Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy

Donald R. Honeycutt, Mahsan Nourani, Eric D. Ragan
University of Florida, Gainesville, Florida
dhoneycutt@ufl.edu, mahsannourani@ufl.edu, eragan@ufl.edu

Abstract

Mixed-initiative systems allow users to interactively provide feedback to potentially improve system performance. Human feedback can correct model errors and update model parameters to dynamically adapt to changing data. Additionally, many users desire the ability to have a greater level of control and fix perceived flaws in systems they rely on. However, how the ability to provide feedback to autonomous systems influences user trust is a largely unexplored area of research. Our research investigates how the act of providing feedback can affect user understanding of an intelligent system and its accuracy. We present a controlled experiment using a simulated object detection system with image data to study the effects of interactive feedback collection on user impressions. The results show that 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. These results highlight the importance of considering the effects of allowing end-user feedback on user trust when designing intelligent systems.

Introduction

Bringing human feedback into the development of machine learning models has many benefits. At its simplest, human feedback allows a model to incorporate new annotations for unlabeled data to increase performance by improving the training set. A common method for introducing human feedback is active learning, where the selection of data to obtain labels for is left to the model (Cohn, Ghahramani, and Jordan 1996). Alternatively, a more human-centered approach has the labeler choose which instances to be labeled, relying on human intuition to decide what feedback would be most relevant to improve the model based on observations of its performance (Tong and Chang 2001). Developers can also allow for further involvement by giving the human participant feature-level control over model parameters, such as allowing direct modification of the feature space and its associated weights (Cho, Lee, and Hwang 2019) or prioritizing decision rules used by the model (Yang et al. 2019).

Frequently, the human-in-the-loop approaches either involve system developers for development and debugging (Vathoopan, Brandenbourger, and Zoitl 2016) or independent workers on crowd-sourcing platforms (Li 2017). 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. (Geng and Smith-Miles 2009; Yamauchi 2009; Elwell and Polikar 2011). Systems can also update over time by implicitly capturing user behaviors, which is a technique commonly used in recommender systems (Shivaswamy and Joachims 2012; Middleton, Shadbolt, and De Roure 2003). 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 (Stumpf et al. 2008).

While human-in-the-loop 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. (1996) 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 (Korsgaard, Schweiger, and Sapienza 1995). 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 human-in-the-loop 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 (Lee and See 2004; Parasuraman and Riley 1997). Therefore, it is important to understand how providing feedback to an intelligent system affects trust so that it can be accounted for when designing

human-in-the-loop systems.

In this paper, we examine how users perceive system accuracy over time and how their trust in the system changes based on the presence of interactive feedback. We used a simulated object-detection system that allowed users to provide interactive feedback to correct system errors by adjusting image regions for detected objects. Additionally, to explore possible implications of how the system responds to given feedback, our experiment also controlled different types of change in system accuracy over time. The results indicate that by providing human-in-the-loop feedback, user trust and perception of accuracy can be negatively affected—regardless of whether the system improves after receiving feedback.

Related Work

In this section, we consider prior work from the perspectives of human-in-the-loop machine learning and trust in artificial intelligence.

Human-in-the-Loop Machine Learning

While machine learning 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. *Relevance feedback* is a human-in-the-loop 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 (Tong and Chang 2001). Another approach that can be taken in domains where human intuition may not result in optimal selections of what data to label is to choose instances to add to the training set by 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 (Cohn, Ghahramani, and Jordan 1996). 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 (Settles 2011).

However, human input is not limited to merely providing new labels to data. 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 (Teso and Kersting 2019). This helps avoid situations where the model has a flaw that happens to result in the correct prediction by chance. Another form of advanced human feedback is to show the oracle model parameters, such as features and their weights (Cho, Lee, and Hwang 2019) or rules used to make decisions within the model (Yang et al. 2019), and allows for direct modification of those parameters. Being able to control model parameters in this way has been found to be useful for debugging models (Kulesza et al. 2010). While this higher level of control over the model may not be desirable in all applications, Holzinger et al. (2016) showed that human-machine teaming can sometimes result in a closer to optimal model than machine learning alone.

