NALIN CHHIBBER, University of Waterloo, Canada JOSLIN GOH, University of Waterloo, Canada EDITH LAW, University of Waterloo, Canada

Traditional crowdsourcing has mostly been viewed as requester-worker interaction where requesters publish tasks to solicit input from human crowdworkers. While most of this research area is catered towards the interest of requesters, we view this workflow as a teacher-learner interaction scenario where one or more human-teachers solve Human Intelligence Tasks to train machine learners. In this work, we explore how teachable machine learners can impact their human-teachers, and whether they form a trustable relation that can be relied upon for task delegation in the context of crowdsourcing. Specifically, we focus our work on teachable agents that learn to classify news articles while also guiding the teaching process through conversational interventions. In a two-part study, where several crowd workers individually teach the agent, we investigate whether this learning by teaching approach benefits human-machine collaboration, and whether it leads to trustworthy AI agents that crowd workers would delegate tasks to. Results demonstrate the benefits of the learning by teaching approach, in terms of perceived usefulness for crowdworkers, and the dynamics of trust built through the teacher-learner interaction.

CCS Concepts: • Human-centered computing \rightarrow Collaborative interaction; User studies.

Additional Key Words and Phrases: Human-AI Interaction, Learning by Teaching, Interactive Machine Learning, Conversational Interactions, Trusting Conversational AI

ACM Reference Format:

Nalin Chhibber, Joslin Goh, and Edith Law. 2022. Teachable Conversational Agents for Crowdwork: Effects on Performance and Trust. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 331 (November 2022), 21 pages. https://doi.org/10.1145/3555223

1 INTRODUCTION

Recent progress in artificial intelligence and rapidly advancing infrastructure around conversational technologies has resulted in the development of intelligent agents with a wide range of applications, including intelligent personal assistants (like Alexa, Cortana, Siri, and Google Assistant), guides in public places [29, 72], smart-home controllers [68], and even cars [46]. This growing ecosystem of applications supporting conversational capabilities has the potential to affect all aspects of our lives, including healthcare, education, work, and leisure. Consequently, agent-based interactions has attracted a lot of attention in HCI research [12, 45, 47, 50, 68]. The success of these agents will depend on their ability to efficiently learn from non-expert humans in a natural way. Therefore, a

Authors' addresses: Nalin Chhibber, University of Waterloo, 200 University Avenue West, Waterloo, Canada, nalin.chhibber@uwaterloo.ca; Joslin Goh, University of Waterloo, 200 University Avenue West, Waterloo, Canada, jtcgoh@uwaterloo.ca; Edith Law, University of Waterloo, 200 University Avenue West, Waterloo, Canada, edith.law@uwaterloo.ca.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(*s*) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2022/11-ART331 \$15.00

https://doi.org/10.1145/3555223

331

lot of work in Computer Science, Psychology, and Educational Research is focused on the efficacy of teachable agents i.e., agents that can learn from human teaching.

In a similar spirit, trust on intelligent agents has been studied extensively across multiple disciplines of Human-Computer-Interaction, Artificial Intelligence, Robotics, and Psychology. While the notion of trust has been explored differently across these areas, its basic formulation has remained the same—trust is the tendency to be vulnerable to the actions of others while believing in a high chance of positive outcome [36, 64]. Recent research has found that trust can predict the level of reliance on technology as well as the overall technology acceptance [31, 36, 62]. In the context of Human-AI interaction, trust is particularly important because it addresses the challenges associated with the complexity and non-determinism of AI-driven systems. Several models of trust have been explored in multi-agent systems research. However, the notion of trust has not been studied exclusively in the context of teachable agents.

In this paper, we explore how teaching agents to assist with crowdwork affects workers' own learning performance and their willingness to delegate tasks to the agent that they trained. We conducted two experiments where crowd workers were asked to teach text classification to a conversational agent, with the assumptions that they can delegate tasks to the trained agent at a later time. The premise is that such a teachable agent will provide both performance benefits (i.e., workers will be able to perform the task better by teaching the agent) and delegation benefits (i.e., workers will be willing to delegate part of their work to the agent, thus lessening their own workload). These benefits, for course, would depend critically on two questions—whether workers do benefit from teaching, and whether workers would trust the agent that they trained themselves enough to delegate tasks to them. Our experimental results provide insights to these two questions. We provide our code, sample conversations from experiments, analysis data, and trained classifier for transparency, replication and extensions ¹. Overall, this paper makes the following key contributions:

- A study quantifying the extent to which crowdworkers perform better by teaching an agent.
- A study evaluating workers' trust and willingness to delegate tasks to an agent that they taught
- A set of design guidelines for the use of teachable agents in crowdsourcing contexts.

The rest of this paper is organized as follows. Section 2 briefly reviews some related work. In Section 3, we formally introduce our research questions with the corresponding hypotheses. Details about the system and agent's interaction model is described in Section 4. Individual experiments focusing on performance and trust are presented in Section 5 and Section 6 respectively. Finally, we discuss the implication of our work along with directions for future research.

