No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML

Alison Smith-Renner,[†] Ron Fan,[‡] Melissa Birchfield,[‡] Tongshuang Wu,[‡]
Jordan Boyd-Graber,[†*] Daniel S. Weld,[‡] and Leah Findlater[‡]

[†]University of Maryland
College Park, MD, USA
amsmit@umd.edu,
jbg@umiacs.umd.edu

[‡]University of Washington
Seattle, WA, USA
rfanhu@gmail.com,
{mbirch2,wtshuang,weld,leahkf}@uw.edu

ABSTRACT
Automatically generated explanations of how machine learning (ML) models reason can help users understand and accept them. However, explanations can have unintended consequences: promoting over-reliance or undermining trust. This paper investigates how explanations shape users' perceptions of ML models with or without the ability to provide feedback to them: (1) does revealing model flaws increase users' desire to "fix" them; (2) does providing explanations cause users to believe—wrongly—that models are introspective, and will thus improve over time. Through two controlled experiments—varying model quality—we show how the combination of explanations and user feedback impacted perceptions, such as frustration and expectations of model improvement. Explanations without opportunity for feedback were frustrating with a lower quality model, while interactions between explanation and feedback for the higher quality model suggest that detailed feedback should not be requested without explanation. Users expected model correction, regardless of whether they provided feedback or received explanations.

Author Keywords
Interactive machine learning; explainable machine learning

CCS Concepts
•**Human-centered computing** → *Interactive systems and tools; Empirical studies in HCI;* •**Computing methodologies** → *Machine learning;*

INTRODUCTION
Complex machine learning (ML) models can be incomprehensible for end users who are not ML experts. While a model may have high accuracy on held-out test sets, users may also want to know *why* the model is making its predictions; if the model is right for the right reasons, they can be more confident that it will generalize or is operating without bias [18]. Automatic model explanations—such as "why" and "why not" justifications [42] and feature visualizations [33]—can provide intuition and increase user confidence and trust [48, 11], human task performance [61, 21, 54], satisfaction [10], and system acceptance [27]. Ongoing government research programs [26], focused academic conferences,[1] and recent legislation on the "right to explanation" [25] have also fueled a general push for ML system transparency.

Explanations are not an unmitigated good, however. Complex explanations may promote over-reliance when they are convincing [62] or lower user satisfaction when they are confusing [44]. Explanations that expose system uncertainty or algorithmic limitations may negatively affect users' perceptions [43, 61, 12], and users may ignore explanations entirely if the benefit to attending to them is unclear [35].

This paper investigates two additional complications of explanations. First, if explanations increase users' understanding of ML models and the errors they make, can this insight in turn increase users' desires to "fix" them and therefore reduce satisfaction if they cannot? Interactive ML allows user feedback, which *can* improve model accuracy [20, 49, 57], but not always [2, 65]. Here, explanations can help: by improving mental models, explanations can improve user feedback [33, 52]. Still, researchers have only begun to examine the relationship between explainability and interactivity in ML.

Second, intuitively, users who provide feedback should expect model improvement, but what about those who do not give feedback; might explanations also cause users to expect model improvement over time? Humans expect that others who are capable of explaining their mistakes will self-reflect and learn from those mistakes [59]. Explanations reveal why the model was incorrect for particular instances, so ML novices, in particular, may similarly expect introspective behavior and learning from the experience, even without user feedback.

To study how explanations and supports for user feedback affect users' experience with a ML model, we conducted two crowdsourced experiments with 180 participants each. Both experiments use a common classification task: is a message about "hockey" or "baseball" [57, 33]? Because we expected explanations and feedback would be particularly salient when the model could be improved, the first experiment used a

[*]Now at Google Research Zürich.

[1]ACM Conference on Fairness, Accountability, and Transparency (https://fatconference.org/)

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376624

lower quality model ($\sim$ 75% accuracy), trained on a handful of training documents. Participants reviewed predictions made by the classification model with or without *explanations*, and with one of three levels of user *feedback* to the model: none, instance-level (correcting or confirming the model's prediction), and feature-level (telling the model *how* to predict). We measured participants' subjective post-task satisfaction, including frustration and trust, as well as how they expected the model to change. The second study experiment was exactly the same as the first, but with a *higher quality* model ($\sim$ 95% accuracy) to understand the effects of model quality.

Our findings contribute the following observations to the nascent understanding of interactive and explainable machine learning: (1) users wanted the *opportunity* to provide feedback, regardless of model quality or whether they received explanations; (2) for the low-quality model, feedback reduced frustration and increased trust and acceptance, but explanations had the opposite effect; therefore, explanations without the opportunity for feedback resulted in an especially negative user experience; (3) for the high-quality model, users were not as frustrated, yet requesting feature-level feedback without an explanation reduced trust; (4) regardless of model quality, when users provided detailed feedback, they expected more improvement; yet, users generally expected model improvement even for conditions without any user feedback, demonstrating possible misconceptions of ML models by end users.

Despite the constrained setting (i.e., a classical, binary text classification task, with a simple explanation), we see this work as an important step in illustrating a key relationship between explanations and feedback. We conclude this paper by discussing extensions to more complex tasks and models with more sophisticated explanation and feedback mechanisms.

RELATED WORK

We review related work in interactive and explainable ML, separately, and then describe prior studies on their relationship.

Interactive Machine Learning

Compared to classical supervised ML's focus on static labels and datasets, interactive machine learning (IML) trains a model through rapid end *user* interaction [4]. IML commonly produces higher quality models [56, 49], personalized recommendations [6, 24] or models that are better aligned with users' understanding [29, 39]. However, user feedback can have negative effects: decreased system performance [2, 65] or inconsistent mental models [8].

This "human-in-the-loop" approach has been applied across machine learning, including in supervised ML [20, 56], unsupervised ML [7], and reinforcement learning [31, 53]. We focus on supervised ML as it provides intuitive mechanisms for non-ML expert, end user feedback, such as providing training examples [22], preferences [50], or by reacting to model predictions with instance-level (i.e., correcting or confirming predictions [20, 17]) or feature-level feedback (i.e., denoting features indicative of each class [57, 33, 49]). Our focus on end users providing feedback, and in particular feature-level feedback aligns with "Machine Teaching", where non-ML experts build models from more than just labeled data [63].

Explainable Machine Learning

Explainability (or intelligibility) in ML has received growing attention as ML models take on more important responsibilities in society. More complex models are often more accurate. Thus, intelligibility research both develops global explanations, such as more transparent models [14, 3, 36, 58] or black-box explanations [37], and local explanations of individual algorithm decisions, which can include input evidence [40, 21], localizations [55, 45], natural language explanations [13, 19, 23], or local approximations [51]. We focus on local explanations (i.e., highlighting important words).

Explanations can support fairness and bias assessments [18], improve perceived understanding [32], promote system acceptance [27], engender trust [48], and convince users to accept recommendations [16]. However, explanations can decrease users' perceptions when algorithmic limitations or uncertainty are portrayed [12, 61, 43]. Explanations can have other negative effects, such as over-reliance [62] or inability to detect mistakes [47]. ML-based systems can set expectations by exposing accuracy [66] or anticipated system mistakes [32]. This work explores whether such insight in turn increases users desire to fix mistakes and improve systems.

