The Seventh AAAI Conference on Human
Computation and Crowdsourcing (HCOMP-19)

The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems

Mahsan Nourani,[1] **Samia Kabir,**[2] **Sina Mohseni,**[2] **Eric D. Ragan**[1]
[1]University of Florida, Gainesville, Florida
[2]Texas A&M University, College Station, Texas
{mahsannourani, eragan}@ufl.edu, samia.kabir@tamu.edu, sina.mohseni@tamu.edu

Abstract

Machine learning and artificial intelligence algorithms can assist human decision making and analysis tasks. While such technology shows promise, willingness to use and rely on intelligent systems may depend on whether people can trust and understand them. To address this issue, researchers have explored the use of explainable interfaces that attempt to help explain why or how a system produced the output for a given input. However, the effects of meaningful and meaningless explanations (determined by their alignment with human logic) are not properly understood, especially with users who are non-experts in data science. Additionally, we wanted to explore how explanation inclusion and level of meaningfulness would affect the user's perception of accuracy. We designed a controlled experiment using an image classification scenario with local explanations to evaluate and better understand these issues. Our results show that whether explanations are human-meaningful can significantly affect perception of a system's accuracy independent of the actual accuracy observed from system usage. Participants significantly underestimated the system's accuracy when it provided weak, less human-meaningful explanations. Therefore, for intelligent systems with explainable interfaces, this research demonstrates that users are less likely to accurately judge the accuracy of algorithms that do not operate based on human-understandable rationale.

Introduction

Intelligent systems have drawn the attention of many researchers and scientists and are widely used in different domains and applications, such as classification, recommendation, and decision-support systems. While intelligent systems are used for various purposes, they share and follow the same motivation: they are designed to help users achieve their goals more conveniently by optimizing and automatizing parts of the process and reducing mental and physical efforts for users.

By taking advantage of machine learning and artificial intelligence, intelligent systems provide users with model-generated outputs of different types, but the rationale behind how and why these outputs are generated is not always clear to the users of such systems. This is a major concern for intelligent systems, as users might need to know how reliable the outputs are (Siau and Wang 2018; Goodall et al. 2018) and when they should—or should not—trust their judgment. Lack of trust in an intelligent system could be problematic even if it produces accurate results, as it might lead to users' reluctance to rely on the system, leading to reduced overall efficiency and human-machine performance. Furthermore, once trust in an automated system is lost, it may be hard to reestablish (Hoffman et al. 2013). In other cases, users may tend to over-rely on a system's results and outputs—regardless of their correctness or relevance—as they believe a computing system is more knowledgeable and "intelligent" than they are (Ribeiro, Singh, and Guestrin 2016)—a phenomenon known as *automation bias* (Goddard, Roudsari, and Wyatt 2011). This could be especially dangerous for systems that help users with critical tasks and decisions. There is no intelligent system that performs with 100% accuracy for meaningful real-world tasks for which people truly need machine assistance. Hence, there are cases where an intelligent system might produce false positive or false negative results. Knowing these situations, we realize how risky over-relying on an intelligent system could be (Parasuraman and Riley 1997).

In order to address these issues, researchers and designers look to *explainable* systems to support machine transparency and human understanding of system functionality (Došilović, Brčić, and Hlupić 2018). Offered in various formats (e.g., textual, visual, or numerical), explanations can help clarify the rationale behind an intelligent system's outputs and judgments to help users better understand how it is working.

While users could potentially benefit greatly by having explanations for intelligent systems, little empirical evidence exists to inform exactly how explanations affect users' behaviors, perception of the system accuracy, and trust in the machine outputs. It is also important to understand the implications of explanation design choices. For instance, system designers must make important decisions on the type of explanation, the level of detail, and use cases for when to present explanations as part of the system workflow. A poor judgment from a system designer in selecting proper explanations could backfire on the user's understanding, decision-making, and reliance on an intelligent system.

Since explanations are introduced to make intelligent systems more understandable, one important consideration is the extent to which explanations are meaningful, especially when users are not knowledgeable in machine learning or artificial intelligence. Many algorithms function according to rules and layers that do not align with human logic, such as in cases where influential features are learned from coincidences in a training set during supervised approaches (e.g., (Ribeiro, Singh, and Guestrin 2016)). There is a lack of empirical knowledge about how perceived meaningfulness of explanations influences the perception of an intelligent system's performance. User trust and understanding of intelligent systems are complex issues and could be potentially affected by multiple factors. Our research focuses on user perception of model accuracy in intelligent systems. Generally, we would expect higher perception of system accuracy to result in higher user trust (Yin, Vaughan, and Wallach 2019), but trust may also be influenced (positively or negatively) by user perception of the appropriateness of the system's rationale for decisions. Our main goal for this research is to understand the relationship between observed accuracy and the meaningfulness of explanations in influencing user perception of accuracy.

We present an experiment with controlled levels of simulated system accuracy and explanation meaningfulness using a binary image classification scenario. Through this work, we contribute novel empirical results demonstrating the importance and influence of meaningfulness of explanations in transparent and explainable intelligent systems for user perception of system accuracy.

