_AF*:

> (P163) “It kept picking up the word trash a lot, maybe because it has "as" in it, and "ass" is already in the system. Also it was completely ignoring the word bitch, or maybe the word isn’t considered bad anymore.”

or:

> (P617) “The system selected mostly rude words (like "bitch") and a possible racial insult ("yellow", but not "black"), and also links to websites. But some of it choices were rather surprising.”

Familiarity. The sub-concept of familiarity is measured using the items “I already know similar systems” and “I have already used similar systems”. Overall, participants confronted with *C_A* reported the lowest familiarity with “similar systems”, while users of *C_H* indicated a higher familiarity. Due to the sample size of each condition (N ≈ 98), we assume that there is no significant difference in the average background knowledge and familiarity with decision-support systems of participants. The difference in scores can then be explained by how each group interpreted the notion of “similar systems” in the questions. We conclude that the results show that the systems are perceived quite differently—different enough to shift the participants’ reference to “similar systems”.

Propensity. The sub-concept of propensity to trust does not focus on the one system being experienced, but on automated systems in general, with the items “I rather trust a system than I mistrust it”, “Automated systems generally work well”, and “One should be careful with unfamiliar automated systems” (negative item). Considering the sample size (N ≈ 98), we assume that the average propensity to trust an automated system was similar across conditions. If this assumption is correct, the interaction with the systems changed the participants’ perception of their own propensity to trust: The two-way ANOVA reveals a significant difference in scores ($p < 0.001$), with the highest propensity being reported in *C_HF* and the lowest in all three conditions of *C_A*.

4.4 RQ3: Do Behavioural Trust Measures and Self-reported Trust Measures Correlate?

The scores of behavioural trust and self-reported trust show multiple discrepancies. If user’s perception (manifested in the self-reported scores) would align with their actions (behavioural scores), the correlation between both measures would be high. Table 11 presents the correlation coefficients

Table 11. Correlation of Behavioural and Self-reported Trust Measures in Terms of Person's Rho and R-squared

conditions	r	r-squared	p-value	interpretation
all	.300	.0899	<.001	weakly positive
C_H	.025	.0006	.676	very weakly positive
C_M	.245	.0602	<.001	weakly positive
C_A	.311	.0966	<.001	weakly positive
F	.403	.1626	<.001	moderately positive
R	.245	.0602	<.001	weakly positive
N	.253	.0640	<.001	weakly positive
reliability / competence	.327	.1072	<.001	weakly positive
predictability / understanding	.246	.0603	<.001	weakly positive
familiarity	−.017	.0003	.624	very weakly negative
intention of developers	.184	.0339	<.001	very weakly positive
propensity to trust	.163	.0267	<.001	very weakly positive
trust in automation	.303	.0918	<.001	weakly positive

(Pearson's rho, interpretation according to Evans [13]) between behavioural and self-reported trust in various condition groups. Overall, the correlation between the two measures is weakly positive (r = 0.2998). A closer look into groups of conditions—either grouped by classifier accuracy or explanation type—reveals that the correlation is the strongest for faithful explanations (moderately positive correlation, r = 0.4033), while the lowest correlation can be found for the high-accuracy classifier (zero correlation, r = 0.0254). The high-accuracy system (*C_H*) provides only few possibilities to show the switching towards the classifier's prediction, as it shows the correct prediction in most cases. This could reduce the chances to observe a correlation due to an insufficient amount of data. However, the correlation between the observational and the self-reported trust measure is also weak for *C_M*, where enough possibilities to observe switching existed. We conclude that the weak correlation between the two measures show that the one is not a good proxy to measure the other.

Besides analyzing the behavioural measure with the self-reported ratings as a whole, we also compare the six sub-concepts of Körber's trust in automation questionnaire with the behavioural measure (six bottom lines in Table 11). The strongest (yet only weak) correlation can be seen with the sub-concept reliability / competence and the lowest correlation with the sub-concept of familiarity. There does not seem to be a sub-concept that represents the behavioural measure much better than the overall trust concept consisting of all six sub-concepts.