2 BACKGROUND AND RELATED WORK

2.1 Improving Crowdworkers Performance

One of the interesting problems that crowdsourcing researchers have actively explored is improving the performance of workers and their engagement in the tasks. Studies have found that while task performance often depends on monetary factors such as magnitude of reward [35], and change in payment level in subsequent tasks [77], it also depends on crowdworkers' perceived fairness in the pay [76]. In contrast, many researchers have also focused on non-monetary mechanisms to support crowdworkers. Morris et al. suggested the use of implicit mechanisms to induce observable changes in behavior and improve worker performance in microtask crowdsourcing environments [55]. Dow et al. found that the quality of work in content creation tasks can be improved with self-assessment and external assessment [26]. In a similar spirit, Doroudi et al. showed that providing

¹Supplementary material available at: https://teachable-conversational-agents.github.io/#cscw

expert examples and asking workers to validate the contributions of peer workers, are also effective forms of improving workers' performance [25]. Gadiraju et al. investigated the effect of implicit and explicit training on worker performance, and removal of unreliable workers. They found that instructing workers on the overall task (explicit training) is as effective as training them only on erroneous responses (implicit training) [30]. Law et al. explored the use of curiosity as a new type of intrinsic motivational driver to incentivize crowd workers [41]. Mavridis et al. showed that crowd workers perform microtasks more effectively when they interact with a text-based chatbot, compared to the traditional web interface in a variety of typical microtasks [53]. Likewise, in this work, we investigate whether the act of teaching an agent would help crowdworkers improve their own performance at the task.

2.2 Humans as Teachers of Machines

Our work is closely related to the research on teachable agents, which are known to produce Protégé effect—a psychological phenomenon that suggests that teaching others can be more effective than self-work [14]. Previous work in cognitive science and education research demonstrates the presence of the Protégé effect in reciprocal teaching [60], peer-assisted tutoring [21], small-group interaction [74] and self-explanation [19]. Studies focusing on the cognitive benefits of teaching suggest that preparing to teach may produce more organized cognitive structures than learning the material for oneself [5]. Biswas et al. has shown that the expectation to teach others helps in self-reflection, builds a sense of responsibility, and is useful for meaningful structuring of information [6, 7]. This has been confirmed in later studies that demonstrate the effectiveness of the Protégé effect for cognitive [58], meta-cognitive [57] and motor learning skills [37]. Despite its recognized benefits, Protégé effect has not been explored in the context of crowdsourcing, and hence one of the primary focus of our work. Further, while the usefulness of Protégé effect has been studied in several psychological studies, it is unexplored how this experience of teaching an agent is actually perceived by the humans. In this work, we aim to address this question by investigating whether crowdworkers find the task of teaching an agent enjoyable and useful.

Beyond teachable agents, there is a wealth of literature on Human-AI collaboration that focuses on teaching and learning aspect of the interaction. *Interactive machine learning*, for instance, involves users in the process of training and optimizing the machine learning models through rapid, focused and incremental updates to the model. Within the HCI community, interactive machine learning was first explored by Fails and Olsen [28], who studied the difference between classical and interactive machine learning and introduce an interactive feature selection tool for image recognition. Porter et al. [63] formally breaks down the interactive machine-learning process into three dimensions: task decomposition, training vocabulary, and training dialogue. These dimensions define the level of coordination, type of input, and level/frequency of interaction between the end-users and machine learners. Later, some researchers examined the role of humans in interactive machine learning, and highlighted various areas where humans have interactively helped machine learning systems to solve a problem [3]. Their case study covered various situations where humans were seen as peers, learners, or even teachers while engaging with interactive systems across different disciplines like image segmentation and gestured interactions.

A special case of interactive machine learning is *active learning* which focuses on improving machine learner's performance by actively querying a human oracle and obtain labels [70]. However, several studies reveal that active learning can cause problems when applied to truly interactive settings [8, 9, 34]. Therefore, an emerging research in this area has started to look at humans as teachers and not just annotators. Simard et al. formalize the role of teachers as someone who transfer knowledge to learners in order to generate useful models [71]. Past work on algorithmic teaching shows that while human teachers can significantly improve the learning rate of a machine

learning algorithm [4, 32, 51], they often do not spontaneously generate optimal teaching sequences as human teaching is mostly optimized for humans, not machine learning systems. Cakmak et al. examined several ways to elicit good teaching from humans for machine learners [10]. They proposed the use of teaching guidance in the form of algorithms or heuristics.

It is important to note that these forms of "machine teaching" is different from prior work on "teachable agents" in the tutoring context where the main goal is to help students learn better. While machine teaching have been extensively studied, it has not been explored within the context of conversational systems that interactively learns a task from humans.

2.3 Human-Computer Trust

The concept of trust between humans and computers with different embodiments has been studied across multiple disciplines. McAllister et al. define Human Computer Trust as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" [54]. This definition of trust not only encompasses user's confidence in the system, but also their willingness to be vulnerable towards the system's decisions, hinting towards the idea that while these two bases of trust may seem different, they are still related to one another. Madsen et al define a hierarchical model for human computer trust with five basic constructs (understandability, technical competence, reliability, attachment and faith) that can be divided in two general categories: cognition-based trust and affect-based [49]. While the cognition-based trust is based on the user's intellectual perceptions of the system's characteristics, affect based trust is based on the user's emotional responses to the system. Beyond this, trust has also been widely regarded as a dynamic concept that is prone to changes based on the behaviour of the trusted agent [23, 67]. Due to its dynamic nature, researchers have investigated the factors that may influence the level of trust that users are willing to place in the agents. Lim et al. showed how different kinds of explanations can influence Human Computer Trust [43]. They studied the effect of Why, Why-not, How-to and What-if explanations on trust and understanding system's actions or reactions. Chattaraman, et al. found that adding visual presence to an agent on a shopping website increased participants' trust and intention to visit the website [15]. Wang et al. found that agent's transparency has a greater effect on cognitive trust whereas adding a visual representation only influences the emotional trust [73]. Despite this, the notion of trust has not been studied in the context of teachable agents. Our work fills this gap by investigating whether workers are willing to delegate tasks to an agent that they themselves teach.