Related Work

Though intelligent systems have proven to be useful in many scenarios and contexts, shortcomings have encouraged researchers to think outside the box to come up with solutions to address issues with system failures and user trust. Model failures and false positives in certain applications and domains (e.g., medical diagnostic systems) could be disastrous (Akata et al. 2018). To avoid misunderstandings and mistrust, many researchers have been interested in approaches that help make machine functionality more transparent for users. In discussion of such concepts, researchers use different related terminology, with examples including *interpretable systems*, e.g., (Abdul et al. 2018), *explainable systems*, e.g., (Core et al. 2006), and *transparent systems*, e.g., (Amershi et al. 2014). For simplicity, we refer to this area as *explainable artificial intelligence* (XAI) (Doran, Schulz, and Besold 2018), and substantial research efforts have been explored in this area, e.g., (Van Lent, Fisher, and Mancuso 2004; Core et al. 2006; Ribeiro, Singh, and Guestrin 2016).

Despite being an inspiring and growing research area, researchers are dealing with major challenges while attempting to advance XAI for today's algorithms. One challenge, for instance, is that deep-learning methods are black-boxes by nature; in other words, they are not human-interpretable (Akata et al. 2018). Another challenge is visualizing the explanations in a comprehensive way for better user understanding, e.g., Alsallakh et al. (2014) and Samek et al. (2017).

Despite many promising directions for explainable intelligent systems, in their paper, Miller, Howe, and Sonenberg (2017) argue that many researchers and designers are falling into the unforeseen pitfall of designing explanations for themselves (i.e., AI researchers) rather than the end users who are often less knowledgeable of data science. Understanding human factors and basic human reactions to intelligent systems is crucial for the design of explainable systems. As a result, numerous research efforts have focused on various aspects of human-centered designs for XAI systems, e.g., Dodge et al.(2018) and Zhu et al. (2017). Such examples help demonstrate the breadth of interests and approaches in human-centered research in XAI systems, but further progress and empirical testing is needed to better understand the implications of explanations on human understanding and behavior.

As explanations are helpful for users in understanding and accepting system output (Cramer et al. 2008) and building a mental model of how it works (Kulesza et al. 2013), it is also important to study the effectiveness of explanations. Adadi and Berrada (2018) argue that in spite of an increasing body of work to implement and produce XAI systems and techniques, there have been few projects focused on evaluating these methods, with only 5% of all the papers in this community focused on user evaluation. Given this limited focus, a need for human evaluation of XAI systems is essential. Human-subjects studies allow researchers to not only evaluate the quality of the system and a particular explanation but also to study more nuanced properties of explanation design.

The current body of work in human-centered evaluation of XAI systems covers various factors. For example, Mohseni and Ragan (2018) presented a benchmark from user annotations of image and text samples that demonstrates how the human-centered evaluations of a system might be used to qualify the local machine learning explanations. In another study, Kizilcec (2016) tested trust in a system with three levels of explanation details. The results show that both explanations with few details and too much detail can cause users to lose trust in the system. Hence, balancing the level of detail is important while designing an XAI system. A final example is the work demonstrated by Roy et al. (2019), where a preliminary evaluation of their proposed explainable deep-learning approach studied how helpful users find different types of explanation representations in their visual interface. Their results show that users preferred more simplistic visual annotations over more detailed probabilistic explanations about the model components.

In addition, several studies have focused primarily on user trust in automated systems. For instance, Holliday, Wilson, and Stumpf (2016) studied how user trust evolves over time in an intelligent system *with* and *without* explanations. Their results showed participants in the *with-explanation* condition tended to trust the system more in the beginning with a more accurate mental model of the system, while the participants in the *no-explanation* condition lost trust over time and had difficulty building a mental model of the system.

According to them, trust in an intelligent system is a multidimensional and highly complex phenomenon that can be hard to assess by directly questioning users about a single factor. This raises a concern of finding potential approaches to assess trust.

Researchers suggest trust can be quantified using other measures and indirect metrics. In other words, instead of measuring trust, we can measure concepts that are believed to affect it. One such approach is to assess user understanding of system accuracy. In recent relevant work, Yin, Vaugh, and Walach (2019) conducted three user studies with controlled accuracy to learn how model accuracy can affect user trust. Their findings show that trust was significantly affected by the observed accuracy compared to the reported accuracy. This finding backs up our hypothesis that user perception of model accuracy can affect understanding of and trust in an intelligent system. While our research similarly studies how observed machine performance can influence perception of accuracy, it focuses on the influence of explanations and explanation meaningfulness in conjunction with observed accuracy.

In other relevant work, Lim, Dey, and Avrahami (2009) studied the effects of two types of *why* and *why not* explanations on user trust, performance, and understandability of the underlying model. Their results show that explanation type influences user understanding and trust in the system. They also found inclusion of explanations lead to better understanding of the XAI system and the underlying model. Our research also compares the presence and absence of explanations, but we quantitatively evaluate perception of system accuracy, and we compare different types of explanations based on how meaningful they are for human subjects.

Experiment

We conducted a controlled experiment to study how the meaningfulness of explanation affects human perception of model accuracy. The experiment is based on an image classification task where participants review and judge the accuracy of machine classification.