5 DISCUSSION

In the following section, we summarise and discuss the findings from the user study with respect to our research questions (RQ1: To what extent is *behavioural trust* influenced by the accuracy and explanation presence and faithfulness of a decision-support system?, RQ2: How do system accuracy and explanation presence and faithfulness impact the user's perception of trust, i.e., the *self-reported trust*?, RQ3: Do the measures of behavioural trust and self-reported trust *correlate*?) and come to the conclusions that:

It is complicated because faithful explanations do not necessarily promote user trust and understanding. The analysis of participant's classification performance clearly shows that a decision-support system of any kind influences the decision behaviour. Participants of our experiment showed an increase in classification performance when being supported by a high-accuracy system. Since the users' performance is influenced by a decision-support system, we also investigate how the user's trust in the system changes with varying conditions. Faithful explanations for any classifier do not have a significantly higher trust rating than the same classifier without any explanations. This holds true for both measures of trust (behavioural and self-reported). One case is especially noteworthy: While giving no explanation has scores that do not differ significantly from the scores of faithful explanations within the same classifier (i.e., *C_AN* compared to *C_AF*, or *C_MN* compared to *C_MF*), *C_HN* even yields significantly higher self-reported trust scores than *C_HF* on the self-reports. The same can be seen in the results of the sub-concept of predictability/understanding. Not giving an explanation leads to higher understanding scores than giving a faithful explanation. Our findings contradict our assumption on research questions **RQ1** and **RQ2**. In our experiment, a system with faithful or random explanations does not yield higher understanding scores than the same system without explanation. The self-reported trust in the developers' intentions likewise did not improve significantly with faithful explanations. The expectation mismatch problem [16] offers a possible explanation for this phenomenon. If lay-users base their mental model of the classifier on their pre-existing knowledge of human reasoning, they might assume that the classifier likewise uses causal information. Repeated interaction with the system and the (non-)offensive Tweets could also lead to the experience of illusory causation [34]. However, the system bases decisions on statistical relations rather than causal information since we use faithful explanations. As mentioned in [32], faithfulness does not necessarily imply meaningfulness in the eyes of the user. Subsequently, users might experience the expectation mismatch, leading to a decrease in trust. For conditions without explanations, users might not experience the expectation mismatch and could therefore end with a relatively high level of trust. Rader, Cotter, and Cho [41] likewise found in their experiment that their users expressed high levels of surprise for any explanation they offered, which could also originate from the expectation mismatch. A comparable observation was made by Cramer et al. [9], who suspect that disclosing system boundaries by giving explanations shows the user that the system is not flawless. The mental model and expectation of causal explanations should be investigated further in future work.

Our results further suggest that the findings of Langer, Blank, and Chanowitz [27] do not fully translate to human-machine interaction. In their experiment in human-human interaction, the level of informativeness of an explanation did not matter, while not giving an explanation resulted in less compliance with a request. Eiband et al. [12] found an indication that a similar trend can be seen in explanations for computer systems. However, our results do not support their conclusions, rejecting hypotheses H1.1–H2.2. In our experiment, not giving an explanation resulted in better or equal trust than when giving an explanation. In some cases, giving an explanation, even a faithful one, leads to worse results. The conclusion from Langer, Blank, and Chanowitz would encourage using explanations, as any explanation improves compliance compared to not giving an explanation. Contrarily, if aiming to improve user trust, our findings could discourage the usage of explanations similar to ours, as they do not necessarily improve the trust levels. Whether our explanations help to bring users' trust to a level appropriate for the respective system is to be explored in future work.

