Human-in-the-Loop User Feedback Affects Perceived Accuracy and Trust, but Task Subjectivity Matters

DONALD R. HONEYCUTT, University of Florida, USA
MAHSAN NOURANI, Northeastern University, USA
ERIC D. RAGAN, University of Florida, USA

While machine learning can produce complex models beyond those that a human could produce manually, incorporating human input can often improve intelligent system performance beyond purely data-driven models. These human-in-the-loop approaches can also be used to dynamically update a model in the presence of shifting data or system goals. While this feedback could come from system designers or domain experts, in many cases, the end users who regularly use the system will naturally develop an understanding of its flaws and desire the ability to change the system's behavior based on their knowledge. While soliciting feedback from end users can result in significant model improvement over time, introducing these feedback techniques can also affect several human factors—such as trust or perception of system accuracy—that are not yet fully understood and have different effects reported in the existing literature. Therefore, we sought to build on the existing research to further explore how the act of providing feedback can affect user understanding of an intelligent system and its accuracy in different contexts. We present three controlled experiments that study the effects of interactive feedback collections on user impressions in domains with objective and subjective feedback. The results show that in a context where there is an objectively correct answer, providing human-in-the-loop feedback lowered both participants' trust in the system and their perception of system accuracy, regardless of whether the system accuracy improved in response to their feedback. However, when the feedback being provided involved subjective opinion, no such negative bias was observed. Furthermore, in the objective context, participants distrusted the system over time, whereas participants in the subjective context mistrusted the system over time. These results highlight the importance of considering the effects of allowing different types of end-user feedback on user trust when designing intelligent systems.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; Interactive systems and tools.

Additional Key Words and Phrases: trust in automation, human-in-the-loop, machine learning

ACM Reference Format:
Donald R. Honeycutt, Mahsan Nourani, and Eric D. Ragan. 2026. Human-in-the-Loop User Feedback Affects Perceived Accuracy and Trust, but Task Subjectivity Matters. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2026), 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

There are many reasons designers of intelligent systems may want to incorporate human input into the training of their machine learning (ML) models. Domain experts may have pre-existing understanding of complex application areas that may be usable to address the unreducible error present when learning from data alone. This process may be more efficient than purely data-driven ML, particularly when there does not already exist sufficient labeled training data and

Authors' addresses: Donald R. Honeycutt, dhoneycutt@ufl.edu, University of Florida, Gainesville, Florida, USA; Mahsan Nourani, m.nourani@northeastern.edu, Northeastern University, Boston, Massachusetts, USA; Eric D. Ragan, eragan@ufl.edu, University of Florida, Gainesville, Florida, USA.

obtaining such labels is prohibitively expensive. Additionally, human modification can be used to tune model behavior to correct errors and patterns learned from the training data that do not generalize to patterns in the data as a whole. ML systems can support this by utilizing *human-in-the-loop* (HITL) approaches that modify the model to be closer to the feedback provided by a human. The most basic approach that is prevalent in the existing body of work—new annotation of unlabeled data—uses traditional ML on an updated set of training data to make predictions that are closer to the user's labels, often in the form of corrective feedback where the user identifies and corrects instances where the model makes mistakes or labels data that is identified as statistically likely to result in the greatest model improvement [9, 36, 40].

Feedback is of course not limited to mere annotation, which may be an inefficient method of achieving specific behavioral changes within a model. Direct modification of model behavior without reliance on retraining with new labeled data allows for a greater level of control in which the user can mold the models behavior to their liking. While directly modifying feature weights is a simple approach that can be effective in some scenarios, this requires either a simple and inherently interpretable model or a strong understanding of the ML model in use and how to modify it. Explainability techniques that provide information about the model's behavior can not only help the user understand the model they are trying to modify but can also serve as a method of giving a user intuitive control over model behavior without explicit modification of model parameters. This approach—*explanatory interactive learning*—has the user provide modifications to the explanations to show how they think the model should behave differently [19, 25, 33, 42, 67]; by producing explanations that more closely adhere to the provided modifications, the updated model should more closely represent the mental model of the user.

Frequently, the HITL approaches either involve system developers for development and debugging [71] or independent workers on crowd-sourcing platforms [41]. By taking advantage of end-users' periodical feedback upon noticing errors, these models can stay updated in the presence of shifting data or changing goals [15, 18, 74]. Systems can also update over time by implicitly capturing user behaviors, which is a technique commonly used in recommender systems [44, 61]. While this feedback is not provided explicitly, users can still observe the system directly reacting in response to their actions, decisions, and feedback. Furthermore, end users of intelligent systems may want the ability to correct observed model errors. When engaged with the outcomes of a system, many users desire the ability to influence those outcomes by providing feedback beyond simple error correction [64].

While HITL systems can have improved model accuracy and provide users control over the systems they rely on, there may also be unexplored consequences to allowing end users to provide feedback. For instance, Van den Bos et al. [70] observed that when interacting with human teams, the ability to provide feedback has been observed to have a positive effect on the perceived fairness of team decisions. In their study, users who felt their feedback was considered reported higher levels of trust in the decision-making process and were more committed that the correct decision was made. They also observed the inverse effect, with a decrease in trust in the team if feedback was provided but ignored [30]. Since providing feedback to an automated decision-making system is similar to providing feedback to a human-based decision making system, it is possible that similar effects could be observed in HITL systems.

If providing feedback does affect user trust, it could lead to people misusing the systems they provide feedback to. When experiencing a higher level of trust than is appropriate based on the system performance, users may over rely on the system. On the other hand, having a lower level of trust could result in not using the system at all [38, 53]. Therefore, it is important to understand how providing feedback to an intelligent system affects trust so that it can be accounted for when designing HITL systems. Prior research has shown that the presence of HITL elements in intelligent systems can have both positive and negative effects on user perception and trust. Users are consistently shown to have both stated preferences towards systems with HITL elements [29, 54, 64] and are more willing to use and rely on HITL systems than non-HITL

intelligent systems [12, 54]. However, such mechanisms can result in increased cognitive load that causes decreased engagement with the system [14, 29]. In systems where the HITL feedback is purely correcting objective system errors, it can result in underestimation of accuracy and distrust [26, 27]. While prior research has clearly established the diametric nature of HITL feedback, further research is needed to understand the reasons behind the differences in observed effects in different contexts. Understanding the effect that providing end users with HITL feedback mechanisms will have in a given system and context is important to ensure it does not have an adverse effect on their perception and use of the system.

In this paper, we present the results of three experiments that build on our prior study [26] that showed participants who provided HITL feedback underestimated system accuracy and had reduced trust compared to those who did not. These experiments aim to explore the effects of providing HITL feedback in both subjective and objective contexts. This paper is a direct extension of the previously published conference paper on the previous study. In this extension, the previously published experiment from [26] was expanded in scope with new experimental conditions to allow a deeper analysis; the expanded study is presented as Experiment 1 of this paper. In addition, Experiments 2 and 3 are entirely new.

Experiment 1 seeks to isolate the difference between the level of feedback interactivity and the belief that the system is updating based on user feedback. We used a simulated object-detection system—the same system as used in prior work [26]—that tasked users with providing interactive feedback via adjusting bounding boxes for detected objects in images. Additionally, this experiment also controlled for whether participants were led to believe that their feedback was resulting in dynamic system changes throughout the task—despite no such updates being made. We found that belief that the system was updating had no significant effect on perception of system accuracy or user trust. However, those who provided interactive feedback perceived the system to be significantly less accurate than those who only provided basic feedback about whether the classification was accurate or not. Similarly, trust in the system was lower among those who provided interactive feedback compared to participants who did not. Regardless of condition, all participants found the system to become less accurate over time, despite the true accuracy of the system remaining constant.

Our second and third studies recreate the experimental design of the first study, but this time in the context of text classification instead of object detection. In this context, the interactive feedback comes from adjusting the list of words which are the most relevant towards identifying the text sample's topic. Where adjusting bounding boxes in images has an objectively correct answer—either there is an object in a location or there is not—the most relevant words to identify the topic of a paragraph is somewhat subjective. We hypothesized that this difference would re-contextualize the act of providing interactive feedback from having to correct system errors to making the system's model behave more closely to the user's mental model. Since shaping system behavior to be more similar to one's own mental model is a reason users have provided for preferring HITL systems in prior research [54, 56], we hypothesized that this re-contextualization would help alleviate the negative biases observed in the image classification study. Unlike the image classification context, we did not observe any effects on trust or perception of accuracy based on level of feedback interactivity. Also contrary to the results in the image classification context, participants in all conditions perceived the system to increase in accuracy over time, even though no such accuracy change occurred.

Additionally, in this paper we discuss the existing literature showing both positive and negative effects of HITL feedback on user perception and the differences in system and task context present in each of those studies, particularly as it relates to feedback subjectivity. This discussion—in combination with the results of the studies presented in this paper—contributes an improved understanding of why different studies of the effects of providing HITL feedback have produced contrasting results.

2 RELATED WORK

The presented research builds on prior work from the perspectives of HITL ML and trust in artificial intelligence.

2.1 Human-in-the-Loop Machine Learning

While ML can be used to train models based purely on data without direct human guidance, there are many scenarios where incorporating human feedback is beneficial. In many cases, this feedback is simply having a human annotate new data to be incorporated into the model. While data annotation can be viewed as merely a requirement to obtain the large pool of labeled data necessary to train a traditional supervised learning model, data annotation also serves as an effective medium for people to transfer their knowledge into a complex modeling process. This approach allows people who may be completely inexperienced with ML to provide feedback to update a model using their understanding of the domain. As data annotation can be time consuming and expensive, much research has been performed on how to optimize the selection of which data to label. *Relevance feedback* is a HITL method where a human reviews the pool of unlabeled data alongside the current model's predictions on that data, choosing when to provide new labels to the system based on their own intuition [68]. As human intuition does not result in optimal selections of what data to label in many domains [6], another approach is to choose instances to add to the training set using objective metrics based on the model. *Active learning* selects relevant instances to show to a human—referred to as an *oracle*—based on which unlabeled data are most likely to represent information missing in the current version of the model [10]. While theoretical active learning research treats the oracle as merely being a way to obtain the true labels for selected data, in practice, active learning models need to account for the fact that the oracle is a human and therefore not infallible [60].