Prior work explores the effect of explanations on mental models, in particular on *predictability*, or the users' ability to predict model behavior [47, 15, 11], finding conflicting results. Explanations improved predictability for apartment pricing [47] and GUI customization [11], but did not have an effect for a visual question answering [15]. This discrepancy could be because that users expected the ML model to change and therefore were less successful at predicting future model behavior. Our studies measure expected change by asking users whether they think the system they evaluated will have higher, similar, or lower accuracy on new data.

Relationship of Explainability and Interactivity in ML

Users need to understand how models work [22, 5, 34] to best fix them, and how models are explained changes user feedback [52, 33]. Kulesza et al. [33] introduced EluciDebug, based on the concept of "explanatory debugging", in which a classifier explains binary predictions to users in the form of important input words and proportion of the data labeled as each class. Users in turn inform the classifier by correcting the prediction—*instance feedback*—or saying which words are important for each class—*feature feedback*. Users both better understood and corrected EluciDebug's mistakes compared to a system without explanations or feature-level feedback. While these studies tell us that explanations foster better feedback, prior work has not investigated how user perceptions—such as frustration and trust—are shaped by the presence or (sometimes more importantly) absence of IML feedback and explanations. Therefore, we address this using a similar data set, task, explanation, and feedback mechanisms.

STUDY 1: UNDERSTANDING EXPLANATIONS AND FEEDBACK WITH A LOW QUALITY MODEL

With a crowdsourced, between-subjects experiment, we explored how explanations and support for feedback affect satisfaction and expectation of change with a low quality model.

Method

Simple models and tasks are a useful starting point to examine the intersection of explanations and feedback. Therefore, in this study, participants reviewed a simple text classification model's predictions with or without explanations and with one of three options for providing user feedback to the model: no feedback, correcting or confirming the model's predictions (instance-level feedback), or suggesting important words to the model (feature-level feedback).

Task, Model, Feedback, and Explanations

We chose a simple model and task that a large population of non-ML experts could use to interact with and evaluate ML models. Specifically, we chose text classification as it is prevalent in real-world use cases, such as document recommendation and search. Borrowing from prior work [33, 57], we used a text classification algorithm to predict the category of emails from a data set of 2,000 "hockey" and "baseball" emails from the 20 Newsgroups corpus [38].

Specifically, we used a Naïve Bayes model with unigram features [41]—the multinomial Naïve Bayes (`MultinomialNB`) classifier from the `scikit-learn` library [46]. We performed standard pre-processing procedures on the emails.[2] For this experiment, we trained the classifier on 16 (of the 1197) labeled training emails (eight per class), with 76.5% classification accuracy on the 796 emails in the test set.

The participants were to imagine that they had been assigned to sort their boss's email inbox. They were told that they would evaluate a ML model designed to help them. Would it be worthwhile to add the model to their workflow?

For this task, we built an interface where participants review emails with the model's "hockey" or "baseball" prediction (Figure 1).[3] The interface either displays an explanation of the model's prediction (or not) and supports either no user feedback, feature-level feedback, or instance-level feedback.

Our explanations tell users what the model regards as important for prediction: we highlight the three words that are most influential to the class prediction for a given email, $abs(p(w|baseball) - p(w|hockey))$. This method is purposefully simple and truthful to the classifier's methodology, two guidelines for good explanations [35, 44]. Additionally, we choose exactly three words as explanations should include *sufficient*, but not *extra*, low-level context [52].

For *instance-level* feedback, participants correct or confirm each classification by telling the model whether the email is about "hockey" or "baseball." For *feature-level* feedback, participants tell the model what should be important by providing the top three words they think would be most useful in classifying a given email and specifying the class with which those words should be associated.

[2]We removed non-alphabetical characters, lowercased all words, tokenized by whitespace, and dropped *From:* lines from the emails to prevent the model from training on email addresses.

[3]https://github.com/rococode/bh-classifier

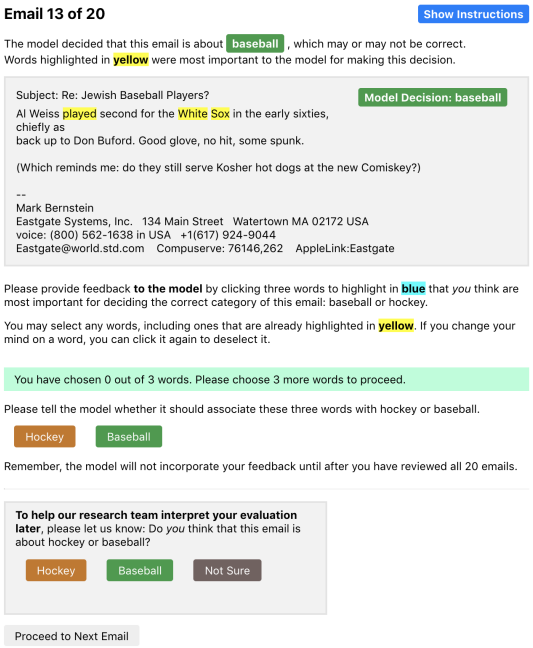

Figure 1. Screenshot of an email in the "interaction phase" for a participant in the feature-level feedback and explanation condition (E-F).

Participants

We recruited 180 unique participants (77 male, 102 female, and one unspecified) from Mechanical Turk,[4] requiring participants with the "Masters" qualification, located in the United States, and having completed more than 500 HITs with approval rate 98% or higher. Two participants were 18–24 years old, 62 aged 25–34, 60 aged 35–44, 30 aged 45–54, 22 aged 55–64, and 4 aged 65–74. Participants rated their prior knowledge on five-point Likert scales for ML (65 had none, 67 had a little, 44 had some, four had a lot, and none had expert), hockey (15 had none, 78 had a little, 65 had some, 18 had a lot, and four had expert), and baseball (two had none, 43 had a little, 68 had some, 57 had a lot, and 10 had expert).

Procedure

Remote study sessions took on average 22.6 minutes ($SD = 15.3$). Participants completed three phases: (1) introduction, (2) "interaction" with the model, and (3) "evaluation" of the model. To motivate quality work, participants were told that at least the top 50% of participants would be given a $2 bonus based on the thoroughness of their evaluations; unbeknownst to them, all ultimately received the bonus.

During the "interaction phase", participants reviewed 20 emails,[5] in randomized order per participant. The model provided a prediction ("hockey" or "baseball") for each email. Participants in the *explanation* conditions saw the model's top three words highlighted. Participants in the *instance-level feedback* conditions corrected or confirmed the model's prediction for each email, and participants in the *feature-level*

[4]http://mturk.com

[5]We randomly select these 20 emails for Study 1, requiring even distribution between hockey and baseball predictions, five incorrect and 15 correct, and that emails be between 30 and 120 characters; we use the same set of emails for Study 2.

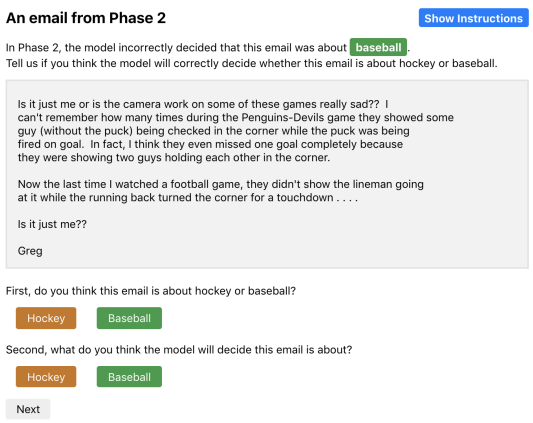

Figure 2. Screenshot of an email in the "evaluation phase," where participants predicted how the model would label an email that it had previously labeled incorrectly in the "interaction phase."

feedback conditions specified their three important words for predicting the correct class. To determine whether participants knew the correct labels, as this might affect their evaluation, all participants also told us (not the model) the correct email label, with an option for "not sure".

