Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience

Bhavya Ghai
Stony Brook University
Stony Brook, USA

Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy
IBM Research AI
Yorktown Heights, USA

Klaus Mueller
Stony Brook University
Stony Brook, USA

ABSTRACT

Active Learning (AL) is a human-in-the-loop Machine Learning paradigm favored for its ability to learn with fewer labeled instances, but the model's states and progress remain opaque to the annotators. Meanwhile, many recognize the benefits of model transparency for people interacting with ML models, as reflected by the surge of explainable AI (XAI) as a research field. However, explaining an evolving model introduces many open questions regarding its impact on the annotation quality and the annotator's experience. In this paper, we propose a novel paradigm of explainable active learning (XAL), by explaining the learning algorithm's prediction for the instance it wants to learn from and soliciting feedback from the annotator. We conduct an empirical study comparing the model learning outcome, human feedback content and the annotator experience with XAL, to that of traditional AL and coactive learning (providing the model's prediction without the explanation). Our study reveals benefits–supporting trust calibration and enabling additional forms of human feedback, and potential drawbacks–anchoring effect and frustration from transparent model limitations–of providing local explanations in AL. We conclude by suggesting directions for developing explanations that better support annotator experience in AL and interactive ML settings.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Computing methodologies → Machine learning**; **Active learning settings**.

KEYWORDS

Active learning, explanation, XAI, human-AI interaction, annotation

ACM Reference Format:
Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2020. Explainable Active Learning (XAL): An Empirical Study of How Local Explanations Impact Annotator Experience. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Although Machine Learning (ML) technologies have become ubiquitous, the challenge to obtain labeled data is a frequent barrier to harness their power, especially for those using data-hungry deep learning models. To tackle this challenge, Active Learning (AL) came to be a vivid research area in recent years. AL could largely reduce the labeling effort by having the model take control of the data, selectively querying a human annotator (or *teacher* as in the machine teaching literature [60]) for labels of instances that it wants to learn from. A growing number of AL techniques have been produced by the research community, addressing various querying settings and sampling strategies to select the queries.

Most AL algorithms assume human annotators as simple oracles [54]. This algorithm-centric view has been criticized for several reasons. First, AL work often ignores the needs and preferences of the human annotators. By being mechanically queried with an opaque model, the annotation work is not only tedious but also unsatisfying [9, 25]. For example, without understanding the competence and learning progress of the model, the annotator may not be able to establish confidence and trust in the model to be deployed. Second, AL algorithms are often developed with the assumption that the oracle will provide error-free labels. However, in reality, annotation errors and biases are commonplace, and can be systematically impacted by a particular AL setting and the interface. By understanding these behavioral patterns, one can either prevent the problems or mitigate their impact with algorithmic interventions. Lastly, from a knowledge transfer point of view, learning from labeled instances is extremely inefficient. To address this, a small set of research explored asking humans for direct input to the learning model, such as the key features it uses to make decisions [20, 49]. However, this is challenging for domain experts without ML expertise as it requires an understanding of a model's inner working, which would inevitably impede the quality of their input.

Outside the AL domain, improving ML model transparency to enable understanding, trust and quality feedback from model developers and end users has spurred wide-spread interest, which motivated the research field of explainable AI (XAI) [26]. Many XAI techniques aim to generate easy-to-consume explanations that do not require understanding of the entire model. For example, *local explanation* (e.g. [43, 51]) is a cluster of XAI techniques that explain the prediction for a particular instance, often by how features of the instance contribute to the model's prediction.

In this work, we explore a novel paradigm of *explainable active learning* (XAL), by providing a local explanation of the model's current prediction as the interface to query an annotator's input. We foresee several potential benefits of providing explanations. First, by making the learner's beliefs and logic transparent, explanations could improve annotator satisfaction, especially by establishing trust in the final model. By enabling the annotator to witness the progress of the model logic as it learns, explanations could potentially support stopping criteria–knowing when to stop the labeling effort, which is a known challenge in AL paradigms [2]. Moreover, local explanations accompanying a specific instance could lead to a better understanding by an annotator and potentially improve the quality of feedback. For example, it may reduce the knowledge required for labeling–even if an annotator cannot always make an accurate independent judgment when labelling a specific instance, he or she may be able to recognize flaws in the model's logic. Explanations could also enable richer forms of knowledge transfer beyond instance labels by eliciting an annotator's feedback for the explanation itself.

However, there are also counter-arguments for introducing explanations in an AL paradigm. First of all, an XAL setting entails the model presenting its prediction accompanied by its explanation, which could potentially anchor the annotator's feedback. It is in fact closer to *coactive learning* (CL) [59], a sub-paradigm of AL, in which the model presents its predictions and the annotator is only required to make corrections if necessary. CL is favored over traditional AL for reducing annotator workload, especially when the feedback availability is limited. While anchoring judgment is not necessarily counter-productive if the model predictions are competent, we recognize that the most popular sampling strategy of AL–uncertainty sampling–focuses on querying those instances the model is most uncertain of. Moreover, unlike conventional XAI work, explanations in XAL would be applied to early-stage, naive models with highly flawed logic, and will undergo drastic changes during the learning process. How people react to flawed and changing explanations remains an open question.

We conduct a case study to empirically explore the feasibility, opportunities and challenges of this new paradigm of XAL. It also offers an opportunity to understand how to design explanations that support annotators interacting with ML models in broader contexts. By comparing the annotation, learning outcome and subjective experience of annotators across three settings: active learning (AL, querying labels), coactive learning (CL, querying feedback for the model predictions) and explainable active learning (XAL, querying feedback for the model predictions accompanied by explanation), we explore the following research questions:

- **RQ1**: How do local explanations impact annotation and the learning outcome of active learning?
- **RQ2**: How do local explanations impact annotators' experience, specifically trust in the model, satisfaction, engagement, cognitive workload and credit attribution?
- **RQ3**: How does the impact of explanations on annotation and annotator experience differ at different stages (early v.s. late stage) of an active learning process?
- **RQ4**: What kind of annotator feedback beyond instance labels can be harnessed with local explanations?