While data annotation and error correction are a simple medium for providing HITL feedback to the model, it may be more efficient to give richer feedback that allows for more direct modification of model parameters, allowing the oracle to mold the ML system to align with their own mental model of the task and data set. The most direct method is to allow for direct control of model parameters, such as features and their weights [7] or the rules that are used to make decisions within the model [75]. Being able to control model parameters in this way has been found to be useful for debugging models [33]. However, while direct parameter modification can be effective, the complex nature of black-box models often causes those techniques to be ineffective or inefficient, particularly among users who are inexperienced in ML. When the level of user control is too complex for the user, it can result in an increase in cognitive load and lower engagement with providing feedback than if they had less control [14, 29, 56]. A potential approach to allow for rich user control through an understandable medium is to incorporate explainable AI techniques and have feedback be based on those explanations. Explanatory interactive learning has the oracle not only provide the appropriate label for the data point but also provides an explanation of the current model prediction and asks the oracle to correct the reasoning in the explanations [67]. This helps avoid situations where the model has a flaw that happens to result in the correct prediction by chance. Examples of explanation types that can be used for interactive feedback include highlighting relevant words for text classification [32, 33, 78], ranking features in tabular data [7], and bounding boxes for image classification and object detection tasks [17]. While this higher level of control over the model may not be desirable in all applications, Holzinger et al. [25] showed that human-machine teaming can sometimes result in a closer to optimal model than ML alone.

Another approach to creating interactive explanations to support human feedback is the use of interpretable surrogate models to generate approximate explanations of model behavior that are applicable to any model architecture [8, 59, 69, 79]. The main challenge in using surrogate explainer models for HITL systems is in updating the original model to be closer to the explanations provided by the oracle when the explanations do not directly correspond to the format of model inputs.

A simple strategy is to use the updated surrogate model to generate a robust set of counterexamples to the data set to retrain the original model with new data that are representative of the changes made by the oracle [55, 75]. By adding these counterexamples to the training data, both performance and explanation quality can be improved when compared to regular active learning [67].

While the person providing feedback is not necessarily the end user for many HITL systems, there are advantages to bringing end users into the loop. Stumpf et al. [64] found that users of intelligent systems largely want to provide feedback to systems they are using, particularly when it gives them a feeling of being able to control some aspect of the model. Similarly, people are more likely to use an imperfect intelligent system when they have the ability to correct its errors [12]. Additionally, end users may notice when an already deployed system begins to falter. Even if a model was very accurate at the initial time of training, the training data may become less representative of the actual population of data as trends shift over time. This is a phenomenon known as concept drift [80]. A HITL approach to dealing with this problem is known as *incremental learning*, where the model periodically obtains labels as they become available to update the model while it is in use [18]. These techniques have been shown to effectively address the problem of concept drift in ML systems [15, 74].

Providing input has been shown to affect trust and perception of fairness in the field of psychology. In decision-making teams, people were observed to place more trust in a team-leader who actively considered their input, and they were also more confident that the correct decision was made after the fact [30]. A similar effect was observed in procedural decision making systems, with people having a higher level of trust and perception of fairness in a decision-making system that they were able to give input to [70]. An interesting result from both of these studies was that providing feedback had a negative effect on trust if the feedback was ignored. Since feedback affects interpersonal trust by improving trust if feedback is considered and decreasing trust if it is ignored, a similar effect may be observed in human-computer interactions. When users of HITL systems cannot directly see how their feedback has changed the model, they can become frustrated and be unsure of what to do next [35, 64]. This is not only an issue of a lack of detection of model changes, but also that users may expect large changes to happen faster than is feasible for a given architecture. Providing information about how a person's contributions are being used has been shown to increase participation in social groups [57], which may motivate similar information being provided to explain how feedback is being used in HITL systems to both explain effects and improve labeling quality [1]. Additionally, when users are treated as an active learning oracle and are asked to correct objective system errors, it can result in them distrusting the system [26, 27]. However, HITL features that allow direct control over model parameters to better match personal preferences can result in increased satisfaction and willingness to use the system [12, 54, 56]. While many papers have shown that the presence of HITL feedback can have both positive and negative effects on user trust and perception of intelligent systems, they use a variety of contexts and the reasons behind the differing results have not been sufficiently established.

2.2 Trust in Artificial Intelligence

User trust in artificial intelligence systems has been studied for many years and is of value since it is directly associated with usage and reliance [53, 62]. As a result, users need to place an appropriate amount of trust in a system based on its performance in different contexts. Reliance and trust in automated systems are not binary (i.e., to trust or not) and are generally more complex [38, 39]. Desired behavior is for a user to examine a system's outputs and decide whether to rely on the system based on the accuracy of results [24]. This behavior has been observed to be more prevalent among users who are domain experts than novice users [50]. Sometimes, however, users might trust a system completely without checking the outcomes, i.e., over-reliance or *automation bias* [20]. This situation can be caused by a user's lack of

confidence or when the system seems more intelligent than they are based on their initial preconceptions [24, 38, 48]. In contrasting scenarios, *mistrust* [53] and *distrust* [38] can cause users to rely more on themselves and avoid using a system they perceive to be more flawed than it is [11]. Both of these situations can be dangerous, especially for systems with critical tasks where decisions can be fatal. For example, wrong decisions in criminal forecast systems can wrongfully convict an innocent person [3] and miscalibrated trust in an AI healthcare system can result in misdiagnosis [73]. The required level of trust and the trust calibration process also differ greatly by task; cooperative tasks such as decision support systems require the user to calibrate how often they utilize the system on a case-by-case basis, whereas delegative tasks such as autonomous driving systems require the user to choose whether or not to trust the system as a whole [72].

To help calibrate users' trust to an appropriate level and provide more information to aid them in their decision-making process, researchers have explored the use of explainability in artificial intelligence systems [13, 58]. Perceived transparency due to system explanations contributes to an increase in user trust [63, 65] and can encourage usage of an artificially intelligent system [22]. Studies of HITL paradigms have shown explainability can help users understand and build trust in the algorithms in order to provide proper feedback and annotations [19, 66]. Explanations of system behavior improve users' mental models of the system's inner workings [34], which is strongly linked with developing an appropriate level of trust in a system [23]. However, if the provided explanations are insufficiently meaningful to the user it can result in an underestimation of system accuracy [49]. Humans are additionally prone to several cognitive biases when interacting with intelligent systems [47, 49]; system explanations can both help mitigate or exacerbate the effects of such cognitive biases [4].

Researchers use different methods to measure trust and reliability in ML and artificial intelligence systems. Trust in automation is a multi-faceted concept, defined by Lee and See [38] as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." To capture this attitude, there has been much research into survey scales that measure different aspects of trust [5, 23, 28, 43]. For example, Madsen and Gregor [43] break down trust in automation into five constructs: perceived reliability, perceived technical competence, perceived understandability, faith, and personal attachment. In addition to measuring trust subjectively through questionnaires, there are many behavioral metrics that are highly correlated with trust. For example, some researchers utilize user's agreement with the system outputs as a measure of reliance and trust; specifically, identifying when the user agrees with the system outputs that are not correct [52]. Yu et al. [77] propose a *reliance rate* based on the number of times the users agreed with the system answers out of all their decisions. In recent work, Yin et al. [76] found that trust is directly affected by user's estimations of the system's accuracy, where underestimation of accuracy can cause mistrust in the system, and vice versa. As a result, a user's estimated or observed accuracy can be used as an indirect measurement for user trust, and we use these methods in the study reported in this paper.

3 EXPERIMENT 1: OBJECT DETECTION IN IMAGES

In this section, we discuss the first experiment based around understanding the differences in user trust and perception of system accuracy among users of HITL systems with interactive feedback mechanisms. We note that part of this experiment was previously published in [26]. We have since expanded the study design by adding additional experimental conditions that will be explained further later in this section. The expanded study is reported here with new analyses and results.

3.0.1 Research Objectives. With the goal of understanding the relationship between belief that the system is dynamically updating based on user feedback and the level of feedback interactivity, we present the following research questions:

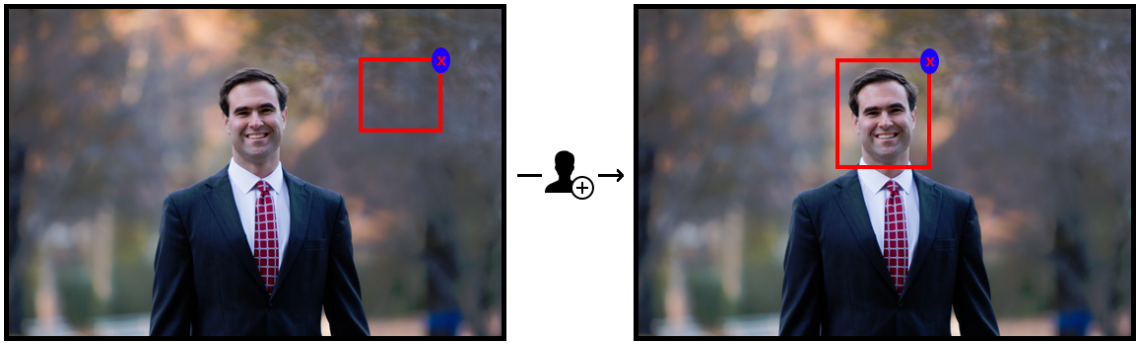


Fig. 1. In the *interactive feedback* conditions of Experiment 1, participants could delete existing bounding boxes or click-and-drag to create new ones. In this example, the left image shows a system error, and the right shows a version after interactive correction.[1]

RQ1: Does belief that user feedback is being used to dynamically update system behavior affect user trust or perceived system accuracy?

RQ2: Does providing more interactive feedback result in a difference in user trust or perceived system accuracy when compared to only providing accuracy-based feedback?

To address these research questions, we designed a controlled experiment using a simulated object detection system with varying levels of feedback interactivity and a representation of feedback usage. Based on the results of prior research [26], we hypothesized that the more interactive feedback method would result in a more negative perception of system accuracy and lower levels of trust due to an increased saliency of system errors by spending more time correcting them. We also hypothesized that no effect would be observed based on the participants being told the system was updating based on their feedback as the provided corrections were of an objective nature rather than being about making the model behave closer to the user's own mental model. The original study design from [26] had change in accuracy as an independent variable, where the system accuracy either improved, stayed the same, or degraded across the three rounds of the task. As the change in accuracy condition did not result in any significant findings, we chose to eliminate this variable to have a more focused study design. All study conditions used a constant 80% system accuracy across all three rounds of the task.

3.1 Method

In this sub-section, we present details of our experimental design and study procedure.

3.1.1 System Design. We note that the system design for Experiment 1 had no changes to the system design from [26]. We provided participants with a series of images with classifications from a simulated model with detection of human faces as the classification goal. To allow for interactive feedback, the simulated system outputs include bounding boxes over the location of each detected human face. For conditions that involved interactive feedback, bounding boxes could be deleted by clicking an X attached to each bounding box, and new bounding boxes could be drawn by clicking and dragging on the image in the interface as shown in Figure 1. While participants were told the classifications came from an artificially intelligent system, the classifications were actually hand-crafted for experimental control purposes.