While the person providing feedback is not necessarily the end user for many human-in-the-loop systems, there are advantages to bringing end users into the loop. Stumpf et al. (2008) 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 (Dietvorst, Simmons, and Massey 2018). 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 (Žliobaitė 2010). A human-in-the-loop 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 (Geng and Smith-Miles 2009). These techniques have been shown to effectively address the problem of concept drift in machine learning systems (Yamauchi 2009; Elwell and Polikar 2011).

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 (Korsgaard, Schweiger, and Sapienza 1995). 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 (Van den Bos, Vermunt, and Wilke 1996). 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.

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 (Parasuraman and Riley 1997; Siau and Wang 2018). 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 (Lewicki, McAllister, and Bies 1998; Lee and See 2004). 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 (Hoffman et al. 2013). This behavior has been observed to be more prevalent among users who are domain experts than novice users (Nourani, King, and Ragan 2020). Sometimes, however, users might trust a system completely without checking the outcomes, i.e., over-reliance or *automation bias* (Goddard, Roudsari, and Wyatt

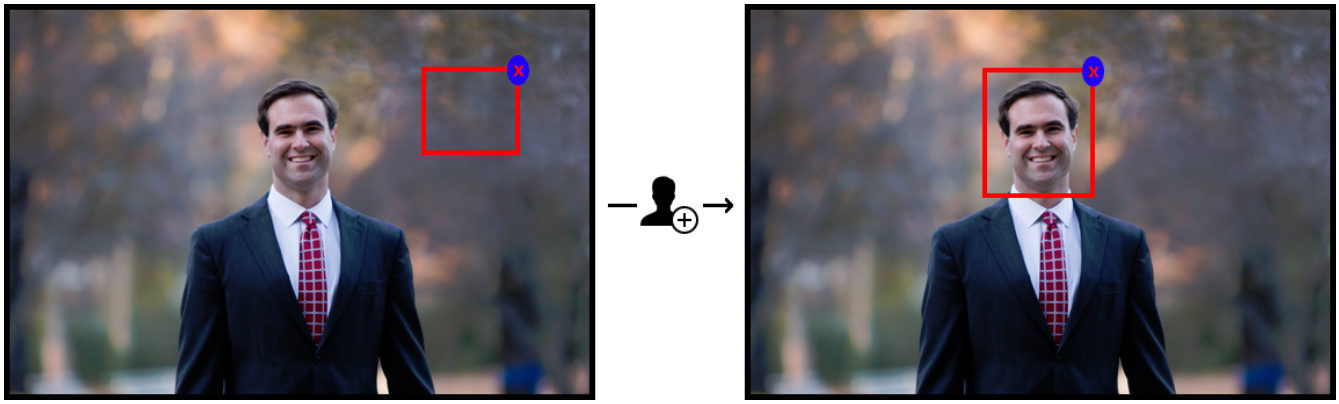

Figure 1: In the *with interaction* conditions, participants could delete existing bounding boxes or click-and-drag to create new ones. In this example, the left image shows a system error, and the right shows a version after interactive correction. [1]

2012). 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 (Lee and See 2004; Hoffman et al. 2013; Nourani et al. 2020a). In contrasting scenarios, *mistrust* (Parasuraman and Riley 1997) and *distrust* (Lee and See 2004) can cause users to rely more on themselves or under-rely on the system. 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 (Berk and Hyatt 2015).

To raise users' trust and provide more information to aid them in their decision-making process, researchers have explored the use of explainability in artificial intelligence systems (Ribeiro, Singh, and Guestrin 2016). Studies of human-in-the-loop paradigms have shown explainability can help users understand and build trust in the algorithms in order to provide proper feedback and annotations (Ghai et al. 2020; Teso and Kersting 2018).

Researchers use different methods to measure trust and reliability in machine learning and artificial intelligence systems. 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 (Nourani et al. 2020b). Yu et al. (2019) 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. (2019) 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.

[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/

Method

In this section, we discuss our research objectives based around understanding the differences in user trust among users of human-in-the-loop systems and non-interactive systems. We also present details of our experimental design and study procedure.

Research Objectives

With the goal of understanding the effects of providing human-in-the-loop feedback on user perception of artificially intelligent systems, we identified the following research questions:

RQ1: Does user trust in an intelligent system change if the user provides feedback to the system?

RQ2: Does providing feedback to an intelligent system affect user ability to detect changes in system accuracy over time?