2.4 Evaluating Conversational Interactions

Our teachable agent is conversational in nature—it prompts for teaching and receives feedback through dialog. In existing literature, user interactions with conversational agents have been measured through both objective and subjective metrics, and across various dimensions such like task performance, user attitudes, perceived usability, system usage, and cognitive load. Task performance is measured using metrics such as the total number of conversational turns [42], percentages of tasks completed correctly [59], and task completion time [61]. Attitude of users towards conversational agents involves measuring likeability and human likeness [20]. Perceived usability has been mostly examined through scale based questionnaires on perceived ease of use and learnability [27]. Prior work has used system usage as an indicator of what people use conversational interfaces for [22] and how they used them [66]. Finally, some studies measured cognitive load to identify the physical, mental and temporal demands on users while they are interacting with conversational systems. The evaluative measures used in our studies are informed by these prior works.

3 RESEARCH QUESTIONS

In this work, we conducted two experiments in which crowd workers individually teach a conversational agent. We investigate whether this learning by teaching approach improves the performance of the human teacher and leads to trusted AI agents that crowd workers would use. We address the following research questions through our studies:

RQ1: Does teaching an agent impact crowdworkers performance in the task?

Despite the comprehensive work on crowdsourcing, improving performance of crowdworkers is challenging because they often have different skills with varying level of competence [30, 52]. Moreover, feedback on the workers' progress is limited, and informative features for data is usually not available [13, 44, 56]. Therefore, we augment this area of research by exploring the use of teachable agents as a mechanism to improve the performance of crowdworkers. The core idea behind this is inspired from Protégé Effect which states that teaching a task to someone else is more beneficial than doing it for oneself [14]. Traditionally, this effect has been studied in peer-to-peer interaction scenarios in classrooms or controlled laboratory settings. In the first experiment, we intend to evaluate the benefits of learning by teaching paradigm within the context of a conversational agent. We hypothesize that:

[H1a]: Participants who teach the agent will perform better than those who do not.

[H1b]: Participants who teach the agent will find the task more enjoyable and useful than those who do the task themselves.

RQ2: Does teaching an agent impact trust and influence crowdworkers' willingness to delegate tasks?

Establishing trust is a key pre-requisite for the adoption of AI systems. Previous work has explored the use of algorithmic transparency (explainability), robustness, bias, privacy, reproducibility and accountability to build trust [11]. In the second experiment, we explore the effects of involving end-users in the training/teaching process and how various factors impact their trust of the agent and their decision to delegate tasks to the agent.

[H2a]: Participants who teach their agent will delegate more tasks to the agent than participants who will not teach.

[H2b]: Participants who teach their own agent will subjectively experience higher levels of trust than those who will not teach an agent.

4 SYSTEM DESCRIPTION

In this work, we introduce a teachable agent that learns to classify text using human feedback through conversational interactions. This section introduces the task environment for our experiments, describes the architecture of the teachable agent including its dialog system and learning mechanism.

4.1 Task Environment

Our teachable agent was deployed as a textual conversational bot embedded into a task environment provided by Curiosity Notebook [40], a web-based learning platform that supports human-teachers learning while teaching conversational agents. In the task interface, participants read an article and converse with a conversational agent to teach it how to classify that article. There are two modes, teaching and testing, as described in Figure 1. In the teaching mode, while reading the article, participants could enter or highlight words to explain why an article should be classified in



(a) Teaching Interface



Fig. 1. Task Environment: Curiosity Notebook

a particular way (Figure 1a). The agent asked questions to the human-teacher and revealed what it did not understand about the topic, or what else it wanted to know. In answering the agent's questions, the human teachers were prompted to reflect on their own knowledge. The assumption was that through this process, human teachers may gain a better understanding about how to perform the classification task themselves. In the testing mode, participants could present new articles to the teachable agent, and ask them to classify articles in real-time based on what they have learned from the conversational interaction (Figure 1b). After the agent's prediction, correctly classified articles were coloured green by the system, whereas incorrectly classified articled were coloured red. During the entire interaction, participants were encouraged to test the agent often to assess their teaching performance and how well the agent is handling unseen examples. In both experiments, each human-teacher taught their own agent.

Heuristic	Description	Conversational Guidance
Externally relevant words	Words 'outside' the text that will	Can you tell me few more words
	most likely describe the <i>category</i>	that should describe the <i>category</i>
		but are not in the text?
Internally relevant words	Words from the text that are most	I wonder which words are most rel-
	relevant to the <i>category</i>	evant while categorizing this text to
		the <i>category</i> ?
Internally irrelevant words	Words from the text that are least	Which words are least relevant
	relevant to the <i>category</i>	while categorizing this text to the
		category?