[1]Image from "Josh McMahon Portraits - 2517" by John Trainor (used under CC BY 2.0) with annotations added by the authors. License available at https://creativecommons.org/licenses/by/2.0/

Prior research showing user preferences for systems with HITL user control have the limitation that the version of the system that allows for user control could be preferred simply because feedback resulted in the system performing better, not necessarily because of the presence of HITL feedback mechanisms [29, 54, 56]. To avoid this potential confound, we chose to use a system that deceives participants to believe that it is updating based on their feedback when in reality all participants viewed the same system outputs regardless of their feedback.

Since this study uses estimation of system accuracy as one of the main metrics, we needed task rounds that were sufficiently long to ensure participants saw enough system outputs to both make an informed estimation and not just remember the exact percentage of system outputs that were correct. We additionally needed enough rounds of the task to be able to detect change over time. However, these needs had to be balanced by the time constraints of an online user study where participants may lose engagement and provide worse feedback over time. In consideration of balancing these tradeoffs, we selected a task consisting of reviewing three rounds of images, with 30 images in each round. The images used in our simulated model were taken from the Open Images dataset [31, 37] with our own manually generated annotations. Each round of 30 images contained 20 pictures of people, with the remaining 10 images containing things such as animals or empty scenery. For images where we chose to simulate system errors, we used a roughly equivalent mix of false positives (bounding boxes placed on objects that were not human faces) and false negatives (unidentified human faces).

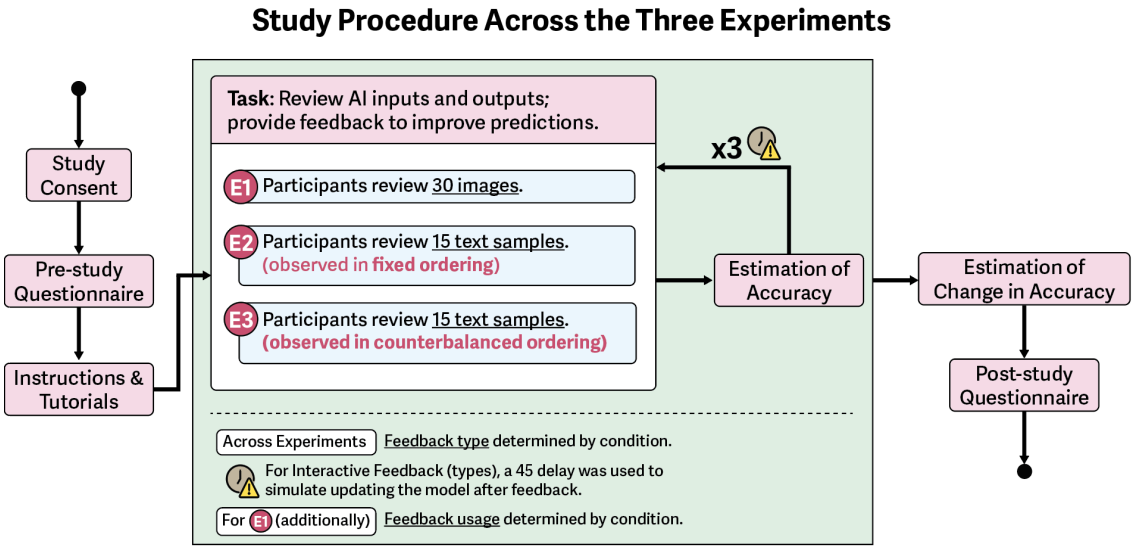


Fig. 2. An overview of the experimental procedure shared across the three experiments. Each participant interacts with only one experiment. After each round of the task, participants rate system accuracy to capture how perception of accuracy changes over time. This is repeated for three task rounds before moving on to post-study questionnaires.

3.1.2 Experimental Design. Because our main metrics—perception of model accuracy and user trust—are based on participants' experiences with the system, we decided that each participant should only see one version of the system so as not to be biased by their experience with the previous system versions. For this reason, we used a 2x2 between-subjects design for the experiment. The first independent variable—*feedback type*—pertains to the feedback mechanism used by

the participants during the task. Those with *binary feedback* responded simply whether the image was classified correctly or not, while those with *interactive feedback* additionally corrected any errors by adding and removing the bounding boxes on top of the images as shown in Figure 1. In this context, the feedback being provided is objective—there is either a human face present in a location or there is not; there is no subjective interpretation of the correct classification. Our second independent variable—*feedback usage*—was based on the information provided to the participants about what was done with the feedback they provided. This variable did not change the nature of how participants interacted with the system at all, but rather changed only the explanation of how the system would be using their responses. For participants in the *without update* condition, they were told that their responses would be saved and sent to researchers after the task for quality control purposes. The *with update* participants were explicitly informed that the system would update in-between each of the three task rounds based on the feedback they provided before making the next set of classifications. As before, the system accuracy was actually constant and did not update based on feedback for the purposes of experimental control. To remind the participants and add to believability, a 45 second pause was present between each round with a message telling participants that the system needed time to update its parameters before they could move on. As a method of confirming that participants in *with update* conditions actually believed their feedback was being used, we additionally asked them to rate their agreement with the statement "The system correctly uses my feedback" on a 7-point Likert scale at the end of the task. The mean response to this question was 4.71 (slightly positive) with a standard deviation of 1.56, confirming that on average participants in *with update* conditions believed the system was actually updating based on their feedback.

For the conditions that mirrored conditions from [26]—*binary feedback without update* and *interactive feedback with update*—the original data was retained and no new data was collected. New participant data was collected for the new conditions—*binary feedback with update* and *interactive feedback without update*—and analysis was performed on the combined dataset from the original study and the newly collected data.

3.1.3 Participants. Participants were recruited from Amazon Mechanical Turk with a requirement for participants to have the Masters qualification, an approval rate of greater than 90%, and 500 or more prior tasks completed successfully. Participants ranged from ages 26–66 and lived in the United States at the time of study completion. To ensure the quality of participant responses, we measured the percentage of responses for which participants correctly identified whether an image corresponded to a system error or not. As a quality check, participants were not included in the results if they had less than 75% accuracy for either correct instances or system errors. Our study had a total of 111 participants, and 4 were removed based on the accuracy criteria. The remaining 107 participants consisted of 45 females and 62 males. Participants took approximately 11 minutes on average to complete the study.

3.2 Experiment 1 Results

In this section, we present the measures of the follow-up study and its empirical results. We report statistical test results along with generalized eta squared (η^2_G) for effect sizes of ANOVA tests and Cohen's d (d_s) for effect sizes of post-hoc tests.

3.2.1 User-Perceived Model Accuracy. Between each round of the task, participants estimated how accurate they thought the system was as a percentage. They were instructed to make their best estimate of how accurate the system was based on the prior round of the task without directly counting the correct and incorrect instances. Estimation of system accuracy has been shown to directly affect trust in ML models [76] and is an established method for estimating general user understanding of an intelligent system's performance and capabilities [45, 48, 52]. We performed a three-way

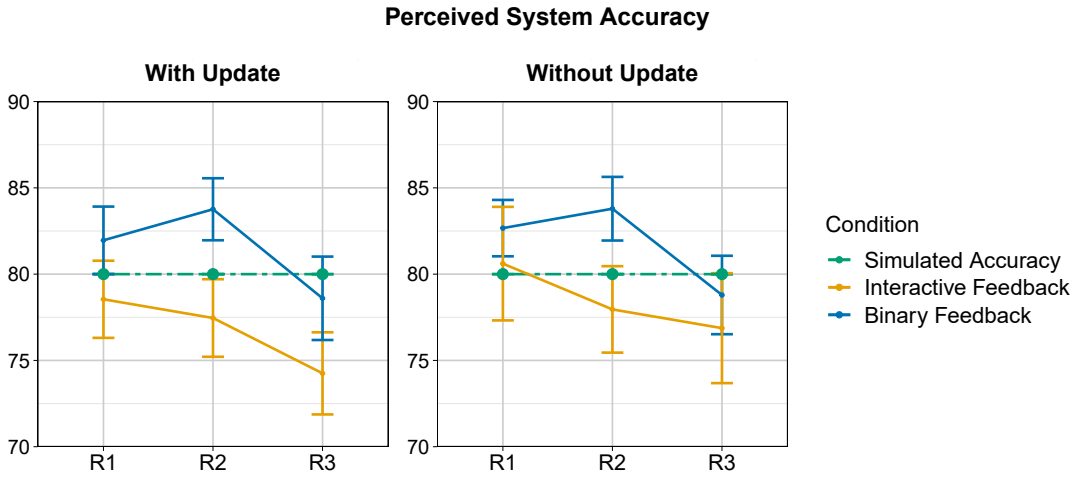


Fig. 3. Perceived system accuracy (Percentage) across the three task rounds (R1, R2, R3)

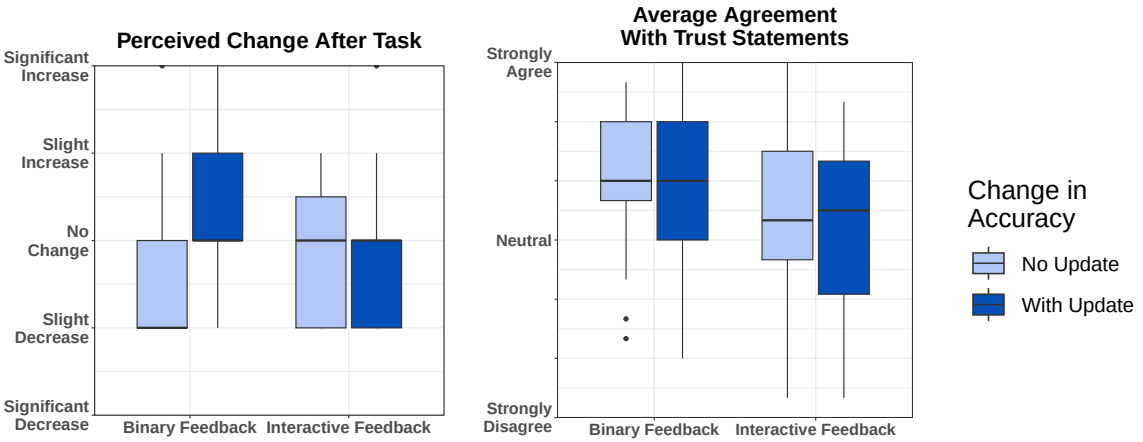
mixed-design ANOVA on the estimated accuracy, with *feedback type* and *feedback usage* as between subjects factors and task round (i.e., first, second, or third round) as a within subjects factor. The analysis showed no significant effects based on *feedback type* ($F(1, 103) = 3.71, p = 0.050$) or *feedback usage* ($F(1, 103) = 0.17, p = 0.676$). *Task round* was found to have a significant effect on estimated model accuracy with $F(2, 206) = 11.88$, $p < 0.001$, $\eta_G^2 = 0.022$. Post hoc pairwise t-tests with Bonferroni correction revealed a significant difference between the first and third round with $p < 0.050$, with the perceived accuracy in the third round being lower than the perceived accuracy in the first round. No significant interaction effects were observed between any variables. Results are shown in Figure 3.