Research Goals and Hypotheses

The primary goal of this research was to evaluate how human-perceived meaningful and non-meaningful explanations influence user perception of model accuracy. We studied how people perceive the accuracy of an explainable intelligent system after a period of system use.

To do this, we controlled two levels of human-meaningfulness for explanations in an image classification scenario. We designed *strong* meaningful explanations to focus on the key features of primary subjects in the images and the *weak* meaningful explanations to highlight background and nondescript features of the image. As we would expect users to trust an intelligent system with a higher observed accuracy, the hypotheses and results for the perceived accuracy are expected to have implications for user trust. We hypothesized that providing different levels of explanation meaningfulness for a classification output would directly affect the perception of accuracy—even when the observed system accuracy is held constant.

We predicted that people provided with human-meaningful explanations would tend to predict a higher accuracy for the system, and hypothesized an underestimation of system accuracy would occur when users see explanations that do not make sense in terms of meaningful human logic or classification. It is important to clarify that our research addresses the effects of explanation differences while maintaining consistency in actual observed system accuracy.

Experimental Design

To test our hypotheses, we designed an experiment with participants reviewing images along with the output from a simulated classification system. For the purpose of evaluating human perception of machine accuracy, we sought a task easy enough for non-expert users to quickly assess the correctness of the classifications. Furthermore, we favoured a straightforward task that did not require any particular domain expertise to complete. To this end, the experiment used an image classification distinction task between images of cats and dogs.

Moreover, we sought an understandable explanation format that did not require a computational background in machine learning and artificial intelligence, so it would be suitable for non-experts. For this reason, we opted for visual explanations where highlighted regions were used to mark key areas used by the model to determine classification. This method is commonly used in explainable classification systems, e.g., Riberio, Singh, Guestrin (2016), Samek, Wiegand, and Müler (2017), and Samek et al. (2017). Most often, heatmaps are used to indicate the relative impact of different image regions, though approximated bounding boxes can also be used for simplicity, e.g., Alsallakh et al. (2018) and Zhang and Zhu (2018). For our study, it was important that participants could easily and quickly interpret the explanations; thus, our implementation used bounding regions to reduce explanation complexity and facilitate faster review.

Our study consisted of two tasks: a *review task* and a *prediction task*. The review task was designed to give participants a chance to observe classification performance over a period of 40 trials, each including an image with its corresponding label and explanation (as appropriate for the study condition). Participants would first review and complete these trials and then continue with the prediction task, a method suggested by others (Hoffman et al. 2018), where subjects were asked to anticipate whether the model would correctly classify items in a new set of images. This task contained 50 trials based on a new set of images without showing explanations and classification labels.

The experiment controlled two independent variables. The first was *explanation type*, which refers to the meaningfulness of an explanation, aligned with human rationale, present in each trial in the review task. We tested three levels of explanation type: *none*, *strong*, and *weak*. Figure 1 shows examples of explanations given for the different explanation types for two input images used in the study. The *none* type was the control condition where no explanation was provided.

The distinction between explanation types in our study is directed by our focus on explanations that are logical and understandable for non-expert end users rather than experts or detailed explanations of machine models. Thus, we designed *strong* explanations to contain and focus on key features of its subject animal that humans find meaningful in order to distinguish that animal from the others (e.g., eyes, ears, snout, and paws). In contrast, we designed *weak* explanations to contain more background content and non-deterministic portions of an image (e.g., nondescript patches of fur, trees in the back, and etc.) rather than the most prominent characteristics of the foreground target.

In addition, the experiment controlled *simulated system accuracy*, which controlled the balance of false and correct classifications from the trials observed in the review task. Controlling the exact accuracy observed by participants was done to prevent confounds due to different participants observing different accuracy levels during the trials of the study. We used 95% as for one accuracy level and 70% as the other; for ease of reference, we refer to these levels as *high* and *low* (respectively). The reason for selecting these accuracy levels was to have a noticeable difference in observed incorrect classifications (i.e., the number of false classifications observed by participants) while still maintaining enough correct classifications that participants could easily see that the system was right more often than not. Note that, in practice, the level of accuracy that can be considered "high" or "low" depends entirely on the difficulty of the classification task and data set, and we use these terms only for the convenience of reference.

The experiment followed a 2x3 between-subjects design, where each participant completed the study in exactly one assigned condition. The between-subjects design was important because participants needed to review multiple inputs and outputs to develop an understanding of the system's accuracy, and we wanted to avoid any potential confusion or learning effects that might have occurred if participants reviewed outputs for multiple conditions.

Dataset, Explanation Creation, and Verification

As the source data for the classification scenario, the study used images from the publicly available Kaggle *Dogs Vs. Cats* dataset [1](a dataset compiled for research on classification algorithms). For the study, we selected a set of 40 images (20 for each animal) to use in the review task and also, a set of 50 images (25 for each animal) to use in the prediction task.