During the "evaluation phase", participants responded to closed- and open-ended questions on satisfaction and model change expectations, including rating scales as shown in Table 1 paired with the follow up of "Why do you feel this way".[6] After these questions, participants were shown four "evaluation" emails and asked to predict how the model would classify them (Figure 2). These emails included two of the 20 from the "interaction phase" and two new ones that were similar to emails in the first 20, as measured by cosine similarity [30]. For each email type (repeat or similar), we selected one that was previously labeled correctly and one that was previously labeled incorrectly by the model. These four emails allowed us to assess whether participants would expect the model's labels to change following the "interaction phase".

Importantly, feature- and instance-level feedback was **not** incorporated into the model during the "interaction phase"; we reminded the feature- and instance-level participants of this with each email. This design choice isolates perceptions of explanations and feedback from how well the model incorporated that feedback. Instead, we told these participants that their feedback would be incorporated into the model *after* they had reviewed all 20 emails, so they would expect an updated model for the "evaluation phase".[7]

Study Design

This study used a 2×3 between-subjects experimental design, with factors of *Explanation*—feature (E), none (N)—and *Feedback*—feature (F), instance (I), none (N). An equal number of participants were randomly assigned to each condition.

[6] An additional two rating scales of acceptable accuracy and expectations of learning are not reported on here due to space constraints and not being as directly related to our research questions.

[7] However, we never incorporate feedback during the study protocol, but users were unaware as we did not show model predictions or explanations during the "evaluation phase."

Measure	Statement
frustration	"I would feel frustrated if I were to use this model to automatically sort my boss's emails"
trust	"I would trust this model to correctly categorize my boss's emails that are about hockey or baseball"
accuracy	"The model is able to distinguish between hockey and baseball emails"
understanding	"I understand how this model makes decisions"
acceptance	"I would use this model to help me sort my boss's emails"
feedback importance	"If I were to use this model, it would be important to have the ability to provide feedback to improve it"
expected change	"If the model were now shown another set of emails, how well do you think it would categorize them?"

Table 1. Seven-point rating scale statements for seven subjective measures. All are on a scale from *strongly disagree* to *strongly agree* aside from expected change, which is from *much worse* to *much better.*

Measures and Hypotheses

We report on seven main subjective measures, collected using seven-point rating scales (Table 1): three *user satisfaction* measures (frustration, trust, model acceptance), three *user perception* measures (expected model improvement, perceived model accuracy, perceived understanding of how the model works), and *desire to provide feedback* (feedback importance).

While we explore the effects of feedback and explanation on user satisfaction in general, our primary user satisfaction hypothesis relates to **frustration**, as we hypothesize that users are frustrated without the ability to fix model errors exposed by explanations.

- **H1.1**: Feedback (instance- or feature-level) reduces frustration compared to no feedback.
- **H1.2**: Explanations without feedback increase frustration compared to no explanation without feedback.

While prior work has explored effects of explanation on mental models and perceptions of quality [9, 42], we explore a new concept, **expected improvement**, or how users expect ML models to improve with or without explicit feedback. Intuitively, providing feedback should increase this expectation. Based on human behavior [59], we also hypothesize that explanations might suggest a model is being introspective and could therefore *learn from its mistakes*.

- **H2.1**: Feedback (instance- or feature-level) increases the user's expectation that the model will improve compared to no feedback.
- **H2.2**: Explanations increase the user's expectation that the model will improve compared to no explanation.

Data and Analysis

After disqualifying one participant who only filled out the demographics survey and another who skipped part of the post-task survey, our dataset includes 178 participants. We used separate 2×3 (*Explanation*×*Feedback*) ANOVAs with Aligned Rank Transforms for each main subjective measure—a test more appropriate for Likert scale data than a standard ANOVA [64]. For significant main effects of feedback we used post-hoc Wilcoxon rank sum tests with continuity correction and Holm-Bonferroni adjustments. We report on all significant results, including pairwise comparisons.

We qualitatively coded the open-ended responses related to our primary measures: frustration and expected improvement.

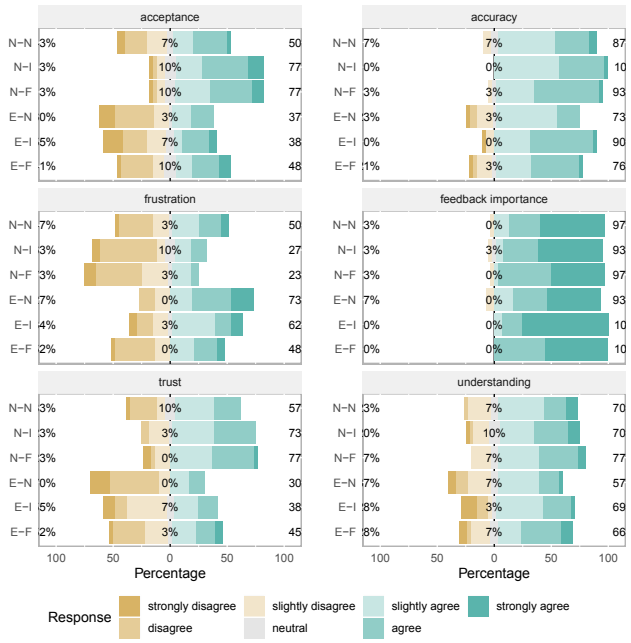

Figure 3. Study 1 seven-point rating scale responses for the main subjective measures (except expected change) from "strongly disagree" to "strongly agree". Responses reported by condition. For each measure, no explanation (N-) conditions are on the top (-N is with no feedback, -I is with instance-level feedback, and -F is with feature-level feedback) and feature explanation (E-) conditions are below Feedback (-I, -F) positively, and explanation (E-) negatively impact satisfaction measures (left).

Two annotators individually read a subset of the responses to identify emergent codes, followed by a discussion period to generate a codebook. Then, the two annotators independently coded a random subset of 20 of the 178 responses; agreement was scored using Cohen's κ: $\kappa = .93$ (raw agreement: 95%) for frustration responses and $\kappa = .88$ (90%) for expected improvement responses. We refer to participants in this experiment with a lower quality model as LP1–LP178.

Results

Figures 3 and 4 show the rating scale responses for the seven main subjective measures by condition. Participants expected the model to improve, and they expected more improvement with feedback. Participants also thought the ability to provide feedback was important. Explanations hurt subjective satisfaction (frustration, trust, and acceptance ratings), while feedback helped. Participants were commonly frustrated by the model's low quality, and this was accentuated by explanations.

To judge user comfort with the task and dataset, we asked participants to tell us (i.e., the researchers . . . not the model) whether they thought each email was about hockey, baseball, or whether they were unsure. Participants did well: 91% of the 3,580 answers reported to us were correct, while 8% were "not sure" and only 1% were incorrect. In the following sections, we provide detailed results regarding satisfaction, expectations, perceptions, feedback quality, and users' desire to provide feedback.

User Satisfaction

Participants were neutral on average, but with high variability across conditions, for each of the user satisfaction measures: frustration ($M = 3.9$ of 7, $SD = 1.8$), trust ($M = 4.1, SD = 1.7$), and whether they would use the system (acceptance) ($M = 4.3, SD = 1.9$). Feedback significantly improved satisfaction, but explanations hampered it. Open-ended responses suggest that the low model quality—highlighted by explanations—frustrated participants.