3.2.2 Perception of Model Change. Upon completion of the task, participants rated how much they thought the system had changed across the rounds on a five point Likert scale. We performed an independent two-way factorial ANOVA on the Likert scale responses, which revealed that those who provided *interactive feedback* thought the system had changed significantly more negatively than those who provided *binary feedback*, with $F(1, 103) = 3.95$, $p < 0.050$, $\eta_G^2 = 0.037$. No significant effect was observed based on *feedback usage*, nor were any significant interaction effects. Results are shown in Figure 4a.

3.2.3 User Trust. To measure participants' trust in the system, we asked them to rate their agreement to several statements about different aspects of trust based on an adjusted version of scales proposed by Madsen and Gregor [43]. The following three statements were shown to participants:

- The system performs reliably.
- The outputs the system produces are as good as that which a highly competent person could produce.
- It is easy to follow what the system does.

We asked participants to rate their agreement with these trust statements on a seven-point Likert scale. We analyzed the aggregated responses to the three trust statements with a two-way factorial ANOVA and observed no significant results for either *feedback type* or *feedback usage*, as well as no interaction effects. Results are shown in Figure 4b.



(a) Responses to the question "Did you notice any changes to system accuracy across the rounds of the task?" on a 5-point Likert scale.

(b) Averaged responses to the three trust scale statements described in Section 3.2.3 on a 7-point Likert scale.

Fig. 4. Experiment 1 post-study questionnaire results.

3.3 Summary of Experiment 1 Findings

Our goal in this experiment was to better understand how *feedback type* and *feedback usage* contributed to the results observed in the initial study [26]. The main finding was that those who gave *interactive feedback* had a more negative perception of the system's change in accuracy over time than those who only gave *binary feedback* (**RQ2**). No significant results were observed based on the *feedback usage* condition for any of the measures (**RQ1**). The primary finding in the original work that this study was based on was that providing interactive HITL feedback lowered user trust and caused underestimation of system accuracy regardless of actual changes in system performance [26]. This result was hypothesized to be caused by an increase in memorability of system errors due to having to spend more time correcting them, resulting in participants disproportionately remembering those system errors. Our new findings support this hypothesis, as an effect was observed based on the act of providing feedback and not based on perception of that feedback being used.

In the context of this study, the outputs of the system—and consequently the feedback being provided—are of an objective nature; there are a defined number of faces in the image in defined locations. Because of this, providing feedback in this context is essentially just correcting obvious mistakes from the perspective of the participants. In prior research that found an increase in trust and usage rates for systems with HITL components over those without, the feedback being provided has involved more subjective system outputs where providing feedback is contextualized as having the ability to make the system behave closer to their own mental model or to their preferences [12, 54, 56]. This difference in context could be why the presence of interactive feedback resulted in a negative bias in this study and in [26]. Additionally, we hypothesize that this context is the cause of the overall decrease in perceived accuracy over time regardless of round. Since the observed system errors were obviously wrong with no nuance or openness of interpretation, seeing such errors resulted in increased distrust over time. In the next section, we present two research studies that explore how a similar study design plays out in a more subjective context.

4 EXPERIMENT 2: TEXT CLASSIFICATION (FIXED ORDERING)

This section reports a follow-up experiment that mirror the experimental design of the Experiment 1 study but with a text classification context instead of an image context.

4.1 Research Objectives

Although we observed a negative bias among those who provided more interactive feedback in both our original object detection study [26] and the extended Experiment 1 presented in Section 3, other research has found that users of intelligent systems prefer to use systems that allow them to provide more interactive feedback and have increased levels of trust in such systems [12, 54, 56]. One notable difference between the systems used for these studies that show a positive effect associated with providing feedback and the two experiments that observed a negative bias from providing feedback is the objectivity of the feedback being provided. In our studies that observed a negative effect, the feedback being provided corresponded to an objectively correct answer; an image region either contains an object or it does not. Providing interactive user feedback in this context is correcting objective system errors by providing the only correct answer, without nuance or subjectivity. However, the studies that observed positive effects associated with providing interactive feedback were conducted in contexts where the correctness of an output was more subjective, and the feedback being provided was more about adjusting the behavior of the model to be closer to the user's own mental model. By there not being a single objectively correct answer, the act of providing explanation-based feedback is less about having to correct system errors and more akin to making the system's model closer to the user's mental model.

With this difference in mind, we looked to create an experiment with the same experimental design as in the first experiment presented in this paper, but in a context with a more subjective task. We chose text classification—specifically working with text samples from computer science and history textbooks—as it would be familiar to the study populations to be able to provide subjective feedback without requiring any more advanced domain-specific knowledge.

With the goal of building upon the core concepts of the prior work, we aimed to design a new study that addressed similar research questions in a different context. Based on the results of [26] and Experiment 1, we eliminated the independent variable of *change in accuracy over time* and chose to focus instead on solely the presence or absence of different types of feedback. The relevant research question from Experiment 1 is as follows:

RQ2: Does providing more interactive feedback result in a difference in user trust or perceived system accuracy when compared to only providing accuracy-based feedback?

Additionally, since the integration of HITL approaches into explainable AI systems can be beneficial for improving users' mental models [2, 46] as well as providing a mechanism for rich interactive feedback [16], we wanted to explore the relationship between effects on user trust when working with explanation-based user feedback. To that end, we expanded our goals with these additional research questions:

RQ3: Does user perception of explanation relevance change if the user provides feedback to the system?
RQ4: Does the more subjective context change the effects on user perception?

We designed a controlled experiment using a system that classified the subject of text samples with no feedback, decision-based feedback, and explanation-based feedback. As per our previous results described in this paper, the presence of HITL feedback could reduce user trust and perception of system accuracy by increasing their saliency and causing an availability bias. Therefore, it could also follow that a similar bias would occur and decrease perceived explanation relevance. However, we hypothesized that using the explanations as a feedback mechanism would offset this effect by

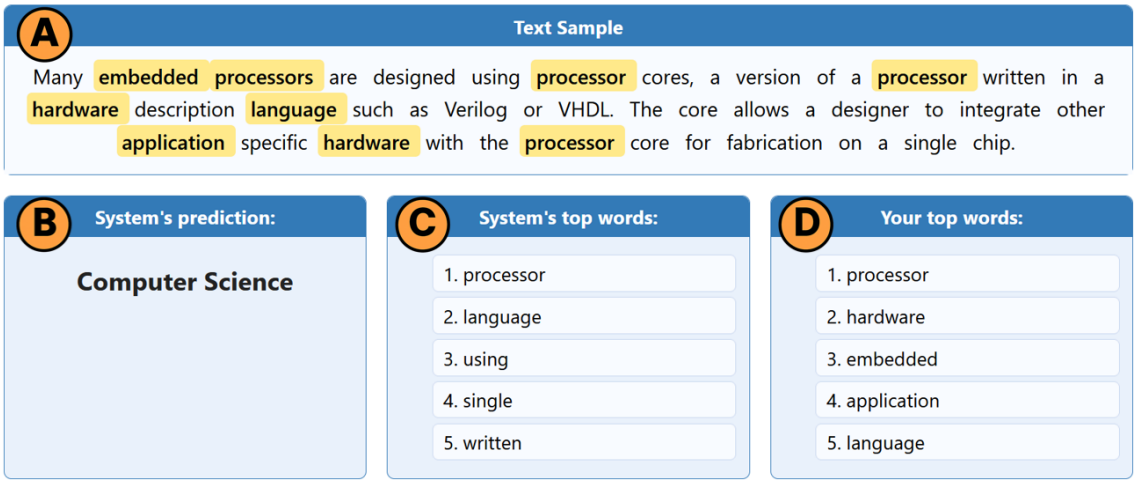


Fig. 5. Overview of the Experiment 2 and 3 interface. The system highlights the top words within the text sample (A) that contribute to the classification (B). Participants in the *explanation-based feedback* condition could change the highlights by clicking on the words. The system's original top 5 words are always visible (C) and the participant's updated list can be manually reordered to put the words they perceive as most important at the top (D).

reframing feedback in the user's mind from telling the system that it was wrong to explaining their own reasoning back to the system.

4.2 Method

This section details of the experimental design and study procedure of Experiment 2.

4.2.1 Experimental Design. For the Experiment 2 user study, we developed a text-classification system using a Naive Bayes classifier that predicted the subject of snippets from textbooks[2]. The subjects of Computer Science and History were selected as they are easily distinguishable to a general audience. In addition to the system's prediction, the top five words that most contributed to the class selection were presented as a list, as well as being highlighted in the text. In order to have a classifier with correctable flaws, we intentionally selected a basic model—Naive Bayes—and did not randomize the training and testing sets so the classifier was trained on the first 70% of each textbook and no text from the latter 30% was included in training. As a result, the system was not aware of words that were only found in the later chapters of the textbook, which was most noticeable for the History textbook as it was presented in chronological order and the subjects being covered changed significantly based on the time period being discussed. Additionally, we intentionally poisoned three words ("hardware", "instruction", and "architecture") by flipping their values for each class. All three words were strongly weighted towards Computer Science in the initially trained model, but were treated as being related to History by the poisoned model. Participants completed three rounds of review, with each round containing 15 text samples. The system prediction accuracy was 80% across all rounds, which was roughly representative of the overall model performance.

To facilitate our research goals, we used a between-subjects study design with the *level of feedback* as our independent variable with three levels: *no feedback*, *decision-based feedback*, and *explanation-based feedback*. Participants in the *no*

[2]https://www.kaggle.com/datasets/deepak711/4-subject-data-text-classification/data

feedback condition were told that they were reviewing the output of a system for quality control and that their responses would be sent to human researchers for later review, with no indication that the system would change based on their input. The feedback provided in the *decision-based feedback* condition consisted of only correcting the system's prediction and rating the relevance of the top words on a 5-point Likert scale, whereas the *explanation-based feedback* condition additionally had participants adjusting the highlighted words to create a new list of top relevant words based on their own judgment. While the classification was objective—each text sample was either about History or Computer Science with no samples that were ambiguous—the feedback about the system's top five words that contributed to the classification was more subjective. The most relevant words are neither purely objective nor purely subjective; certain words may be more or less related to each subject (e.g., "compiler" is more associated with Computer Science than History) and generic words such as articles and prepositions should not be contributing to the classification one way or another. However, the specific top five most relevant words in order are not an objective truth like the location of human faces in images as in Experiment 1. Different people may have different opinions on the most relevant words to explain why a text sample is the topic that it is, whereas the presence and location of a face in an image is not an opinion. In order to keep the system's performance constant for all participants, in the conditions with feedback we chose to only simulate the model updating and incorporating participant feedback rather than using an actual HITL system. However, participants were informed that the system would update using their feedback between each round of the task. As in Experiment 3, this decision to deceive participants to believe that the system was updating based on their feedback when it did not was made to ensure that any observed effects would be due to providing feedback and not because they were interacting with an improved version of the system. Additionally, Experiment 2 also kept the order of the text samples constant to ensure that each participant saw the same system outputs in the same order, thus providing increased experimental control to prevent possible confounds from different participants experiencing instances in different orders [47, 51]. To simulate the updating process for participant believability, a 45 second pause was introduced between each round under the premise of waiting for the model to adjust its classification algorithm based on their responses in the previous round.