The experimental design required *strong* and *weak* explanations for our simulated classification system. In order to maintain experimental control over the style and meaningfulness of the explanations in accordance with our *strong* and *weak* explanation types, we manually created explanations (see Figure 1). By crafting explanations specifically for the experiment, we avoided possible confounds from high variance in explanation relevance or unexpected feature highlighting that sometimes results from real classification and explanation algorithms. To create *weak* explanations, we added red, transparent bounding boxes over areas of the image in background and with little, vague, or ambiguous focus on the animal's face or body. The highlighted areas of *strong* explanations focus mainly on the animal's facial features such as eyes, ears, whiskers, and snout, or they highlight prominent body features such as the tail or feet. As the experimental design required a clear distinction between *strong* and *weak* explanation types, we developed the explanation images with iterative refinements based on preliminary verification studies. These verification studies helped assure that subjective interpretations of the explanations' meaningfulness would align with their intended design. In these preliminary verification studies, a total of 14 participants provided feedback. Of the 14, eight were familiar with the concepts of machine learning through the courses they took and their personal experiences.

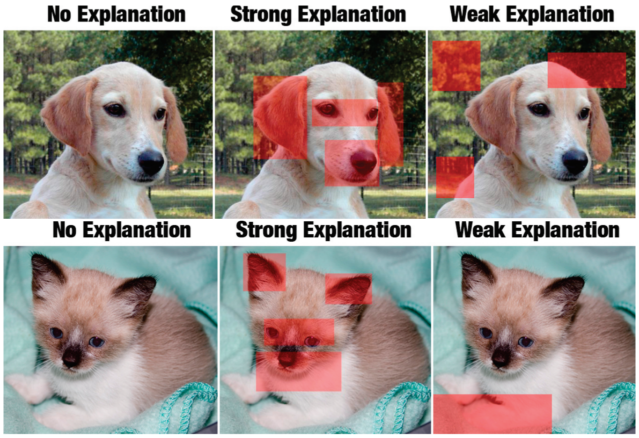


Figure 1: Example of images and explanations used in the experiment. The left shows the original image with no explanation (i.e., the *none* condition). The middle shows a *strong* explanation, focusing the highlighted areas on the animal's facial features. The image on the right is a *weak* explanation, focusing the highlighted areas on the background and non-deterministic areas.

In the preliminary verification study, participants were asked to view each generated explanation image and rate how well they thought the highlighted areas distinguish the animal for a correct classification. They rated the quality of each explanation separately using a five-point Likert scale. We then calculated how many of these ratings matched the intended positive (*strong*) and negative (*weak*) designations for each image. For the purpose of achieving clear differences in meaningfulness between *strong* and weak *groups*, any explanation with a neutral rating was considered as a failed match. Explanations that had less than 80% match were omitted or revised. We iteratively revised explanations and conducted additional user tests to refine a final explanation dataset with clear distinctions for the *strong* and *weak* classifications, such that participant responses matched the intended category. The resulting dataset was comprised of both *strong* and *weak* explanation images for each of 40 base images, giving us a total of 80 highlighted explanation images to use in the review phase of the experiment.

[1] https://www.microsoft.com/en-us/download/details.aspx?id=54765

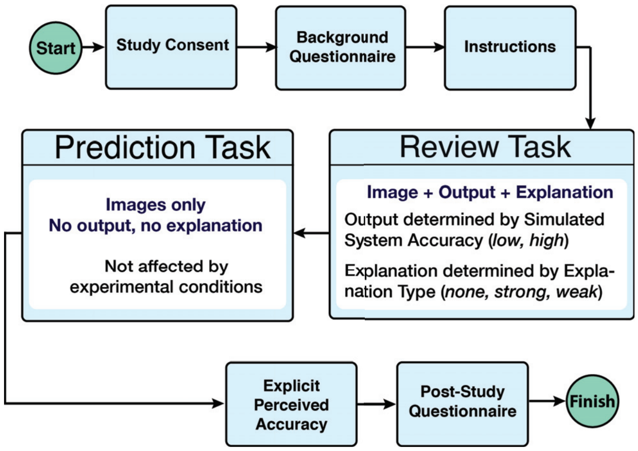

Start
Study Consent
Background Questionnaire
Instructions
Prediction Task
Images only
No output, no explanation
Not affected by experimental conditions
Review Task
Image + Output + Explanation
Output determined by Simulated System Accuracy (low, high)
Explanation determined by Explanation Type (none, strong, weak)
Explicit Perceived Accuracy
Post-Study Questionnaire
Finish

Figure 2: An overview of the experiment's procedure.

Procedure

The experiment was conducted through an in-lab user study using a custom web application and a standard keyboard-and-mouse computer interface. Each participant completed the study in a single session, which lasted approximately 30 minutes. The overall procedure is shown in Figure 2. This research was approved by our organization's institutional review board (IRB).

Participants first completed a questionnaire to obtain basic information about their age, education, and knowledge of using computer systems and machine learning. The experimenter then explained an overview of the cat-or-dog classifier system, how the system offers an output for each image, and the associated explanation image (for the appropriate conditions with explanations). A think-aloud approach was encouraged, but was not mandated to the users.

Participants then completed the *review* task. For this task, each participant would see: (1) an image, (2) a textual output label representing the system's classification of that image, and (3) an additional explanation image based on the explanation type of the experimental condition (participants in the *none* condition did not see an explanation). Participants reviewed 40 images for this task. Images were viewed one at a time and with order randomized per participant. After finishing each trial, they could not go back or change past responses.