To address these research questions, we designed a controlled experiment using a simulated image classification system both with and without feedback. With these different systems, we hypothesized that the effects of participant trust due to interaction presence would be based on system response to their feedback, similar to the effects observed in human-based decision making systems (Korsgaard, Schweiger, and Sapienza 1995; Van den Bos, Vermunt, and Wilke 1996). If the system reacted positively, improving as the participant provided feedback, we expected that participants would feel more invested in the system and as a result, they would trust the system more. However, if the system did not honor the user's feedback and did not improve after taking participant feedback, we expected that they would become negatively biased against the system.

Experimental Design

For our study, we provided participants with a series of images with classifications from a simulated model. To avoid confusion of whether a system prediction was correct or not, we chose to focus on a domain which required no prior experience, which led us to use detection of human faces as

our classification goal. With the goal of increasing participant engagement in the feedback process, we wanted to use a system with more intricate outputs than binary classification alone. Therefore, we decided to simulate a system that detected the location of human faces rather than just their presence, placing bounding boxes over each face in the image. Classifications were hand-crafted, not actually from an artificially intelligent model as participants were told. The task consisted 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 (Kuznetsova et al. 2020; Krasin et al. 2017) 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).

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 2x3 between-subjects design for the experiment. The first independent variable in our experiment was *interaction presence* with two levels: *with interaction* and *without interaction*. Participants in the *with interaction* condition were asked to provide feedback to the model for each image by correcting any errors, or by verifying that the system's classification was correct. To do this, participants could interact with the system to removing any bounding boxes from an image that did not contain a human face, and they could add new bounding boxes over any unidentified faces in the image. Participants in *with interaction* condition were explicitly instructed that their feedback would be used by the model in-between each round of images to update the model's parameters before classifying the next round of images. To maintain a feeling of realism that the model was actually updating, we added a 45 second pause between each round and told participants that they would need to wait for the model to take their feedback into account and update the predictions for future classifications.

The *without interaction* system removed the ability to interact with the bounding boxes to correct erroneous instances. In both conditions, to ensure participant engagement in this condition, we asked participants to respond whether the model's classification was correct or incorrect. Unlike the participants who saw the interactive system, participants in the *without interaction* condition were told that their responses would be sent to the researchers after the completion of the final round of images, with no indication that their responses would be used by the model in any way.

Our second independent variable was *change in accuracy*, which corresponded to the simulated accuracy of the system in each round of images with levels as shown in Table 1:

While the change in accuracy factor influenced the distribution of errors over sections of the study, it is important to note that the *total* number of system errors observed

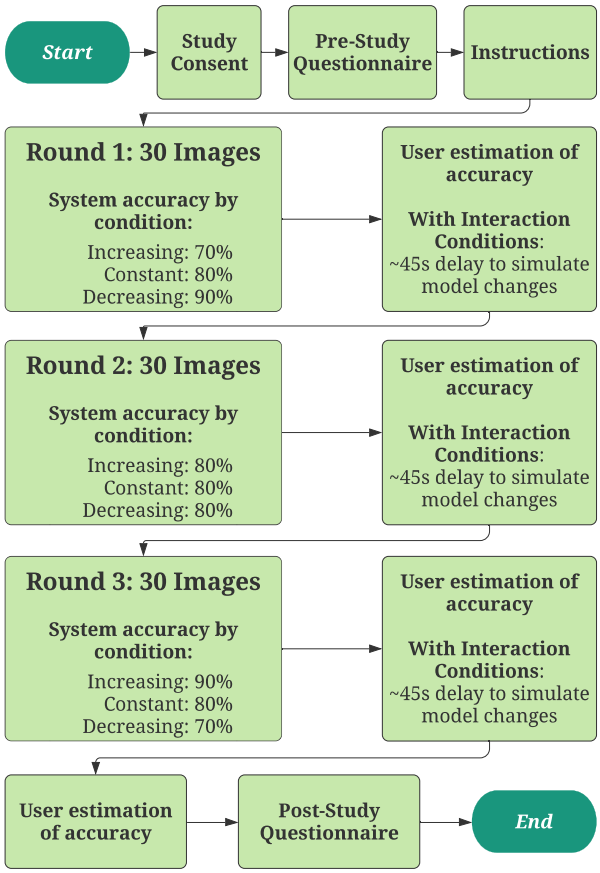

Figure 2: Study procedure overview.

by participants across the entire study was the same in all conditions—a total of 18 of 90 images were shown as classified incorrectly regardless of condition. The only difference among these conditions was when those errors were shown.