Table 1. Three types of heuristic teaching guidance

4.2 Teachable Agent's Design

Our teachable agent's design is inspired from previous work on Teachable Conversational Agents [16, 17], where the authors demonstrated interaction mechanisms of a conversational agent that learns from direct human feedback. They found that performance of the teachable agent improved with the number of times it was tested during the session. This implies that people who are concerned about their agent's performance through repeated testing, are more successful in training smarter agents. Consequently, we implemented two interaction modes to allow the users to teach

Proc. ACM Hum.-Comput. Interact., Vol. 6, No. CSCW2, Article 331. Publication date: November 2022.

and test their agents. In the teaching mode, the agent was focused on learning new features through conversations related to a given topic; whereas in the testing mode, agent predicted the category of unseen articles and asked for more samples from the human teachers. For the learning algorithm, we used a modified version of the Naive Bayes algorithm that incorporates human teaching as additional input. Specifically, the algorithm predicted posterior probabilities of text belonging to a category by amplified priors from conversations. This technique has been shown to improve the classification performance with increasing number of dialogues exchanged between the human and the agent [18].

Agent's dialogue system was structured as a conversational tree with a rule-based approach to identify different intents during the conversational interactions [1]. Further, the agent could switch between teaching or testing interaction modes based on explicit user actions or implicit intent identification from the conversation. In addition, we also developed agent strategies loosely consistent with Speech Act theory [69], that directed the user to ask about content within the agent's dialog system repertoire. In certain cases in which no input was recognized, the agent would default to one of several fallback options like: asking users to paraphrase, repeat or simply ignore and move to next article.

Table 1 summarizes the different types of heuristic teaching guidance that were provided to the human teacher. These heuristics were identified based on Macgregor et al. [48], who proposed teaching heuristics for optimizing the classification algorithms. Features identified through these heuristics were meant to improve classifier's performance by proposing new features (externally relevant words), amplifying relevant ones (internally relevant words), or discounting the irrelevant ones (internally irrelevant words). These heuristics were randomly selected by the agent and used in equal proportion for each article. Table 2 shows a sample conversation between a participant and their teachable agent.

Agent:	Can you please help me figure out the category for this article?	
Agent:	SciTech? Oh wow, I thought so too. I remember seeing the word Google in other SciTech articles.	
Agent:	Can you tell me some other words from the text that are also relevant to this topic?	
Human:	Search Engine, web surfers	
Human:	dot-com	
Agent:	Interesting, I thought surfers may also belong to a Sports category.	
Human:	Yes, but web surfers is related to technology. Individually they may be different, but	
	together they make sense.	
Agent:	Mmm okay I see.	
Agent:	Do you think there may be some other terms in this article that may be irrelevant to	
-	SciTech but relevant to other categories?	
Human:	Investors might belong to business.	
Agent:	Okay.	
Agent:	Can you teach me some additional words outside from this article that are related to	
-	SciTech?	
Human:	Computers, Robots etc.	
Agent:	Awesome.	
Agent:	You can continue teaching more articles to me, or test my knowledge.	

Table 2. Sample conversation between participant and their teachable agent

4.3 Training Dataset

The same dataset was used in both experiments. For text classification, we used a subset of AG News Classification Dataset [78], which has 4 largest classes representing the topics World, Sports, Business and SciTech. Each class contained 30,000 training samples and 1,900 testing samples. The total number of training samples in the dataset is 120,000 and number of test samples is 7,600. We used the standard data pre-processing techniques including tokenization, Stop-words removal and lemmatization. Tokenization was done using *word_tokenize()* function from NLTK that splits the raw sentences into separate word tokens. This was followed by a text normalization step where we converted individual tokens into lowercase to maintain the consistency during training and prediction. Stopwords filtering was also done using NLTK to filter out the words that did not contain vital information for text classification. Finally, we used WordNetLemmatizer with part-of-speech tags to obtain the canonical form (lemmas) of the tokens. Conversion of tokens to their base form was done to reduce the language inflections from words expressing different grammatical categories, including tense, case, voice, aspect, person, number, gender, and mood.

5 STUDY 1: LEARNING BY TEACHING

In this study, we aim to investigate the learning-by-teaching paradigm within the context of crowdsourcing tasks. This was done through a between-subject experiment—in the teaching condition, crowdworkers taught a classification task to an AI agent, whereas in the self-classification condition, crowdworkers were provided more instructions for the task without teaching any agent. We are interested in knowing whether crowdworkers can improve their own performance by teaching the task to a virtual conversational agent, compared to doing the same task themselves with additional task instructions. We compare the pre- and post-interaction performance of crowdworkers in the two conditions to validate the effectiveness of learning-by-teaching technique. Additionally, we also capture participants' opinion on the usefulness of the system and their self-reported enjoyment during the task using the Activity Perception Questionnaire from the Intrinsic Motivation Inventory [65].



Fig. 2. Study procedure for experiment 1

5.1 Participants

We recruited 100 crowdworkers from Amazon Mechanical Turk (38 females, 62 males), 22 to 65 years old (M= 33.74, SD= 9.24). A variety of professions were reflected in the participant pool, including freelancers (42), managers (23), engineers (15), home-makers (8), and designers (3). The remaining 9 participants were self-employed. 94% of the participants were native English speakers, but all reported some prior experience with conversational agents on a 7-point scale (M=5.58, SD=1.56). 37 % of the participants reported prior experience in teaching a classification to someone else, the remaining 63% had no prior experience on teaching. Regarding the prior knowledge on the 4 given news categories, participants rated most for SciTech (M=5.26, SD=1.43), followed by Sports (M=5.09, SD=1.71), World (M=5.01, SD=1.45) and Business (M=4.08, SD=1.52).