We conclude that, overall, equipping a system with explanations by highlighting decisive words in the input text does not necessarily improve user trust. In fact, giving any explanation (both faithful and random) reduced self-reported user trust significantly for high-accuracy systems in our setting. Although explanations are often deemed to have a positive effect on user trust [4, 16, 40, 51], our data do not show such a positive effect, neither for user trust nor for the users' feeling of understanding the system. Whether the explanations help to improve actual understanding (and therefore be a potential mechanism for increasing awareness of the system's workings) should be investigated in future work. Our finding makes adding minimal faithful explanations (e.g., highlighting of words in a text) unattractive for practitioners as to not harm user trust. Practitioners who are legally required to provide information on the mechanisms involved in a decision could be inclined to use persuasive yet unfaithful explanations as to not risk harming user trust, or to construct extensive explanations that elicit over-reliance [6].

In our experiment, a high-accuracy decision-support system did; however, improve the users' classification performance (i.e., how accurately they detected offensive language with the help of the support system). The increase in performance was especially significant for a system with faithful explanations and weakly significant for a system without explanations. The increase in performance was not significant for the system with random explanations. This leaves a mixed view on using explanations for boosting users' classification performance: If practitioners aim at increasing performance, only faithful explanations help achieve this goal.

It is complicated because random explanations confuse the user. Especially a random explanation that does not represent the actual mechanism of the system harms user trust for both classifiers with "realistic" accuracy levels (high-accuracy and medium-accuracy). A user of *C_HR* expressed the confusion about random explanations in the open comments box at the end of the survey:

> (P130) "I did find the automated system a little odd in terms of which words were highlighted in each tweet."

In *C_MR*, a user commented about the selection of words (which are random in *C_MR*):

> (P545) "In some cases it highlighted the word "bitch" and sometimes it did not highlight that word. (...) So I really do not know.."

C_HR received significantly lower self-reported trust scores than *C_HN*, yet the score is not significantly different from *C_MN*. *C_MR* shows the same dynamic: The score is significantly lower than *C_MN* and not significantly different from *C_AN*. Therefore, practitioners that make use of machine learning algorithms in their applications find themselves in a complicated situation: On one hand, they have to comply with the GDPR (if processing personal information) that calls for a right to explanation, or might want to add explanations for other reasons (e.g., increasing understanding, providing accountability, other ethical reasons). On the other hand, they can potentially harm the trust that users put in the system if the explanation is not well-designed. To avoid the risk of harming user trust, practitioners who are not legally required to add explanations might be inclined to keep their systems intransparent.

It is complicated because measuring trust by self-reports or by observation of behaviour is not equivalent to each other. Comparing the self-reported trust scores with the behavioural measure of trust reveals various discrepancies. The differences become especially obvious in the high-accuracy conditions. On the self-reports, *C_HN* yields significantly higher user trust than both *C_HF* and *C_HR*. The behavioural trust measure, however, draws a different picture. Here, *C_HN* and *C_HF* do not differ significantly from each other, while *C_HR* yields significantly lower trust scores than *C_HF*. Moreover, while one can draw the conclusion from the self-reported trust that the level of faithfulness does not play a role in trust, behavioural trust shows that faithful explanations have been rated significantly higher than random explanations in the case of the high-accuracy classifier. The correlation analysis did not show a strong correlation of the two measures, although the correlation differs within the three accuracy levels. While no correlation could be found for high-accuracy systems, a weak positive correlation was found for antagonistic systems. In both systems (*C_H* and *C_A*), the number of possibilities to see switching behaviour is low, which could have influenced the reliability and significance of the correlation test results. In conditions with the medium-accuracy model (*C_M*), a sufficient number of possibilities for observing switching was present and the variance in switching behaviour scores was lower than for *C_H* and *C_A*. Nevertheless, we could not observe a strong correlation in *C_M* conditions either. We conclude that the two measures cannot be used interchangeably. One possible reason for the discrepancy between both measures could be that self-reports are typically used after the interaction, while behavioural measures are recorded throughout the interaction. As trust develops over time [54], behavioural measures could be influenced by the trust-building process, while self-reports reflect only the final trust level. Another possible reason for the discrepancy could be that the behavioural measure of trust is influenced by other interactional factors. Furthermore, the two measures (questionnaire and behavioural measure) could be measuring something else than trust. The questionnaire, on one hand, was built on a long line of research on trust and was validated to function as a measure for trust in automation. The "switching rate" that was proposed as a behavioural measure of trust by Yin, Wortman Vaughan, and Wallach [53], on the other hand, is not yet a validated measure for trust. Similarly, other behavioural measures related to the switching rate have not previously been compared to a validated questionnaire ([54] compared a behavioural measure to a non-validated one-item questionnaire, and [48] used static agreement with the classifier instead of a switching rate). It is therefore possible that switching rates are not a good proxy for user trust. We found the sub-concept of reliability/competence from the questionnaire to have the highest correlation with the behavioural measure (although still being weak). Similar measures have been used to derive insights about the reliance of users on systems. It seems that the behavioural method investigated in this article is a better proxy for reliance than for trust as a whole.