To make sure participants were giving sufficient attention to each image and the system's output, participants were required to answer two questions per trial. First, we asked whether or not they agreed with each classification. Since these were images of cats and dogs, this was an easy question, but it required participants to pay attention to the output label and its correctness. Second, we asked participants to indicate how well the highlighted areas explain the associated label on a five-point Likert scale. This question was included to ensure participants reviewed the explanation for each image (note that this question was excluded for participants without explanations in the *none* conditions). While both questions in the review task were included primarily to promote sufficient participant engagement and attention, we used the results from the second question to assess the connection between meaningfulness of explanations and the observed accuracy.

After the review task, participants completed a prediction task, which was done for the sole purpose of collecting an additional measure of perceived system accuracy. In this task, participants viewed new images of cats and dogs, but this time, no classification labels were given. Instead, they were asked if the same computer system from the review task would classify each image correctly. The prediction task was not influenced by the assigned experimental condition; all participants viewed images of cats and dog without an associated output label and with no explanation. Each participant predicted outputs for 50 randomly-ordered images. We used the percentage of responses where participants predicted correct system classifications as way of estimating participant perception of system accuracy. We refer to this measure as *implicit perceived accuracy*. After the prediction task, participants were asked to provide a numerical estimate of the system's accuracy on the scale of 0 to 100 percent. We refer to this measure as *explicit perceived accuracy*.

Participants

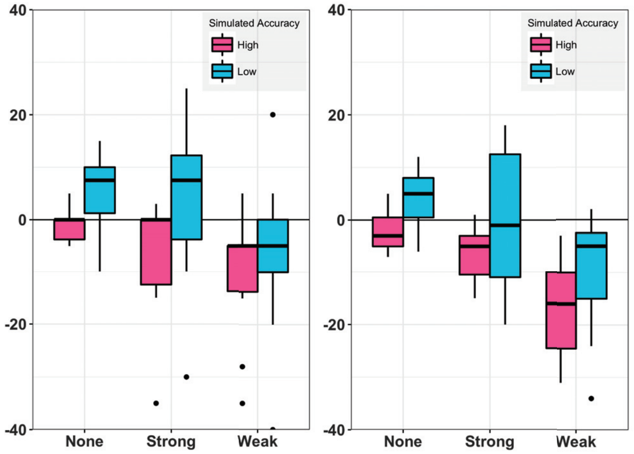
The study had 60 participants, all 18 or older, with 24 females and 36 males. Participants were undergraduate and graduate students. To make sure their familiarity with the data science and machine learning concepts was reasonably similar across the conditions, we used responses to two questions in the background questionnaire about their academic specializations and whether they had taken any courses in machine learning. We categorized the participants into three levels of familiarity and distributed participants among conditions based on familiarity.

Results

We analyzed the perceived accuracy results from the experiment along with user ratings for explanation meaningfulness from the review task.

Perceived Accuracy

We tested the effects of simulated system accuracy and explanation type on both explicit and implicit perceived accuracy. For statistical analysis, we applied an independent two-way factorial ANOVA statistical analysis for each measure. Since measured levels of accuracy are related to the simulated level of system accuracy in the review task, we calculated the error in perceived accuracy as the difference between the actual simulated system accuracy (i.e., the accuracy rate participants observed) and the user perceived accuracy (i.e., the rate measured from participant responses). This approach allows us to easily interpret cases when users overestimate or underestimate the accuracy relative to the observed simulated accuracy. Figure 3a and Figure 3b show the output plots for the calculated difference in perceived accuracy for both implicit and explicit accuracy measures, respectively.



(a) **Implicit Perceived Accuracy (Difference in Percentage)**

(b) **Explicit Perceived Accuracy (Difference in Percentage)**

Figure 3: The perceived accuracy results expressed as the difference from the true observed accuracy determined by the simulated accuracy condition. Higher than 0 indicates overestimating the system's accuracy, while lower than 0 indicates underestimation. Participants underestimated accuracy significantly more with *weak* explanations.

Implicit Perceived Accuracy Implicit perceived accuracy was determined over the course of the prediction task (see Section). Figure 3a shows the results for implicit perceived accuracy by explanation type and simulated system accuracy. We will discuss the analysis of results in terms of error (i.e., difference from the observed accuracy), where zero error indicates a match with the conditions controlled level of accuracy. Note that in the controlled condition with no explanation, error is close to the zero line, being more often overestimated for the *low* simulated accuracy and more often underestimated for the *high* simulated accuracy.

The two-way ANOVA results show a significant main effect for simulated system accuracy on implicit error with $F(1, 54) = 7.51, p < 0.01$, and effect size of $\eta^2 = 0.087$. A post-hoc test via Tukey HSD showed a significant difference between the effect of *low* and *high* accuracy ($p < 0.01$). This means when the simulated system accuracy was *low*, people perceived the system's accuracy higher than when the simulated accuracy was *high*.

The ANOVA also detected a significant effect on error in implicit perceived accuracy due to change in explanation type with $F(2, 54) = 12.29$, $p < 0.001$, and an effect size of $\eta^2 = 0.285$. The post-hoc test via Tukey HSD found *weak* explanations to have significantly lower error in implicit accuracy as compared to both *strong* ($p < 0.01$) and *none* explanation types ($p < 0.001$). This result shows strong evidence that *weak* explanations caused underestimation of system accuracy.