2 RELATED WORK AND BACKGROUND

2.1 Active learning

The core idea of active learning is that if the learning algorithm intelligently selects instances to be labeled, it could perform well with less training data [54]. This idea resonates with the critical challenge in modern ML, that labeled data are extremely time-consuming and expensive to obtain [68]. Active learning can be used in different scenarios like stream based [14] (from a stream of incoming data), pool based [39], membership query synthesis, etc [54]. To select the next instance for labeling, multiple query sampling strategies have been proposed in the literature [16, 17, 22, 28, 40, 56, 58]. *Uncertainty sampling* [5, 16, 40, 56] is one of the most commonly used strategies which selects instances the model is most uncertain about. Different AL algorithms exploit different notions of uncertainty, e.g. entropy [56], confidence [16], margin [5], etc. The *query by committee* sampling strategy selects those instances where there is maximum disagreement among multiple ML models (committee) that have been trained on the same set of currently labeled instances. *Hierarchical sampling* selects instances that are most representative of the unlabeled dataset [17]. More recent approaches like QUIRE [28] selects instances which are informative and representative of the unlabeled dataset.

Although the original definition of AL is concerned with querying the annotator for the label of a single instance, alternative querying approaches have been explored. Settles et al. introduced multiple-instance active learning [57], in which instances are grouped into "bags" to be queried for labels. Several works explored querying feedback for features, e.g., by asking the annotator whether a feature is important or relevant for the target concept [20, 49, 55]. Although conceptually querying input for features or model logic from a domain expert is an efficient way of knowledge transfer, it is challenging for people without ML expertise to comprehend model features. Thus most existing works were conducted in text-based ML contexts where the features (keywords) are relatively intuitive. Other relevant AL paradigms include *active class selection* [42] and *active feature acquisition* [67] which query the annotator for examples and missing features, respectively.

2.2 Interfaces for active learning and interactive machine learning

The human annotators in AL are often treated as simple oracles. How they respond to the learning model's queries is given little attention. To the best our knowledge, few in the human-computer interaction (HCI) community have explored the interfaces of AL or annotators' interaction behaviors. An exception is the sub-field of human-robot interaction (HRI), where AL algorithms are used to develop robots that could continuously learn by asking humans questions [9, 10, 12, 23, 53]. In this context, the robot is the interface for AL algorithms. Instead of asking a stream of instances, HRI work is often interested in enabling the robot to ask diverse types of queries. For example, in a series of studies [9, 10], Cakmak et al. explored robots that ask three types of AL queries: instance queries, feature queries and demonstration queries (analogous to active class selection [42]). The studies found that people were more receptive of feature queries and perceived robots asking feature queries to be more intelligent. More importantly, these studies found that a constant stream of queries were not only perceived to be annoying by the annotators, but also led to a decline in their situation awareness, causing people to "lose track of what they were teaching" [9]. These results highlight the problems with AI's fundamental assumption of treating the annotators as oracles. Without understanding the annotators' behaviors and supporting their needs, one may not be able to obtain high-quality input and thus fail to harness the benefit of AL.

Supporting the needs and productivity of people who build or evolve ML models is the central interest in the research area of *interactive machine learning* [3], and more recently, *machine teaching* [59]. By definition, interactive ML emphasizes a tight interaction loop in which the ML learner takes input from people, often domain experts who lack expertise in ML, and transparently presents how it is impacted by the human input [3, 21]. Effectively designed transparency is not only key to a sound mental model and satisfying user experience [35], but also helps people adapt their input to improve the ML models in a more effective way [21, 52]. Another core theme in the interactive ML literature is the importance of empirically studying how people interact with ML systems [3, 38, 63], by which gaps in the system design as well as the algorithms can be revealed. For example, using paper-based mockups to present a text classification system with explanations, Stumpf et al. [63] explored what types of feedback people naturally want to give. By analyzing the free-form feedback, the authors summarized a variety of feedback types, which point to opportunities for new human-in-the-loop ML algorithms that could better support and harness people's natural feedback.

We note that although AL is achieved through continuous interaction between the human and the ML model, it lacks the two above-mentioned elements in interactive ML. First, the annotator is completely oblivious to the behavior and progress of the model. Second, there is little empirical understanding on how people interact in an AL setting. Our work aims to fill both gaps.

2.3 Explanation for active learning

To provide transparency in AL, our study is also motivated by work in the rapidly growing field of explainable AI (XAI) [24, 26] or interpretable ML [11, 19]. Explanations are sought for various reasons. Most importantly, with the increasing adoption of opaque deep learning models, explanations are considered indispensable for establishing trust and acceptance of AI [51]. Furthermore, explanations allow a model's faulty behaviors to be detected and evaluated for the desiderata, including capability, fairness, safety, etc. [18, 19]. Explanations are therefore increasingly incorporated in ML development tools supporting various debugging tasks such as performance analysis [50], interactive debugging [34], feature engineering [33], instance inspection and model comparison [27]. Several studies also explored supporting ideation or feedback from domain experts to improve ML models by utilizing explanations that are easy to consume for people without ML expertise, such as using local explanation techniques and visualization [7, 13, 32].

The explanation adopted in our study is a type of local explanation with *feature importance* [24, 51], which justifies a particular prediction by how the model weighs the instance's features. Recent work argues that, while model developers and ML experts may desire complete explanation of the model at a global level, lay users may prefer local explanation grounded in specific examples [4, 36]. Local explanation is sought not only to judge whether to trust a particular prediction, but also, through case-based experience, to help people evaluate whether to trust the model in general to behave in a reasonable way [51]. The idea of providing local explanation in AL was considered in a very recent paper by Teso and Kersting [64], with three proposed benefits: 1) To improve the understandability and thus trust in the model; 2) To improve the quality of feedback by directing the annotators' attention to "aspects of the instance deemed important by the model"; 3) To allow annotators to provide feedback for the explanation itself (e.g., "right for the wrong reason"), and thus enable additional human input into the learning model. This work, however, did not conduct empirical studies to validate these claims. We concur these claims and use them to guide some of our research questions.

We further hypothesize that explanations could help the annotators calibrate their trust at different stages of an AL process. As the learning model progresses, explanation could help the annotator establish confidence and trust in the model that they could be responsible for deploying. We will empirically test this hypothesis by comparing annotation experience in two snapshots of an AL process, an *early stage* annotation task with the initial model, and a *late stage* when the model is close to the stopping criteria.