Explanations increased frustration, while support for feedback reduced it. Participants who received explanations were more frustrated than those who did not; this difference was significant (main effect of *Explanation*: $F_{1,172} = 20.05, p < .001$). *Feedback* also significantly impacted frustration (main effect: $F_{2,172} = 7.92, p < .001$). Posthoc pairwise comparisons showed that no feedback resulted in significantly higher frustration than instance-level and feature-level feedback (both comparisons $p < .05$); this supports **H1.1** for frustration, which stated that feedback would reduce frustration. The interaction between *Explanation* and *Feedback* was not significant ($F_{2,172} = .06, p = .094$); thus, **H1.2** is only partially supported by the main effect of *Feedback*.

Many participants were frustrated by low quality, which was highlighted by explanations. We coded participants' open-ended reasons for their frustration ratings, resulting in six codes. Participants felt the model was: "not good enough" (40% of the 178), or "good enough" (27%), would help "save time" (13%), would "require user review" of the decisions (11%), is "able to improve" (3%), or "other" reasons (6%).

Confirming the rating scale data, more participants with explanations (81% of 89) thought the model was "not good enough" compared to those who did not get explanations (only 26% of 89). Participants who got explanations (E-) often expressed their frustration in terms of the important words, e.g., *"I don't think it highlighted the best words in many cases"* (LP3, E-I), while those who did not see explanations (N-) were more likely to comment on the model's shortcomings in terms of accuracy, *"it made too many mistakes"* (LP175, N-N).

Less frustrated participants felt the model was "good enough" or would "save time", saying, for example, *"it would be much easier than sorting through them myself"* LP132 (E-N).

Trust and acceptance were reduced by explanations and increased by feedback. Reflecting the frustration findings, trust was significantly impacted by *Explanation* ($F_{1,172} = 14.57, p < .001$); participants who received explanations trusted the model less those who did not. There was also a significant main effect of *Feedback* on trust ($F_{2,172} = 4.27, p = .015$). Posthoc pairwise comparisons showed that both instance- and feature-level feedback increased trust compared to none (both comparisons $p < .05$). The *Explanation* × *Feedback* interaction was not significant ($F_{2,172} = .15, p = .863$).

Similarly, *Explanation* significantly impacted acceptance ($F_{1,172} = 19.49, p < .001$), where participants who saw explanations accepted the model less than those who did not. *Feedback* also significantly impacted acceptance ($F_{2,172} = 3.76, p = .025$). Posthoc pairwise comparisons showed that

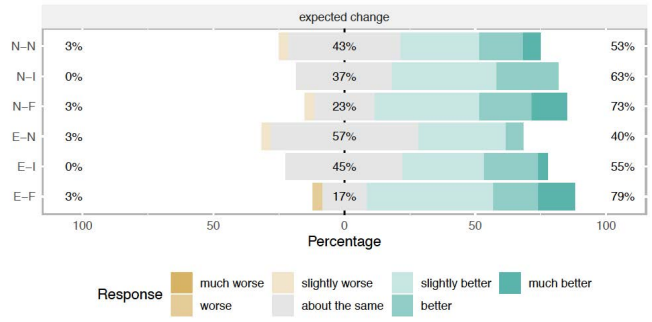

Figure 4. Study 1 participant responses for the subjective expected change measure by condition. Participants expected the model to improve. (Figure 3 describes *y*-axis labels.)

feature-level feedback resulted in higher model acceptance compared to none ($p < .05$). The interaction between *Explanation* and *Feedback* was not significant ($F_{2,172} = .97, p = .38$).

User Expectations for and Perceptions of the Model

Participants provided subjective ratings of their perceptions and expectations (Figure 3 and 4). On average, they expected the model to improve ($M = 5.2, SD = .9$), thought it worked fairly well ($M = 5.2$, $SD = 1.1$), and were neutral regarding whether they understood how it works ($M = 4.7$, $SD = 1.6$). We also examined expectations through participants' *simulated model predictions*: how they thought the model would label the four evaluation emails at the end of the study. As detailed below, feedback caused participants to think the model was more accurate and would improve, but explanation did not. Moreover, some participants who did not provide feedback thought the model would *self-correct*.

Feature-level feedback increased expected improvement compared to no feedback. *Feedback* significantly increased users' expectations ($F_{2,172} = 5.29, p = .006$); post-hoc comparisons showed that feature-level feedback raised expected improvement compared to no feedback ($p < .05$), partially supporting **H2.1**. The main effect of *Explanation* was not significant ($F_{1,172} = 1.28, p = .259$; opposing **H2.2**), nor was the *Explanation* $\times$ *Feedback* interaction ($F_{2,172} = .42, p = .656$).

A substantial portion of participants expected model corrections, even without feedback. Across all three feedback conditions, about half or more of participants expected the model to improve (rating > 4): 76.3% of 59 who had feature-level feedback, 59.3% of 59 who had instance-level feedback, and even 47% of 60 who had no feedback.

Participants' predictions about what the model would do with the four same/similar "evaluation" emails reflected this strong expectation of improvement. For the previously correct emails, participants thought the model would now be incorrect in only 4 of 712 instances, and each of these was for a "similar" email rather than the email that was exactly the "same" as in the initial set of 20.

For the previously incorrect emails, "similar" and "same" follow a similar pattern, so we focus on the "same" email to provide a straightforward assessment of whether participants think the model will improve. Most (82%) of participants who provided feedback ($N = 118$) thought the model would get the previously incorrect email correct (Table 2), which is not surprising given that they had spent time trying to improve the model. More surprising, however, is that 53% of participants in the no feedback condition ($N = 60$) thought the model would somehow correct itself.

	Feedback		
Explanation	None	Instance	Feature
None	63%	80%	90%
Feature	43%	86%	73%

Table 2. Percentage of Study 1 participants (*N*=178) by condition during the "evaluation phase" who thought the model would now correctly label an email it had previously labeled incorrectly. Many participants in the *no feedback* conditions thought the model would self correct.

Participants described the model improving from their feedback or learning from its mistakes. We coded participants open-ended reasons for their expected change ratings, resulting in nine codes. Participants felt the model would improve with "feedback" (29%), was capable of "self learning" (20%), was "high quality" (5%), or showed "evidence of improvement" (1%). Those who felt it would not improve cited that it received "inadequate feedback" (14%), showed "no evidence of improvement" (11%), had "nothing to learn from" (6%) or was of "low quality" (5%). And, 9% of participants gave "other" reasons.

Interestingly, of the 60 participants who did not provide feedback (-N), 17 (28%) still expected the model to learn from its mistakes, such as, *"it would take what it did wrong, learn from it, and apply it in future trials"* (LP141, N-N), or reported other misconceptions, including, *"these programs get better as they function and learn algorithms"* (LP154, E-N). In fact, only 13% of the 60 participants who did not provide feedback *correctly* identified that the model would not improve as it had "nothing to learn from", like, *"if it still used the same words to try to identify the correct sports emails, then it would still make the same amount of errors"* (LP87, E-N).