Irrespective of the reasons for the discrepancies, we conclude that the results of studies that use different measures for user trust cannot easily be compared to each other. Considering the differences in results for behavioural trust and self-reported trust, the findings are in disagreement with hypothesis H3 for **RQ3**. Agreeing with Williams et al. [52], we suggest using both behavioural and self-report measures to reach a holistic picture of the user's perception of decision-support systems. Especially practitioners who aim at adding explanatory mechanisms to their decision-support systems should make use of both types of measures to evaluate not only the self-reported opinion of users but also inspect the practical (behavioural) consequences of the explanations.

It is complicated because the system's accuracy is not the only factor affecting user trust. Comparing conditions without explanations (*C_HN*, *C_MN*, and *C_AN*) shows that users adapt their trust and their perception of understanding to the classifier's accuracy. All three conditions have strong significant differences in means, with the high-accuracy classifier having the highest and the antagonistic classifier having the lowest self-reported trust. These findings confirm the results of Yu et al. [54], who conclude that their participants fit their trust to the accuracy of a system, and Dzindolet et al. [11], who found a correlation between trust and accuracy. Adding an explanation; however, influences this relationship. A random explanation added to a system even decreases self-reported trust significantly, leading to scores as low as the antagonistic system (*C_HR* is comparable to *C_MN* on the self-reports). Even a faithful explanation can decrease trust to the level of a system with lower accuracy (*C_HF* has equal self-reported trust scores as *C_MN*). The type of explanation, therefore, seems to impact the user's perception so that the general rule "trust adapts to accuracy" does not always hold true if an explanation is added.

It is complicated because interacting with the system might change how people evaluate their own ability to trust. With a sample size of at least 91 participants per condition, we assumed that there was no significant difference in the average propensity to trust across conditions. However, the two-way ANOVA on the sub-concept of propensity to trust showed a significant difference in scores. Propensity to trust is a sub-concept that focuses on the respondent's own (perceived) abilities rather than the system's. Seeing a significant difference across conditions might mean that the interaction with the system changed how strongly users believe in their own ability to trust. Although this should be investigated thoroughly in an experiment with pre- and post-task measurements, we see an indication that the interaction does not only change how the users see the system, but also how the users see themselves (i.e., how much they believe they are capable of trusting in general). In the open comments, one user of *C_AN* expressed mistrust in his own skills after being repeatedly confronted with decisions that diverted from his own:

> (P997) "I thought I knew about inappropriate language now I'm not so sure!"

Another participant (*C_MR*) ascribed her perception of improperly highlighted words on her age:

> (P831) "I am really sorry but due to the age I am I had difficulty understanding some of the words used."

The comments were recorded in an optional comment box—that is, participants were not prompted to comment on their skills, but still felt the need to express their thoughts. We used the comments to find hints for the low propensity to trust scores from the structured questionnaire. If a system could indeed change how users see themselves, the impact of decision-support systems would go beyond single decisions. A follow-up study based on this observation should investigate to what extent a decision-support system changes the users' perception of skills in more detail. Furthermore, it should be investigated whether the (potential) change in the ability to trust endures beyond the immediate interaction with the system.