3 METHODOLOGY

3.1 Prediction task

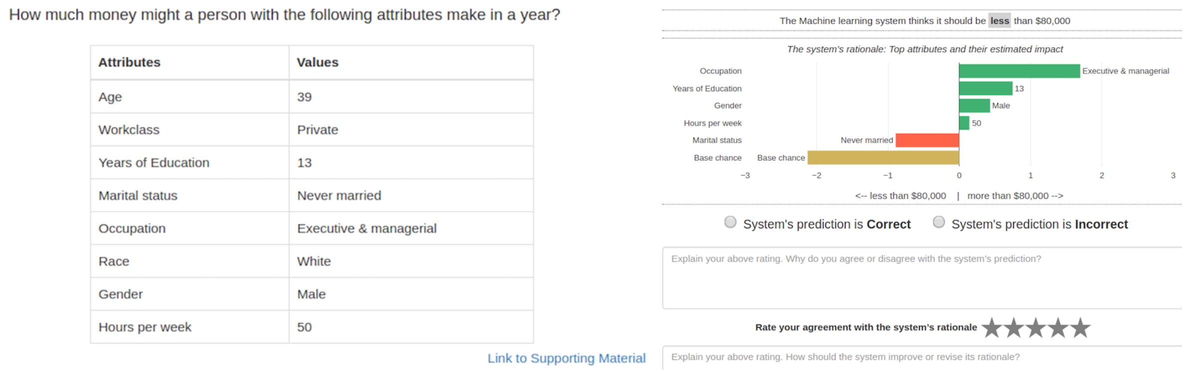
We aimed to design a prediction task that would not require deep domain expertise, where common-sense knowledge could be effective for teaching the model. The task should also involve decisions by weighing different features so explanations could potentially make a difference (i.e., not simple perception based judgment). Lastly, the instances should be easy to comprehend with a reasonable number of features. With these criteria, we chose the Adult Income dataset [30] for a task of predicting whether the annual income of an individual is more or less than $80,000 [1]. The dataset is based on a Census survey database. Each row in the database characterizes a person with a mix of numerical and categorical variables like age, gender, education, occupation, etc., and a binary annual income variable, which was used as our ground truth.

In the experiment, we presented the participants with a scenario of building an ML classification system for a customer database. Based on a customer's background information, the system aimed to predict the customer's income level for a targeted service. The task of the participants was to judge the income level of instances that the system selected to learn from, as presented in Figure 1a,

3.2 Active learning setup

To choose the ML model for active learning, we had two constraints. First, the model should not be computationally expensive. AL requires the model to be retrained after new labels are fetched, so the model needs to train fast to avoid latency in the experiment. Second, the model should be interpretable, i.e. it should be easy to generate local explanations for each prediction. Although techniques exist for generating post-hoc explanations for non-interepretable models [51], they are not perfectly faithful to the model's logic

[1] After adjusting for inflation (1994-2019) [1], while the original dataset reported on the income level of $50,000



(a) Customer profile presented in all conditions for annotation **(b) Explanation and questions presented in the XAL condition**

Figure 1: Experiment interface

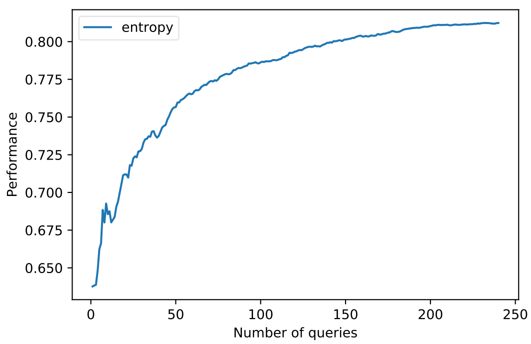


Figure 2: Accuracy as a function of number of queries in the simulation experiment

and hence not desirable for the experiment. Considering both constraints, we opted for logistic regression (with l2 regularization) which has been used extensively in the AL literature [54, 66].

Building an AL pipeline involves the design choices of sampling strategy, batch size, the number of initial labeled instances and test data. For this study, we used entropy based uncertainty sampling to select the next instance to query, as it is the most commonly used sampling strategy [66] and also computationally inexpensive. We used a batch size of 1 [6], meaning the model was retrained after each new queried label. We initialized the AL pipeline with two labeled instances. To avoid tying the experiment results to a particular sequence of data, we allocated different sets of initial instances to different participants, by randomly drawing from a pool of more than 100 pairs of labeled instances. The pool was created by randomly picking two instances with ground-truth labels, which were kept in the pool only if they produced a model with initial accuracy between 50%-55%. This was to ensure that the initial model would perform worse than humans. 25% of all data were reserved as test data for evaluating the model learning outcomes.

As discussed, we are interested in the effect of explanations at different stages of AL. We took two snapshots of an AL process–an early-stage model just started with the initial labeled instances, and a late-stage model that is close to the stopping criteria, i.e., convergence of accuracy on the test data. To determine where to take the late-stage snapshot, we ran a simulation where AL queried instances were given the labels in the ground truth. The simulation was run with 10 sets of initial labels and the mean accuracy is shown in Figure 2. Based on the convergence pattern, we chose the late stage model to be where 200 queries were executed. To create the late-stage experience without having participants answer 200 queries, we took a participant’s allocated initial labeled instances and simulated an AL process with 200 queries answered by the ground-truth labels. The model was then used in the late-stage task for the same participant. This also ensured that the two tasks a participant experienced were independent of each other i.e. a participant’s performance in the early-stage task did not influence the late-stage task. In each task, participants were queried for 20 instances. Based on the simulation result in Figure 2, we expected an improvement of 10%-20% accuracy with 20 queries in the early stage, and a much smaller increase in the late stage.

3.2.1 Explanation method. Figure 1b shows a screenshot of the local explanation presented in the XAL condition, for the instance shown in Figure 1a. The explanation was generated based on the coefficients of the logistic regression, which determine the impact of each feature on the model’s prediction. To obtain the *feature importance* for a given instance, we computed the product of each of the instance’s feature values with the corresponding coefficients in the model. The higher the magnitude of a feature’s importance, the more impact it had on the model’s prediction for this instance. A negative value implied that the feature value was tilting the model’s prediction towards less than $80,000 and vice versa. We sorted all feature by their absolute importance and picked the top 5 features responsible for the model’s prediction.

The selected features were shown to the participants in the form of a horizontal bar chart as in Figure 1b. The importance of a feature was encoded by the length of the bar where a longer bar meant greater impact and vice versa. The sign of the feature importance was encoded with color (green-positive, red-negative), and sorted to have the positive features at the top of the chart. Apart from the top contributing features, we also displayed the intercept of the logistic regression model as an orange bar in the bottom. Because it was a relatively skewed classification task (the majority of the population has an annual income of less than $80,000), the negative base chance needed to be understood for the model's decision logic. For example, in Figure 1, Occupation is the most important feature. Martial status and base chance are pointing towards less than $80,000. While all others are tilting positively, the model prediction for this instance is still less than $80,000.