Feature-level feedback reduced perceived accuracy compared to no feedback. Overall, participants thought the system worked fairly well, giving it an average accuracy rating across all conditions of 5.2 out of 7 ($SD = 1.1$). However, counter to our other user experience measures, feature-level feedback had a negative effect on perceived accuracy. There was a significant main effect of *Feedback* on perceived accuracy ($F_{2,172} = 4.72, p = .010$), with posthoc pairwise comparisons showing that feature-level feedback reduced perceived accuracy compared to no feedback ($p < .05$). Neither the main effect of *Explanation* nor the *Explanation* $\times$ *Feedback* interaction effect were significant (respectively: $F_{1,172} = 1.59, p = .209$; $F_{2,172} = 2.20, p = .114$).

Quality of and Desire for User Feedback

Participants thought being able to provide feedback was important ($M = 6.4$ out of 7, $SD = .9$), regardless of condition (Figure 3); there were no significant main or interaction effects on this measure. However, do the experimental conditions impact feedback *quality*? To answer this question, we applied participants' feedback to the model after the study.

Feedback improved the model, regardless of explanation. We incorporated instance-level feedback by including the 20 emails labeled by the participant as additional training emails. To incorporate the feature-level feedback, we adjusted the classifier's weight for each word provided by the participant: the word weight was both increased by 20% for the specified class and decreased by 20% for the opposite class.

The feature-updated models were 86.2% accurate on average ($SD = 2.7\%$), which is a 9.7 percentage point improvement over the initial low quality model. In comparison, the instance-updated models were 83.6% accurate ($SD = 1.4\%$)—a 7.1 percentage point improvement. Instance and feature model improvements were similar regardless of whether the participants saw an explanation (difference in accuracy $< .2\%$).

Participants did not agree with the words the model thought were important. The 59 participants who gave feature-level feedback highlighted a total of 3,533 words. Regardless of whether explanations were shown or not, we compared the model's top three words for each email (i.e., the words the model would have highlighted) to the three words selected by the participant. Most (76.9%) of the participants' words were not in the model's top set. This disagreement is likely due both to the model's low quality and because the explanation method can highlight words that are probable for the non-predicted class (see Limitations). Participants with explanations were more likely to reuse the model's words (28% of selected words overlapped with the model's) than the 30 participants who did not see explanations (21%).

Summary

Explanations significantly increased frustration, while feedback—especially feature-level—significantly decreased it (partial support for **H1.1** and **H1.2**). There were similar patterns for other user satisfaction measures (trust and acceptance). Therefore, the worst combination was explanation without feedback, and the best was no explanation with feedback. Open-ended responses suggested that frustration was primarily due to the low model quality exposed by explanations and *not* inability to provide feedback, as we had hypothesized. Although ability to provide feedback did temper some of the frustration. This general dislike for explanations confirms prior work where user perceptions were negatively impacted by explanations that exposed flaws and limitations [12]. While this may seem inconsistent with our hypothesis at first blush, an alternate interpretation is that explanations can improve satisfaction *so long as users have a means for feedback*.

Feedback also significantly increased expectations of model improvement, as hypothesized in **H2.1**, but particularly for feature-level feedback opposed to none. Explanation did not impact expected change, in contrast to **H2.2**. Also, somewhat surprisingly, most participants expected the model to improve, including many who had not provided feedback.

STUDY 2: UNDERSTANDING EXPLANATIONS AND FEEDBACK WITH A HIGH QUALITY MODEL

In Study 1, expectations rose with feedback but not explanations and satisfaction fell with explanations but rose with feedback. As the Study 1 model's low quality appeared to overwhelm participants' subjective ratings, an additional study had a higher quality model. While we expected participants to be more satisfied with the higher quality model (e.g., observed and stated model accuracy can affect users' trust [66]), we retained the Study 1 hypotheses regarding our primary measures (frustration and expected change).

Method

This experiment was exactly the same as Study 1 with the exceptions described here. We trained the `MultinomialNB` classifier on 200 labeled training emails (100 from each class), with 94.4% accuracy on the test set. This model predicted the correct label for 18 of the 20 emails in the interaction phase. As in Study 1, we chose four emails for the evaluation phase (two "same" and two "similar"), but because of the higher accuracy of the model in Study 2 there were no available emails that were "similar" to ones the model labeled incorrectly in the evaluation phase; thus, both of the "similar" emails were similar to previously correct ones.

As in Study 1, we recruited 180 participants (99 female, 78 male, 3 unspecified). Two participants were aged 18–24 years old, 46 aged 25–34, 66 aged 35–44, 43 aged 45–54, 16 aged 55–64, and 6 aged 65–74. Participants had varied prior knowledge of machine learning (63 participants had none, 65 had a little, 50 had some, two had a lot, and none had expert), hockey (23 had none, 64 had a little, 58 had some, 25 had a lot, and none had expert), and baseball (12 had none, 37 had a little, 66 had some, 54 had a lot, and 11 had expert).

Study sessions took on average 22.8 minutes ($SD = 14.6$), and we used the same measures and data analyses as in Study 1. Our dataset included all 180 participants. We used the Study 1 codes to code the open-ended responses for frustration and expected change. We refer to participants as HP1–HP180.

Results

Figure 5 and 6 show the rating responses for the seven main subjective measures by condition. Overall, participants were less frustrated with the high quality model than the low quality one (Figure 3). The interaction between explanation and feedback was significant for other subjective measures: trust and acceptance. As in Study 1, feedback impacted expected change but explanation did not, and participants expected the model to improve and wanted the ability to provide feedback.

Regarding task difficulty, participants again performed well: 92% of their 3,600 answers to us were correct, while 7% were "not sure" and only 1% were incorrect. We provide detailed results regarding satisfaction, expectations and perceptions, and quality and desire for feedback in the following sections.

User Satisfaction

Overall, frustration was lower ($M = 2.64$ of 7, $SD = 1.54$) compared to the low quality model in Study 1 ($M = 3.90$, $SD = 1.85$). Perhaps accordingly, there were no significant main or interaction effects on frustration. Open-ended responses suggest explanations exposed the high quality model's *good* behavior. Trust and acceptance ratings were also relatively high compared to Study 1: 5.1 out of 7 on average for trust ($SD = 1.5$) and 5.4 for acceptance ($SD = 1.5$) here

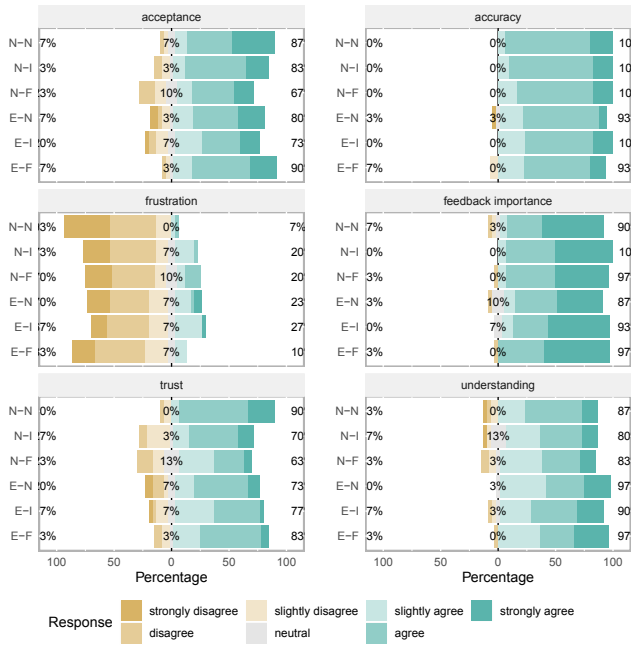

Figure 5. Study 2 responses by condition for the main subjective measures (except expected change). Participants were more satisfied, but trust suggests nuance (e.g., E-N vs. N-N, without feedback, explanation has a negative impact). (Figure 3 describes *y*-axis labels.)

compared to 4.1 for trust ($SD = 1.7$) and 4.3 for acceptance ($SD = 1.9$) in Study 1. The interaction between explanations and feedback on these measures was significant.