Procedure

Participants completed the experiment using an online web application without intervention or live communication with the researchers. The study began with a pre-study questionnaire, asking basic demographic information including age, gender identification, and educational background. Additionally, we asked participants to self-report their experience with machine learning and artificially intelligent systems to ensure there was no significant difference between the experience of the populations for each condition. Participants then received instructions on completing the task, including examples of correct and incorrect classifications of images.

	Round 1	**Round 2**	**Round 3**
Increasing accuracy	70%	80%	90%
Constant accuracy	80%	80%	80%
Decreasing accuracy	90%	80%	70%

Table 1: System accuracy by round.

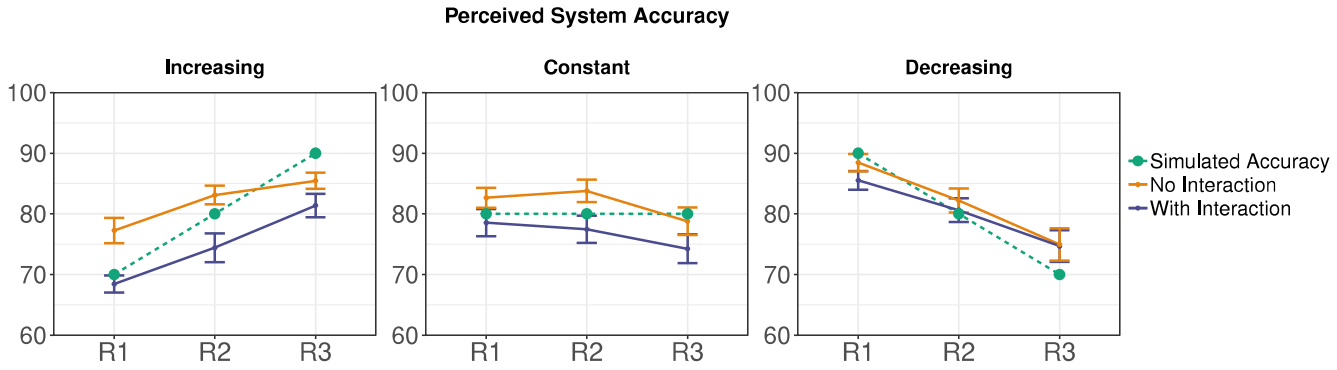

Figure 3: Perceived accuracy across rounds (error bars show standard error). Participants *with interaction* (purple) rated the system as less accurate than those with *no interaction* (yellow).

To avoid ambiguity as to what constituted a correct classification, we instructed participants to consider any bounding box that contained a portion of a human face to be correct. Additionally, when designing the outputs of the simulated model we avoided placing any bounding boxes that only partially contained a face. Participants in conditions with interaction also saw a tutorial on how to edit the bounding boxes to provide feedback to the model which reminded them that their feedback would be used to update the model between rounds.

After finishing the instructions and tutorial, participants moved on to the main task, which consisted of three rounds of reviewing 30 images with bounding boxes corresponding to the system classifications. Between each set of images, participants were asked to estimate how accurate the system was during the previous set of images. Participants in the *with interaction* conditions were required to wait for an added time delay before being able to continue to the next round (simulating the time required for the system to update based on participant feedback). A notification about the reason for this delay was also shown to remind participants that their feedback was being used dynamically (although the actual system remained static regardless of their feedback). After all rounds of images were completed, participants filled out a post-study questionnaire to evaluate their level of trust in the system.

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 24–68 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 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 157 participants, and 4 were removed based on the accuracy criteria. The remaining 153 participants consisted of 83 males, 69 females, and one non-binary response. Participants took approximately 14 minutes on average to complete the study.

Results

In this section, we present the measures of our study and 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.

User-Perceived Model Accuracy

To examine the effects of providing interactive feedback on user-perceived system accuracy, the participants numerically estimated the system accuracy after each round of image review. To account for differences in observed accuracy controlled by the *change in accuracy* factor, we analyzed estimated accuracy as the error of participant responses compared to the actual simulated accuracy. Results are shown in Figure 3. A three-way mixed-design ANOVA was performed on the error of estimated accuracy, with *change in accuracy* and *interaction presence* as between subjects factors and image set (i.e., first, second, or third round) as a within subjects factor. The analysis showed a significant effect of interaction presence on error of estimated accuracy. Participants who provided system feedback estimated the system as being less accurate than those who did not provide feedback to the model, with $F(1, 147) = 6.99$, $p < 0.01$, $\eta^2_G = 0.035$. No significant interaction effects were detected.