3.3 Experimental design

We adopted a 3×2 experimental design, with the learning condition (AL, CL, XAL) as a between-subject treatment, and the learning stage (early v.s. late) as a within-subject treatment. That is, participants were randomly assigned to one of the conditions to complete two tasks, with queries from an early and a late stage AL model, respectively. The order of the early and late stage tasks was randomized and balanced for each participant to avoid order effect and biases from knowing which was the "improved" model.

We posted the experiment as a human intelligence task (HIT) on Amazon Mechanical Turk. We set the requirement to have at least 98% prior approval rate and each worker could participate only once. Upon accepting the HIT, a participant was assigned to one of the three conditions. The annotation task was given with a scenario of building a classification system for a customer database to provide targeted service for high- versus low-income customers, with a ML model that queries and learns in real time. Given that the order of the learning stage was randomized, we instructed the participants that they would be teaching two configurations of the system with different initial performance and learning capabilities.

With each configuration, a participant was queried for 20 instances, in the format shown in Figure 1a. A minimum of 10 seconds was enforced for each query. In the AL condition, the participants were presented with a customer's profile and asked to predict whether his or her annual income was above 80K. In the CL condition, the participants were presented with the profile and the model's prediction. In the XAL condition, the model's prediction was accompanied by an explanation revealing the model's "rationale for making the prediction" (the top part of Figure 1b). In both the CL and XAL conditions, the participants were asked to judge whether the model prediction was correct and optionally answer an open-form question to explain that judgement (the middle part of Figure 1b). In the XAL condition, the participants were further asked to also give a rating to the model explanation and optionally explain their ratings with an open-form question (the bottom part of Figure 1b). After the participants submitted each query, the model was retrained, and performance metrics of accuracy and F1 score (on the 25% reserved test data) were calculated and recorded, together with the participant's input and time stamps.

After every 10 trials, the participants were told the percentage of their answers matching similar cases in the Census survey data, as a measure to help engaging the participants. An attention-check question was prompted in each learning stage task, showing the customer's profile in the prior query with two other randomly selected profiles as distractors. The participants were asked to select the one they just saw. Only one participant failed both attention-check questions, and was excluded from the analysis.

After completing 20 queries for each learning stage task, the participants were asked to fill out a survey regarding their subjective perception of the ML model they just finished teaching and the annotation task. The details of the survey will be discussed in Section 3.3.2. At the end of the HIT we also collected participants' demographic information.

3.3.1 Domain knowledge training. We acknowledge that MTurk workers may not be experts of an income prediction task, even though it is a common topic. Our study is close to *human-grounded evaluation* proposed in [19] as an evaluation approach for explainability, in which lay people are used as proxy to test general notions or patterns of the target application (i.e., by comparing outcomes of proxy participants between the baseline and the target treatment).

To improve the external validity, we designed a practice task to help the participants gain domain knowledge. First, throughout the study, we provided a link to a supporting document with statistics of personal income based on the Census survey. Specifically, chance numbers–the chance of people with a feature-value to have income above 80K–were given for all feature-values the model used (by quantile if numerical features). Participants were then given 20 practice trials of income prediction tasks and encouraged to utilize the supporting material. The ground truth–income level reported in the Census survey–was revealed after they completed each trial. The participants were told that the model would be evaluated based on data in the Census survey, so they should strive to bring the knowledge from the supporting material and the practice trials into the annotation task. They were also incentivized with a $2 bonus if the consistency between their predictions and similar cases reported in the Census survey were among the top 10% of all participants.

After the practice trials, the agreement of the participants' predictions with the ground-truth in the Census survey for the early-stage trials reached a mean of 0.65 (SD=0.08). We note the queried instances in AL using uncertainty-based sampling are challenging by nature. The agreement with ground truth by one of the authors, who is highly familiar with the data and the task, was 0.75.

3.3.2 Survey design. To understand how explanation impacts annotators' subjective experiences (**RQ2**), and comparing them between the early and late stage of an AL process (**RQ3**), we designed a survey for the participants to fill after completing each learning stage task. We asked the participants to self report the following (all based on 5-point Likert Scale):

Trust in deploying the model: We asked participants to assess how much they could trust the model they just finished teaching to be deployed for the target task (customer classification). Trust in technologies is frequently measured based on McKnightâĂŹs framework on Trust [45, 46], which considers the dimensions of *capability*, *benevolence*, *integrity* for trust belief, and multiple action-based items (e.g., "I will be able to rely on the system for the target task") for trust intention. We also consulted a recent paper on trust scale for automation [31] and added the dimension of *predictability* for trust belief. We picked and adapted one item in each of the four trust belief dimensions (e.g., for benevolence, "Using predictions made by the system will harm customersâĂŹ interest") , and four items for trust intention, and arrived at an 8-item scale to measure trust (3 were reversed scale). The Cronbach's alpha is 0.89.

Satisfaction of the annotation experience, by the 3-item After-Scenario Questionnaire [41] to measure user satisfaction in usability studies (e.g. "I am satisfied with the ease of completing the task"). The Cronbach's alpha is 0.84.

Engagement of the annotation experience, by selecting two applicable items from the User Engagement Scale [47] (e.g., "It was an engaging experience working on the task"). The Cronbach's alpha is 0.89.

Cognitive workload of the annotation experience, by selecting two applicable items from the NASA-TLX task load index (e.g., "How mentally demanding was the task: 1=very low; 5=very high"). The Cronbach's alpha is 0.86.

Attribution of credit, as prior work suggests it could be influenced by explanations [44]. We used an item proposed in [8] to ask participants to report: "The systemâĂŹs performance is totally due to : 1= the quality of my input, 5=the systemâĂŹs learning capability."

3.3.3 Participants. 37 participants completed the study. One participant did not pass both attention-check tests and was excluded. The analysis was conducted with 12 participants in each condition. Among them, 27.8% were female; 19.4% under the age 30, and 13.9% above the age 50; 30.6% reported to have no knowledge of AI, 52.8% with little knowledge ("know basic concepts in AI"), and the rest to have some knowledge ("know or used AI algorithms"). In total , participants spent about 20-40 min on the study and was compensated for $4 with a 10% chance for additional $2 bonus, as discussed in Section 3.3.1

4 RESULTS

We report the results based on the research questions introduced in the beginning. We will first report the statistics and then summarize the take-away messages at the end of each sub-section.

