Human Expectations and Perceptions of Learning in Machine Teaching

Swati Mishra
Cornell University
United States of America
swati@infosci.cornell.edu

Jeffrey M Rzeszotarski
Cornell University
United States of America
jeffrz@cornell.edu

ABSTRACT

Interactive interfaces in tandem with Machine Learning (ML) models support user understanding of model uncertainty, build confidence, improve predictive accuracy and enable users to teach application-specific concepts that are difficult for the model to learn otherwise. These systems offer empirically proven benefits due to tightly coupled feedback loops and workflow scaffolding. However, deployment with ML non-experts who cannot manage the complex, expertise-heavy process remains challenging. Through deployment with non-expert users in a common classification task, we investigate the impact of human factors of machine teaching interfaces such as user expectations, their perceptions of the learning process and user engagement with respect to teaching process and outcomes. We measure how affective and performance attributes shape the success or failure of the process. Finally, we reflect on how intelligent user interfaces can be designed to accommodate these factors for successful deployment with a broad spectrum of human adjudicators.

KEYWORDS

Interactive Machine Teaching, User Experience, Expectations

ACM Reference Format:
Swati Mishra and Jeffrey M Rzeszotarski. 2023. Human Expectations and Perceptions of Learning in Machine Teaching. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3565472.3595612

1 INTRODUCTION

Interactive interfaces transformed static command line interfaces into systems which actively responded to user requests in human-friendly ways. As new interaction paradigms such as direct manipulation [75] emerged to serve users with diverse expertise, researchers increasingly realized the importance of understanding the human factors underpinning these interfaces such as the expectations users brought to the system and their sense of efficacy. A similar transformation is now occurring in machine learning (ML), where interactive tools are transforming the ML model-building pipeline by providing rich user interactions with a learning model, supporting understanding of model uncertainty and engaging them in a tightly coupled feedback loop [27]. For instance, in the past a healthcare practitioner anonymizing patient records for research might transact with an ML expert to build a model that accelerates the laborious annotation process. In a modern interactive machine learning (iML) system, the practitioner themselves might iteratively refine the ML model, observing its progress and adjusting their feedback [54]. As ML systems reach more diverse users, it is important to understand the human factors that guide this teaching process to develop more effective interactive tools [90, 95].

In a typical Machine Teaching (MT) system, the user : 1) defines the teaching curriculum and learning objectives (e.g. data instances, features and concepts to teach), 2) engages in the act of teaching, 3) evaluates the learned information, 4) refines the teaching curriculum or revisits the teaching process, 5) arrives at a satisfactory outcome and 6) repeats the process when new curriculum is established [34, 40, 51, 76]. In above process, the user (or the teacher) engages in many cognitively demanding judgements to teach the underlying ML model (or the learner). While significant research has been conducted in improving the statistical processes involved in MT and improving the learner performance [34, 90, 95], there are still unanswered questions about how best to account for the underlying human factors involved in the teaching process. This is primarily because user interface design for MT poses different challenges than traditional user interfaces, in the sense that the underlying information structure with which teachers interact evolves iteratively as the session progresses [30, 63, 88]. This iterative refinement acts as a forcing function for the teacher to constantly update their mental models of the past, present, and future state of their system. Intelligent user interfaces need to factor in these changes in order to prevent the teaching process from becoming inefficient and cumbersome. If they do not, they risk adverse outcomes like undesirable user experiences, fatigue and even reluctance to system use [42]. In this paper, we provide a nuanced view of the MT process from a human-centered perspective and investigate how teacher expectations, as grounded in their perceptions of the learner's performance, influence their teaching behavior.

User expectations play a critical role in how an interactive system is perceived and successfully used [66], and they are grounded in a wide range of socio-technical, cultural, and experiential contexts of the user [3, 38, 67]. In our work we explore teacher expectations from the lens of *task goals* and *system performance*. This perspective has been widely and successfully adopted in the user-centered design of information systems research [11, 41]. With respect to task goals, *expectations* can be defined as teachers' "representation of the consequences of an action or sequence of actions in terms of

UMAP '23, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9932-6/23/06…$15.00
https://doi.org/10.1145/3565472.3595612

the appearance of one or more objects on the display," [45] which manifests as teachers' expected changes in model learning as a consequence of teaching (i.e. whether they expect significant or very little learning from a specific iteration). With respect to system performance which changes when the learner learns new concepts (or unlearns old ones) [35, 68], the teachers adapt their mental models of how their input is impacting the learner's output [22, 53]. These *perceptions of the learning process* are defined by "how teachers perceive the effective and iterative improvements in the model performance in context to a desired goal" and serve as a proxy on which to base future expectations. Machine teachers develop an understanding and expectation of learner performance based on early observations and interactions with the system. If a model shows early promise, teachers may perceive it as highly "teachable". This may motivate them to provide more challenging concepts with higher expectations of learning gain. If a model fails to learn, teachers may become pessimistic and offer up less valuable samples, hindering chances for improvement.

We perform a deep dive into this complex relationship between an interactive MT system and its users, including: teachers' perceptions of the learner's progress, how teaching modalities shape their perceptions over iterations, teacher expectations of future performance, satisfaction, and willingness to deploy the learner in the future. The role of these human factors in MT systems can be investigated only by engaging human teachers in the MT process and capturing their interactions. To serve this goal, we designed a fully functional MT system that helps to interactively build a text classification model using online learning [91]. We designed a methodology to capture the teacher expectations and perceptions of the learner's performance, to study the following research questions:

- *RQ1* As the MT session unfolds, how will teachers' perceptions of learner performance and the teaching modalities offered by the system affect their expectations and system engagement?
- *RQ2* How do teachers' prior experiences with AI systems impact their expectations of model performance in a MT session, regardless of learner performance?
- *RQ3* How do teachers' expectations and their perceptions of learning performance impact their overall satisfaction with the MT session?

Through an empirical study, we observe that the teachers expected the learner to perform better when they determined the teaching curriculum themselves as opposed to providing feedback on instances selected by the learner. These initial expectations were adjusted as the teaching session progressed and teachers updated their mental models of learner performance. We also observed that teachers' expectations have a strong influence on their engagement with the system and their overall satisfaction with the teaching process. In the following sections, we describe our system design, experimental methodology and detailed analysis of results to answer the above research questions. We further synthesize our observations as design implications for future teaching interfaces.

2 RELATED WORK

Machine Teaching was originally introduced in 1996 by Goldman and Mathias [34] as an idea of building computational teachers that are tailored towards a specific machine learner. These computational teachers are ML algorithms that teach another ML algorithm, by learning how the later responds to a teaching instance. They proposed a theoretical framework for building MT systems that eliminates any collusion between the learner and the teacher and modeled how careful selection of teaching samples leads to efficient learning. Following this work, there have been several attempts to enhance various aspects of the MT paradigm [94, 95]. These include recent efforts focusing on how to design teaching algorithms for learners that are linear, black boxed, and iterative [29, 51, 52], as well as building interfaces that enable ML users to provide rich nuanced feedback and integrate their domain expertise [27, 57, 89]. These mechanisms have proven useful for solving challenging real-world problems through ML [14]. However, these interfaces are primarily designed to improve learner efficiency with little focus on improving teacher efficiency, despite empirical evidence that involving nuanced human input in the algorithm can lead to substantial improvements on algorithm performance [5, 79, 86], even on computationally NP hard problems [39]. In this paper we bridge this gap and explore factors which might influence teacher efficiency and enable them to provide continued valuable information to the ML system.