Additionally, the ANOVA test found participant error to be significantly different based on round with $F(2, 294) = 10.29, p < 0.001, \eta^2_G = 0.016$, as well as an interaction effect between change in accuracy and round with $F(4, 294) = 38.19$, $p < 0.001$, $\eta^2_G = 0.108$. As the simulated accuracy in each round was different based on the change in accuracy condition, these results are not surprising.

Perception of Model Change

In addition to the reported accuracy during the task, we asked participants to rate how much they thought the system

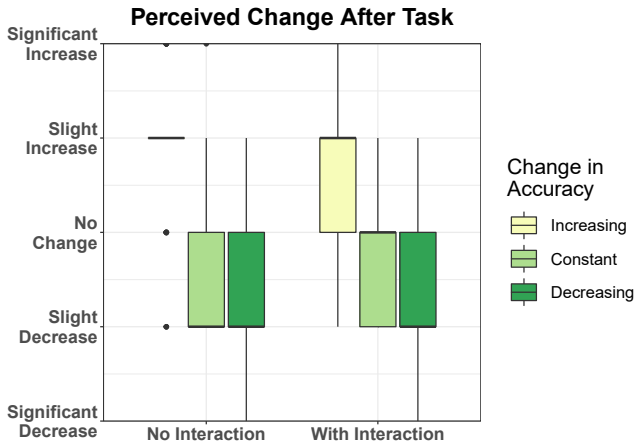

Figure 4: Perceived change in system accuracy from the first round to the last. Participants who saw an increase in accuracy reported a significantly more positive perceived change in accuracy than those who observed constant accuracy, and those who saw constant accuracy reported a significantly more positive change than those who saw decreasing accuracy.

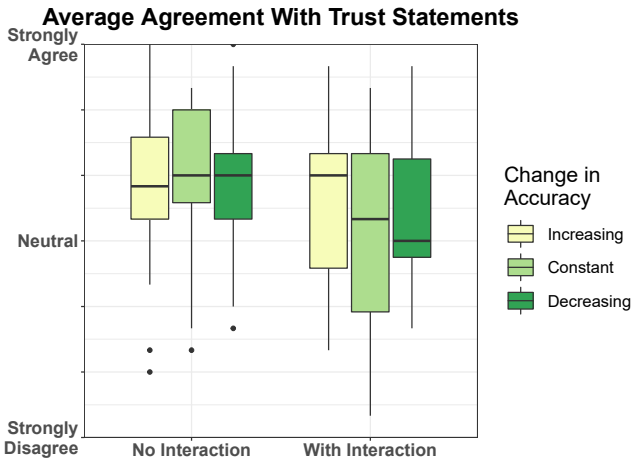

Figure 5: Average agreement with the three trust statements. Participants with interactions had significantly lower trust, regardless of their observation of change in accuracy.

had changed across the different rounds on a five-point Likert scale. Figure 4 shows the distribution of participant responses to this measure. We performed an independent two-way factorial ANOVA on participant responses that showed no significance based on interaction presence. However, we did observe that change in accuracy was significant with $F(2, 147)$, $p < 0.001$, $\eta^2_G = 0.405$. A Tukey posthoc test showed that each pair was significantly different. Participants who saw increasing accuracy rated the system as having changed significantly more positively than both constant accuracy ($p < 0.001$, $d_s = 1.366$) and decreasing accuracy ($p < 0.001$, $d_s = 1.946$). Those who saw constant accuracy thought that the system had a more positive rate of change than participants who observed decreasing accuracy ($p < 0.05$, $d_s = 0.498$).

User Trust

To measure participants' trust, we asked participants to rate their agreement using a series of scales proposed by Madsen and Gregor that focus on capturing different aspects of human-computer trust (Madsen and Gregor 2000). Participants rated each item on a seven-point Likert scale. Because the scales were developed for intelligent systems that aid in user decision making, we selected the following subset that applied most to our system. The following three statements were shown to all participants, and the aggregate rating was used as a measure for trust:

- 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.