6 LIMITATIONS AND FUTURE WORK

Our research work focuses on the evaluation of explanations for text input by highlighting individual words in the text. While this satisfies the definition of a minimum explanation [17], a richer explanation revealing more information would be possible. Poursabzi-Sangdeh et al. [39] showed that explanation depth itself does not improve user trust; however, using a combination of multiple explanation types (e.g., showing similar Tweets from the training set and their assigned labels in addition to word highlighting in the text) is possible.

Richardson and Rosenfeld [46] argue that explanations are needed for applications where faulty behaviour leads to severe consequences. As our use case involves only mild consequences (in the worst-case publishing a text with offensive language that can be read by children and teenagers), users might have been less careful when classifying. A use case that convincingly introduces a scenario where a misclassification would have severe consequences might deliver different results. We believe, however, that our study setup inflicted enough risk and uncertainty to be suitable to measure the impact of explanations on trust: (1 - risk) Participants did not know what the goal of the classification task was beforehand. As crowd workers are often used to train systems, participants did not know whether their decisions would become important training data for a system that protects the youth from bad language. (2 - risk) Participants were paid a fair compensation for their time, with the option to withhold payment for low-quality replies. Participants did not know whether their performance on detecting offensive language would be later used to analyse the response quality. (3 - uncertainty) The data points chosen for the study were not always easy to classify, which shows in the classification accuracy of participants (around 79% in block 1). Assuming that participants were eager to make correct classifications (given points (1) and (2)), they find themselves in a situation where the correct answer is not obvious and they potentially rely on the system and its explanations.

Another limitation of our approach is a difference in measurement time for the two trust measures (behavioural and self-reported). While the behavioural measure is recorded throughout the whole interaction, the questionnaire is used after the interaction only. We, therefore, assign the higher variances in the behavioural trust scores partially to the fact that trust builds up and develops over time, which could reflect in the behavioural measure but not in the self-reports. An appropriate counter-measure would be a training phase in which the user can get accustomed to the system and get to a stable trust level (if existing). We assume that adding a training phase would eventually lower the variances for the behavioural measures and make a more thorough analysis possible. After the training phase, an extended main phase could follow. In our experiment, participants interacted 15 times with the system. The experiment could be extended to allow for more interactions, either by extending the main phase or by introducing multiple sessions over a longer period of time. Although having other drawbacks (e.g., respondent fatigue or fading memory effects), more interactions would allow to record more possibilities to see switching behaviour and investigate the long-term usage and trust in the system.

We further focus on the practical consequences of adding an explanation; that is, whether users adapt their behaviour (in our use case their judgement of the offensiveness of a text) and their attitude towards a support system when being confronted with explanations of the system's mechanism. We report quantitative findings from closed questionnaires and behavioural measures. In our experiment, participants had the possibility to express their thoughts on the system in an open comment box. Many participants made use of the comment box (although being optional), showing that users are open to discuss their perception and ideas about the system, as seen in the comments cited in this article. A qualitative follow-up study could investigate the participants' mental models with qualitative methods and examine the actual understanding of the systems (rather than only the perceived understanding). We believe it would be interesting to investigate what brought about the differences in trust scores across conditions, what indications users rely on to build their trust, and how the user's attitude towards themselves get influenced by explanations.