One must consider that the teaching process is not only defined by time/instances, feedback accuracy and learner performance as in traditional ML pipelines, but also by human factors such as engagement, agency, sensemaking, cognitive load, perceptions of learner behavior, and users' expectations [10, 25, 80]. For instance, researchers observed that high levels of user engagement can motivate users to continue to teach a learning model [28, 68] and seek to provide more meaningful feedback [5]. Empirical observations that human-in-the-loop systems perform better than fully automated systems [39] has motivated researches to identify various factors involved in the Human-AI relationship and delineate their impact by designing experimental studies in context to system usage [24]. For instance, the HAII-TIME model [80] attempts to capture the human experience of algorithms and identify various human factors that influence how AI non-experts user AI systems. Similarly, the Information Systems (IS) Continuance model [11] (a model that extends Expectation-Confirmation Theory (ECT) [66]) provides theoretical foundations to study how expectations and their confirmation may lead to continued usage of digital products like knowledge sharing systems [15], e-commerce websites[82], smart-wearables [36] and video streaming services [26]. While these models provide some insights into how human factors like expectations and feedback agency impact use of AI systems, they do not characterize their specific impacts on the teaching process. We build upon this body of work by experimentally investigating human expectations in context to MT environments, where users iteratively refine a model by engaging with its parameters.

User expectations with an interactive interface can be grounded in a variety of factors such as knowledge of the task [45], overall experience with other "similar" agents [53], emotional state, product attributes like aesthetics, or via formative evaluations like value for money [92]. The subjectivity and polysemic definition of user expectations makes them challenging to quantify. However, by contextualizing them within user goals in the interface, researchers have studied their impact on user behavior [19, 60, 72]. For instance,

by grounding user expectations in the functional attributes of machine explanations, researchers explored the usability of factual and counterfactual explanations in a text classification task [72]. For MT interfaces, the user goals are to guide the model towards a target performance level - as indicated by quantitative metrics of accuracy, precision and recall [84] or more nuanced metrics like predictive fairness, trust, security [85]. We study teacher expectations from the lens of perceived learner performance, measuring *how well do the users expect the model performance to change by learning from the samples they are teaching.*

An important challenge in investigating how these expectations unfold in the teaching process is to provide faithful representations of an ML model's learning process and its outcome [78]. While complex ML algorithms like Stochastic Bandits [1], Transformers [87], and Deep Neural Networks [74], might be more efficient at learning data patterns, interpreting their learning process requires significant expertise and is a topic of ongoing research [55, 58, 73]. The black-boxed nature of these performant models makes it difficult to influence their learning in interpretable ways and hence, are not well suited for this work. More importantly, these models require high training data and computation time to provide any meaningful results. Therefore, we adopt Logistic Regression [47] model for our study which is simple, performant and interpretable. It is a popular choice for many critical applications like credit scoring [8], loan approvals [81] and disease prediction [65] and has also been widely adopted to investigate human factors in MT interfaces like studying annotator experience [31], usability of explanations [32] and feature selection pitfalls [88]. In this study we employ crowd sourcing [46] which has been widely adopted for ML research in data annotation and evaluating ML performance [54], building advanced models [2] and studying explanability [56]. In the realm of MT, crowdsourcing has been used to evaluate the performance of human-in-the-loop systems [43], improve feature selection for building less biased models [83] and even understand the perceptions of learning for teaching tasks [40]. Our work furthers this line of investigation by providing insights on human behavioral and cognitive features during MT sessions.

3 METHODOLOGY

In this work we seek to investigate how user (teacher) expectations and perceptions of learner performance impact satisfaction, confidence and continued use of an MT system. More specifically, we study how user expectations grounded primarily in system performance change over the course of teaching sessions [11]. While learner performance can be directly measured using performance metrics like accuracy, precision, recall and predictions, user expectations are much more subjective and highly difficult to quantify. Such expectations may even fall below the conscious level of consideration for users. For an MT system, successful outcomes are characterized by both the efficacy of the teacher in providing the intended feedback as well as the efficacy of the learner in delivering the desired improvements. The pace and time with which outcomes are achieved also shapes how success is evaluated. Due to the highly mutable nature of MT relationships as they unfold over multiple iterations, synthetic experimental evaluation of factors such as expectations risk low environmental validity and consistency. Instead, we opt to design an interactive system that supports different levels of teacher and learner efficacy so as to study human factors in situ. As generally the performance of a ML learner will be asymptotic (i.e. approach a point of diminishing returns) [70], the goal of our system design is to help users assess whether they are approaching the asymptote where the learner's gains outweigh their own limited time and effort (a form of satisficing [77]).

3.1 Task

Machine Teaching environments are generally customized to a specific task [24]. To begin assembling our experimental study, we selected a complex-yet-achievable task of building a model that can identify clickbait headlines. Clickbaits are content (text or images) designed to attract user attention by sensationalizing an information piece. Automatically identifying clickbaits is notoriously difficult and prior research has investigated techniques ranging from feature engineering [69] to deep neural networks [4] and is still an open challenge in ML research [4, 12, 61]. There are three reasons why news headline clickbait detection is an especially ideal candidate to for the MT framework: Firstly, stylistic attributes of clickbaits vary significantly across different publishers and are constantly evolving, thus requiring constant maintenance well suited to MT. Secondly, clickbaits often try to masquerade as regular news headlines and can share significant feature similarities with non-clickbait headlines. Therefore, web user domain expertise is critical to identify them [20] which fits well within the MT warrant. Lastly, users of different digital media platforms have varied levels of tolerance for clickbait headlines. For instance, news headlines like "21 items people are talking about this summer" may be ruled as clickbait for users of a financial webpage, but may not be a clickbait for an entertainment news web page. The MT framework provides a mechanism to easily adopt and customize a model for a specific news environment. We used a Clickbait dataset containing 32,000 headlines with balanced classes [7], where each news headline is pre-processed into a vector that is understandable by the learner using tf-idf features. This is a popular approach in clickbait detection and generally provides results acceptable for this task[23].

3.2 System Design

We designed and implemented an interactive MT system linking an online active learning environment (a logistic regression model with a stochastic gradient decent optimization function [13]) with the graphical user interface shown in Figure 1. We refined our design through multiple rounds of pilot testing with non-expert users (not integrated into our final analysis). Our dashboard (Figure 1) visualizes all of the available clickbait samples as a sorted, colored grid based on the model's confidence levels (Figure 1.C). At the left of the grid are blue samples which the model currently believes are not clickbait (predicted confidence (PC) <=0.4) Grey, low-confidence samples occupy the middle ground (PC >0.4 and <0.6). High confidence clickbait are highlighted in pink at the right (PC > =0.6) This grid-based representation provides quick snapshot of the model's current state and visually estimate the epistemic