Additionally, as a simple measure of participants' thoughts on the model updating with feedback, the following was only shown to participants in conditions with interaction:

- The system correctly uses the information I enter.

Aggregated responses for the first three trust items were analyzed with a two-way factorial ANOVA testing the effects of interaction presence and accuracy change. This test showed that the *with interaction* condition had significantly lower trust than the *without interaction* condition with $F(1, 147) = 7.61$, $p < 0.01$, and $\eta^2_G = 0.049$. The test did not detect a significant effect of *change in accuracy* on participant trust. The distribution of average participant agreement with the first three trust statements is shown in Figure 5.

The results from the fourth statement about agreement that the system correctly used their feedback are shown in Figure 6. Since this measure was only relevant and collected for participants in the *with interaction* conditions, we performed a one-way ANOVA test with *change in accuracy* as the only factor. The test revealed a significant effect with $F(2, 75) = 3.263$, $p < 0.05$, $\eta^2_G = 0.080$. From a Tukey posthoc test, participants who observed an increase in system accuracy had a higher level of agreement that the system was updating correctly than those with constant accuracy, with $p < 0.05$, $d_s = 0.725$. Thus, participants believed their feedback was being used when they corrected the image detection and observed increased accuracy over trial rounds. We did not observe any significant effect between participants who saw a decrease in system accuracy and participants in either of the other conditions.

Discussion

This section discusses the results of the experiment in the context of our research questions and hypotheses. We also consider limitations of our experiment and opportunities for further work on this subject.

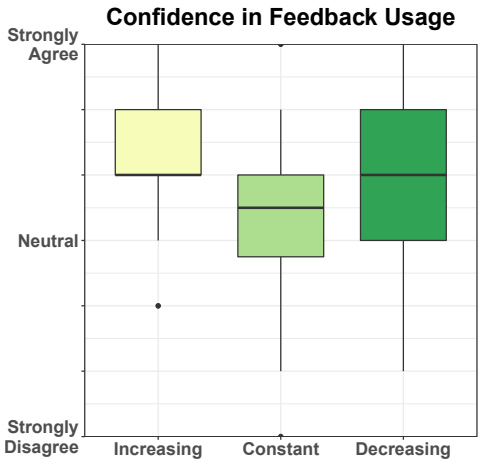

Figure 6: Participant agreement that the system correctly used their feedback. Participants with increased accuracy were significantly more confident in correctness of feedback usage than those with constant accuracy.

Interpretation of Results

Our goal for this study was to explore the effects that providing feedback to an automated system has on both user trust and perception of system accuracy. In our experiment, we controlled for both presence of interaction and change in system accuracy. We expected that participants who saw a positive response to their input—an increase in accuracy—would experience an increase in trust and perceived accuracy compared to participants who did not provide feedback. For participants who did not observe a positive response—constant accuracy or a decrease in accuracy—we expected the opposite. However, while our analysis did detect a significant effect for presence of interaction on both perceived accuracy and trust, the effect did not depend on the observed change in accuracy as expected. Rather, participants who provided feedback to the system perceived the system as less accurate and had less trust compared to those who did not provide feedback, regardless of the observed change in accuracy.

This leads us to believe that the observed decrease in user trust may be due to an increase in the salience of system errors. By correcting the mistakes made by the system, those who provided feedback spent more effort on system errors than those using the non-interactive system. Since disagreement resulted in participants taking action while agreement did not (and was therefore reviewed faster), memory of disagreements may have been reinforced in the participants' minds. That is, participants may have remembered the system's mistakes more strongly than the instances where they agreed. Participants may have also considered the act of providing feedback as an inconvenience, as correcting the system required more time and effort than simply observing whether the system was right or wrong. The increased memorability of system errors might help explain why participants who used the interactive system trusted it less than those who used the non-interactive version. This interpretation is also supported by the results of the responses to the trust questionnaire, as trust can be strongly influenced by observed accuracy (Yin, Wortman Vaughan, and Wallach 2019).

We also expected a positive change in system accuracy would correspond to higher confidence in feedback usage. We observed this effect between the increasing and constant conditions, where participants who observed an increase in system accuracy were more confident that their feedback was used correctly than those who observed a constant level of accuracy. However, no significant difference was observed when comparing participants from increasing and decreasing conditions. One possible interpretation of this result is that providing more feedback also increases perception of feedback usage. Participants in the decreasing condition provided the most feedback in the last trial of the task. As a result of the trial being the closest to the final questionnaire, these participants may have remembered giving the system more feedback than participants in the other conditions, negating the effect that seeing a decrease in accuracy could have.