No interaction effects were detected between explanation type and simulated accuracy.

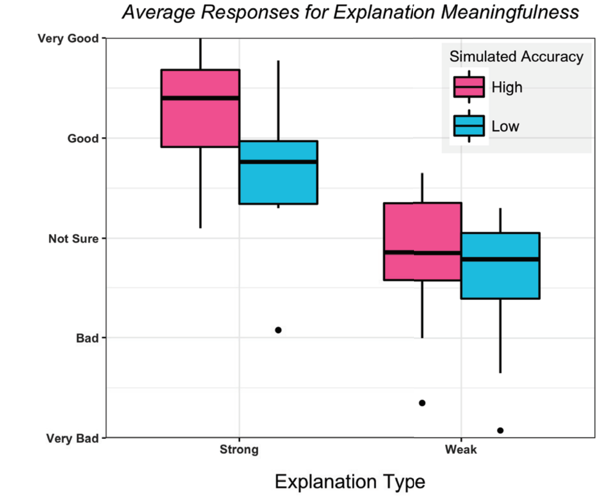


Figure 4: Average responses for explanation meaningfulness from the question "How well does the highlighted area explain the computer's answer?" in the review task on a five-point Likert scale.

Explicit Perceived Accuracy The measures for explicit perceived accuracy was the direct numeric estimation from participants at the end of the study. Figure 3b shows the results from explicit perceived accuracy. The results are similar to the findings from the implicit accuracy measure from the prediction task.

For *high* simulated system accuracy, the results exhibit near accurate or slight under estimation of the actual accuracy. On the other hand with low simulated system accuracy, overestimation was more often observed for *none* and *strong* explanations (Figure 3b). A two-way independent factorial ANOVA found simulated system accuracy to have a significant effect on explicit error with $F(1, 54) = 5.31, p < 0.05$, and $\eta^2 = 0.077$. The post-hoc test via Tukey HSD showed a significant difference between the effect of *low* and *high* accuracy ($p < 0.05$). Participants estimated higher accuracy when the simulated accuracy was *low*, and they estimated lower accuracy when the actual simulated accuracy was *high*.

A two-way ANOVA test result also shows a significant effect on explicit error due to change in explanation type with $F(2, 54) = 4.48$, $p < 0.05$, and $\eta^2 = 0.130$. A post-hoc Tukey HSD test showed accuracy was estimated significantly lower in the *weak* than the *none* group ($p < 0.05$).

The statistical analysis indicated no evidence of an interaction effect between explanation type and simulated accuracy for explicit error.

Ratings for Explanation Meaningfulness

In the conditions that included explanations, the review task included questions asking participants to rate the perceived quality of the explanations. Figure 4 shows the distribution of responses across conditions with explanations. We analyzed user responses for differences between these conditions (*strong* and *weak* explanations) and the simulated accuracy using a two-way ANOVA. The test showed a significant main effect for explanation type, showing *strong* explanations were judged as more meaningful than *weak* explanations, with $F(1, 35) = 35.37$ and $p < 0.001$. This was not unexpected since the dataset was designed with these differences for the sake of the experimental conditions, but the results add further verification for the clear differences between our *strong* and *weak* explanations.

While the results from a two-way ANOVA test did not show a significant effect at $\alpha = 0.05$ for simulated accuracy on meaningfulness rating, the results are close to significant with $F(1, 35) = 3.27$ and $p = 0.08$. This observation might suggest weak evidence that observed system accuracy could potentially influence perceived meaningfulness in explanations. However, we emphasize that the current results do not support this claim, and further study would be needed to directly investigate this hypothesis.

The statistical analysis did not yield evidence of an interaction effect between explanation type and simulated accuracy for these explanation ratings, with $F(1, 35) = 0.86$ and $p = 0.36$ (NS).

Discussion

In this section, we will discuss the results of this paper and how they are supporting our research questions and hypotheses. We also discussed the limitations and future directions of this study.

Results Interpretation

In this research, we aimed to study the influences of explanations on user perception of accuracy for an observed classification system. Without explanations, the only factor users have available to determine system capabilities and build trust on it is the accuracy of its outputs. Even then, many people are reluctant to accept output when they do not understand how or why it was generated.

In our experiment, we not only studied the presence and absence of explanations in an image classification scenario, but we also accounted for how explanation meaningfulness, in terms of alignment with human rationale and judgment, can affect users understanding of the system accuracy and performance. The results demonstrate that differences in explanation human-understood meaningfulness significantly affected user perception of system accuracy. The results provide evidence in support of the hypothesis that non-meaningful explanations can reduce perceived accuracy. That is, compared to *strong* or *no* explanations, participants estimated system accuracy to be significantly lower after using the system with *weak* explanations that do not align with natural human logic (see Figures 3b and 3a). This effect held true across both levels of simulated system accuracy. The significant difference between *strong* and *weak* explanations highlights that people tended not to trust what they did not understand. The results of the prediction task suggest that without human-meaningful explanations, participants expect a greater number of failures than observed in the past. This implies that understanding of processing logic is more important for user trust than the history of observed results alone.