4.2.2 Procedure. The study was conducted remotely through an online web application with no contact between the researchers and the participants during the task. Participants were first asked to complete a demographic questionnaire that covered their age, gender identification, educational background, and self-reported experience with ML and artificial intelligence to ensure that there was not a significant difference in the study population between conditions. This was followed by instructions on how to complete the task and included examples of accurate and inaccurate system predictions. All participants were instructed that the goal of the classifier was to predict whether a text sample was from a History or Computer Science textbook, and the classifier provided its top five words in order that contributed to its classification. For participants who would provide *explanation-based feedback* an additional tutorial was provided to explain how to modify the top words for a given text sample, with a test at the end that required them to demonstrate that they understood how to use the word highlighting system.

Following the instructions and tutorial (if applicable), the main task consisted of three rounds of reviewing 15 text samples with corresponding system classifications and the system's top five words for the sample. The number of samples in each round was reduced from the 30 images in Experiment 1 to 15 text samples due to the increased amount of time required to process each sample due to the change in domain. After each round, participants were asked to estimate how likely they thought the system was to correctly classify a sample of text from 0-100%. For the *decision-based feedback* and *explanation-based feedback* conditions, participants were made to wait 45 seconds and were provided with a reminder that they were waiting for the system to update its classification algorithm based on their previous responses, although

no such update was occurring. After finishing the final round of text samples, participants were asked to complete a questionnaire asking for their level of agreement on a 7-point Likert scale with several statements regarding their level of trust in the system and were given the opportunity to provide any comments they had regarding the study.

4.2.3 Participants. We recruited participants through Amazon Mechanical Turk with a requirement for participants to have the Masters qualification, an approval rate of greater than 90%, and a history of 500 or more successfully completed tasks. Participants ranged from ages 23–68 and lived in the United States at the time of study completion. To ensure the quality of participant responses, we measured the percentage of responses for which participants correctly identified whether the classification of a text sample corresponded to a system error or not. Participants were not included in the results if they had less than 75% accuracy for either correct instances or system errors. This study had a total of 149 participants, 5 of whom were removed based on the accuracy criteria. The remaining 144 participants consisted of 72 males and 72 females. Participants took approximately 19 minutes on average to complete the study.

4.3 Experiment 2 Results

In this section, we present our study's measures and corresponding statistical test results. We report statistical test results along with generalized eta squared (η^2_G) for effect sizes of ANOVA tests.

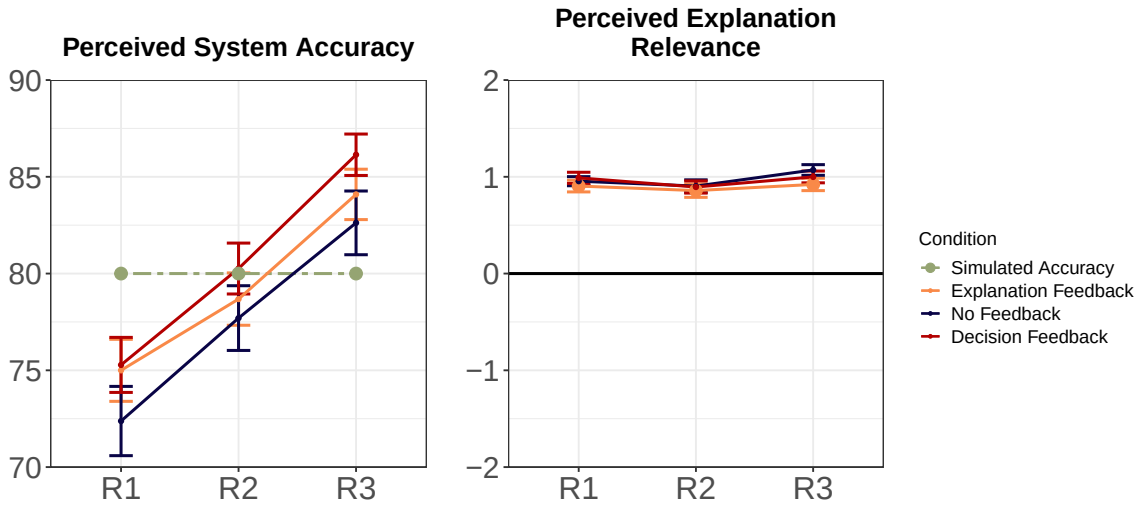
4.3.1 User-Perceived Model Accuracy. Between each task round, we asked participants to estimate as a percentage how accurate they thought the system's classifications were. We performed a two-way mixed-design ANOVA with *level of feedback* as a between subjects factor and *task round* (R1, R2, R3) as a within subjects factor.

No significant results were found for the *level of feedback* condition, with $F(2, 137) = 0.951$, $p = 0.389$. *Task round* was found to have a significant effect, with $F(2, 274) = 99.328$, $p < 0.001$, $\eta^2_G = 0.166$. Post hoc pairwise t-tests with Bonferroni correction revealed significant differences between the task rounds. Each pair yielded a p-value of $p < 0.001$, with the perceived accuracy being higher in each consecutive round (R1 < R2 < R3). No significant interaction effects were observed between any variables for this measure. Results are shown in Figure 6a.

4.3.2 User-Perceived Explanation Relevance. For each text sample, participants rated the system's top words on a five-point Likert scale. These responses were then averaged for each task round, and we performed a two-way mixed-design ANOVA with *level of feedback* as a between subjects factor and *task round* (R1, R2, R3) as a within subjects factor.

The *level of feedback* condition was not found to have a statistically significant effect, with $F(2, 140) = 0.481$, $p = 0.619$. The *task round* had a significant effect observed with $F(2, 280) = 6.325$, $p = 0.002$, $\eta^2_G = 0.011$. However, post hoc pairwise t-tests with Bonferroni correction did not reveal significant differences between task rounds. No significant interaction effects were observed between any variables. Results are shown in Figure 6b.

4.3.3 User Trust. At the end of the session, participants rated their agreement with each of a set of trust statements on a seven-point Likert scale. The same three statements were used as in Experiment 1 (see Section 3.2.3), and the aggregate rating was used as a measure for trust. For analysis, we performed a one-way ANOVA with *level of feedback* as the between-subjects factor. No statistically significant result was observed for this measure, with $F(2, 141) = 0.846$, $p = 0.431$. Additionally, no significant interaction effects were observed between any variables. Results are shown in Figure 7a.



(a) Responses to "How likely do you think the system is to classify a text sample successfully?" (Percentage).

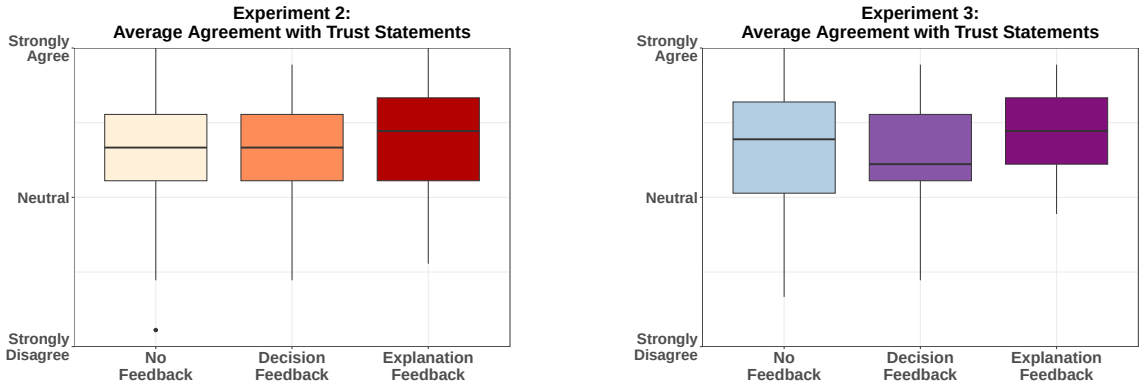
(b) Responses to "How relevant were the system's top words?" on a 5-point Likert scale averaged across each task round.

Fig. 6. Experiment 2: Participants' perceived accuracy and ratings of perceived explanation relevance over time, as grouped by the three task rounds (R1, R2, R3) of the study period.

4.4 Summary of Experiment 2 Findings

Our goal for Experiment 2 was to explore the effects of providing interactive feedback in a context where the feedback being provided is of a somewhat subjective nature, as contrasted to the more objective detection task of Experiment 1. Although the Experiment 1 image study described in Section 3 and the results from our initial study [26] both showed a negative bias toward the perception of system accuracy and a reduced level of trust in the system from providing interactive feedback (**RQ2**), no such effect was observed in the context of text classification in Experiment 2—despite the experimental design approximating the earlier studies. Additionally, no effect was observed on perceived explanation relevance (**RQ3**) As the experimental design was retained across the object detection and text classification studies, we hypothesize that the level of subjectivity in the context of the feedback mechanism is responsible for the effects observed in [26] and Experiment 1 in Section 3 (**RQ4**).

We hypothesize that when the feedback being provided is of a subjective nature—such as which words are the most relevant to identifying a text snippet's subject—the act of providing feedback is contextualized as modifying the system to behave like one's own mental model. This is much closer to the contexts of the existing research which has found positive effects on trust and willingness to rely on HITL systems [12, 54, 56], leading us to believe that this is why we did not observe the same negative effect as in Experiment 1. Similarly, we hypothesize that this is also the reason for the observed increase in perceived system accuracy across the rounds of the task. In the object detection context, providing feedback meant the user was spending most of the time and mental energy associated with providing feedback to correcting objective system errors, resulting in a negative bias towards the system due to those errors being more salient. In contrast, the more subjective context of the text classification studies had participants actively thinking about their mental model for how to classify the text samples and comparing it to the system's model. Since the system's accuracy did not change across the rounds of the task, participants perceiving the system to improve in accuracy could indicate that



(a) Experiment 2

(b) Experiment 3

Fig. 7. Averaged responses to the three trust scale statements described in Section 3.2.3 on a 7-point Likert scale.

they were adjusting their mental model to be more accepting of the system's reasoning as they spent more time interacting with it.