Implication for Human-in-the-Loop Systems

The findings of this study highlight the importance of thoughtful design of feedback systems. While these systems can benefit from user feedback to improve model performance, they can also negatively affect trust and perception of accuracy for their users. This distrust of the system can lead to humans self-relying for critical decisions, which in many cases will result in a higher rate of human error. For instance, Parasuraman et al. (1997) report a case where train operators disabled automated alarms due to distrust, which resulted in a significant increase in accidents. Therefore, designers of human-in-the-loop systems should consider ways to avoid biasing the trust of their users who give feedback.

One potential way to reduce bias due to over-emphasizing attention to errors could be to capture user feedback implicitly. This technique is commonly found in recommender systems, where the system updates its recommendations based on prior user behavior to improve the relevance of their results. For example, users perceive search engine results as more relevant if the search engine prioritizes results that were clicked on by prior users (Shivaswamy and Joachims 2012). Similarly, Middleton et al. (2003) designed a system that recommended research papers based on identifying similar users and found that it was effective at making relevant recommendations. However, they also found that their recommendations were significantly improved when users provided explicit feedback of the topics they were interested in, suggesting that implicit feedback may not always be an adequate replacement for explicit feedback. While implicit feedback might address the issue of feedback making system errors more salient, it is important to note that systems which operate with implicit updates may open new possibilities for other forms of bias. By focusing on recommendations that are similar to those which were previously used, the scope of system recommendations for each user can become increasingly narrow over time (De Gemmis et al. 2015). Furthermore, if the user of such systems is unaware that this is hap-

pening, they can become biased by only being shown content that matches their current beliefs (Knijnenburg, Sivakumar, and Wilkinson 2016).

Another approach developers have taken to increase user trust and facilitate an accurate understanding of system accuracy is to introduce explanations of system behavior. Adding system explanations to interactive systems results in a more appropriate level of user trust, increasing trust when the system is accurate and lowering trust for inaccurate systems (Teso and Kersting 2018). Ribeiro et al. (2016) also showed that by explaining the features used by a classifier in making a prediction, users identified the system accuracy more precisely. They also found that model accuracy was improved by having users provide feedback to the explanations by removing features that they deemed unimportant. This suggests that explanations can not only help offset distrust caused by providing in human-in-the-loop systems, but that explanations may provide further interaction modes and improve the quality of feedback given.

Finally, it may be beneficial to directly make users aware of their potential biases. Wall et al. (2017) proposed a series of metrics to detect potential biases by focusing on patterns in what data has been observed by the user. This can be used to help users identify potential biases towards their trust or understanding of a system based on how much attention they have given to different types of system outputs. For example, human-in-the-loop systems could be accompanied by visual analytics tools that notify users when they spend a disproportionate amount of time and effort on a certain type of data (e.g., a system weakness). By making users aware of the potential mistrust caused by providing feedback, they may be able to adjust their level of trust accordingly.

Limitations and Future Work

This research contributed empirical evidence that providing feedback can negatively effect user trust and perception of accuracy, but our findings also motivate the need to explore different kinds of feedback systems. While our study focused on mandatory feedback to ensure that all participants engaged with the system equally, it is also possible to allow users to provide feedback optionally—an approach used in many intelligent systems. In many such systems, users only provide feedback when they are already inclined to. It is possible that removing the requirement to provide feedback could change the effects it has on user trust. Therefore, more research is needed to fully understand and compare the impacts of mandatory and optional feedback on user trust and perception of accuracy.

Another potential direction for further research in this field would be to explore different methods of collecting human feedback. Implicit feedback systems detect user behavior to update models without directly asking for input (Shivaswamy and Joachims 2012). However, with systems using this feedback technique, users might be aware that their behaviors are being recorded. It may be interesting to see how this knowledge can affect trust and whether it is similar to our findings and observations. Furthermore, it might be worthwhile to study whether this type of feedback can improve users' understanding of system accuracy and the impact of the their feedback usage for the model. Additionally, while participants in our study provided feedback for individual system outputs, explanatory interactive learning systems can also allow users to modify model features directly (Teso and Kersting 2019). The differences in these systems suggest that further research can study whether our findings extend to systems that use feature-based feedback. Along these lines, future research may also consider whether the role of algorithmic transparency and system explanation might influence the user biases and perception of accuracy. Thus, continued studies may also incorporate evaluation measures for explainable systems and understandability (Mohseni, Zarei, and Ragan 2018) as an essential element of human-in-the-loop experiences.