We must note, however, that having weak or non-meaningful explanations, as studied in our experiment, does not necessarily mean that the explanations themselves are wrong or bad with respect to accurately representing a model. Regions of an image that humans do not consider meaningful might be the most relevant for the computational model if they accurately present the most important regions for the classification algorithm. In other words, explanation veracity with respect to the model is separate from human meaningfulness with respect to human rationale. While users might have a hard time making sense of classification based on, for example, a patch of the grass in the background or a indistinct patch of fur, it is certainly possible that the model could use such features to make its decision. From this stance, the results further demonstrate the value of and need for human-interpretable explanations. The results provide strong evidence that user understanding of computational processing can influence trust or perception of the accuracy of machine outputs. Users consider how a system draws a conclusions in order to best make decisions on whether to use and rely on its output for real tasks, and machine rationale that does not make sense to a person can cause users to systematically perceive accuracy as lower than observed over a series of observed cases.

On the other hand, the opposite effect was not observed for human-meaningful explanations. That is, the study did not contribute evidence for the hypothesis that the addition of meaningful explanations improves perception of accuracy; levels of perceived accuracy were relatively similar for the *none* and *strong* conditions. Accuracy estimations were relatively accurate in the baseline *none* conditions, and it seems the addition of explanations did not cause participants to overestimate the machine's abilities. It would be considered problematic if participants trusted in the system's accuracy more than they should, but it is important to study the possibility of such issues for different types of systems and with different explanation formats.

Additionally, aside from the explanation types, the results show a significant difference between *high* with *low* simulated accuracy conditions for perceived accuracy. As is evident in Figures 3a and 3b, accuracy was generally estimated higher for the *low* accuracy conditions and often estimated as lower for the *high* accuracy conditions. This may be related to expectations of accuracy for computational systems, where one possibility is that participants might have thought that the actual 70% in the *low* conditions was lower than researchers would use for testing, so they might have estimated higher. For the high accuracy cases, since the simulated accuracy was held at 95%, and the maximum possible accuracy is 100%, it maybe have been a type of a ceiling effect since the allowable range had more possible values below the true observed percentage.

Implications for Explanation Interfaces

Further research is needed to collect evidence about how explanation meaningfulness affects perception of system accuracy and reliability for other types of applications and algorithms, but the study outcomes using the simplistic image classification scenario provide clear results for the general effect. Assuming the same effect observed in our scenario apply to other cases, the findings from this experiment are important for designers who are interested in providing explanations for the purpose of improving trust in the system. If an explanation is not easily understandable or meaningful to the end users, the addition of explanations could have the opposite effect and actually reduce the user's trust. This highlights the importance of researchers and system designers considering the quality and format of explanations.

Determining an appropriate design for an explanation can be challenging, particularly for systems meant to support more complex tasks or those that use advanced or deep models. Designers have near limitless options for crafting explanations from different formats, styles, and levels of detail. Even assuming an understandable explanation format and sound machine reasoning, providing too much detail about the model or computational processing could potentially confuse or overwhelm users, e.g., (Kizilcec 2016). In many cases, it may be hard to know ahead of time how well users will be able to make sense of any given design. This fact motivates the need for iterative design and user testing to help identify and resolve issues in explanation design.

User interpretations of what makes explanations meaningful may also depend on a user's backgrounds and levels of knowledge, which can vary greatly. It may therefore be helpful to consider dynamic levels or styles of explanation that are customized based on user meta-data or even interaction behavior. The range of design possibilities for explanation design and current limitations in knowledge opens a number of opportunities for future research.

Limitations and Opportunities for Future Work

Though the experiment contributed novel findings about the impact of *strong* and *weak* meaningful explanations on perceived accuracy, the specifics of any experimental design introduces limitations for interpretations of the results. For instance, the experiment used a hypothetical interpretable classifier to maintain experimental control over both the classification accuracy and explanation quality. Real machine learning systems and data analysis scenarios are often much more complex and may have greater variation or noise in explanation output for a given dataset. Therefore, it will be important to conduct similar evaluations using systems with other tasks and using actual, genuine models. Also along these lines, it would be useful to study systems based on other data types (e.g., video or text) or to involve more advanced decision-making tasks. Studying more complex situations may make it more difficult to assess differences in measures such as trust, task performance, and perceived accuracy, but it may also allow opportunities to observe more nuanced behaviors and to learn more about the development of mental models.

Similarly, while our work focuses on users' perception of accuracy using a system that supports image classification and bounding boxes as local explanations, it would be valuable to investigate other explanation types such as natural language explanations, analytical explanations based on model metrics, or global explanations about the model as a whole (e.g., a decision tree). We expect that the nature of the explanation could influence the user's understanding of an intelligent system, and it would be interesting to assess whether results similar to those of this study might be observed with alternative presentations.

Finally, in addition to explanation meaningfulness, it would be valuable to investigate potential relationships between meaningfulness and user understanding of computational functionality in affecting perceived accuracy. Perhaps if users can develop an understanding of a system's underlying models, they might be able to distinguish between explanatory elements that are meaningful from the standpoint of normal human logic and those that are accurate with respect to computational processes. We are interested in studying whether user understanding of the model might allow users to more accurately assess system accuracy even when explanations are not obviously human-meaningful.