Participants received \$0.5 USD for the pre-study questionnaire on demographics, \$2 for both baseline tasks, \$2 for completing the condition task (teaching the classification or self-classification with instructions), and \$0.5 USD for the post-study questionnaire. The two baseline tasks were approximately 5 minutes long, interaction phase was designed to last for 10 minutes and the preand post-study questionnaires took 2-5 minutes to complete. Specifically, all the participants were compensated with \$5.00 USD for an estimated completion time of 20 minutes for the entire task that follows a rate consistent with an hourly wage of \$15.00 USD. At the time of our studies, U.S. federal minimum wage was \$7.25/hour.

5.2 Procedure

Crowdworkers were first given a series of text-classification tasks to capture their baseline performance before teaching the agent. Then, they enter the second phase of the experiment and were told that the purpose of this phase was to classify articles while providing more granular details supporting their decision. During this phase, participants were asked to mark certain words from the text that helped them choose a specific category for the overall article. Participants were divided into two experimental conditions. Participants in self-classification condition were given a set of rubric-instructions to follow while annotating the text to classify. These rubric-instructions asked them to further specify which words in the text were most and least relevant to the category belonging to the article. In teaching condition, workers interacted with a conversational agent (named Kai) who elicits teaching guidance from the participants, asking them to teach the classification task. The task interface for the two conditions is shown in Figure 3. Finally, participants from both conditions were asked to perform the post-interaction baseline task by labelling few more news articles without explicit cues from the interface.

Both rubric instructions and teaching guidance were different forms of the same underlying heuristics as described in Table 1. The *amount* of information elicited in the self-classification and teaching condition were held relatively constant to make the two conditions comparable. Effectively, workers were exposed to the same amount of information, but they assumed different responsibilities during the training process. Crowdworkers in self-classification condition were supposed to do the task themselves, whereas crowdworkers in the teaching condition were asked to teach the task to a virtual conversational agent. Participants in this condition could also test their agent's classification accuracy by switching to the testing mode as described in Figure 1. Study procedure for the experiment is described in Figure 2

5.3 Analysis

In this section, we describe the metrics and methods to investigate **RQ1**: whether teaching a task to an agent is beneficial in comparison to a situation when crowdworkers do a task on their own. To compare the two experimental conditions, we analyzed various metrics, including the amount of information the crowdworkers taught, differences between their pre- and post-interaction task performance, and perceived usefulness and perceived enjoyment of the interaction. The amount of information gathered was measured through the number of words taught ($N_{WordsCaptured}$). For better granularity and completeness in results, $N_{WordsCaptured}$, was analyzed across three dimensions: (a) words that were relevant to the topic and present in the given article (internally relevant), (b) words that were irrelevant to the topic, but present in the given article (internally irrelevant), and (c) words that were relevant to the topic and not present in the given article (externally relevant). Performance of the crowdworkers was measured by comparing the average time spent ($T_{TimeSpent}$) and average classification accuracy ($P_{Accuracy}$) in the pre/post-interaction baselines. Finally, perceived usefulness and enjoyment of the interaction was measured from the responses of Activity Perception Questionnaire in Intrinsic Motivation Inventory[65]. Continuous



(a) Interface for Teaching Classification



Fig. 3. Study conditions in experiment 1 with interfaces for (a) teaching-classification, and (b) self-classification

variables such as participants' accuracy ($P_{Accuracy}$) and average time spent ($T_{TimeSpent}$) on the task were analyzed using linear models, and discrete variables like total number of words captured ($N_{WordsCaptured}$) were analyzed using Poisson regression models [2]. For completeness, we also investigated the effect of these factors with respect to participants' background. The fit of the model was checked using graphical residual analysis. To account for over-dispersion in the Poisson regression model, quasi-likelihood estimation was applied. Statistical power of the study was found to be 91.5% using post-hoc power analysis on $P_{Accuracy}$, $\mu_0 = 1.75$, $\mu_1 = 10.25$, $\sigma = 18.03$, $\alpha = 0.05$.

5.4 Results

5.4.1 Participant's Performance. We investigated participant's performance in pre- and post interaction tasks across both conditions. A total of 2,443 words were taught by all the crowdworkers throughout the study, with an average of 24.43 (SD = 19.34). On average, the crowdworkers taught 11.14 (SD = 10.81) internal relevant words, 4.84 (SD = 10.81) internal irrelevant words and 8.45 (SD = 8.78) external relevant words. $N_{WordsCaptured}$ from the interaction was found to be significantly affected by the experimental condition that participants were assigned to. In particular, those assigned to the teaching condition taught more words that were internally relevant ($\beta = 0.86$, t(94) =5.32, p < .001), or externally relevant to the topic ($\beta = 1.18$, t(97) = 6.82, p < .001). Additionally, participants with prior teaching experience and higher knowledge scores were found to teach fewer number of words ($\beta = -0.63$, t(97) = -2.75, p = .01 and $\beta = -0.13$, t(97) = -2.24, p = .02respectively). Other factors do not show significant impact on the amount of information gathered from the interaction.

Each participant spent an average of 75.48 minutes (SD = 146.35) on post interaction tasks. The large standard deviation is due to the wide range of time (8.77 - 1450.66 minutes) the participants took to complete the tasks. Surprisingly, these figures were significantly different than the estimated task completion time of 20 minutes so we investigated further to determine the cause. No significant effect of experimental conditions (teaching vs self-classification) was found on the average $T_{TimeSpent}$ by participants in the pre- and post- interaction tasks. To ensure the goodness of fit, we model the log of the time post interaction using the linear model, and found that the only significant factor is the time of pre-interaction, $\beta = 0.36$, t(98) = 4.08, p < .001. Participants who have longer pre-interaction time are more likely to spend more time on the post interaction tasks. Together, these results indicate towards task switching and multitasking behaviour of crowdworkers: a well explored phenomenon in crowdsourcing research [33, 75].