4.1 Labels and learning outcomes (RQ1, RQ3)

First, we looked into the model learning outcomes in different conditions. In Table 1 (the third to sixth columns), we report the statistics of performance metrics (accuracy and F1 scores) after the 20 queries in each condition and learning stage. We also report the performance improvement, as compared to the initial model performance before the 20 queries. For each of the performance and improvement metrics, we ran a repeated measures ANOVA with *Condition* as a between-subject variable and learning *Stage* as a within-subject variable. As reported in Table 2, we found only significant main effect of *Stage* for all performance and improvement metrics. The results indicate that participants were able to improve the early-stage model significantly more than the later-stage model, but the improvement did not differ across learning conditions.

We then looked into the labels given by the participants, by comparing their agreement with the model's predictions (agreement) and the ground-truth (human accuracy) respectively. The statistics are reported in the last two columns in Table 1. We ran similar repeated measures ANOVA as above. Interestingly, we found a significant main effect of *Condition* on participants' agreement with the model's predictions (Table 2). We conducted post-hoc analysis on the effect of *Condition* with Tukey's Test, and found that the

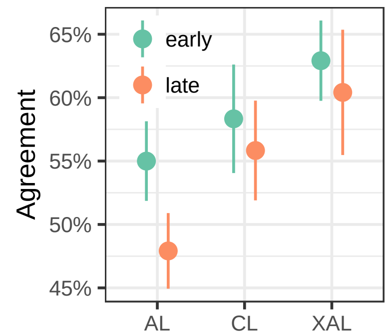


Figure 3: Agreement percentages across conditions. All error bars represent +/- one standard error.

Table 1: Results of model performance and labels

Stage	Condition	Acc.	Acc. improve	F1	F1 improve	%Agree	Human Acc.
Early	AL	67.0%	13.7%	0.490	0.104	55.0%	66.7%
	CL	64.2%	11.7%	0.484	0.105	58.3%	62.1%
	XAL	64.0%	11.8%	0.475	0.093	62.9%	63.3%
Late	AL	80.4%	0.1%	0.589	0.005	47.9%	54.2%
	CL	80.8%	0.2%	0.587	0.007	55.8%	58.8%
	XAL	80.3%	-0.2%	0.585	-0.001	60.0%	55.0%

Table 2: Effects of stage and condition on model performance and labels

Measure	Stage F	Stage p	Condition F	Condition p	Stage × Condition F	Stage × Condition p
Acc.	143	<.001*	.474	.627	.610	.550
Acc. improve	91.1	<.001*	.261	.772	.271	.764
F1	73.5	<.001*	.138	.872	.064	.938
F1 improve	49.5	<.001*	.159	.853	.010	.990
%Agree	1.84	.184	3.32	.049*	.265	.769
Human Acc.	15.5	<.001*	.141	.869	1.68	.201

significant difference of agreement existed between the AL condition and the XAL condition ($p < 0.01$). The difference is illustrated in Figure 3: compared to the control condition of AL where the participants made independent judgment of the instance labels, seeing the model's prediction and explanation increased participants' agreement with the model.

To summarize, we found that presenting a model's prediction accompanied by the local explanation had an *anchoring effect* on the annotators' judgment. However, we did not find this anchoring effect significantly impaired the accuracy of the annotators' judgment (compared to the ground truth) nor the model learning outcomes. We also showed that with uncertainty sampling of AL, both the model improvement and the annotation task itself (human accuracy) became more challenging as the model matured.

4.2 Annotator experience (RQ2, RQ3)

We investigated how participants' self-reported experience differed across conditions by analyzing the following survey scales (measurements discussed in Section 3.3.2): trust in the model, satisfaction, engagement, cognitive workload and attribution of credit. Table 3

Table 3: Survey results

Stage	Condition	Trust	Satisfaction	Engagement	Workload	Attribution
Early	AL	3.14	4.14	4.38	4.08	4.42
	CL	2.80	3.69	4.04	3.29	4
	XAL	2.42	3.15	3.55	2.91	3.36
Late	AL	3	4.11	4.29	4.04	4.42
	CL	2.71	3.5	3.83	3.21	3.67
	XAL	2.99	3.18	3.59	3	3.27

reports the mean ratings in different conditions and learning stage tasks. For each scale, we ran a repeated measures ANOVA with *Condition* as a between-subject variable and *Stage* as a within-subject variable. The statistics are shown in Table 4[2].

First, we found a significant interactive effect between *Condition* and *Stage* on trust. We ran pairwise comparison and found this interactive effect to be significant for XAL and AL ($p = 0.02$) and marginally significant for XAL and CL ($p = 0.08$). Compared to the other two conditions, participants in the XAL condition had significantly lower trust in deploying the early stage model, but enhanced their trust in the later stage model. The results confirmed our hypothesis that explanation could help calibrate annotators' trust in the model at different stages of the AL process, while showing model predictions alone (CL) was not able to have that effect.

We also found the effect of *Condition* on satisfaction and attribution. To our surprise, there was a decrease in the task satisfaction in the XAL condition Turky's test ($p < 0.01$), and an increase in the credit attribution to oneself instead of the model Turky's test ($p < 0.01$), as compared to the baseline AL condition (no pairwise effect of CL was found). We speculate that the reason was the close feedback loop of explanation exposed the model's limitations and learning capability, which was less than desirable for the participants.

We found a marginally significant effect of *Condition* on cognitive workload. Post-hoc analysis with Turky's Test showed that in both the XAL ($p < 0.01$) and CL ($p = 0.04$) conditions, participants reported lower cognitive workload than those in the AL condition, even though in these conditions most of them answered the additional open-form questions. It suggests that annotators found it easier to judge the model's prediction than making their own judgement, confirming the proposed benefit of co-active learning for the annotators over traditional active learning [59].

To summarize, participants' self-reported subjective experience confirmed the benefit of explanations to help annotators calibrate trust and judge the maturity of the model. Thereby we postulate that explanations can potentially be used to help annotators form stopping criteria. We also found an unexpected effect of explanations in reducing annotator satisfaction and their credit attribution to the model. It suggests that transparency could create frustration in an AL setting with naive models. It may be necessary to provide additional support to help the annotators manage their expectations. Lastly, we found evidence that for the annotators, judging model's prediction imposed less cognitive workload than making their own judgment, regardless of whether explanations were provided.