Trust and acceptance were affected by the combination of explanations and feedback. Neither explanation nor feedback had a clear effect on trust; the main effects of *Feedback* ($F_{2,174} = 2.59, p = .078$) and *Explanation* ($F_{1,174} = 2.00, p = .159$) were not significant. However, the interaction between *Explanation* and *Feedback* was significant ($F_{2,174} = 5.69, p = .004$), meaning that certain combinations of explanations and feedback impact trust.

From the responses (Figure 5), when feature-level feedback is requested, not providing an explanation might decrease trust (N-N compared to N-F). And, without feedback, explanation might decrease trust (N-N compared to E-N). After a Holm-Bonferroni correction, only the former posthoc pairwise comparison was significant: participants trusted the model more with neither feedback nor explanation compared to a model with feedback but no explanation ($p < .05$).

Acceptance shows a similar pattern: while there is no clear effect of either explanation or feedback, some combinations do; the *Explanation* × *Feedback* interaction was significant ($F_{2,174} = 4.11, p = .018$), while the main effects of *Feedback* ($F_{2,174} = 1.23, p = .295$) and *Explanation* ($F_{1,174} = .036, p = .850$) were not. While Figure 5 shows similar trends for acceptance as for trust, no posthoc pairwise comparisons were significant after a Bonferroni correction, so further work is needed to explore this relationship.

Explanations may have shown participants that the model was behaving properly. Participants gave lower frustration

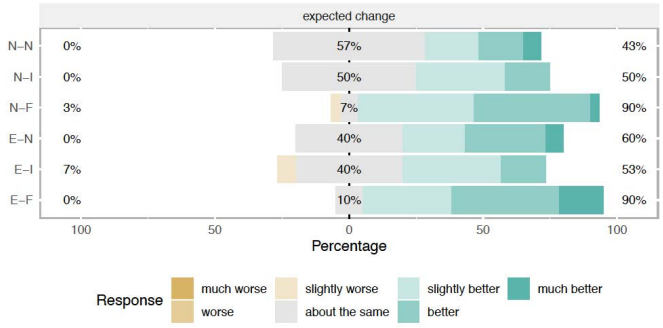

Figure 6. Study 2 responses for the expected change measure by condition, showing that in general participants expected improvements (green bars), but more in feature-level feedback conditions (E-F and N-F). (See Figure 3 for a description of y-axis labels).

ratings than in Study 1 (Figure 6); they said the model was "good enough" (49% of all participants) or would "save time" (23%). Only 15% of participants felt the model was "not good enough", that is, not of an acceptable accuracy for the task.

In Study 1, explanations exposed issues with the model's highlighted words, resulting in 81% of the 89 participants who had received explanations in that study thinking the model was "not good enough." Study 2 responses were the opposite: 80% (of 90) participants who saw explanations thought the model was "good enough", and explicitly described good model behavior, such as *"...I was able to see the reasoning from the machine and I agreed with it most of the time"* (HP139, E-F). For Study 2 participants who did not see explanations, only 65% (of 90) felt the model was "good enough", emphasizing how explanations can improve perceptions of model quality with a higher quality model.

User Expectations for and Perceptions of the Model

Figure 5 and 6 show responses for subjective rating scales regarding expectations and perceptions of the model. On average, participants expected improvement ($M = 5.0$, $SD = 1.0$), thought they understood the model ($M = 5.5$, $SD = 1.2$), and thought it worked well ($M = 5.9$, $SD = .76$).

As detailed below, feature-level feedback caused participants to think the model would improve, and explanation yielded higher perceived understanding. Neither explanation nor feedback had an impact on perceived accuracy. Open-ended responses suggest misconceptions regarding how ML models evolve, providing further explanation for why a substantial portion of participants, regardless of condition, expected the model to improve (Figure 6).[8]

Feature-level feedback increased expected improvement. As in Study 1, *Feedback* significantly impacted expected change ($F_{2,174} = 15.84, p < .001$). Posthoc pairwise comparisons showed that feature-level feedback resulted in higher expected improvement than instance feedback or none (both comparisons $p < .05$). *Explanation* did not have a significant impact on expected change ($F_{1,174} = .79, p = .375$) nor did the *Explanation* × *Feedback* interaction ($F_{2,174} = 1.41, p = .246$).

[8]We do not report on participants' simulated model predictions due to space and because trends are in line with the rating data.

Participants described misconceptions for how ML changes over time. Participants gave similar reasons for expecting model change as in Study 1. 27% of all participants credited the "feedback" they provided while 19% suggested the model was "self learning." Many participants noted similar misconceptions, including, *"my understanding is these sorts of things just get better at what they do the more they do them"* (HP84, E-N) and, *"it learns with each new experience, and I choose the word 'experience' intentionally as the machine gains consciousness"* (HP62, N-I).

Similar to Study 1, 21 (12%) participants thought their feedback was "inadequate" (whether in quality or quantity). Of these, 17 provided instance-level feedback (compared to three who provided no feedback and three who provided feature-level feedback), and suggested that they would have preferred to tell the model *why* it was wrong. For example, HP128 (E-I) said, *"simply telling it that it was wrong may make it less accurate, but it is unlikely to make it more accurate without knowing how it made its mistake."*

Explanations increased perceived understanding. *Explanation* significantly impacted perceived understanding ($F_{1,174} = 3.92, p - 0.49$). Participants thought they understood the model more when given an explanation (Figure 5). Neither the main effect of *Feedback* ($F_{2,174} = .13, p = .876$) nor the *Explanation* $\times$ *Feedback* interaction effect were significant ($F_{2,174} = .53, p = .591$).

Quality and Desire for User Feedback

Like in Study 1, participants wanted the ability to provide feedback ($M = 6.3$ of 7, $SD = 1.0$), regardless of condition (Figure 5). There were no significant main or interaction effects on this measure. But how useful is their feedback for the high quality model?

Feedback provided only minor improvement. We incorporated participant's feature-level and instance-level feedback into the model. While the updated models in Study 1 greatly improved, in Study 2 they did not. The feature updated models averaged 95.8% accuracy ($SD = .8\%$), only a 1.4 percentage point improvement over the initial high quality model. The instance updated models had 95.1% accuracy ($SD = .5\%$; a .7 percentage point improvement). As in Study 1, instance and feature model improvements were similar regardless of whether the participants saw an explanation (difference in accuracy $< .2$).

Participants agreed more with the high quality model's words. Participants provided 3,589 words as feature-level feedback. Participants were similarly likely to provide new words (1,942) as reuse model words (1,647), unlike Study 1 participants who reused less than 25% of the model's words. The 30 participants shown explanations reused provided words (52% overlap of their words to the model's important words), more than the 30 who did not see explanations (40%).

Summary

Neither feedback nor explanation impacted frustration, which was generally lower than in Study 1. For other user experience measures, there were no main effects either, although significant interaction effects on trust and acceptance suggest nuance in how explanations and feedback impact each other. As with the low quality model, feature feedback significantly increased expected change, this time over both instance and no feedback (confirming partial support for **H2.1**), but explanation did not have an effect. Again, participants generally thought the model would improve.