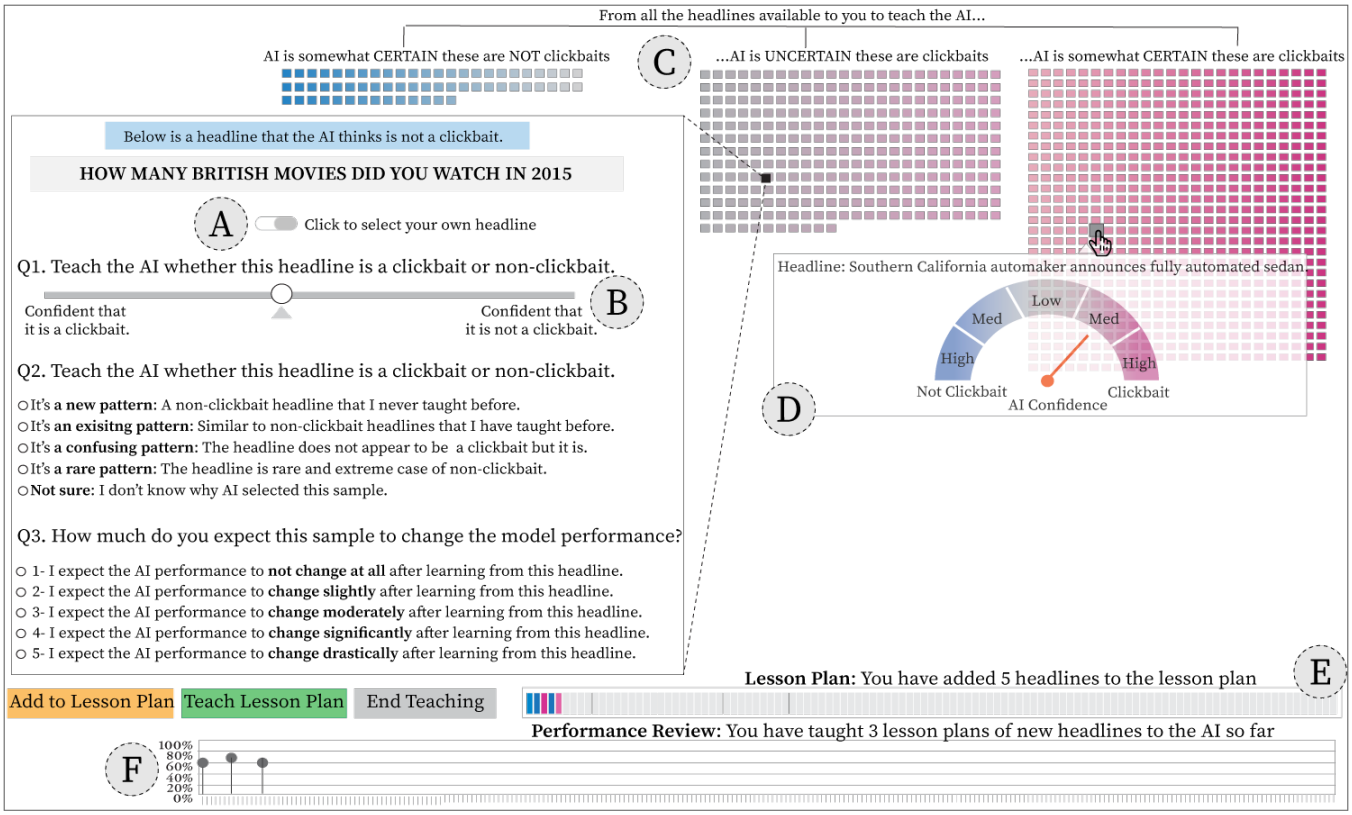

Figure 1: Interactive Machine Teaching Interface. A) The learner selects a news headline from the pool; B) teacher responds by selecting confidence on the headline; C) the entire pool of headlines is visualizes learner confidence on predicted class (clickbait or not clickbait); D) the teacher can explore confidence scores for each headline, click, and select them to teach; E) the progress bar shows headlines that are added to the lesson plan so far; F) the current accuracy of the model is presented by evaluating on a held out test set.

uncertainty of the learning algorithm [93]. For every teaching iteration, a headline is selected and presented to the user. They can provide an affirmative (clickbait) or disapproving (not a clickbait) response, they may override the selected headline with their own selection, or they may entirely curate their own headlines to teach without any assistance from the learner. After every 4 instances, the model re-trains on the feedback provided so far, the grid is re-evaluated and an animation plays to update the each box with new confidence values of retrained model. The animation draws attention to the updated data. Users can mouse-over each headline instance represented by the box to see a simplified "confidence meter" and get more details (Figure 1.D). The data instance selected by the learner (or the teacher, depending on the experimental condition that are described in detail in the next section) becomes part of the lesson plan, and are represented in the progress bar (Figure 1.E). This bar was scaled to a number of samples which is just beyond the amount necessary to reach the asymptotic best performance of the model. We determined the necessary number by evaluating our model's performance during pilot testing. Users may strategically reach a reasonable performance level without exhausting the maximum limit of instances allowed to teach. The model is retrained after every 4 instances, and its performance is calculated on a held out test set that is not available to the users. Model performance in terms of accuracy is updated on a lollipop chart which displays the gradual progression of the model over the teaching session.

Our system offered 3 learning strategies that impact system performance. As the learning strategy result in different performance outcomes, this allows us to manipulate perceptions of the learning process as an independent measure.

(1) *A Random Learner*: This learning approach is based on random sampling where a data sample is randomly chosen and presented to the teacher. The learner in this setting acts unpredictably and improves very slowly.
(2) *A Gradual learner*: In this mode, the learner converges steadily in incremental steps towards a target accuracy. We implement an entropy-based uncertainity sampling technique [49] where the learner selects the instance it is most uncertain about for feedback. Since entropy-based methods are prone to selecting redundant information, the resultant performance curve takes longer to converge, giving an impression of a slow machine learner.
(3) *A Brisk Learner*: In this mode, the learner's performance curve reaches an asymptotic value in short duration and plateaus. We implement a diversity sampling technique [62]

Table 1: Experimental Conditions

Experimental Condition	Learning Strategy	Teaching Strategy
Gradual	Gradual	Active
Brisk	Brisk	Active
Random	Random	Active
Hybrid	Gradual	Hybrid
Traditional	-	Traditional

where the learner selects data instances that represent a wider range of features for feedback. It does not distinguish well between data points close to decision boundaries. As a result the teacher might perceive it to be strategic (or scatterbrained).

Most iML systems incorporate limited user engagement to accelerate the learning process [44]. However, research in iML interface design shows that users quickly give up on using an interactive system when they are required to provide only "yes" and "no" responses [18]. It was important for our users to be engaged in order to continue using the system and offer insights on their expectations during the course of the MT session. Therefore, we also implemented 3 teaching modes in our system with varied levels of engagement in curating the teaching set.

(1) *Active Teacher*: In this mode, the teacher only reviews data instances selected by the learner, providing an appropriate label or vetoing them if necessary while also providing a confidence level of their review (i.e. Figure 1. A is unchangeable other than to skip an instance). In this mode, the teacher has *low efficacy* as the teacher engagement is limited in terms of the feedback they can offer.
(2) *Traditional Teacher*: This mode provides full control to the teacher to select any headline they want to teach, with no support from the learner (i.e. Fig 1.A is not initially populated). While this teaching style may provide *high efficacy* in terms of feedback control , it is likely to yield low system performance due to high effort, fatigue and limited ML expertise.
(3) *Hybrid Teacher*: In this teaching style, the teacher has *moderate efficacy* where they can review the instances selected by the learner and/or select their own by examining the pool (i.e. Fig 1.A is shown, but can be overridden). This mixed-initiative system design provides the teacher slightly more engagement and reduced effort than Active Teaching and may reach a "best of both worlds" for teaching set curation.

3.3 Experimental Conditions

Our system can exhibit three learning strategies and three teaching strategies, each involving varied learner performance levels and teacher effort. While a 3x3 study design might be expected in this case, some of the experimental conditions do not comport with each other, resulting in invalid combinations. For instance, a *Traditional* teaching strategy is entirely driven by the teacher, rendering the presence of learning strategies redundant. Furthermore, in our pilot experiments with 6 in-person participants and 30 crowdsourced workers, we identified experimental conditions that did not yield meaningful results. For instance, a random learner with a hybrid teaching style was so unpredictable for users that they fixated on the chosen samples to the degree that they defaulted to either always intervening or never intervening in suggested samples. Following these pilots, we developed a set of 4 conditions which capture a broad spectrum of potential MT environments with varied levels of system performance as shown in 1. We use a between-subjects methodology for our study, where each participant is exposed to a single pair of learning and teaching strategies, to avoid expertise development and fatigue impacting our observations.

3.4 Study Procedure and Data Gathering

Participants for our study were recruited remotely via the Amazon Mechanical Turk platform. Prior to conducting the study, we compared pilot data from in-person sessions and pilot MTurk recruitments, finding that they had largely the same quality if we observed common best practices for data cleaning and crowdsourced recruitment. This included recruiting high reputation "Master" workers, eliminating participants who did not train a model through several samples, and making the task largely as challenging to exploit as it would be to complete normally [46]. Participants were compensated $2.5 for completing an expected 10 minute duration session (approximating a $15.0 hourly wage; with a bonus incentive if they invested up to 20 minutes).