5 EXPERIMENT 3: TEXT CLASSIFICATION (COUNTERBALANCED ORDERING)

Following the unexpected results of Experiment 2, we sought to deepen the investigation through an additional text classification study following the experimental design of Experiment 2 (Section 4).

5.1 Research Objectives and Experimental Design

As the results of Experiment 2 showed that participants erroneously believed the system became significantly more accurate across task rounds regardless of condition (Figure 6a), we conducted a follow-up study for confirmation. Experiment 3 repeats the study design of Experiment 2 but with a counterbalanced ordering of system outputs to ensure that our observations were not caused by the choice of constant ordering of instances in Experiment 2. Though the constant ordering in Experiment 2 was an intentional decision based on prior work showing differences and biases due to seeing the same system outputs in different orderings [47, 51], there was a possibility that the specific tested ordering of observed text instances could have (by chance) contributed to the finding of increased perceived accuracy over time. To address this potential confound, Experiment 3 counterbalanced the three groups of system outputs so that all possible orders of the text samples and their outputs were represented equally. Other than this change to how participants experienced ordering, the other aspects of the experimental design (the research objectives, study design, and experimental procedure) remain identical to Experiment 2 as described in Section 4. Participants again provided feedback to the outputs of a text-classification system that predicted the subject of snippets from History and Computer Science textbooks[3]. Experiment 3 used the same between-subjects study design with the *level of feedback* as our independent variable with three levels: *no feedback*, *decision-based feedback*, and *explanation-based feedback*, as in Experiment 2.

[3]https://www.kaggle.com/datasets/deepak711/4-subject-data-text-classification/data

5.2 Participants

For Experiment 3, participants were recruited from the University of Florida Department of Computer and Information Science and Engineering, consisting of both undergraduate and graduate students. Participants ranged from ages 26–66 and lived in the United States at the time of study completion. To ensure the quality of participant responses, we measured the percentage of responses for which participants correctly identified whether the classification of a text sample corresponded to a system error or not. As a quality check, participants were not included in the results if they had less than 75% accuracy for either correct instances or system errors. This study had a total of 104 participants, and 10 were removed based on the accuracy criteria. The remaining 94 participants consisted of 32 females, 60 males, and 2 non-binary. Participants took approximately 24 minutes on average to complete the study.

5.3 Experiment 3 Results

In this section, we present our study's measures and corresponding statistical test results. We report statistical test results along with generalized eta squared (η^2_G) for effect sizes of ANOVA tests.

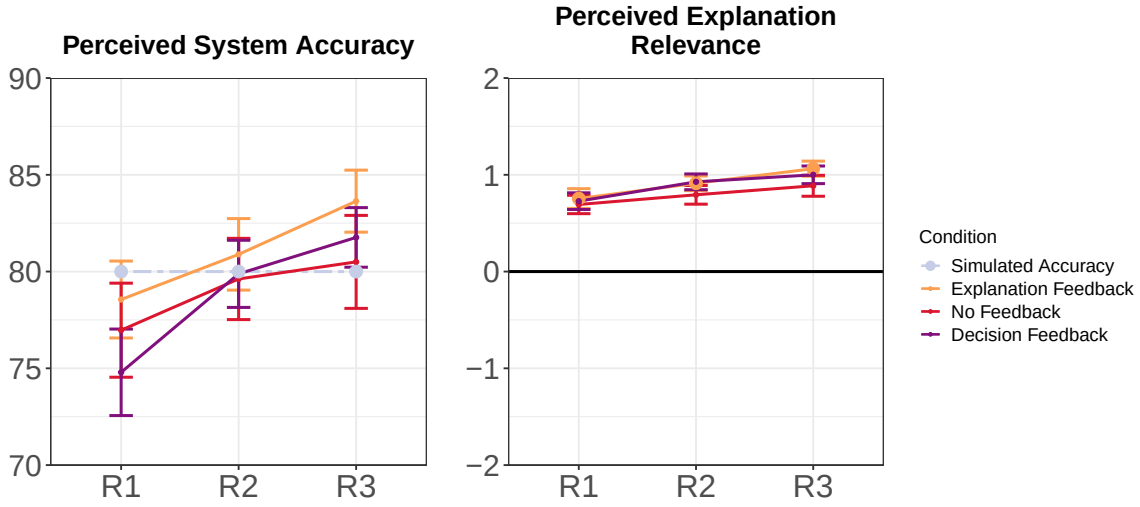
5.3.1 User-Perceived Model Accuracy. Between each task round, we asked participants to estimate as a percentage how accurate they thought the system's classifications were. We performed a two-way mixed-design ANOVA with *level of feedback* as a between subjects factor and *task round* (R1, R2, R3) as a within subjects factor.

The *level of feedback* condition did not have significant results, with $F(2, 91) = 0.871$, $p = 0.422$. A significant effect was observed based on *task round*, with $F(2, 182) = 14.194$, $p < 0.001$, $\eta^2_G = 0.045$. Post hoc pairwise t-tests with Bonferroni correction revealed significant differences between the task rounds. Specifically, the comparison between R1 and R3 yielded a p-value of $p < 0.0015$, with user perception of system accuracy in R3 being higher than in R1. The other paired comparisons did not yield statistically significant results. Additionally, no significant interaction effects were observed between any variables. Results are shown in Figure 8a.

5.3.2 User-Perceived Explanation Relevance. For each text sample, participants rated the system's top words on a five-point Likert scale. These responses were then averaged for each task round, and we performed a two-way mixed-design ANOVA with *level of feedback* as a between subjects factor and *task round* (R1, R2, R3) as a within subjects factor.

Level of feedback was not found to have a statistically significant effect, with $F(2, 87) = 0.853$, $p = 0.430$. A significant effect was observed based on *task round*, with $F(2, 174) = 13.079$, $p < 0.001$, $\eta^2_G = 0.057$ Post hoc pairwise t-tests with Bonferroni correction revealed significant differences between the task rounds. Specifically, the comparison between R1 and R3 yielded a p-value of $p < 0.001$, with user perception of explanation relevance in R3 being higher than in R1. The other paired comparisons did not yield statistically significant results. No significant interaction effects were observed between any variables. Results are shown in Figure 8b.

5.3.3 User Trust. At the end of the session, we asked participants to rate their agreement with each of a set of trust statements on a seven-point Likert scale. The same three statements were used as in Experiments 1 and 2 (described in Section 3.2.3), and the aggregate rating was used as a measure for trust. We performed a one-way ANOVA with *level of feedback* as the between-subjects factor. No statistically significant result was observed for this measure, with $F(2, 101) = 1.509$, $p = 0.226$. Additionally, no significant interaction effects were observed between any variables for this measure. Results are shown in Figure 7b.



(a) Responses to "How likely do you think the system is to classify a text sample successfully?" (Percentage)

(b) Responses to "How relevant were the system's top words?" on a 5-point Likert scale averaged across each task round.

Fig. 8. Experiment 3: Participants' perceived system accuracy and ratings of explanation relevance over time across the three task rounds (R1, R2, R3).

5.4 Summary of Experiment 3 Findings

The main goal of Experiment 3 was to ensure that the perceived increase in accuracy over time observed in Experiment 2 was not due to something specific with the fixed ordering of text samples. Additionally, this study used Computer Science students as participants whereas the study population from Experiment 2 was crowdsource workers on Amazon Mechanical Turk.

There was no significant effect based on *level of feedback* for any metrics. However, participants in all conditions perceived the system to become more accurate over time and have more relevant explanations by the final round of the task than they did for the first round. The only difference from the Experiment 2 results is the significant result for perceived explanation relevance. This result is similar to the result for perceived accuracy in that participants perceived the system to be improving over time when it was not. These results support that the increase in perceived accuracy across the task rounds in Experiment 2 was not due to the fixed ordering. Additionally, the lack of an effect based on *level of feedback* contributes more evidence that there is a difference due to task context between the object detection and text classification studies. The only major difference between the results of Experiment 2 and Experiment 3 was the average task length—19 minutes on average for Experiment 2 and 24 minutes on average for Experiment 3. This is likely due to the differences between the study populations. The crowdsource workers from Amazon Mechanical Turk used in Experiment 2 were paid a flat rate based on the expected length of the task, so the more tasks they can complete in a given period of time the higher their effective hourly pay rate becomes. While the students in Experiment 3 were also given a flat amount of extra credit, the total amount of extra credit they could receive was fixed regardless of how quickly they completed the task.

In the next section, we will discuss the results of all three experiments together along with a selection of the most relevant related work.

6 DISCUSSION

In this section, we discuss the results of the three experiments in the context of the existing body of research on the effects of providing users with HITL mechanisms. We also consider limitations of our work and opportunities for continued research.

6.1 Effects of Interactive User Control on User Perception

Several studies have observed positive effects on user behavior and system perception by providing the opportunity to give HITL feedback. Multiple studies have shown participants to have a stated preference towards versions of intelligent systems that allow them to adjust model behavior to be closer to how they would complete the system's task themselves [21, 29, 54, 56, 64]. This stated preference is reflected in users' behavior; while users will stop using intelligent systems they perceive to be flawed—even when their use would still improve performance [11]—this aversion can be alleviated through the presence of even a minimal level of user control [12]. When using an intelligent citation recommendation system, users who were allowed to adjust the weights of the recommender perceived the same recommendations as more relevant and displayed a pattern of accepting recommendations immediately after making a HITL adjustment [54].

However, the presence of these HITL feedback mechanisms is not without cost. Several studies have shown an increase in cognitive load resulting from an increased level of user control, which can result in decreased levels of engagement with the HITL mechanisms despite their effectiveness [14, 29, 56]. When the HITL feedback mechanism is purely corrective—treating the user as an active learning oracle as we did in Experiment 1—it can result in users underrating the accuracy of the system and trusting the system less [26, 27]. Even when the HITL mechanisms are viewed as a positive addition by the user, those feelings can be reversed if they feel that their feedback is not being reflected in meaningful behavioral change throughout the system, resulting in a higher level of frustration with the system than if the system had no HITL options at all [64].

6.2 Importance of Feedback Objectivity

Both the results of the experiments presented in this paper and the existing body of literature on the effects of providing HITL feedback on user perception suggest that the context of the system that feedback is being provided to contributes to user perception of system accuracy. In the context of object detection in images from Experiment 1, participants were providing feedback of an objective nature; a region of a picture either contains a human face or it does not. There was no room for individual preference, but rather the system is either demonstrating a correct output or an incorrect output and the participants were tasked with correcting the errors. The main finding of this experiment was that participants who gave *interactive feedback* perceived the system to have a greater decrease in accuracy across the rounds of the task than those who only gave *binary feedback* (Figure 4a) Additionally, we observed that participants across all conditions perceived a significant decrease in system accuracy throughout the task, even though the system accuracy remained constant (Figure 3). In contrast, the text classification context used in Experiments 2 and 3 provide a more subjective form of feedback. Although the true classification of a text sample's topic is objectively known, identifying which words most contribute to that classification is somewhat subjective. These studies did not observe a negative bias based on feedback interactivity, and participants in all conditions believed that the accuracy of the system improved across the duration of the task (Figure 8a and Figure 6a).