DISCUSSION

We relate our findings to prior work and provide design recommendations for interactive and explainable ML systems. We also discuss limitations and extensions to more complex tasks, models, explanations, and feedback mechanisms.

Users want the opportunity to provide feedback, and in particular, provide more than just labels. In both studies and all conditions, participants felt strongly that the *opportunity* to provide feedback was important; however, this does not tell us how often or whether users will provide such feedback in practice. Although, successful commercial projects, such Common Voice,[9] exemplify that users might be willing to spend time improving models.

Our studies provide additional evidence for how different levels of feedback impact user behavior and subjective response. In particular, we confirmed Amershi et al.'s [4] recommendation that "people naturally want to provide more than just data labels" to ML models. With both the low and high quality models, only those participants who told the model what words were important (i.e., provided feature-level feedback) and not those who corrected or confirmed the model's predictions (i.e., instance-level), expected the model to improve more than participants in the no feedback condition. Similarly, some participants who provided instance-level feedback described their feedback as inadequate in open-ended responses. Finally, not only was feature-level feedback better received by participants, for the low quality model it also improved accuracy more than instance-level feedback. This ability of non-ML expert participants to improve the models in our study beyond just labeling data supports the goals of machine teaching [63].

Explanations can reveal model flaws, which users desire to fix. Displaying uncertainty scores for model predictions negatively impacts users' perceptions [43]; similarly, for the low quality model, explanations were frustrating, precisely because they exposed flaws, including *uncertainty* in the model's reasoning. Because feedback reduced frustration, the most frustrating combination of explanations and feedback for the low-quality model was thus a situation with explanations but no opportunity for feedback. Indeed, no explanations and no feedback may be the least frustrating design option; however, this combination would inherently limit the model's *potential* performance, and likely result in disuse over time. In such cases, explanations provide insight to how to solve model errors [33]. Therefore, for similar models and tasks, when the model quality is low, feedback should be supported alongside explanations.

Explanations and feedback complement each other. For the high quality model, explanations increased understanding and may have exposed model strengths. But, models are rarely

[9] https://voice.mozilla.org/en

perfect, and participants wanted the opportunity to provide feedback to improve models. Therefore, providing explanations without means for feedback may reduce satisfaction. Future work should explore this relationship between explanations and feedback in more detail. Feedback alone is not always positive either: asking participants for feature-level feedback without providing explanations reduced trust compared to when explanations were provided. Users may not want to provide detailed feedback without understanding why it is needed or how best to help the model. Therefore, to improve satisfaction, similar systems should neither request detailed feedback without explanation nor provide explanation without some means for feedback.

Preconceived ML expectations should be managed. Whether from prior experience or general misunderstanding, users may have misconceptions about whether and how much models can improve. In our experiments, many participants expected the model to improve regardless of whether they provided feedback. Open-ended responses provide insight: participants described their understanding that ML models *"get better as they function and learn algorithms"* (LP154, E-N), or even *"gain consciousness"* (HP62, N-I).

Interactive ML designers must ensure that these expectations are managed, such as by clarifying how model feedback is treated or what accuracy the model could achieve. Or if feedback is not supported, designers should ensure users do not think they are in some way providing feedback to the model.

Limitations & Future Work

Generalization from a tightly scoped domain. Our aforementioned findings are made in a tightly scoped domain, with a simple model and task (categorizing sports' emails). While this constrained setting provides a necessary first step in illustrating the relationship between explanations and feedback—it is simple enough to support a controlled experiment for non-expert users, and common enough in IML research to be compared to past studies—our findings should be generalized with caution. For example, explanations and feedback mechanisms in our studies were simple and intuitive. However, explanations in other domains, such as image classification, can be confusing or misleading [1], and interaction with more complex models, such as topic models, exposes users to other challenges, such as instability and latency [60]. These differences would likely affect satisfaction with and expectations of these systems.

We hypothesize that even for more complex models or subjective tasks, if users understand how models work and how they can better improve them, they will want the *opportunity* to do so and may be frustrated if such feedback is restricted. However, the *degree* of their frustration would likely vary along with their actual desire and ability to provide feedback in more realistic settings. All are likely affected by task and model complexity, task importance (and therefore user motivation), and domain expertise. Would users be eager to provide feedback (in lieu of abandonment) in an imperfect self-driving car? Would they be less able to detect systems' mistakes for more subjective tasks? Future studies should further explore the relationship between feedback and explanation.

The effect of explanation and feedback mechanisms. Motivated by prior work [44, 35], our simple and truthful method chooses the top three overall important words for classification. This method inherently exposes system uncertainty in the low quality model, as words that are probable for both classes may be highlighted. This does not occur as often in the high quality model as it is more certain about most of the emails. Therefore, this could explain some of the additional frustration in the lower quality model. Future work could explore the effects of different, more advanced, explanation types and feedback mechanisms. For example, global explanations (e.g., differential explanations [37]) might be equivalently faithful, while better counteracting the user experience concerns. "Human-like" explanations may increase expectations of improvement, as human-like characteristics in ML systems can cause users to believe systems will act rationally or take responsibility for their actions [28]. Furthermore, explanations that expose when models are right for the wrong reason might further increase frustration if adequate feedback is not allowed, as users would be unable to rectify apparent mistakes. For this case, to align the information received by the model and the user, feedback mechanisms should be changed accordingly.

CONCLUSION

We present two controlled experiments to understand how the combinations of explanation and feedback affect users' satisfaction and expectations of improvement of high and low quality ML models. We conclude that, for the simple models and task of our studies, when possible explanations and feedback should be provided together: (1) while explanations negatively impacted user satisfaction with the low quality model, they can show users how to fix models, and support for feedback had positive effects; and (2) for the higher accuracy model, requesting detailed feedback without explanations reduced trust. Additionally, regardless of model quality, feature-level feedback increased expectations that models would improve, yet users generally expected model correction, regardless of whether they provided feedback or received explanations.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful and constructive comments, and we thank Kevin Seppi, Simone Stumpf, and James Fogarty for their feedback on early ideas for this study. This work was partially supported by NSF Grants IIS-1822494 and IIS-1409287 (UMD), ONR grant N00014-18-1-2193, the WRF/Cable Professorship, and the Allen Institute for Artificial Intelligence (UW). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Proceedings of Advances in Neural Information Processing Systems*.

[2] Jae Wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. 2007. Open User Profiles for Adaptive News Systems: Help or Harm?. In *Proceedings of the World Wide Web Conference*.

[3] David Alvarez-Melis and Tommi S. Jaakkola. 2018. Towards Robust Interpretability with Self-explaining Neural Networks. In *Proceedings of Advances in Neural Information Processing Systems*.

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* (2014), 105–120.

[5] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. 2010. Examining Multiple Potential Models in End-user Interactive Concept Learning. In *International Conference on Human Factors in Computing Systems*.

[6] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. ReGroup: Interactive Machine learning for On-demand Group Creation in Social Networks. In *International Conference on Human Factors in Computing Systems*.

[7] Maria-Florina Balcan and Avrim Blum. 2008. Clustering with Interactive Feedback. In *International Conference on Algorithmic Learning Theory*.

[8] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Association for the Advancement of Artificial Intelligence* (2019).

[9] Mustafa Bilgic and Raymond J Mooney. 2005. Explaining Recommendations: Satisfaction vs. Promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at IUI*.