Fig. 4. Average words captured per article during the interaction phase in experiment 1



Fig. 5. Comparison of pre-interaction and post-interaction tasks across both conditions with respect to (a) average time taken, and (b) accuracy of the participants

In general, participants with higher accuracy in the pre-interaction tasks had significantly higher accuracy in the post interaction tasks, $\beta = 0.40$, t(96 = 5.87, p < .001. Interestingly, $P_{Accuracy}$ was also found to be significantly impacted by experimental condition as the participants who were assigned to the teaching condition, tend to have higher accuracy in the post interaction tasks as compared to the participants who did the task themselves, $\beta = 6.63$, t(96) = 2.74, p = .01. All other participants' background did not have significant impact on the accuracy of the post interaction tasks.

5.4.2 Survey Responses. After the experiment, participants were asked to fill a post-study questionnaire on Activity Perception from Intrinsic Motivation Inventory (IMI). The IMI activity perception questionnaire is a collection of 25 questions containing 7-point scale type questions. We selected a subset of the scale containing 17 questions that measured participants' interest/enjoyment in the activity as well as the perceived value/usefulness. Overall, we collected a total of 100 survey responses corresponding to 50 participants in each experiment condition.

The average scores of the enjoyment rating is 5.26 (SD = 1.31) with a range of 1.50 and 6.88, while the average usefulness rating is 5.33 (SD = 1.08), with a range of 1.67 and 7. Although many factors were included in the model selection process, only two factors were found to have significant impact on the perceived enjoyment and usefulness ratings, namely prior knowledge

score and the condition the participants were assigned to. Participants who have higher knowledge score are expected to have significantly higher enjoyment rating ($\beta = 0.60, t(97) = 6.50, p < .001$) and higher usefulness rating ($\beta = 0.42, t(97) = 5.75, p < .001$). Those who were assigned to the teaching condition also have significantly higher enjoyment rating ($\beta = 0.67, t(97) = 3.02, p = .003$) and higher usefulness rating ($\beta = 0.94, t(97) = 5.33, p < .001$).

5.5 Summary

Participants who taught the classification to an agent exchanged more information and performed better in the post-interaction tasks than those who self-classified the articles. This supports our hypothesis **H1a** which states that teaching a task to an agent is positively associated with participants' performance in the task, both in terms of task accuracy, as well as the amount of useful information exchanged during the interaction. These results also indicate that teaching, or preparing to teach an information to an agent is beneficial for the human-teachers implying a presence of Protégé Effect. Further, responses from the post-study questionnaire that focuses on subjective experience of participants during the interaction, reveals teaching to be more useful and valuable than the individually performing the task. Findings from this study draws broader implications on how different roles assumed by humans are perceived differently. It was observed that humans who identify themselves as teachers find a task more useful and valuable than those who identify themselves as annotators. This finding offers support for our hypothesis **H1b**.

6 STUDY 2: DYNAMICS OF TRUST

The previous experiment was focused on evaluating the effects of a teachable conversational agent on participant's performance and enjoyment. In this experiment, we investigate the perception of participants trust and their willingness to delegate tasks to an agent in the context of crowdsourcing. This was done using a between-subject experiment—in the treatment condition, participants taught news classifications to an AI agent, whereas in the control condition they did not teach and performed the task in the presence of an agent that was trained beforehand. Testing mode was enabled for both the conditions. After the primary task, all participants were given an optional bonus task where they were asked to classify a few more articles either by themselves or by delegating a portion of it to the agent they interacted with.

The goal of this experiment is to determine if crowdworkers would prefer incorporating teachable agents in their workflow if they diligently teach them certain human intelligence tasks, and what factors might influence their decision to delegate tasks in a positive or negative way.

6.1 Participants

We recruited 80 new crowdworkers from Amazon Mechanical Turk (33 females, 47 males). 21 to 69 years old (M= 36.5, SD= 10.6). Participants from the first experiment were excluded in this study. This new participant pool represented a variety of professions including managers (10), IT technicians (10), engineers (6), teachers (5), nurse (3) and designers (3). The remaining 3 were self employed. 89% of the participants were native English speakers, but all reported some prior experience with conversational agents on a 7-point scale (M=5.63, SD=1.31). 30 % of the participants reported prior experience in teaching a classification task to someone else, the other half had no prior experience on teaching (70%). Regarding the prior knowledge on the 4 given news categories, participants rated most for SciTech (M=5.225, SD=1.28), followed by World (M=5.1, SD=1.81), Sports (M=4.75, SD=1.51) and Business (M=4.62, SD=1.41).

The HITs were posted with the title "Teach How to Classify News Articles to a Chatbot" and "Read the text and classify news articles while talking to an agent" for treatment and control conditions respectively. Eligible participants received \$0.5 USD for the pre-study questionnaire on demographics, \$2 for the first task, and \$0.5 USD for the post-study questionnaire. Furthermore, for every correctly labelled sample in the bonus task, they received an amount of \$0.17 USD. The maximum possible payment for the bonus task was \$2 USD (12 articles * \$0.17 USD per correctly labelled article). The first task was approximately 10 minutes long, bonus task was designed to last for 0-5 minutes depending on whether the crowdworker decided to delegate the task or do it all by themselves. Pre- and post-study questionnaires took 2-5 minutes to complete. Similar to the previous study, total payout was \$5.00 USD for the second study, and with an estimated completion time of 20 minutes, the rate approximates to an hourly wage of \$15 USD/hr. Federal minimum wage in U.S. at the time of our studies was \$7.25/hour. To prevent crowdworkers from task switching and discourage multitasking, we applied special interventions such as preventive alerts and session expiry upon extended inactivity.