Each participant was first provided with a set of detailed instructions on what to expect in the session, after which, they completed a survey assessing their prior knowledge about AI and ML and their news reading habits. Participants then watched a short video tutorial (accessible throughout the session and customized to the experimental condition) that provided step-by-step instructions on how to use the system, at their own pace. There were also written instructions presented, after which participants started the task of teaching a machine learner how to identify clickbaits in news headlines. We situated the participants in a real world scenario where they fine-tune a pre-trained model [96] (held-out test set accuracy 60.9%) towards a marginally increased target of 70% accuracy (a common method in natural language processing). This end goal was designed to avoid inducing fatigue. Participants received a 100% bonus reward if they managed to achieve higher than the target. Participants could opt to end the teaching after reaching 70% accuracy or continue on without prompting. On end participants completed a post-task questionnaire which captured their overall experience with the process and their satisfaction with the machine learner.

We gathered data from participant responses on the pre-task survey that requested information about their age, education, domain

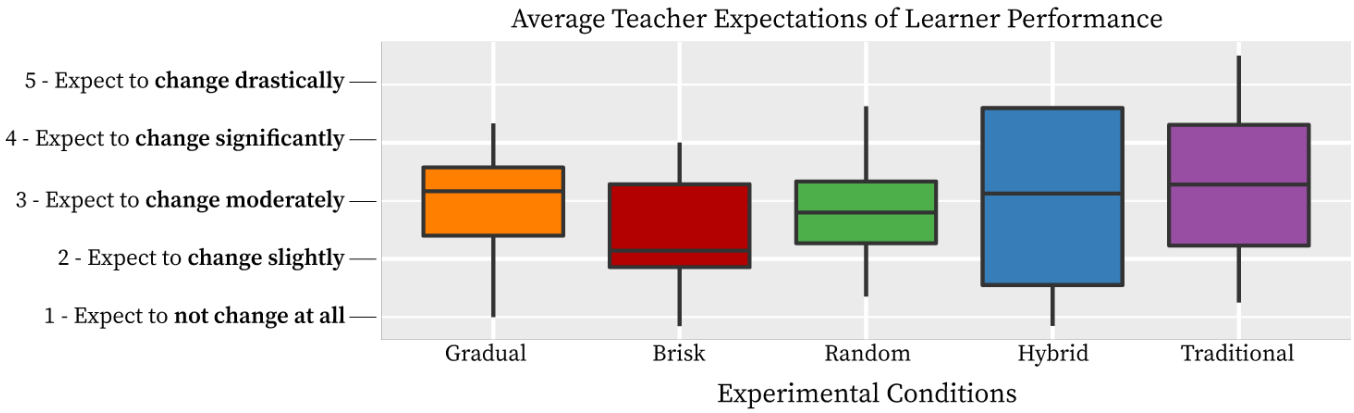

Figure 2: Distribution of average teacher expectations across teaching iterations (batch of 4 samples) for each experimental conditions. The *Traditional* and *Hybrid* experimental conditions both had significantly higher expectations than the *Gradual* and *Brisk* conditions ($p < 0.01$), but *Traditional* had only marginally higher expectations than *Random* and *Hybrid* ($p < 0.10$) and no effect could be detected from *Hybrid* to *Gradual* and *Brisk*

knowledge about news reading, and AI knowledge. During the task, we captured their feedback on each news headline across all conditions (Fig 1.B). We further requested participants to provide an estimate for two additional parameters during the task. They first estimated the informativeness of the headline by labelling what kind of information it might provide to the learner and assessing how informative they *expected* the headline to be for the learner based on their experiences so far. They rated expected learning by estimating how the model's future performance would change after learning from the sample. We also integrated an experience sampling modal dialogue [48] which intermittently captured how participants felt about the learner performance (after every 5 iterations) as the teaching session progressed. This modal asked the users to rate from 1- to 5-stars how well it has learned so far, and how well do they expect it to perform going forward. Finally, once the teaching session was complete, we administered a post-task survey to capture how challenging participants found the task, the cognitive load they experienced [37], and whether they were likely to use and recommend their model for an actual task.

4 RESULTS

92 individuals participated in our study via Amazon Mechanical Turk (not including the 36 participants in the pilot studies and not reported here) under supervision of an Institutional Review Board. Of those 92 participants, 75 of them successfully performed *at least* one iteration (4 headlines) to train the model. The remaining 17 participants were excluded from our analysis. 75 participants created 75 different clickbait identification models using a pool of about 1,500 news headlines (average of 15 participants in each condition) The participants followed a wide age distribution, with 36 participants reporting as below 35 and remaining above 36 years of age. 21 participants identified as female, 50 as male, 0 as non-binary, and remaining did not disclose their gender identity. A majority of our participants reported having college degrees (N=43), some had advanced Master's degrees (N=17), a small proportion were High School graduates (N=13), and two participants reported having earned a PhD. Each participant spent an average of 20 minutes on the task (min=6, max=54), and they curated a dataset of an average of 25 news headline samples (min=6, max=211, std=33.45) during each session. With regards to the final models produced by the participants, there was variability in their performance depending on the type and number of data instances they taught. While some participants, generally those who only trained a few samples, did not do much better than random chance, 12 participants broke the 80% accuracy threshold. On average participants taught models with mean accuracy of 69.8% (std=8.48).

4.1 RQ1 - Expectations and Learner Performance

For each data instance participants selected to teach, they estimated on a 5-point Likert scale how likely they thought the learner performance was to change as a result of learning this new instance. This estimation is a reasonable proxy of their expected future behavior of the learner: if they expected it to learn a lot from the sample, then they are likely to expect the performance to change drastically. We compared the individual judgments of each participant using an ANOVA, observing a significant difference between how each participant rated their expectations based on their teaching strategy condition assignment ($F(4, 1576) = 14.39$, $p < 0.001$). Further pairwise analysis with a Tukey HSD test showed that there was significant difference in participants' mean expectations between treatments pairs with opposing teaching strategies. The *Brisk* experimental condition had significantly worse expectations than all other conditions (all $p < 0.01$), perhaps because its learning strategy demonstrated rapid initial improvements which led to greater confidence in future results which did not ultimately bear fruit as it approached the asymptote.

When we grouped each expectation judgment by teaching style (bundling all conditions with *Active* teaching strategy), an ANOVA indicated a significant difference between *Hybrid* and *Traditional* teaching ($F(2, 1578) = 16.59$, $p < 0.001$). A Tukey HSD test was consistent with the previous findings, showing that *Traditional* had significantly higher than *Hybrid* which was also significantly higher than the *Active* teaching strategy (all pairs $p < 0.05$). This

Table 2: Effect of Teaching Time on Expectations

	Coefficient	t value	P value
(Intercept)	1.4150	23.11	<0.001
Current Iteration*	0.0037	5.53	<0.001
Last Iteration (lagged)*	0.3815	16.05	<0.001
Hybrid Condition (vs. Active)	-0.0763	-1.33	0.183
Traditional Condition (vs. Active)*	0.1877	3.37	<0.001

suggests that the teacher might have higher expectation if they have higher user engagement in determining teaching curriculum, as the *Traditional* condition, which required the participant to choose all samples, also resulted in the greatest expectations.

A similar phenomenon was observed in *Hybrid* where average expectations reported by the teacher on *self-selected samples* was also significantly higher than samples that were not selected by them ($t(608) = 2.356$, $p = 0.019$). As our work was conducted primarily with non-expert users, it remains to be seen whether experts might underestimate performance when provided agency as prior literature indicates [17].