When the feedback being provided is of an objective nature—treating the user as an active learning oracle as in the image context of Experiment 1 and active learning literature [27]—the act of providing feedback is contextualized as having to correct a faulty system and results in perceiving the system to be worse than it actually is due to the increased focus on system errors. In contrast, when the feedback is of a subjective nature—as in the text context of Experiments 2 and 3 where the most relevant words are open to interpretation—it can be contextualized as having the ability to make the system behave closer to the user's own mental model, which is closer to the reasons for preferring systems with HITL components observed by prior research [54]. In studies where users have been found to report increased satisfaction or exhibited preferences towards systems with HITL feedback, the feedback mechanisms have predominantly been collaborative systems framed as providing user control based on their preferences [12, 54, 56]. The feedback in these systems is even more subjective than the annotative feedback in Experiments 2 and 3, and stronger effects are observed in those contexts. We hypothesize that this distinction is responsible for the differences observed across the different research in this field.

Furthermore, we hypothesize that the differences in the explanations provided across the different study contexts contributed to participants in the image classification study perceiving the system to get worse over time, while participants in the text classification studies found the system to become more accurate over time, despite the accuracy of the systems remaining constant in all three studies. In the image classification context, the objectivity of the presence of faces within the bounding boxes contributed to system errors seeming obvious. The location of faces in an image is relatively obvious to identify as a human, and bounding boxes alone do not provide much insight into the reason behind model errors. In the text classification studies, we provided participants with the top five words that contributed to the classification of each text sample. Since there is no objectively correct set of five words which most contribute to a text sample's topic, there is room for both the user to provide feedback that feels like they are making the system behave closer to their mental model and for the user to adjust their mental model to be closer to the system's. This could be the cause of participants finding the system to be more accurate over time in the more subjective context of our text classification studies.

6.3 Design Implications

A potential reason to include HITL components is to crowd-source model improvements, giving end users the ability to correct objective system errors. Although this may be a beneficial feature for both model improvement and user control, this research suggests that treating the end user as an active learning oracle in this manner could actually result in a negative bias towards their perception of the system's accuracy over time. Overly distrusting an intelligent system in this manner can result in self-reliance for critical decisions, resulting in higher rates of failure than if they trusted the system at an appropriate level [53]. This could be offset by including types of feedback that users feel allow them to influence model behavior rather than just having them correct objective mistakes. Additionally, providing robust explanations of model behavior can help calibrate an appropriate level of trust [58, 66]. System explanations can also act as a natural mechanism for more detailed user control, helping the act of providing feedback to be more about changing model behavior rather than simply labeling mistakes. However, over-complicating the feedback mechanisms can result in an increase in cognitive load that may cause less experienced users to engage with the system less [14, 29, 56]. Finding the appropriate level of explanations and control to provide to different types of users is important to ensure an appropriate level of engagement and reliance.

Designers of intelligent systems should also consider whether they are in a context where users will be able to objectively identify system errors or whether errors are more subjective and difficult to identify. In the image classification study where system outputs were fully objective, we observed a negative bias towards the system over time; in the

text classification studies where system outputs were more subjective we saw the opposite effect. The ideal behavior would be for a user of an intelligent system to base their decision to rely on a system off of examining the system's outputs and making an evaluation based on its performance over time [24]. However, users do not always engage in this behavior, particularly if they are not experts in the domain of the system [50]. The findings of this paper suggest that when system errors feel objectively wrong, it results in overly distrusting the system. On the other hand, there is also a risk of over-reliance—also known as automation bias [20]—if the user has difficulty identifying system errors [24, 38, 48]. Ensuring that even less experienced users can make informed decisions about the trustworthiness of AI systems is critical to avoid improper levels of reliance.

6.4 Limitations and Future Work

The main limitation of this work is that neither system studied in this paper actually utilized the participants' feedback to update the system, instead relying on deceiving the participants into thinking it was. Although this provided a necessary level of experimental control, it increases the gap between this research and the real world scenarios our findings could be applied to. Other studies have utilized systems containing actual HITL updates, but they often fail to distinguish between the effects of providing feedback and the effects of interacting with the resulting improved system [29, 54]. There is a clear need for further exploration into studying the effects of providing HITL feedback on system perception and user behaviors in increasingly realistic contexts. Additionally, both studies presented in this paper used mandatory feedback, while most related work focuses on optional feedback—a more common real-world scenario. A potential avenue for further research would be to explicitly explore if there is a difference in the observed effects when working with mandatory versus optional feedback.

This research contributed empirical evidence that asking users of intelligent systems to provide corrective feedback in the style of an active learning oracle can result in underrating system accuracy and distrust towards the system. We also showed that this effect was not observed when repeating the experiment in the context of asking the user to provide more subjective feedback. However, these were separate studies and subjectivity was not a variable that was directly controlled for. Additionally, the more subjective text case has more of an element of explainability than the more objective image case—a bounding box around a face is less of an explanation than it is part of the classification—which may have also contributed to the differences in observed effects. While our experiments and prior work support the interpretation that level of subjectivity is contributing to the differences in observed results across studies, future research that directly explores the effects of different levels of subjectivity on the user perception is needed.

The studies presented in this paper all observed significant changes in perception of accuracy over time, which was not what the studies were directly designed to explore. Future research about the change in perception of intelligent system accuracy over time could consider more task rounds or longer durations, potentially focusing on extended usage over multiple sessions. They could also utilize more involved tasks or observe real world intelligent system use where the participants are more invested in the system outcomes.

7 CONCLUSION

Including HITL feedback in an intelligent system can be beneficial, both in terms of improved model performance and providing users with agency over system behavior. When it comes to its effects on user satisfaction, perception of system performance, and user behaviors, results can wildly vary. We presented three experiments that explored how different feedback contexts can change the effects that providing such feedback has on users of HITL intelligent systems. From Experiment 1, we found that in cases where the accuracy of a given system output feels objective, users distrust the system

through continued use. However, in a context where the outputs of a system are judged more subjectively (Experiments 2 and 3), participants perceived that the system improved over time despite no change in its true accuracy. This suggests the perceived objectivity of system accuracy could contribute to both distrust and mistrust. Furthermore, the results showed that when participants are treated as an active learning oracle and asked to provide objective feedback about the accuracy of the system and correct its errors, they distrust the system and underrate its accuracy. However, no distrust was observed when participants were given a similar task but asked to provide feedback that was subjective in nature. These results suggest that developers who wish to incorporate HITL mechanisms without into their systems without negatively biasing their users should focus on using it to enhance user agency rather than as a method of objective data collection.

8 ACKNOWLEDGEMENTS

Anonymized for review

REFERENCES

[1] AMERSHI, S., CAKMAK, M., KNOX, W. B., AND KULESZA, T. Power to the people: The role of humans in interactive machine learning. Ai Magazine 35, 4 (2014), 105–120.

[2] ARGALL, B. D., CHERNOVA, S., VELOSO, M., AND BROWNING, B. A survey of robot learning from demonstration. Robotics and autonomous systems 57, 5 (2009), 469–483.

[3] BERK, R., AND HYATT, J. Machine learning forecasts of risk to inform sentencing decisions. Federal Sentencing Reporter 27, 4 (2015), 222–228.

[4] BERTRAND, A., BELLOUM, R., EAGAN, J. R., AND MAXWELL, W. How cognitive biases affect xai-assisted decision-making: A systematic review. In Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (2022), pp. 78–91.

[5] CAHOUR, B., AND FORZY, J.-F. Does projection into use improve trust and exploration? an example with a cruise control system. Safety science 47, 9 (2009), 1260–1270.

[6] CAKMAK, M., AND THOMAZ, A. L. Eliciting good teaching from humans for machine learners. Artificial Intelligence 217 (2014), 198–215.

[7] CHO, M., LEE, G., AND HWANG, S.-W. Explanatory and actionable debugging for machine learning: A tableqa demonstration. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (2019), pp. 1333–1336.

[8] ČÍK, I., RASAMOELINA, A. D., MACH, M., AND SINČÁK, P. Explaining deep neural network using layer-wise relevance propagation and integrated gradients. In 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) (2021), IEEE, pp. 000381–000386.

[9] COHN, D., ATLAS, L., AND LADNER, R. Improving generalization with active learning. Machine learning 15, 2 (1994), 201–221.

[10] COHN, D. A., GHAHRAMANI, Z., AND JORDAN, M. I. Active learning with statistical models. Journal of artificial intelligence research 4 (1996), 129–145.

[11] DIETVORST, B. J., SIMMONS, J. P., AND MASSEY, C. Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of experimental psychology: General 144, 1 (2015), 114.

[12] DIETVORST, B. J., SIMMONS, J. P., AND MASSEY, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science 64, 3 (2018), 1155–1170.

[13] DOSHI-VELEZ, F., AND KIM, B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017).

[14] EHRMANN, D. E., GALLANT, S. N., NAGARAJ, S., GOODFELLOW, S. D., EYTAN, D., GOLDENBERG, A., AND MAZWI, M. L. Evaluating and reducing cognitive load should be a priority for machine learning in healthcare. Nature medicine 28, 7 (2022), 1331–1333.

[15] ELWELL, R., AND POLIKAR, R. Incremental learning of concept drift in nonstationary environments. IEEE Transactions on Neural Networks 22, 10 (2011), 1517–1531.

[16] ESTIVILL-CASTRO, V., GILMORE, E., AND HEXEL, R. Constructing explainable classifiers from the start—enabling human-in-the loop machine learning. Information 13, 10 (2022), 464.

[17] FAILS, J. A., AND OLSEN JR, D. R. Interactive machine learning. In Proceedings of the 8th international conference on Intelligent user interfaces (2003), pp. 39–45.

[18] GENG, X., AND SMITH-MILES, K. Incremental learning., 2009.

[19] GHAI, B., LIAO, Q. V., ZHANG, Y., BELLAMY, R., AND MUELLER, K. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. arXiv preprint arXiv:2001.09219 (2020).

[20] GODDARD, K., ROUDSARI, A., AND WYATT, J. C. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 19, 1 (2012), 121–127.

[21] HE, C., PARRA, D., AND VERBERT, K. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. Expert Systems with Applications 56 (2016), 9–27.

[22] Hoff, K. A., and Bashir, M. Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors 57, 3 (2015), 407–434.

[23] Hoffman, R., Mueller, S., Klein, G., and Litman, J. Measuring trust in the xai context.