6.2 Procedure



Fig. 6. Study procedure for experiment 2

Crowdworkers were first given a series of 8 text classification tasks and asked to teach a virtual teachable agent (again, named Kai) that delivered the same conversational interventions as described in experiment 1. In the treatment condition, participants were told that their future compensation may depend on how successfully they teach Kai during the task. In control condition, they were informed that Kai was trained by other trustworthy workers who have similar traits like them. In both the conditions, participants were also expected to validate the classification performance by observing the agent's accuracy in test mode, similar to the first experiment. In the second part, they were presented with a bonus task to label 12 more news-snippets. The bonus task was structured such that workers could choose to either do the tasks themselves, or delegate a portion of the task to an agent. Crowdworkers were given a 12-point slider to specify which portion of task they want to do themselves and which portion they want to delegate to the agent they recently interacted with. For either of the options, they were also asked to provide a reason for their choice in an open-form text box.

After the bonus task, they were asked to fill a questionnaire, which investigated their level of trust on the agent and factors that influence their trust level.

6.2.1 Agent's Learning. Unlike previous experiment where the agent was learning autonomously from the conversational interaction, the agent in the treatment condition of this experiment was only *simulating* its learning based on the number of news articles covered. This was achieved by training the agents beforehand and adjusting the error-rate based on the proportion of tasks completed. For control condition, the agent demonstrated a consistent accuracy based on its pre-training. For treatment condition, simulated learning of the agent was also based on task progression percentage (number of articles covered/total number of articles). Note that the effective accuracy still depend

Read the text and teach news classification agent by suggesting words.	on to the	Bonus Task
Article 1/20 Topic: Web Serie Barne Enter Google Confounds European Investors (Reuters): Reuters - Web suffers from Beijing to Berkelp have made Google their search engine of choice, but some of Europe's institutional investors have gone sour on the dot-com on the eve of its initial public offering.	entred to random average entred to random average for the second averag	There are 12 more news snippets to label Boxus with be given to correctly liabeled news nippets irrespective of who labels them (you verse in the said to the gent). You can delegate the task to the gent you recently integrity of you can obtain the said to the gent you recently integrity of you can obtain the said to the gent you recently integrity of you can obtain the said to the gent you recently integrity of you can obtain the said to the gent you recently integrity of you can obtain the said to the gent you recently select the portion of delgate Agent will label 6/12 articles You will used 6/12 articles
Nexts staught from this ancient 0 Nexts staught from this ancient 0 Xick 🕑 to see the taught words and tune learning parameters	SEND	
	Test Next	Note: Non devision to delaparti-re-delapara file task ell have ne effect on how you have performed so tal: This will gray demand the boxe.
(a) Teaching mode		(b) Delegation mode

Fig. 7. Task interface for experiment 2 while (a) teaching the agent, and (b) delegating the task.

on the pre-trained accuracy of the classifier along with the proportion of tasks completed. For instance, if 4 out of 8 articles are covered in the task (4/8 = 0.50), teachable agent with a pre-trained accuracy of 90% would show an effective accuracy of 45% (0.50*0.90) with our simulated learning technique. The reason behind simulating agent's learning rather than the actual online learning was to ensure that agents within the treatment condition are consistent with each other and agents across the two conditions are equally accurate towards the end despite any differences in how participants performed the teaching. This is important because the experiment was designed to measure the trust of participants while delegating the tasks that involve monetary compensation. Since a part of this decision may come from their perceived accuracy of the agent, it was important to account for teachers who may not succeed in teaching, resulting in less-accurate agents. As it was difficult to quickly and accurately distinguish between ineffective and effective teachers, we decided to simulate agent's learning and keep its performance consistent across all participants. This helped in isolating the effects of teaching and variable accuracy on participants' willingness to delegate the tasks.

6.3 Analysis

We analyzed the proportion of tasks delegated to the agent in both experimental conditions in order to address **RQ2**, and whether the act of teaching influences the dynamics of trust between Human-AI collaboration. Particularly, we evaluated the impact of teaching on participants' decision to delegate the tasks to the agent. Qualitative data such as reasoning for delegation was studied using thematic analysis. Dependent variables such as portion of delegated tasks, general trust, task efficacy and general self-efficacy were analyzed using Linear models [2]. The independent variable included in the selection process was based on whether the participant interacted with a teachable or non-teachable agent. The Binomial regression models were used to model the proportion of task allocated to the conversational agent. Similar to previous experiment, the fit of linear models was checked using graphical residual analysis, and Hosmer-Lemeshow and Stukel's tests were applied to test the fit of the Binomial regression models. Post-hoc power analysis of general trust scores was used to obtain the statistical power of 92.7% for the study, $\mu_0 = 4.13$, $\mu_1 = 4.68$, $\sigma = 1.02$, $\alpha = 0.05$.