However, these results operate on an instance level which may neglect time-dependent effects as a session unfolds and participants construct a mental model. In order to examine how past expectations influenced future expectations during a session, we constructed a lagged linear regression model that predicted participants' expectation ratings using 1) the rating's current iteration count (i.e. the order in which the rating was given, 1 for the first sample provided, 2 for the second provided, and so on), 2) a lagged expectation rating from 1 iteration before (to account for autoregressive effects over time), and 3) the simple teaching condition (Active, Hybrid, or Traditional). The resulting model identified an time-dependent effect on expectations (Table 2), and we observed that the past expectations played a role in present expectations. As before, the *Traditional* modality had higher expectations, though the signal was too weak to detect the intermediate expectations of *Hybrid*. Perhaps as non-experts became more familiar with the teaching environment and the learner, they grow more optimistic which is an encouraging sign for user engagement with MT systems.

We conducted an additional regression analysis to explore the relationship between participant expectations and the learner accuracy they observed. First, we extracted all of the instances where participants trained a batch of 4 instances and observed an accuracy change. We then eliminated the first observed accuracy number, as it had no basis for comparison. For the rest of the accuracy observations, we averaged their past 4 and future 4 expectation ratings (since training batch size is 4). We also computed the difference in accuracy from the present to the past, reflecting whether the model improved or reduced its performance. This permits us to explore the potential change of expectations as a result of accuracy shifts in the model. Our regression results for a model that predicts future expectation ratings based on accuracy shifts and past expectations is shown in Table 3.

The results of this regression suggest a relationship between past and future expectations, aligning with our RQ1 findings without any observable significance of changes in accuracy or interaction effect between change in accuracy and past expectations. Therefore, we cannot definitively conclude that whether participants adjusted expectations in response to model performance in general (the independent effect) or it matching or mis-matching their expectations (the interaction effect). Using this same dataset, we also examined whether prior expectations were predictive of accuracy changes following those expectations with no detectable signal for such relationship. In this case, we speculate that there may be other factors playing into expectation which may be masking any observable relationship.

As the participant progressed through the teaching session, we also gathered experiential data as participants rated how well they thought that the learner learned in the past and how well was it learning recently on a 5-point Likert scale. An ANOVA detected significant differences among the conditions ($F(4, 298) = 23.02$, $p < 0.001$). On an average, they rated the learner very poorly in the *Traditional* condition (M=2.8,STD=1.16). This could perhaps be because users in this condition saw unpredictable accuracy improvements due to relatively naive teaching strategies.

4.2 RQ2 - Prior Experience

In the pre-task survey we asked participants to report their overall experience with AI tools and software programming in general on a 7-point Likert scale. On an average a majority of the participants reported some level of knowledge of AI but had never used it (mean=2.9, std=1.32). 52% of the participants (N=48) did not have any formal knowledge of AI but had heard about it from news media or other sources. About a third (N=32) of the users had were familiar with the basic concepts of AI due to their own reading or research. 8 participants identified as advanced users of AI, and 3 reported being AI experts. We observed that the overall mean expectations reported by participants did not have any detectable relationship to prior AI experience ($t(75) = 1.1165$, $p = 0.268$). While prior knowledge and experience with ML tools did not have any observable impact on participant's initial expectations, participants with little or no prior knowledge of AI reported significantly lower expectations compared to those with some prior knowledge ($t(1760) = 2.9559$, $p = 0.003$). We also observed that expectations reported by the participants with minimal prior AI knowledge had lower deviation (mean=2.4, std=0.86) and variance (var=0.74) as compared to the other groups, perhaps due to floor effects. This suggests that expectations might be harder to manipulate within MT systems for complete novices.

Table 3: Effect of Accuracy Change on Expectations

	Coefficient	t value	P value
(Intercept)	0.1163	9.11	0.0000
Average Past Expectations*	0.0451	12.72	<0.001
Change in Accuracy	0.0233	1.08	0.282
Interaction Effect	0.0082	-1.05	0.294

4.3 RQ3 - Overall Satisfaction

In the post-survey we asked the participants to reflect on the teaching process and assess whether they would use this model for identifying clickbaits from their own news feeds and recommend it to others. We use teachers' self-reported measures of the willingness to use and recommend the model they trained as a reasonable proxy measure of how satisfactory they deemed the final outcome to be. Our goal was to capture a snapshot in the short term following the completion of a MT session, though we acknowledge that this may not fully capture satisfaction. We instead use the basis that a teacher whose expectations are met is more likely to use and recommend the system [11]. In this task context, we define the teacher's expectations to be met in two ways. Firstly, we consider the experience sampling data provided by the teacher through a quick 5-star rating on the question *How well is the AI understanding and learning from your feedback now?*. Using the self-reported variable, teacher's expectations are likely to be met if the average per-instance expectation in the previous training batch is lower than the 5-star rating. Secondly, we consider if there were any changes in accuracy after a lesson plan is taught. If the accuracy increased, the teacher's expectations are likely met. Since the goal assigned to the teacher is to improve the learner's accuracy, this is a reasonable assumption.

Using the experience sampling response as the independent variable, we observed that participants whose expectations were met had a significantly higher probability of recommending ($p = 0.002$) and using ($p < 0.001$) the learner they built. The null hypothesis for this analysis is that there is no relationship between expectations being met and the high likelihood of using/recommending the clickbait model they built. While the significance provides some supporting evidence, we also observed that the converse was not true; if the expectations were not met, there was a not a significant likelihood that the teacher will not recommend and use the model. We next considered whether the increase in accuracy had any significant impact, and surprisingly we observed no meaningful relationship between relative change in accuracy the teacher's overall satisfaction.

5 DISCUSSION

In this study we explored how teacher expectations impact the teaching process in an MT environment and are grounded in their perceptions of learner's capabilities and the teaching modalities afforded to them. In this section we will discuss the implications of our work and how it connects to the unique design challenges posed by interactive MT systems.

5.1 Managing impacts of user engagement in intelligent MT interfaces

Our system design facilitated engagement at various levels to encourage users to provide continued feedback to the learner. We observed that their expectations were strongly related to the level of engagement afforded by the interactive teaching modality. They expected the model to learn more when they selected the instances by themselves as opposed to when the interface assisted them in deciding the best candidates for learning. This reflects both perhaps a sense of optimism in participants' own abilities as well as a potential attitudinal shift as a result of improved agency in determining the teaching curriculum. An implication of this observation is in teaching environments where the data are nuanced and the learner has a gradual learning curve (e.g., identifying news articles with specific sentiments regarding a topic). The large feature space of how a sentiment can be defined results in a multitude of different potential teaching strategies. It is incredibly useful to have MT interfaces in these applications because they readily allow users to leverage their domain expertise to tune models. However, if the interface does not provide a matching level of agency and engagement in its teaching modality, then it may reduce the benefit of MT interfaces.

We observed a significant increase in participant expectations when they had no assistance from the learner (*Traditional* condition). This builds upon findings from prior research that people do not prefer to repeatedly respond to "yes" and "no" tasks , as has been a common practice in ML annotation tasks [59]. However, until this study, there was no empirical evidence that the "yes" and "no" tasks (as we implemented through *Active* condition) also negatively impact teacher's expectations of system performance. An important implication of this observations is that while promoting user engagement is beneficial, it also means increased user expectations. This may be harmful in some interactions like feature selection [88], resulting in teaching goals that are either too direct or too complex. Intelligent MT interfaces should carefully scrutinize of the impact of system's expressibility on teacher expectations. Consider for instance, few shot learners like GPT-2 [16] and ChatGPT, which can learn complex tasks through natural language instructions. In this setting, the teacher's pool of possible data instances to teach from is potentially infinite and significantly impact teachers' initial expectations and their continued usage of the systems. This is another reason why we did not select extremely performant learners for our study.