[10] Or Biran and Kathleen McKeown. 2017. Human-centric Justification of Machine Learning Predictions. In *International Joint Conference on Artificial Intelligence*.

[11] Andrea Bunt, Joanna McGrenere, and Cristina Conati. 2007. Understanding the Utility of Rationale in a Mixed-Initiative System for GUI Customization. In *International Conference on User Modeling*.

[12] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-based Explanations in a Machine Learning Interface. In *International Conference on Intelligent User Interfaces*.

[13] Oana Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. E-SNLI: Natural Language Inference with Natural Language Explanations. In *Proceedings of Advances in Neural Information Processing Systems*.

[14] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. 2015. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Knowledge Discovery and Data Mining*.

[15] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do Explanations make VQA Models more Predictable to a Human?. In *Proceedings of Empirical Methods in Natural Language Processing*.

[16] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-based Art Recommender. *User Modeling and User-Adapted Interaction* (2008).

[17] Aron Culotta, Trausti Kristjansson, Andrew McCallum, and Paul Viola. 2006. Corrective Feedback and Persistent Learning for Information Extraction. *Artificial Intelligence* (2006).

[18] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K.E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *International Conference on Intelligent User Interfaces*.

[19] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *International Conference on Intelligent User Interfaces*.

[20] Jerry Alan Fails and Dan R Olsen. 2003. Interactive Machine Learning. In *International Conference on Intelligent User Interfaces*.

[21] Shi Feng and Jordan Boyd-Graber. 2019. What can AI do for me? Evaluating Machine Learning Interpretations in Cooperative Play. In *International Conference on Intelligent User Interfaces*.

[22] Rebecca Fiebrink, Dan Trueman, and Perry R Cook. 2009. A Metainstrument for Interactive, On-the-fly Machine Learning. In *Proceedings of New Interfaces for Musical Expression*.

[23] Dimitra Gkatzia, Oliver Lemon, and Verena Rieser. 2016. Natural Language Generation Enhances Human Decision-making with Uncertain Information. In *Proceedings of the Association for Computational Linguistics*.

[24] Dorota Głowacka, Tuukka Ruotsalo, Ksenia Konyushkova, Kumaripaba Athukorala, Samuel Kaski, and Giulio Jacucci. 2013. Directing Exploratory Search: Reinforcement Learning from User Interactions with Keywords. In *International Conference on Intelligent User Interfaces*.

[25] Bryce Goodman and Seth Flaxman. 2017. European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". *AI Magazine* (2017).

[26] Dave Gunning. 2016. Explainable Artificial Intelligence (XAI). (2016). https://www.darpa.mil/program/explainable-artificial-intelligence

[27] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Conference on Computer Supported Cooperative Work and Social Computing*.

[28] Kristina Höök. 2000. Steps to Take before Intelligent User Interfaces become Real. *Interacting With Computers* 12, 4 (2000), 409–426.

[29] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith. 2014. Interactive Topic Modeling. *Machine Learning* 95, 3 (6 2014), 423–469.

[30] Anna Huang. 2008. Similarity Measures for Text Document Clustering. In *New Zealand Computer Science Research Student Conference*.

[31] W. Bradley Knox and Peter Stone. 2012. Reinforcement Learning from Human Reward: Discounting in Episodic Tasks. In *IEEE International Workshop on Robot and Human Interactive Communication*.

[32] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you Accept an Imperfect AI? Exploring Designs for Adjusting End-user Expectations of AI systems. In *International Conference on Human Factors in Computing Systems*.

[33] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *International Conference on Intelligent User Interfaces*.

[34] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *International Conference on Human Factors in Computing Systems*.

[35] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too Much, too Little, or Just Right? Ways Explanations Impact End Users' Mental Models. In *IEEE Symposium on Visual Languages and Human-Centric Computing*.

[36] Isaac Lage, Andrew Slavin Ross, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2018. Human-in-the-loop Interpretability Prior. In *Proceedings of Advances in Neural Information Processing Systems*.

[37] Himabindu Lakkaraju, Rich Caruana, Ece Kamar, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Conference on AI, Ethics, and Society*.

[38] Ken Lang. 1995. NewsWeeder: Learning to Filter Netnews. In *Machine Learning Proceedings*.

[39] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum* 31 (2012), 1155–1164.

[40] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of Empirical Methods in Natural Language Processing*.

[41] David D Lewis. 1998. Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. In *Proceedings of European Conference of Machine Learning*.

[42] Brian Lim, Anind Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. (2009).

[43] Brian Y. Lim and Anind K. Dey. 2011. Investigating Intelligibility for Uncertain Context-aware Applications. In *Proceedings of the international conference on Ubiquitous computing*.

[44] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-interpretability of Explanation. (2018). http://arxiv.org/abs/1802.00682

[45] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2017. Attentive Explanations: Justifying Decisions and Pointing to the Evidence. In *Computer Vision and Pattern Recognition*.

[46] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[47] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and Measuring Model Interpretability. (2018). http://arxiv.org/abs/1802.07810

[48] Pearl Pu and Li Chen. 2006. Trust Building with Explanation Interfaces. In *International Conference on Intelligent User Interfaces*.

[49] Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active Learning with Feedback on both Features and Instances. *Journal of Machine Learning Research* (2006).

[50] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John T. Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Conference on Computer Supported Cooperative Work and Social Computing*.

[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why Should I Trust You? Explaining the Predictions of any Classifier. In *Knowledge Discovery and Data Mining*.

[52] Stephanie Rosenthal and Anind K Dey. 2010. Towards Maximizing the Accuracy of Human-labeled Sensor Data. In *International Conference on Intelligent User Interfaces*.

[53] William Saunders, Andreas Stuhlmüller, Girish Sastry, and Owain Evans. 2018. Trial without Error: Towards Safe Reinforcement Learning via Human Intervention. In *International Joint Conference on Autonomous Agents and Multiagent Systems*.

[54] Philipp Schmidt and Felix Biessmann. 2019. Quantifying Interpretability and Trust in Machine Learning Systems. (2019). `http://arxiv.org/abs/1901.08558`

[55] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *International Conference on Computer Vision*.

[56] Burr Settles. 2010. Active Learning Literature Survey. *Machine Learning* 15, 2 (2010), 201–221.

[57] Burr Settles. 2011. Closing the Loop: Fast, Interactive Semi-supervised Annotation with Queries on Features and Instances. In *Proceedings of Empirical Methods in Natural Language Processing*.

[58] Zhangzhang Si and Song Chun Zhu. 2013. Learning and-or Templates for Object Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013).

[59] Robert S. Siegler. 2002. *Microgenetic Studies of Self-explanation*. Cambridge University Press, 31–58.

[60] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-centered Design and Evaluation of a Human-in-the-loop Topic Modeling System. In *International Conference on Intelligent User Interfaces*.

[61] Kimberly Stowers, Nicholas Kasdaglis, Michael Rupp, Jessie Chen, Daniel Barber, and Michael Barnes. 2017. Insights into Human-agent Teaming: Intelligent Agent Transparency and Uncertainty. In *Advances in Intelligent Systems and Computing*.

[62] Simone Stumpf. 2016. Explanations Considered Harmful? User Interactions with Machine Learning Systems. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.

[63] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching. In *IFIP Conference on Human-Computer Interaction*.

[64] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using only ANOVA Procedures. In *International Conference on Human Factors in Computing Systems*.

[65] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. 2019. Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis. *ACM Transactions on Computer-Human Interaction* (2019).

[66] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *International Conference on Human Factors in Computing Systems*.