7 CONCLUSION

In this article, we detail how explanation fidelity and model accuracy of an automated decision-making system impact user trust. We contribute results from an online user study with 959 participants using the use case "automatic detection of offensive text in Tweets". Applying a between-subject design, we let participants decide whether a Tweet is offensive or not, first without any system support and subsequently with the support of a system with various accuracy and explanation types. We found that the relationship between explanations and user trust is more complicated than reported in prior literature: In our experiment, adding explanations had a different impact on user trust depending on the level of accuracy. For a high-accuracy system, any explanation reduced self-reported user trust. However, for a medium-accuracy system, only random explanations received lower self-reported trust scores than a system without explanations. Furthermore, we compared a measure of self-reported trust to a measure of behavioural trust. Furthermore, complicating the research about trust in automated decision making, our findings reveal differences between the "switching rate" measure of behavioural trust and the self-reported trust measure. We conclude that behavioural and self-report measures cannot and should not be used interchangeably. In future work, we aim at investigating the user-system trust relationship with a more complex use case in which the costs of erroneous decisions are higher and take a closer look at the mental model the user builds during the interaction.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their time and effort taken to provide feedback on the manuscript. Their comments were greatly appreciated and improved our work.

REFERENCES

[1] P. S. Altares, A. R. I. Copo, Y. A. Gabuyo, A. T. Laddaran, L. D. P. Mejia, I. A. Policarpio, E. A. G. Sy, H. D. Tizon, and A. M. S. D. Yao. 2003. *Elementary statistics: A modern approach.* Rex Bookstore Inc. Manila, Philippines.

[2] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. "What is relevant in a text document?": An interpretable machine learning approach. *PloS One* 12, 8 (2017), e0181142. DOI: https://doi.org/10.1371/journal.pone.0181142

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. DOI: https://doi.org/10.1016/j.inffus.2019.12.012

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

[5] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces.* Association for Computing Machinery, New York, NY, 454–464. DOI: https://doi.org/10.1145/3377325.3377498

[6] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics.* IEEE Computer Society, 160–169. DOI: https://doi.org/10.1109/ICHI.2015.26

[7] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. 2018. Learning to explain: An information-theoretic perspective on model interpretation. In *Proceedings of the 35th International Conference on Machine Learning.* Jennifer Dy and Andreas Krause (Eds.), Vol. 80, PMLR, Stockholmsmässan, Stockholm Sweden, 883–892. DOI: http://proceedings.mlr.press/v80/chen18j.html.

[8] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, 12 pages. DOI: https://doi.org/10.1145/3290605.3300789

[9] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455. DOI: https://doi.org/10.1007/s11257-008-9051-3

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

[11] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. DOI: https://doi.org/10.1016/S1071-5819(03)00038-7

[12] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, 6 pages. DOI: https://doi.org/10.1145/3290607.3312787

[13] James D. Evans. 1996. *Straightforward Statistics for the Behavioral Sciences.* Thomson Brooks/Cole Publishing Co, Belmont, CA.

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

[15] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics.* 80–89. DOI: https://doi.org/10.1109/DSAA.2018.00018

[16] Alyssa Glass, Deborah L. McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th International Conference on Intelligent User Interfaces* . Association for Computing Machinery, New York, NY, 227–236. DOI: https://doi.org/10.1145/1378773.1378804

[17] Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38 (2016). DOI: https://doi.org/10.1609/aimag.v38i3.2741

[18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys* 51, 5 (2018), 42 pages. DOI: https://doi.org/10.1145/3236009

[19] Nigel Harvey and Ilan Fischer. 1997. Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes* 70, 2 (1997), 117–133. DOI: https://doi.org/10.1006/obhd.1997.2697

[20] Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. In *Proceedings of the NIPS 2017 Symposium on Interpretable Machine Learning*.

[21] Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research* 4, 2 (2008), 267–288. DOI: https://doi.org/:10.1515/JPLR.2008.013

[22] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. DOI: https://doi.org/10.1207/S15327566IJCE0401_04

[23] Audun Jøsang and Stéphane Lo Presti. 2004. Analysing the relationship between risk and trust. In *Proceedings of the International Conference on Trust Management*. Springer, 135–145.

[24] Moritz Körber. 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Proceedings of the 20th Congress of the International Ergonomics Association*. Sebastiano Bagnara, Riccardo Tartaglia, Sara Albolino, Thomas Alexander, and Yushi Fujita (Eds.), Springer International Publishing, Cham, 13–30.