Despite several projects that suggest high user engagement is the key to success of a user interface [64], we argue that intelligent machine teaching interfaces should be cautious in providing levers of control and maximum controllability in learning strategies, especially when the data are rich and nuanced. As more non-expert participants use MT systems, there is a greater risk that misconceptions could shape the success of the process more than neutral assessments of performance. Likewise, the impact may be even greater as MT extends to applications which can learn with more intuitive interactions like natural language (e.g. ChatGPT). For highly efficient learning algorithms, this might lead to overestimating learners' abilities and placing too much trust in them [9]. A useful approach for intelligent MT interface might be to capture

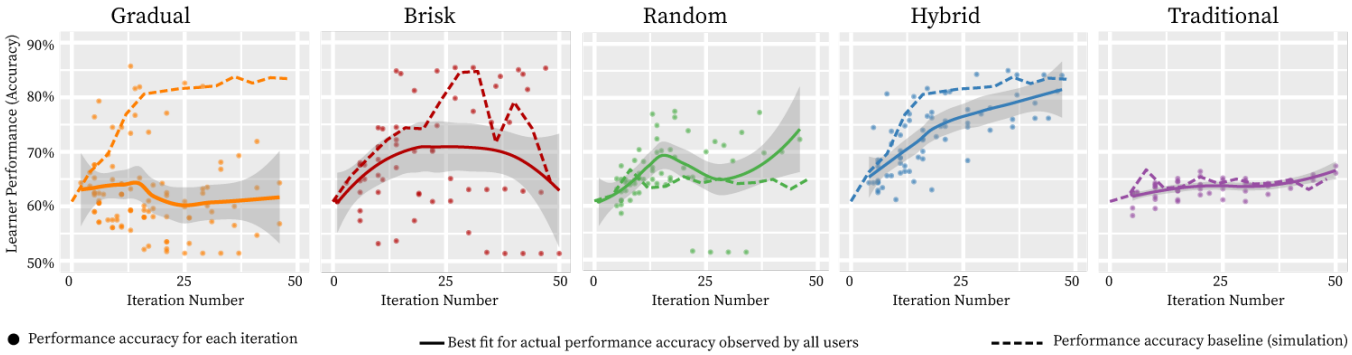

Figure 3: Accuracy achieved by the learner under different experimental conditions during 50 teaching iterations (mean + 1 std). The dotted line shows the learner accuracy achieved under simulated environment where labeled instances are drawn from our datasets [7] without any active human teacher.

the teacher's intent behind the concept they want to teach, but not necessarily demand micro-level control of all teaching decisions. MT interfaces can then use these models of intent to direct user's attention towards data instances or features that are representative of the user's decision boundary and also have high information entropy from which the model can learn.

5.2 Adapting to evolving teacher expectations

The goal of machine teaching is to effectively integrate a teacher's understanding of the task to the machine learner. We observed that prior knowledge had no significant impact on the teacher's expectations in an MT interface. However, as the session progressed, their initial expectations evolved. The teacher expectations serve as a baseline on which they evaluate the learner's performance [11]. For an MT system, this means that incremental improvements in learner's performance may not be perceived uniformly over a teaching session. For instance, a teacher whose expectations continuously increase over a teaching session may perceive even a high performing learner to be average. As a result, the teacher may try to introduce more difficult concepts that might be beyond the learner's capabilities, thus abandoning the system if it fails to learn, or re-calibrating their own expectations. This behavior is particularly problematic for learners that claim to be highly generalizable and can learn from only few instances (e.g., large language models like GPT-2 [16]), since they might quickly plateau in terms of their perceived performance improvements. For MT systems, thus it is critical to establish accurate teacher expectations early on in the process.

These findings present an opportunity for ongoing research in ML model explainability [33] and auditing [71] to guide the teacher towards what to expect. While significant research in explainability focuses on improving the transparency of the algorithmic decision-making process, research remains to be done in identifying what specific explanation features may lead to *reasonable* user expectations of model outcomes. This is especially critical in MT systems whose progress is guided by the feedback. For instance, while requesting for user feedback, the system might also present an explanation that *predicts* the potential costs and benefits of the user response thus grounding the teacher feedback in realistic expectations of the learner's abilities. We also call for research into higher fidelity techniques for visualizing ML model behavior [21] that are grounded in people's conceptual understanding of the task [56], perhaps offering a fairer assessment of learner performance at every step of the teaching process.

5.3 Facilitating accurate user perceptions of learner performance

Our study incorporated a singleton metric of performance assessment: accuracy on a held out test set. We did not provide the test set to the participants in order to reduce gaming of the metric. We anticipated that the changes in accuracy would guide the teacher as to how well the model is doing. However, we found that the teacher's willingness to use the model they trained and their overall satisfaction with the process was generally not impacted by the accuracy measure. Even users who attained greater than 80% accuracy did not feel sufficiently satisfied with the process to use the learner they built. There are a number of possible explanations for this observation. One is the weakness of accuracy with regards to interpretability and generalizability, especially for non-experts. Statistical metrics like accuracy provide a quantifiable representation of system performance which can serve as achievable teaching goals. However, when exposed to the learner over time, teacher's might develop other implicit assessments that are difficult to quantify and uncover empirically, like assessing model success in context to their every day lives.

Additionally, this observation suggests that the user experience in teaching has more to do with the process than the final outcome. If the teacher has to spend frustrating amounts of time to improve the model, they are more likely give up. For example, one participant mentioned that *"[It] seemed like my choices had no real effect unless I was really meant to do 100+ lessons"*. Therefore, metrics not only have to accurately inform the teacher about performance, but also potentially offer encouragement to the teacher so that they feel motivated to proceed during a long session. This desire for improved transparency to interpret system performance was also echoed in another participant's feedback: *"It would be very helpful if I could see more than just the headline."* (condition *Random*). Designing systems that assist teachers to better interpret the performance criteria, or even establish their own criteria, would help to mitigate the effects of evolving expectations on performance assessment.

With respect to the machine learning strategy employed in our study (a logistic regression model), we provided a fairly intuitive interaction paradigm: teach by selecting a group of positive and negative examples [6, 18]. We observed that when the interface provided no guidance on samples that are likely to yield better performance, the users had high expectations. One cause for this might be that when the teacher's instructions are based entirely on their judgement and domain knowledge, they are likely to expect the model to learn more. This may have a devastating effect on the success of MT system with free-form interactions where the space of possible ways to teach a concept is vast. Perhaps the most relevant example of this phenomenon can be observed when designing prompts for large language models[50]. Thus it is also important to note in this regard that when the interactions support providing detailed instructions, teachers who rely primarily on their domain knowledge and have little understanding of the task may struggle to teach.

6 LIMITATIONS

There are a number of limitations for our work. Foremost, our use of accuracy as an assessment technique and limited methods for feedback might have shaped participants' perceptions of both the learner as well as the MT process as a whole. While clickbait is a familiar task for many, it may not have been a completely familiar domain task for all participants, which could play a role in both the accuracies of end models as well as affective metrics we observed. Even though, we did not explicitly study the impact of task expertise on teacher expectations, teacher's ability to identify a clickbait might impact their expectations and perceptions of the learner performance. In future studies, we intend to factor in task expertise and other usability dimensions like cognitive load and teacher expressibility and application context. Including these factors will help build more nuanced user models of MT process. Likewise, while we observed several setpoints for agency on part of the teacher, there are a wide variety of potential interventions which might help to shed more light on the issue of teacher expressability. For instance, conducting comprehensive user interviews can help to understand the teacher's sensemaking process in great detail. Overall, while we were able to make some inroads in understanding some of the human factors at play in MT, our investigation is by no means exhaustive. In future work, we intend to expand our research to examine how different feedback mechanisms, measures for assessing model performance, and teaching strategies affect MT outcomes.

7 CONCLUSION

In this paper, we implemented an interactive system in order to gain a qualitative and quantitative understanding of the human factors and human-teacher/machine-learner relationship in MT. Through deployment with a variety of non-expert machine teachers in a common classification task, we demonstrated how different affective and performance attributes shaped the overall success or failure of the MT process. We showed that participant expectations of learning affected both the progress and outcomes of an MT session, and that specific software affordances that provided teachers with additional teaching tools changed the affective landscape of the session. By examining the human factors of MT systems, our work uncovers the rich interplay between learners and teachers in MT. Our work sheds light on broader questions about the role of stakeholders and models in the development of Machine Teaching systems.