[2]We consider $p < 0.05$ as significant, and $0.05 \leq p < 0.10$ as marginally significant, following statistical convention [15]

Table 4: Effects of stage and condition on survey results

	Stage		Condition		Stage × Condition	
Measure	F	p	F	p	F	p
Trust	.624	0.435	.789	.463	3.15	.050*
Satisfaction	.615	.439	2.66	.085·	.629	.540
Engagement	.927	.343	1.83	.177	.667	.520
Workload	.025	.874	3.00	.064·	.730	.490
Attribution	1.47	.235	3.21	.050*	1.68	.201

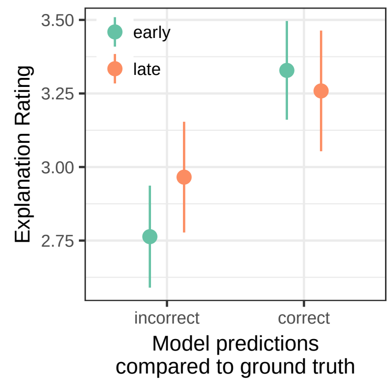


Figure 4: Explanation ratings for correct and incorrect model predictions

4.3 Feedback for explanation (RQ4)

In the XAL condition, participants were asked to rate the system's rationale based on the explanation. In the XAL and CL conditions, participants were asked an optional question to explain their judgment for accepting or rejecting the model's prediction. An additional optional question to explain their explanation ratings was asked in the XAL condition. Analyzing answers to these questions allowed us to understand what kind of feedback was given to the explanations.

First, we inspected whether participants' explanation ratings could provide useful information for the model to learn from. Specifically, if the ratings could distinguish between correct and incorrect model predictions, then they could provide additional signals. Focusing on the XAL condition, we calculated that for each participant, in each learning stage task, the *average explanation ratings* given to instances where the model made correct and incorrect predictions (compared to ground truth). The results are shown in Figure 4. By running an ANOVA on the *average explanation ratings*, with *Stage* and *Model Correctness* as within-subject variables, we found the main effect of *Model Correctness* to be significant, $F(1, 11) = 14.38$, $p < 0.01$. This result indicates that participants were able to distinguish the rationales of correct and incorrect model predictions, in both the early and late stages, confirming the utility of annotators' feedback on the explanations for improving the model.

One may further ask whether explanation ratings provided additional information beyond the judgement in the labels. For example, among cases where the participants disagreed (agreed) with the model predictions, some of them could be correct (incorrect) predictions, as compared to the ground truth. If explanation ratings could distinguish right and wrong disagreement (agreement), they

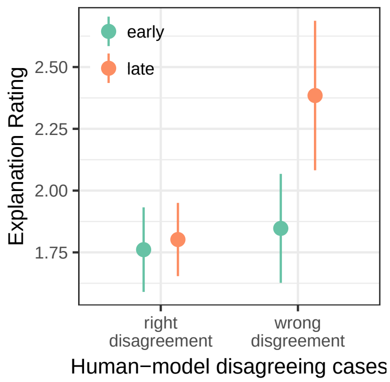


Figure 5: Explanation ratings for disagreeing instances

could serve as additional signals that supplement instance labels. Indeed, as shown in Figure 5, we found that among the *disagreeing instances*, participants' average explanations given to *wrong disagreement* (the model was making the correct prediction and should not have been rejected) was higher than those to the *right disagreement* ($F(1,11) = 3.12$, $p = 0.10$), especially in the late stage (interactive effect between *Stage* and *Disagreement Correctness* $F(1,11) = 4.04$, $p = 0.07$). We did not find this differentiating effect of explanation for agreeing instances. In short, annotators' ratings for model explanation could help distinguish "strong rejection" and "weak rejection". It could potentially be utilized to improve the learning outcome, for example, with AL algorithms that can consider probabilistic annotations [61].

4.3.1 Open form feedback. We also conducted content analysis on participants' open form answers to provide feedback, especially by comparing the ones in the CL and XAL conditions. All 24 participants in the two conditions provided at least one open-form answer, with a mean of 80.0% (SD=36.3) queries having feedback provided by the participant.

In the CL condition, participants' feedback almost exclusively focused on the top features that they believed should determine the prediction, since they only had access to the model's prediction but not how the predictions were made. For example: "*I looked at occupation and years of education. These factors make me believe the prediction is correct.*" In contrast, 9 out of 12 participants in the XAL condition commented on the feature and weights presented in the explanation. We summarize these comments in the following categories:

- *Features*: Unlike in the CL condition where participants focused on the top features they considered for the prediction, feedback in the XAL conditions was reactive of the features participants saw in the explanation. It often expressed surprise, e.g."*not sure why females would be rated negatively*", or "*how is divorce a positive thing*". Some also commented on missing features in the explanation, e.g., "*should take age into account*". These patterns echoed observations from prior work that local explanation could heighten people's attention towards unexpected, especially sensitive features such as race and gender [18].
- *Weights*: The majority of feedback focused on the weights bars presented in the explanation, expressing agreement, disagreement and adjustment one wanted to make on the weights. E.g.,"*agree with all ratings, except marital status, which should be weighted somewhat less*". By identifying problematic weights, comments also indicated that the explanation helped participants reason about accepting or rejecting the model's prediction, e.g., "*how would private employment be enough to push him into the 80k bracket?*"
- *Ranking or comparison of multiple feature weights*: Some comments explicitly addressed the ranking or comparison of multiple features, such as "*occupation should be ranked more positively than marital status.*"
- *Reasoning about combination and relations of features*: Consistent with observation in Stumpf et al.'s study [63], which solicited natural feedback with a paper prototype of a explainable text classification system, some participants suggested the model to consider combined or relational effect of features–e.g., "*years of education over a certain age is negligible.*" Such natural feedback is rarely considered in current AL or interactive ML systems.
- *Logic to combine features and weights*: The feature importance based explanation associates the model's prediction with the combined weights of all features. Two participants expressed confusion, e.g. "*literally all of the information points to earning more than 80,000.*" (while the base chance was negative). Such comments highlight the importance of designing user-friendly explanation and also indicate people's natural tendency to provide feedback on the model's overall logic.
- *Changes of explanation*: Even though our study did not test a complete AL process, one participant in the condition seeing the late-stage model before the early-stage model noted the declining quality of the system's rationale. Change of explanation is a unique property of AL setting. Future work could explore interfaces that explicitly present changes or progress in the explanation and utilize the feedback that annotators would give.