Conclusion

Human-in-the-loop machine learning has many benefits, including the potential for increased model performance and providing users a way to control the outcomes of autonomous systems. We conducted an experiment of the effects that providing such feedback has on users of intelligent systems in the presence of differing levels of change in accuracy over time. The results show that regardless of the actual observed changes in system performance over time, participants who provided human-in-the-loop feedback believed the system to be significantly less accurate than participants who did not provide feedback. The study also suggests participants who provided feedback trusted the system less than those who did not. Therefore, developers of autonomous systems may need to consider the effects that allowing end users to provide feedback could have on how people perceive their models.

Acknowledgements

This work was supported by the DARPA Explainable Artificial Intelligence (XAI) Program under contract number N66001-17-2-4032 and by NSF award 1900767.

References

[Berk and Hyatt 2015] Berk, R., and Hyatt, J. 2015. Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter* 27(4):222–228.

[Cho, Lee, and Hwang 2019] Cho, M.; Lee, G.; and Hwang, S.-w. 2019. 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*, 1333–1336.

[Cohn, Ghahramani, and Jordan 1996] Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4:129–145.

[De Gemmis et al. 2015] De Gemmis, M.; Lops, P.; Semeraro, G.; and Musto, C. 2015. An investigation on the serendipity problem in recommender systems. *Information Processing & Management* 51(5):695–717.

[Dietvorst, Simmons, and Massey 2018] Dietvorst, B. J.; Simmons, J. P.; and Massey, C. 2018. Overcoming

algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

[Elwell and Polikar 2011] Elwell, R., and Polikar, R. 2011. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks* 22(10):1517–1531.

[Geng and Smith-Miles 2009] Geng, X., and Smith-Miles, K. 2009. Incremental learning.

[Ghai et al. 2020] Ghai, B.; Liao, Q. V.; Zhang, Y.; Bellamy, R.; and Mueller, K. 2020. Explainable active learning (xal): An empirical study of how local explanations impact annotator experience. *arXiv preprint arXiv:2001.09219*.

[Goddard, Roudsari, and Wyatt 2012] Goddard, K.; Roudsari, A.; and Wyatt, J. C. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1):121–127.

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

[Holzinger et al. 2016] Holzinger, A.; Plass, M.; Holzinger, K.; Crişan, G. C.; Pintea, C.-M.; and Palade, V. 2016. 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*, 81–95. Springer.

[Knijnenburg, Sivakumar, and Wilkinson 2016] Knijnenburg, B. P.; Sivakumar, S.; and Wilkinson, D. 2016. Recommender systems for self-actualization. In *Proceedings of the 10th ACM Conference on Recommender Systems*, 11–14.

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

[Krasin et al. 2017] 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. 2017. 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*.

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

[Kuznetsova et al. 2020] 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. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

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

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

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

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

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

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

[Nourani et al. 2020a] Nourani, M.; Honeycutt, D. R.; Block, J. E.; Roy, C.; Rahman, T.; Ragan, E. D.; and Gogate, V. 2020a. 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*, 1–8.

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

[Nourani, King, and Ragan 2020] Nourani, M.; King, J. T.; and Ragan, E. D. 2020. 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*.

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

[Ribeiro, Singh, and Guestrin 2016] Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

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

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

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

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

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

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

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

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

[Vathoopan, Brandenbourger, and Zoitl 2016] Vathoopan, M.; Brandenbourger, B.; and Zoitl, A. 2016. 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)*, 1–4. IEEE.

[Wall et al. 2017] Wall, E.; Blaha, L. M.; Franklin, L.; and Endert, A. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 104–115. IEEE.

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

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

[Yin, Wortman Vaughan, and Wallach 2019] Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

[Yu et al. 2019] Yu, K.; Berkovsky, S.; Taib, R.; Zhou, J.; and Chen, F. 2019. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 460–468.

[Žliobaitė 2010] Žliobaitė, I. 2010. Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*.