ACKNOWLEDGMENTS

We thank Bloomberg AI for supporting this research through Data Science Fellowship. We also thank the reviewers for their comments and suggestions that helped to improve this work.

REFERENCES

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. 2011. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems* (Granada, Spain) *(NIPS'11)*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.). Curran Associates Inc., Red Hook, NY, USA, 2312–2320.

[2] BS Abhigna, Nitasha Soni, and Shilpa Dixit. 2018. Crowdsourcing–A step towards advanced machine learning. *Procedia computer science* 132 (2018), 632–642.

[3] Hillary Abraham, Bobbie Seppelt, Bruce Mehler, and Bryan Reimer. 2017. What's in a Name: Vehicle Technology Branding & Consumer Expectations for Automation. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 226–234. https://doi.org/10.1145/3122986.3123018

[4] Amol Agrawal. 2016. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. IEEE, India, 268–272. https://doi.org/10.1109/NGCT.2016.7877426

[5] 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.

[6] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive Machine Learning for on-Demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2207676.2207680

[7] Aman Anand. 2020. Clickbait dataset. https://www.kaggle.com/datasets/amananandrai/clickbait-dataset Accessed Oct 2021.

[8] Alejandro Correa Bahnsen, Djamia Aouada, and Björn Ottersten. 2014. Example-Dependent Cost-Sensitive Logistic Regression for Credit Scoring. In *2014 13th International Conference on Machine Learning and Applications*. ICML, USA, 263–269. https://doi.org/10.1109/ICMLA.2014.48

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[10] Francisco Bernardo, Michael Zbyszynski, Rebecca Fiebrink, and Mick Grierson. 2016. Interactive Machine Learning for End-User Innovation. In *Designing the User Experience of Machine Learning Systems*. American Association for Artificial Intelligence (AAAI), USA, 369–375. https://research.gold.ac.uk/id/eprint/19767/

[11] Anol Bhattacherjee. 2001. Understanding Information Systems Continuance: An Expectation-Confirmation Model. *MIS Quarterly* 25, 3 (2001), 351–370. http://www.jstor.org/stable/3250921

[12] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information Sciences* 497 (2019), 38–55.

[13] Léon Bottou. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, 421–436.

[14] Daniel S. Brown and Scott Niekum. 2019. Machine Teaching for Inverse Reinforcement Learning: Algorithms and Applications. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 7749–7758. https://doi.org/10.1609/aaai.v33i01.33017749

[15] Susan A Brown, Viswanath Venkatesh, and Sandeep Goyal. 2012. Expectation confirmation in technology use. *Information Systems Research* 23, 2 (2012), 474–487.

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[17] Katherine A Burson, Richard P Larrick, and Joshua Klayman. 2006. Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration

in relative comparisons. *Journal of personality and social psychology* 90, 1 (2006), 60.

[18] Maya Cakmak and Andrea L. Thomaz. 2014. Eliciting Good Teaching from Humans for Machine Learners. *Artif. Intell.* 217, C (dec 2014), 198–215. https://doi.org/10.1016/j.artint.2014.08.005

[19] Davide Calvaresi, Giovanni Ciatto, Amro Najjar, Reyhan Aydoğan, Leon Van der Torre, Andrea Omicini, and Michael Schumacher. 2021. Expectation: Personalized Explainable Artificial Intelligence for Decentralized Agents with Heterogeneous Knowledge. In *Explainable and Transparent AI and Multi-Agent Systems: Third International Workshop, EXTRAAMAS 2021, Virtual Event, May 3–7, 2021, Revised Selected Papers*. Springer-Verlag, Berlin, Heidelberg, 331–343. https://doi.org/10.1007/978-3-030-82017-6_20

[20] Abhijnan Chakraborty, Rajdeep Sarkar, Ayushi Mrigen, and Niloy Ganguly. 2017. Tabloids in the Era of Social Media? Understanding the Production and Consumption of Clickbaits in Twitter. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 30 (dec 2017), 21 pages. https://doi.org/10.1145/3134665

[21] Angelos Chatzimparmpas, Rafael Messias Martins, Ilir Jusufi, Kostiantyn Kucher, Fabrice Rossi, and Andreas Kerren. 2020. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 713–756.

[22] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438 arXiv:https://doi.org/10.1080/10447318.2020.1841438

[23] Maria D. Molina, S Shyam Sundar, Md Main Uddin Rony, Naeemul Hassan, Thai Le, and Dongwon Lee. 2021. Does clickbait actually attract more clicks? Three clickbait studies you must read. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[24] Tirtharaj Dash, Sharad Chitlangia, Aditya Ahuja, and Ashwin Srinivasan. 2022. A review of some techniques for inclusion of domain-knowledge into deep neural networks. *Scientific Reports* 12, 1 (2022), 1–15.

[25] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 278–288. https://doi.org/10.1145/3025453.3025739

[26] Zhengfang Duanmu, Kede Ma, and Zhou Wang. 2018. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach. *IEEE Transactions on Image Processing* 27, 12 (2018), 6135–6146.

[27] John J Dudley and Per Ola Kristensson. 2018. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–37.

[28] Kirstin Early, Stephen E. Fienberg, and Jennifer Mankoff. 2016. Test Time Feature Ordering with FOCUS: Interactive Predictions with Minimal User Burden. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 992–1003. https://doi.org/10.1145/2971648.2971748

[29] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. 2011. Human Model Evaluation in Interactive Supervised Learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 147–156. https://doi.org/10.1145/1978942.1978965

[30] Wilbert O Galitz. 2007. *The essential guide to user interface design: an introduction to GUI design principles and techniques*. John Wiley & Sons.

[31] 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. *arXiv* (2020), arXiv–2001.

[32] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 235 (jan 2021), 28 pages. https://doi.org/10.1145/3432934

[33] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[34] Sally A Goldman and H David Mathias. 1996. Teaching a smarter learner. *J. Comput. System Sci.* 52, 2 (1996), 255–267.

[35] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and Kevin McIntosh. 2014. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering* 40, 3 (2014), 307–323. https://doi.org/10.1109/TSE.2013.59

[36] Anil Gupta, Neeraj Dhiman, Anish Yousaf, and Neelika Arora. 2021. Social comparison and continuance intention of smart fitness wearables: An extended expectation confirmation theory perspective. *Behaviour & Information Technology* 40, 13 (2021), 1341–1354.

[37] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

[38] Jani Heikkinen, Thomas Olsson, and Kaisa Väänänen-Vainio-Mattila. 2009. Expectations for User Experience in Haptic Communication with Mobile Devices. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Bonn, Germany) *(MobileHCI '09)*. Association for Computing Machinery, New York, NY, USA, Article 28, 10 pages. https://doi.org/10.1145/1613858.1613895

[39] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. 2019. Interactive Machine Learning: Experimental Evidence for the Human in the Algorithmic Loop. *Applied Intelligence* 49, 7 (jul 2019), 2401–2414. https://doi.org/10.1007/s10489-018-1361-5

[40] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2020. Crowdsourcing the perception of machine teaching. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[41] Eva Hudlicka. 2003. To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies* 59, 1 (2003), 1–32. https://doi.org/10.1016/S1071-5819(03)00047-8 Applications of Affective Computing in Human-Computer Interaction.

[42] K. Höök. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* 12, 4 (2000), 409–426. https://doi.org/10.1016/S0953-5438(99)00006-5

[43] Luis-Daniel Ibáñez, Neal Reeves, and Elena Simperl. 2020. Crowdsourcing and Human-in-the-Loop for IoT. *The Internet of Things: From Data to Insight* (2020), 91–105.