To summarize, we identified many opportunities to use local explanations to elicit knowledge from the annotators beyond instance labels. By simply soliciting a rating for the explanation, additional signals for the instance could be obtained for the model to learn better. Through qualitative analysis of the open-form feedback, we identified several categories of input that people naturally wanted to give by seeing and reacting to the local explanation. Future work could explore algorithms and systems that utilize annotators' input based on local explanations for the model's features, weights, feature ranks and relations, and changes during the learning process.

5 CONCLUSIONS AND DISCUSSIONS

Our work is motivated by the desire to create a more human-centered experience for annotators interacting with ML models. We proposed a novel paradigm of explainable active learning (XAL) by introducing explanation features in an active learning setting. We demonstrated the benefits of local explanation as an interface. For the annotators, the transparent interface could help them gain better situation awareness and trust in the models they achieve.

For the model, the interface enables new opportunities to elicit richer forms and higher quality input from humans. Meanwhile, we also uncovered potential drawbacks of the explanation feature and suggest areas for improvement–to reduce the anchoring effect of explanation and mitigate potential frustration due to transparent model limitation, through improved explanation design and additional interaction interventions. Below we discuss directions for developing explanation features that could better support the needs of annotators and harness their input to improve ML models. While this study was conducted in an AL setting, these conclusions are applicable to broader interactive ML contexts.

5.1 Mitigating anchoring effect of explanations

We found evidence that local explanations, which intend to justify a particular prediction, could increase people's inclination to agree with the model. Although we found the anchoring effect was not strong enough to undermine the active learning results, this is still a potential concern when introducing explainability features in active or interactive learning settings. Alternative design of explanations or interventions could be sought to mitigate the anchoring effect. For example, it would be interesting to test the effect of a partial explanation that does not reveal the model's judgment (e.g., only a small set of top features [37]), or having the annotators first making their own judgment before seeing the explanation.

Our study provides another example where the interface of an interactive learning model could systematically bias the human input. By understanding this systematic pattern, algorithmic solutions could possibly account for such a bias. For example, previous HRI work revealed that people have a positivity bias when providing feedback to teachable robots [65]. Researchers then developed a more robust Reinforcement Learning algorithm by weighing down short-term reward signals from people [29].

A recent empirical study [48] found that explanations had little effect on swaying people's judgement in an AI-assisted decision-making, which used a similar setup as our study but a fully deployed model. We note that what we found was a combined effect of showing the prediction and explanation. The anchoring effect of explanation alone might be a weak one as we did not find a significant difference between the CL and XAL conditions. We also note that the effect of explanation could be sensitive to the choice of the sampling strategy. Since uncertainty sampling focuses on the most uncertain cases, feature importance based explanation could appear to be less convincing for those cases [18]. Future work looking into introducing explainability features in AL settings should consider the impact of different sampling strategies and how they pose different design requirements for explanations.

5.2 Explaining evolving models

To the best of our knowledge, our work is among the first to empirically study explanations applied to a naive and evolving model. By making the model logic transparent, explanation could help people calibrate their confidence and trust as the model improves. Enabling people to witness the model's progress, assess satisfaction for stopping criteria, and establish trust in the final model is valuable not only for AL but also in general ML development or debugging to incorporate explainability features.

Meanwhile, we uncovered the unintended effect of explanations in undermining annotators' satisfaction, as the transparent feedback loop can be frustrating if the model progress is less than desired. This may indeed be a persistent problem for an AL model using uncertainty sampling as the annotators would keep seeing uncertain instances. Our results highlight the needs to help the annotators manage their expectation in AL settings. It is important for them to not only anticipate the model's limitations in the learning capability, but also set expectation for the characteristics of the particular annotation task. For example, uncertainty sampling would focus on uncertain instances that may be challenging to judge in nature, and the challenge would keep increasing as the model matures. This pattern was reflected in our results of decreasing human accuracy in the late-stage task.

Several recent empirical studies also highlighted the potential drawback of explanations [48, 62] by creating additional cognitive workload and hampering people's ability to detect model errors. One potential solution proposed was *progressive disclosure* by starting from simplified explanations and progressively provide more transparency [62]. This idea could apply to AL setting as well. Since the early-stage model has obvious flaws, using simpler explanations could suffice and may be less frustrating. Another idea emerged in our study is to *explain model progress*, for example by explicitly showing changes in the model logic compared to prior versions. This could potentially help the annotators better assess the model progress with less frustration.

5.3 Explanations for knowledge elicitation

Lastly, our study suggests the benefit of explanation as an interface in AL and broader interactive ML settings for eliciting knowledge from people. Teso and Kersting proposed that explanations could enable feedback of "right decision for the wrong reason" [64], in which people accept the system's prediction but suggest changes in the model logic. Empirically, we showed that there was a stronger tendency for people to give feedback of "weak rejection", where they deemed the model prediction to be wrong but the rationale "almost got it". Future work should explore AL algorithms that could leverage the kinds of feedback signals enabled by explanations.

We also join the effort of the interactive ML field in studying natural feedback from people to inform opportunities for algorithmic solutions [3, 38, 63]. While most prior works explored eliciting feedback on keywords based features in text-based ML contexts [20, 38, 55, 63], our study showed that with feature-importance based local explanations, people are willing to provide rich forms of feedback on features and weights for a model using tabular data. The types of free-form feedback we observed were mostly consistent with what Stumpf et al. identified for a text-based classification system [63], but they also revealed differences for tabular data. For example, while the majority of the feedback in [63] focused on changing features (keywords), participants in our study were more interested in the weights and ranks of the features, and considered their relations based on real-world knowledge.

Not limited to AL, future work could explore systems that incorporate domain experts' feedback based on *local* explanations. Prior work of feature-querying AL mostly considered querying feedback for model features at a global level [20, 49, 55]. The potential problem is that lay people without ML expertise may find it challenging to understand abstractly how a model weighs different features [36]. Our study illustrated that model explanation for a specific instance could naturally invoke people's reactions and feedback. This is close to the idea of error-driven debugging that has been used to solicit feature ideation from domain experts [7]. Our study suggested many types of feedback that can be harnessed with a model suggesting instances and providing local explanations, often by "surprising" explanations. Future work could explore what are the best strategies to select the instances, design local explanations, and incorporate the feedback they could elicit.