[25] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 126–137. DOI: https://doi.org/10.1145/2678025.2701399

[26] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Proceedings of the IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. DOI: https://doi.org/10.1109/VLHCC.2013.6645235

[27] Ellen J. Langer, Arthur Blank, and Benzion Chanowitz. 1978. The mindlessness of ostensibly thoughtful action: The role of "placebic" information in interpersonal interaction. *Journal of Personality and Social Psychology* 36, 6 (1978), 635. DOI: https://doi.org/10.1037/0022-3514.36.6.635

[28] John D. Lee and Katrina A . See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

[29] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–15. DOI: https://doi.org/10.1145/3313831.3376590

[30] Brian Y. Lim and Anind K. Dey. 2011. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*. Association for Computing Machinery, New York, NY, 415–424. DOI: https://doi.org/10.1145/2030112.2030168

[31] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 2119–2128. DOI: https://doi.org/10.1145/1518701.1519023

[32] Zachary Lipton. 2018. The mythos of model interpretability: In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. *Queue* 16, 3 (2018), 31–57. DOI: 10.1145/3236386.3241340

[33] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*. Citeseer, 6–8.

[34] Leslie Zebrowitz McArthur. 1980. Illusory causation and illusory correlation: Two epistemological accounts. *Personality and Social Psychology Bulletin* 6, 4 (1980), 507–519. DOI: https://doi.org/10.1177/014616728064003

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

[36] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* 11, 3–4 (2021), 45 pages. DOI: 10.1145/3387166

[37] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. 112–121.

[38] Kenya Freeman Oduor and Eric N. Wiebe. 2008. The effects of automated decision algorithm modality and transparency on reported trust and task performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 302–306. DOI: https://doi.org/10.1177/154193120805200422

[39] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Article 237*. Association for Computing Machinery, 52 pages.

[40] Alun Preece. 2018. Asking 'Why' in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72. DOI: https://doi.org/10.1002/isaf.1422

[41] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 13 pages. DOI: https://doi.org/10.1145/3173574.3173677

[42] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95.

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, 1135–1144. DOI: https://doi.org/10.1145/2939672.2939778

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1527–1535.

[45] Mireia Ribera and Agata Lapedriza. 2019. Can we do better explanations? A proposal of user-centered explainable AI. In *Proceedings of the IUI Workshops*.

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

[47] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science* 26, 5 (2002), 521–562. DOI: https://doi.org/10.1207/s15516709cog2605_1

[48] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, 240–251. DOI: https://doi.org/10.1145/3301275.3302308

[49] Philipp Schmidt, Felix Biessmann, and Timm Teubner. 2020. Transparency and trust in artificial intelligence systems. *Journal of Decision Systems* 29, 4 (2020), 260–278. DOI: https://doi.org/10.1080/12460125.2020.1819094

[50] Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, and Mark Neerincx. 2018. Contrastive explanations for reinforcement learning in terms of expected consequences. In *Proceedings of the IJCAI-17 Workshop on Explainable AI (XAI)*.

[51] Eric S. Vorm. 2018. Assessing demand for transparency in intelligent systems using machine learning. In *Proceedings of the 2018 Innovations in Intelligent Systems and Applications*. IEEE, 1–7. DOI: https://doi.org/10.1109/INISTA.2018.8466328

[52] Parker A. Williams, Jeffrey Jenkins, Joseph Valacich, and Michael D. Byrd. 2017. Measuring actual behaviors in HCI research–A call to action and an example. *AIS Transactions on Human-Computer Interaction* 9, 4 (2017), 339–352.

[53] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY 12 pages. DOI: https://doi.org/10.1145/3290605.3300509

[54] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM, New York, NY, 307–317. DOI: https://doi.org/10.1145/3025171.3025219

[55] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li, and Yang Wang. 2019. Effects of influence on user trust in predictive decision making. In *Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, 1–6. DOI: https://doi.org/10.1145/3290607.3312962

Received January 2021; revised July 2021; accepted October 2021