Conclusion

With the recent advances in machine learning, artificial intelligence, and deep learning, it is more important than ever to understand methods for explainable machine models. In this research, we studied how visual explanations influence user trust and perception of the system accuracy for participants experiencing different levels of system accuracy. The results show that participants significantly underestimated the system's accuracy when it provided weak, less-meaningful explanations that did not align with users' logic. Therefore, for intelligent systems with explainable interfaces, the results demonstrate that users are less likely to accurately judge the accuracy of algorithms that do not operate based on human-understandable rationale. It is not sufficient for designers to incorporate explanations into intelligent systems without considering the implications and ethics of user interpretation and trust. Further research is needed to understand how other properties and variations of explanations affect user perceptions and behaviors.

Acknowledgments

This work was supported by the DARPA Explainable Artificial Intelligence (XAI) Program under contract number N66001-17-2-4032.

References

Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B. Y.; and Kankanhalli, M. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 582. ACM.

Adadi, A., and Berrada, M. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access* 6:52138–52160.

Akata, Z.; Stumpf, S.; Kieseberg, P.; and Holzinger10, A. 2018. Explainable ai: The new 42? In *Machine Learning and Knowledge Extraction: Second IFIP TC 5, TC 8/WG 8.4, 8.9, TC 12/WG 12.9 International Cross-Domain Conference, CD-MAKE 2018, Hamburg, Germany, August 27–30, 2018, Proceedings*, 295. Springer.

Alsallakh, B.; Hanbury, A.; Hauser, H.; Miksch, S.; and Rauber, A. 2014. Visual methods for analyzing probabilistic classification data. *IEEE transactions on visualization and computer graphics* 20(12):1703–1712.

Amershi, S.; Cakmak, M.; Knox, W. B.; and Kulesza, T. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35(4):105–120.

Core, M. G.; Lane, H. C.; Van Lent, M.; Gomboc, D.; Solomon, S.; and Rosenberg, M. 2006. Building explainable artificial intelligence systems. In *AAAI*, 1766–1773.

Cramer, H.; Evers, V.; Ramlal, S.; Van Someren, M.; Rutledge, L.; Stash, N.; Aroyo, L.; and Wielinga, B. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18(5):455.

Dodge, J.; Penney, S.; Anderson, A.; and Burnett, M. 2018. What should be in an xai explanation? what ift reveals. *In IUI Workshops*.

Doran, D.; Schulz, S.; and Besold, T. 2018. What does explainable ai really mean? a new conceptualization of perspectives. In *CEUR Workshop Proceedings*, volume 2071.

Došilović, F. K.; Brčić, M.; and Hlupić, N. 2018. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 0210–0215. IEEE.

Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2011. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1):121–127.

Goodall, J. R.; Ragan, E. D.; Steed, C. A.; Reed, J. W.; Richardson, G. D.; Huffer, K. M.; Bridges, R. A.; and Laska, J. A. 2018. Situ: Identifying and explaining suspicious behavior in networks. *IEEE transactions on visualization and computer graphics* 25(1):204–214.

Hendricks, L. A.; Hu, R.; Darrell, T.; and Akata, Z. 2018. Grounding visual explanations. *arXiv preprint arXiv:1807.09685*.

Hoffman, R. R.; Johnson, M.; Bradshaw, J. M.; and Underbrink, A. 2013. Trust in automation. *IEEE Intelligent Systems* 28(1):84–88.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Holliday, D.; Wilson, S.; and Stumpf, S. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 164–168. ACM.

Kizilcec, R. F. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. ACM.

Kulesza, T.; Stumpf, S.; Burnett, M.; Yang, S.; Kwan, I.; and Wong, W.-K. 2013. Too much, too little, or just right? ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*, 3–10. IEEE.

Lim, B. Y.; Dey, A. K.; and Avrahami, D. 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*, 2119–2128. ACM.

Miller, T.; Howe, P.; and Sonenberg, L. 2017. Explainable ai: Beware of inmates running the asylum. In *IJCAI-17 Workshop on Explainable AI (XAI)*, volume 36.

Mohseni, S., and Ragan, E. D. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*.

Parasuraman, R., and Riley, V. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human factors* 39(2):230–253.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 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*, 1135–1144. ACM.

Roy, C.; Shanbhag, M.; Nourani, M.; Rahman, T.; Kabir, S.; Gogate, V.; Ruozzi, N.; and Ragan, E. D. 2019. Explainable activity recognition in videos. In *IUI Workshops*.

Samek, W.; Binder, A.; Montavon, G.; Lapuschkin, S.; and Müller, K.-R. 2017. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems* 28(11):2660–2673.

Samek, W.; Wiegand, T.; and Müller, K.-R. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Siau, K., and Wang, W. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31(2):47–53.

Van Lent, M.; Fisher, W.; and Mancuso, M. 2004. An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence*, 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Yin, M.; Vaughan, J. W.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 2019*.

Zhang, Q.-s., and Zhu, S.-C. 2018. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering* 19(1):27–39.

Zhu, J.; Liapis, A.; Risi, S.; Bidarra, R.; and Youngblood, G. M. 2017. Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation. *In 2018 IEEE Conference on Computational Intelligence and Games (CIG) (pp. 1-8). IEEE.*