REFERENCES

[1] 2019 (accessed July, 2019). US Inflation Calculator. https://www.usinflationcalculator.com/.

[2] Michael Altschuler and Michael Bloodgood. 2019. Stopping active learning based on predicted change of f measure for text classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 47–54.

[3] 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* 35, 4 (2014), 105–120.

[4] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012* (2019).

[5] Maria-Florina Balcan, Andrei Broder, and Tong Zhang. 2007. Margin based active learning. In *International Conference on Computational Learning Theory*. Springer, 35–50.

[6] Garrett Beatty, Ethan Kochis, and Michael Bloodgood. 2018. Impact of batch size on stopping active learning for text classification. In *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*. IEEE, 306–307.

[7] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 105–112.

[8] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 258–262.

[9] Maya Cakmak, Crystal Chao, and Andrea L Thomaz. 2010. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development* 2, 2 (2010), 108–118.

[10] Maya Cakmak and Andrea L Thomaz. 2012. Designing robot learners that ask good questions. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 17–24.

[11] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019), 832.

[12] Crystal Chao, Maya Cakmak, and Andrea L Thomaz. 2010. Transparent active learning for robots. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 317–324.

[13] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 559.

[14] David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning* 15, 2 (1994), 201–221.

[15] Duncan Cramer and Dennis Laurence Howitt. 2004. *The Sage dictionary of statistics: a practical resource for students in the social sciences*. Sage.

[16] Aron Culotta and Andrew McCallum. 2005. Reducing labeling effort for structured prediction tasks. In *AAAI*, Vol. 5. 746–751.

[17] Sanjoy Dasgupta and Daniel Hsu. 2008. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*. ACM, 208–215.

[18] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 275–285.

[19] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[20] Gregory Druck, Burr Settles, and Andrew McCallum. 2009. Active learning by labeling features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, 81–90.

[21] Jerry Alan Fails and Dan R Olsen Jr. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*. ACM, 39–45.

[22] Yoav Freund, H Sebastian Seung, Eli Shamir, and Naftali Tishby. 1997. Selective sampling using the query by committee algorithm. *Machine learning* 28, 2-3 (1997), 133–168.

[23] Victor Gonzalez-Pacheco, Maria Malfaz, and Miguel A Salichs. 2014. Asking rank queries in pose learning. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 164–165.

[24] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2019), 93.

[25] Andrew Guillory and Jeff A Bilmes. 2011. Simultaneous Learning and Covering with Adversarial Noise.. In *ICML*, Vol. 11. 369–376.

[26] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017).

[27] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 579.

[28] Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. 2010. Active learning by querying informative and representative examples. In *Advances in neural information processing systems*. 892–900.

[29] W Bradley Knox and Peter Stone. 2012. Reinforcement learning from human reward: Discounting in episodic tasks. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 878–885.

[30] Ronny Kohavi and Barry Becker. 1996. Adult Income dataset (UCI Machine Learning Repository). https://archive.ics.uci.edu/ml/datasets/Adult

[31] Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*. Springer, 13–30.

[32] Josua Krause, Aritra Dasgupta, Jordan Swartz, Yindalon Aphinyanaphongs, and Enrico Bertini. 2017. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 162–172.

[33] Josua Krause, Adam Perer, and Enrico Bertini. 2014. INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 1614–1623.

[34] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 126–137.

[35] 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 *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10.

[36] 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 *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. IEEE, 3–10.

[37] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 29–38.

[38] Tak Yeon Lee, Alison Smith, Kevin Seppi, Niklas Elmqvist, Jordan Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies* 105 (2017), 28–42.

[39] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIRâĂŽ94*. Springer, 3–12.

[40] David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. In *SIGIRâĂŽ94*. Springer, 3–12.

[41] JR Lewis. 1995. Computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction* 7, 1 (1995), 57–78.

[42] Rachel Lomasky, Carla E Brodley, Matthew Aernecke, David Walt, and Mark Friedl. 2007. Active class selection. In *European Conference on Machine Learning*. Springer, 640–647.

[43] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

[44] Bertram F Malle. 2011. Attribution theories: How people make sense of behavior. *Theories in social psychology* 23 (2011), 72–95.

[45] D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. 2002. Developing and validating trust measures for e-commerce: An integrative typology.

Information systems research 13, 3 (2002), 334–359.

[46] D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review* 23, 3 (1998), 473–490.

[47] Heather L OâĂŹBrien, Paul Cairns, and Mark Hall. 2018. A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies* 112 (2018), 28–39.

[48] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[49] Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research* 7, Aug (2006), 1655–1686.

[50] Donghao Ren, Saleema Amershi, Bongshin Lee, Jina Suh, and Jason D Williams. 2016. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics* 23, 1 (2016), 61–70.

[51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[52] Stephanie L Rosenthal and Anind K Dey. 2010. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 259–268.

[53] Giovanni Saponaro and Alexandre Bernardino. 2011. Generation of meaningful robot expressions with active learning. In *2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 243–244.

[54] Burr Settles. 2009. *Active learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

[55] Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1467–1478.

[56] Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 1070–1079.

[57] Burr Settles, Mark Craven, and Soumya Ray. 2008. Multiple-instance active learning. In *Advances in neural information processing systems*. 1289–1296.

[58] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. 1992. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 287–294.

[59] Pannaga Shivaswamy and Thorsten Joachims. 2015. Coactive learning. *Journal of Artificial Intelligence Research* 53 (2015), 1–40.

[60] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).

[61] Jinhua Song, Hao Wang, Yang Gao, and Bo An. 2018. Active learning with confidence-based answers for crowdsourcing labeling tasks. *Knowledge-Based Systems* 159 (2018), 244–258.

[62] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 107–120.

[63] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 82–91.

[64] Stefano Teso and Kristian Kersting. 2018. " Why Should I Trust Interactive Learners?" Explaining Interactive Queries of Classifiers to Users. *arXiv preprint arXiv:1805.08578* (2018).

[65] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[66] Yazhou Yang and Marco Loog. 2018. A benchmark and comparison of active learning for logistic regression. *Pattern Recognition* 83 (2018), 401–415.

[67] Zhiqiang Zheng and Balaji Padmanabhan. 2002. On active learning for data acquisition. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 562–569.

[68] Xiaojin Jerry Zhu. 2005. *Semi-supervised learning literature survey*. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.