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

[25] Holzinger, A., Plass, M., Holzinger, K., Crişan, G. C., Pintea, C.-M., and Palade, V. Towards interactive machine learning (iml): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In International Conference on Availability, Reliability, and Security (2016), Springer, pp. 81–95.

[26] Honeycutt, D., Nourani, M., and Ragan, E. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2020), vol. 8, pp. 63–72.

[27] Iuzzolino, M. L., Umada, T., Ahmed, N. R., and Szafir, D. A. In automation we trust: investigating the role of uncertainty in active learning systems. arXiv preprint arXiv:2004.00762 (2020).

[28] Jian, J.-Y., Bisantz, A. M., and Drury, C. G. Foundations for an empirically determined scale of trust in automated systems. International journal of cognitive ergonomics 4, 1 (2000), 53–71.

[29] Jin, Y., Cardoso, B., and Verbert, K. How do different levels of user control affect cognitive load and acceptance of recommendations? In Jin, Y., Cardoso, B. and Verbert, K., 2017, August. How do different levels of user control affect cognitive load and acceptance of recommendations?. In Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems co-located with ACM Conference on Recommender Systems (RecSys 2017) (2017), vol. 1884, CEUR Workshop Proceedings, pp. 35–42.

[30] Korsgaard, M. A., Schweiger, D. M., and Sapienza, H. J. Building commitment, attachment, and trust in strategic decision-making teams: The role of procedural justice. Academy of Management journal 38, 1 (1995), 60–84.

[31] Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Kamali, S., Malloci, M., Pont-Tuset, J., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., and Murphy, K. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html (2017).

[32] Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. Principles of explanatory debugging to personalize interactive machine learning. In Proceedings of the 20th international conference on intelligent user interfaces (2015), pp. 126–137.

[33] Kulesza, T., Stumpf, S., Burnett, M., Wong, W.-K., Riche, Y., Moore, T., Oberst, I., Shinsel, A., and McIntosh, K. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In 2010 IEEE Symposium on Visual Languages and Human-Centric Computing (2010), IEEE, pp. 41–48.

[34] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. Too much, too little, or just right? ways explanations impact end users' mental models. In 2013 IEEE Symposium on visual languages and human centric computing (2013), IEEE, pp. 3–10.

[35] Kulesza, T., Wong, W.-K., Stumpf, S., Perona, S., White, R., Burnett, M. M., Oberst, I., and Ko, A. J. Fixing the program my computer learned: Barriers for end users, challenges for the machine. In Proceedings of the 14th international conference on Intelligent user interfaces (2009), pp. 187–196.

[36] Kumar, P., and Gupta, A. Active learning query strategies for classification, regression, and clustering: a survey. Journal of Computer Science and Technology 35, 4 (2020), 913–945.

[37] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., and Ferrari, V. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV (2020).

[38] Lee, J. D., and See, K. A. Trust in automation: Designing for appropriate reliance. Human factors 46, 1 (2004), 50–80.

[39] Lewicki, R. J., McAllister, D. J., and Bies, R. J. Trust and distrust: New relationships and realities. Academy of management Review 23, 3 (1998), 438–458.

[40] Lewis, D. D., and Gale, W. A. A sequential algorithm for training text classifiers. In SIGIR'94 (1994), Springer, pp. 3–12.

[41] Li, G. Human-in-the-loop data integration. Proceedings of the VLDB Endowment 10, 12 (2017), 2006–2017.

[42] Li, N., Adepu, S., Kang, E., and Garlan, D. Explanations for human-on-the-loop: A probabilistic model checking approach. In Proceedings of the IEEE/ACM 15th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (2020), pp. 181–187.

[43] Madsen, M., and Gregor, S. Measuring human-computer trust. In 11th australasian conference on information systems (2000), vol. 53, Citeseer, pp. 6–8.

[44] Middleton, S. E., Shadbolt, N. R., and De Roure, D. C. Capturing interest through inference and visualization: Ontological user profiling in recommender systems. In Proceedings of the 2nd international conference on Knowledge capture (2003), pp. 62–69.

[45] Mohseni, S., Zarei, N., and Ragan, E. D. A survey of evaluation methods and measures for interpretable machine learning. ACM Transactions on Interactive Intelligent Systems (2018).

[46] Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. Human-in-the-loop machine learning: A state of the art. Artificial Intelligence Review 56, 4 (2023), 3005–3054.

[47] Nourani, M., Hashky, A., and Ragan, E. D. User profiling in human-ai design: an empirical case study of anchoring bias, individual differences, and ai attitudes. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2024), vol. 12, pp. 137–146.

[48] Nourani, M., Honeycutt, D. R., Block, J. E., Roy, C., Rahman, T., Ragan, E. D., and Gogate, V. Investigating the importance of first

impressions and explainable ai with interactive video analysis. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (2020), pp. 1–8.

[49] Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (2019), vol. 7, pp. 97–105.

[50] Nourani, M., King, J. T., and Ragan, E. D. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In Eighth AAAI Conference on Human Computation and Crowdsourcing (2020).

[51] Nourani, M., Roy, C., Block, J. E., Honeycutt, D. R., Rahman, T., Ragan, E. D., and Gogate, V. On the importance of user backgrounds and impressions: Lessons learned from interactive ai applications. ACM Transactions on Interactive Intelligent Systems 12, 4 (2022), 1–29.

[52] Nourani, M., Roy, C., Rahman, T., Ragan, E. D., Ruozzi, N., and Gogate, V. Don't explain without verifying veracity: An evaluation of explainable ai with video activity recognition. arXiv preprint arXiv:2005.02335 (2020).

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

[54] Parra, D., and Brusilovsky, P. User-controllable personalization: A case study with setfusion. International Journal of Human-Computer Studies 78 (2015), 43–67.

[55] Qian, K., Popa, L., and Sen, P. Systemer: A human-in-the-loop system for explainable entity resolution.

[56] Rani, N., Chu, S. L., and Mei, V. R. Investigating the effects of different levels of user control on the effectiveness of context-aware recommender systems for web-based search. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (2022), pp. 1–6.

[57] Rashid, A. M., Ling, K., Tassone, R. D., Resnick, P., Kraut, R., and Riedl, J. Motivating participation by displaying the value of contribution. In Proceedings of the SIGCHI conference on Human Factors in computing systems (2006), pp. 955–958.

[58] Ribeiro, M. T., Singh, S., and Guestrin, C. " 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 (2016), pp. 1135–1144.

[59] Roy, C., Nourani, M., Honeycutt, D. R., Block, J. E., Rahman, T., Ragan, E. D., Ruozzi, N., and Gogate, V. Explainable activity recognition in videos: Lessons learned. Applied AI Letters 2, 4 (2021), e59.

[60] Settles, B. From theories to queries: Active learning in practice. In Active Learning and Experimental Design workshop In conjunction with AISTATS 2010 (2011), pp. 1–18.

[61] Shivaswamy, P., and Joachims, T. Online structured prediction via coactive learning. arXiv preprint arXiv:1205.4213 (2012).

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

[63] Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y., Dong, J., Duffy, V. G., Fang, X., Fidopiastis, C., Fragomeni, G., Fu, L. P., et al. Seven hci grand challenges. International Journal of Human–Computer Interaction 35, 14 (2019), 1229–1269.

[64] Stumpf, S., Sullivan, E., Fitzhenry, E., Oberst, I., Wong, W.-K., and Burnett, M. Integrating rich user feedback into intelligent user interfaces. In Proceedings of the 13th international conference on Intelligent user interfaces (2008), pp. 50–59.

[65] Szczuka, J. M., Horstmann, A. C., Szymczyk, N., Strathmann, C., Artelt, A., Mavrina, L., and Krämer, N. Let me explain what i did or what i would have done: An empirical study on the effects of explanations and person-likeness on trust in and understanding of algorithms. In Proceedings of the 13th Nordic Conference on Human-Computer Interaction (2024), pp. 1–13.

[66] Teso, S., and Kersting, K. " why should i trust interactive learners?" explaining interactive queries of classifiers to users. arXiv preprint arXiv:1805.08578 (2018).

[67] Teso, S., and Kersting, K. Explanatory interactive machine learning. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019), pp. 239–245.

[68] Tong, S., and Chang, E. Support vector machine active learning for image retrieval. In Proceedings of the ninth ACM international conference on Multimedia (2001), pp. 107–118.

[69] Torcianti, A., and Matzka, S. Explainable artificial intelligence for predictive maintenance applications using a local surrogate model. In 2021 4th International Conference on Artificial Intelligence for Industries (AI4I) (2021), IEEE, pp. 86–88.

[70] Van den Bos, K., Vermunt, R., and Wilke, H. A. The consistency rule and the voice effect: The influence of expectations on procedural fairness judgements and performance. European Journal of Social Psychology 26, 3 (1996), 411–428.

[71] Vathoopan, M., Brandenbourger, B., and Zoitl, A. A human in the loop corrective maintenance methodology using cross domain engineering data of mechatronic systems. In 2016 IEEE 21st International Conference on Emerging Technologies and Factory Automation (ETFA) (2016), IEEE, pp. 1–4.

[72] Wischnewski, M., Krämer, N., and Müller, E. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In Proceedings of the 2023 CHI conference on human factors in computing systems (2023), pp. 1–16.

[73] Wong, K. K., Han, Y., Cai, Y., Ouyang, W., Du, H., and Liu, C. From trust in automation to trust in ai in healthcare: A 30-year longitudinal review and an interdisciplinary framework. Bioengineering 12, 10 (2025), 1070.

[74] Yamauchi, K. Optimal incremental learning under covariate shift. Memetic Computing 1, 4 (2009), 271.

[75] Yang, Y., Kandogan, E., Li, Y., Sen, P., and Lasecki, W. A study on interaction in human-in-the-loop machine learning for text analytics. In IUI Workshops (2019).

[76] Yin, M., Wortman Vaughan, J., and Wallach, H. Understanding the effect of accuracy on trust in machine learning models. In Proceedings

of the 2019 CHI Conference on Human Factors in Computing Systems (2019), pp. 1–12.

[77] YU, K., BERKOVSKY, S., TAIB, R., ZHOU, J., AND CHEN, F. Do i trust my machine teammate? an investigation from perception to decision. In Proceedings of the 24th International Conference on Intelligent User Interfaces (2019), pp. 460–468.

[78] ZAIDAN, O., EISNER, J., AND PIATKO, C. Using "annotator rationales" to improve machine learning for text categorization. In Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference (2007), pp. 260–267.

[79] ZHAO, X., HUANG, W., HUANG, X., ROBU, V., AND FLYNN, D. Baylime: Bayesian local interpretable model-agnostic explanations. In Uncertainty in Artificial Intelligence (2021), PMLR, pp. 887–896.

[80] ŽLIOBAITĖ, I. Learning under concept drift: an overview. arXiv preprint arXiv:1010.4784 (2010).