[44] Liu Jiang, Shixia Liu, and Changjian Chen. 2019. Recent research advances on interactive machine learning. *Journal of Visualization* 22, 2 (2019), 401–417.

[45] Muneo Kitajima and Peter G. Polson. 1992. A Computational Model of Skilled Use of a Graphical User Interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 241–249. https://doi.org/10.1145/142750.142803

[46] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[47] David G Kleinbaum, K Dietz, M Gail, Mitchel Klein, and Mitchell Klein. 2002. *Logistic regression*. Springer.

[48] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology*. Springer, 21–34.

[49] Mingkun Li and Ishwar K Sethi. 2006. Confidence-based active learning. *IEEE transactions on pattern analysis and machine intelligence* 28, 8 (2006), 1251–1261.

[50] Vivian Liu and Lydia B Chilton. 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 384, 23 pages. https://doi.org/10.1145/3491102.3501825

[51] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. 2017. Iterative machine teaching. In *International Conference on Machine Learning*. PMLR, 2149–2158.

[52] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. 2018. Towards black-box iterative machine teaching. In *International Conference on Machine Learning*. PMLR, 3141–3149.

[53] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[54] Oisin Mac Aodha, Vassilios Stathopoulos, Gabriel J Brostow, Michael Terry, Mark Girolami, and Kate E Jones. 2014. Putting the scientist in the loop–Accelerating scientific progress with interactive machine learning. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 9–17.

[55] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[56] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Crowdsourcing and evaluating concept-driven explanations of machine learning models. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[57] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[58] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

[59] Robert Munro and Robert Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI.* Simon and Schuster.

[60] Christine Murad and Cosmin Munteanu. 2019. "I Don't Know What You're Talking about, HALexa": The Case for Voice User Interface Guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 9, 3 pages. https://doi.org/10.1145/3342775.3342795

[61] Bilal Naeem, Aymen Khan, Mirza Omer Beg, and Hasan Mujtaba. 2020. A deep learning framework for clickbait detection on social area network using natural language cues. *Journal of Computational Social Science* 3, 1 (2020), 231–243.

[62] Hieu T Nguyen and Arnold Smeulders. 2004. Active learning using pre-clustering. In *Proceedings of the twenty-first international conference on Machine learning*. 79.

[63] J. Nielsen. 1993. Iterative user-interface design. *Computer* 26, 11 (1993), 32–41. https://doi.org/10.1109/2.241424

[64] Jakob Nielsen. 1994. Usability Inspection Methods. In *Conference Companion on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 413–414. https://doi.org/10.1145/259963.260531

[65] Simon Nusinovici, Yih Chung Tham, Marco Yu Chak Yan, Daniel Shu Wei Ting, Jialiang Li, Charumathi Sabanayagam, Tien Yin Wong, and Ching-Yu Cheng. 2020. Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology* 122 (2020), 56–69. https://doi.org/10.1016/j.jclinepi.2020.03.002

[66] Richard L Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of marketing research* 17, 4 (1980), 460–469.

[67] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-Shared AI Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376450

[68] Reid Porter, James Theiler, and Don Hush. 2013. Interactive Machine Learning in Data Exploitation. *Computing in Science & Engineering* 15, 5 (2013), 12–20. https://doi.org/10.1109/MCSE.2013.74

[69] Abinash Pujahari and Dilip Singh Sisodia. 2021. Clickbait detection using multiple categorisation techniques. *Journal of Information Science* 47, 1 (2021), 118–128.

[70] Anant Raj and Francis Bach. 2022. Convergence of uncertainty sampling for active learning. In *International Conference on Machine Learning*. PMLR, 18310–18331.

[71] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[72] Maria Riveiro and Serge Thill. 2021. "That's (not) the output I expected!" On the role of end user expectations in creating explanations of AI systems. *Artificial Intelligence* 298 (2021), 103507. https://doi.org/10.1016/j.artint.2021.103507

[73] Dominik Sacha, Matthias Kraus, Daniel A. Keim, and Min Chen. 2019. VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 385–395. https://doi.org/10.1109/TVCG.2018.2864838

[74] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2021. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* 109, 3 (2021), 247–278. https://doi.org/10.1109/JPROC.2021.3060483

[75] BEN SHNEIDERMAN. 1982. The future of interactive systems and the emergence of direct manipulation. *Behaviour & Information Technology* 1, 3 (1982), 237–256. https://doi.org/10.1080/01449298208914450 arXiv:https://doi.org/10.1080/01449298208914450

[76] Patrice Simard, Saleema Amershi, Max Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Chris Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. *Machine Teaching: A New Paradigm for Building Machine Learning Systems.* Technical Report MSR-TR-2017-26. https://www.microsoft.com/en-us/research/publication/machine-teaching-new-paradigm-building-machine-learning-systems/

[77] Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 69, 1 (1955), 99–118.

[78] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martín Pereira-Fariña. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access* 9 (2021), 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

[79] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting meaningfully with machine learning systems: Three experiments. *International journal of human-computer studies* 67, 8 (2009), 639–662.

[80] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (01 2020), 74–88. https://doi.org/10.1093/jcmc/zmz026 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/74/32961171/zmz026.pdf

[81] Ashlesha Vaidya. 2017. Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 1–6. https://doi.org/10.1109/ICCCNT.2017.8203946

[82] Aikaterini C Valvi and Douglas C West. 2013. E-loyalty is not all about trust, price also matters: extending expectation-confirmation theory in bookselling websites. *Journal of Electronic Commerce Research* 14, 1 (2013), 99.

[83] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[84] Cornelis Joost Van Rijsbergen. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation* (1977).

[85] Kiri L. Wagstaff. 2012. Machine Learning That Matters. In *Proceedings of the 29th International Coference on International Conference on Machine Learning* (Edinburgh, Scotland) *(ICML'12)*. Omnipress, Madison, WI, USA, 1851–1856.

[86] MALCOLM WARE, EIBE FRANK, GEOFFREY HOLMES, MARK HALL, and IAN H WITTEN. 2001. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies* 55, 3 (2001), 281–292. https://doi.org/10.1006/ijhc.2001.0499

[87] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[88] 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 (TOCHI)* 26, 4 (2019), 1–27.

[89] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. 2018. Accelerating Human-in-the-Loop Machine Learning: Challenges and Opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning* (Houston, TX, USA) *(DEEM'18)*. Association for Computing Machinery, New York, NY, USA, Article 9, 4 pages. https://doi.org/10.1145/3209889.3209897

[90] Luyao Yuan, Dongruo Zhou, Junhong Shen, Jingdong Gao, Jeffrey L Chen, Quanquan Gu, Ying Nian Wu, and Song-Chun Zhu. 2021. Iterative Teacher-Aware Learning. *Advances in Neural Information Processing Systems* 34 (2021), 29231–29245.

[91] Lijun Zhang, Rong Jin, Chun Chen, Jiajun Bu, and Xiaofei He. 2021. Efficient Online Learning for Large-Scale Sparse Kernel Logistic Regression. *Proceedings of the AAAI Conference on Artificial Intelligence* 26, 1 (Sep. 2021), 1219–1225. https://doi.org/10.1609/aaai.v26i1.8300

[92] Zhigang Zhang, Wangshu Cheng, and Zhenyu Gu. 2016. User experience studies based on expectation dis-confirmation theory. In *International Conference of Design, User Experience, and Usability*. Springer, 670–677.

[93] Jingbo Zhu, Huizhen Wang, Benjamin K. Tsou, and Matthew Ma. 2010. Active Learning With Sampling by Uncertainty and Density for Data Annotations. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 6 (2010), 1323–1331. https://doi.org/10.1109/TASL.2009.2033421

[94] Xiaojin Zhu. 2015. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[95] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. 2018. An overview of machine teaching. *arXiv preprint arXiv:1801.05927* (2018).

[96] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2020. A comprehensive survey on transfer learning. *Proc. IEEE* 109, 1 (2020), 43–76.