6.4 Results

6.4.1 Task Delegation. Each participant was given a total of 12 articles in the bonus task, and a choice to either label all the articles themselves, or delegate a portion of those to the agent they

recently taught. It was clearly mentioned that only the correctly labelled articles in the bonus task will be considered while calculating the bonus amount. In treatment condition, 63% (n = 25) of the participants trusted the agent and delegated a portion of the bonus task. The remaining 37% (n = 15) decided to trust their own skills to classify news articles in order to maximize the bonus amount. Among the 25 participants who delegated the tasks to the agent, 44% (n = 11) decided to assign all 12 articles to the agent, while 40% (n = 10) assigned 6 of the 12 articles in order to compare agent's work with their own. On the other hand, in control condition, 47.5% (n = 19) of the participants delegated their task to the agent whereas 52.5% (n = 21) decided to self-classify news articles in the bonus task. Among the 19 participants who delegated the tasks to the agent, around 21% (n = 4) decided to assign all 12 articles to the agent. Overall, we found that participants in the control condition who worked with a pre-trained agent delegated lesser tasks compared to the participants in the treatment condition who taught their own agent, $\beta = -0.89$, t(74) = -6.25, p < .001. The data also shows that participants' interest with news article classification and knowledge about world news can significantly affect their decision to delegate tasks. In fact, those who indicated more knowledge of World type news article delegated lesser task to the agent, $\beta = -0.87, t(74) = -5.00, p < .001$. However, the effect of their prior knowledge in World news is reduced if the participants had indicated more interest in news article classification, $\beta = 0.10$, t(74) = 3.54, p < .001.

Reasoning for Delegation. We performed thematic analysis on the rationale provided by the 6.4.2 participants in the bonus task while they decided whether to delegate the articles or not. For this, the responses were first highlighted to preserve interesting information. Next, those highlighted segments were clustered into groups and assigned codes to encapsulate potential themes. These themes were then cross-referenced with the corresponding proportion of tasks delegated by the participants. Three themes were identified: (T1) those who did not trust their agent, (T2) those who significantly trusted their agent, and (T3) those who trusted their agent as much as themselves. Participants who did not delegate any portion of the task to the agent, believed more in their own ability to perform the task and thought that their agent was not ready yet. This is evident from statements like: "I am more confident of my ability" (P19), "I'm not confident the agent can identify the articles as well as I can." (P37), and "The agent only seemed to agree with things I said but I'm not sure that the agent would correctly label the news snippets on their own." (P54). On the other hand, participants who delegated all the tasks to their agent, trusted their own teaching abilities more, and consequently, their agent's capability to complete the task. One participant said: "Because I feel like I taught it how to quickly find the high points to categorize" (P13), while participants mentioned: "I think that I taught the agent well and that the agent will deliver solid results" (P33). This trend was particularly missing in the control condition where no teaching was involved. Hence participants from the control condition who still delegated the tasks to the agent did so because of agent's accuracy or to save time. Interestingly, participants who delegated exactly half the portion of their task to the agent reasoned that they wanted to compare agent's performance with their own. For instance: "I taught the agent everything that I know related to the categories. They should be able to label the articles just as well as I can" (P31). Note that for this experiment, agent's learning was simulated so that they all learn at the same pace based on the number of articles covered during the teaching phase. Despite this, and the fact that all agents had same classification accuracy, participants chose different proportions of tasks to delegate and corresponding reasons for why. Interestingly, some participants from the control condition reported lack of trust on the agents due to limited involvement in teaching. This is evident from statements like "i haven't seen it make decisions without my own input. I just don't have any reason to trust it yet." (P47), and "I'm not totally sure if the agent knows which category to choose for the articles, as I haven't seen enough evidence thus far of the agent's ability. I'd rather just do it myself [..]. These

results indicate that the dynamics of trust on teachable agents may depend on user's subjective perception of agent's ability to perform the task.

6.4.3 Survey Responses. Participants completed a post-study questionnaire after the experiment that was designed for the assessment of trust and competence and self-efficacy. The questionnaire was adapted from a combination of three surveys on general self-efficacy [38], task-specific competence [24], and empirically determined scale of trust between people and automation [39].



Fig. 8. Results comparing (a) Proportion of tasks completed by the participants and agent, and (b) Ratings for General Trust, Task Self-Efficacy, and General Self-Efficacy in Control and Treatment conditions.

Overall, the participants indicated their general trust score between 1.67 and 6.67 with an average of 4.41(SD = 1.07). We found that participants in the treatment condition who taught their own agents reported higher level of trust scores in the post-study questionnaire ($\beta = 0.56, t(78) = 2.40, p = 0.02$). Other factors such as participants' background do not appear to be significant.

6.5 Summary

While majority of the crowdworkers preferred delegating a portion of their task to the agent they taught, a significant portion of participants also decided to do the entire task themselves. It was observed that participants in the treatment condition who taught their own agents, were more likely to delegate tasks to the teachable agent as compared to the participants who interacted with a pre-trained agent. This offers support for our hypothesis **H2a** which states that participants who teach their own agent will delegate more over the participants who will not teach. Through the post-study questionnaire, it was observed that participants who delegate more tasks in the treatment condition reported more trust on automated systems and were more confident in themselves and their own ability to do the same task as illustrated in Figure 8. This confirms our hypothesis **H2b** that the dynamics of trust will be higher for the participant group that will teach their own agent.

7 DESIGN CONSIDERATIONS FOR INTERACTIVE HUMAN-AI SYSTEMS: A DISCUSSION

In this paper, we introduce the notion of teachable agents in the context of crowdsourcing, and demonstrate how they fare as an interface to an interactive machine learning system that workers can teach and delegate tasks to. The first experiment provides some interesting evidence that favours the use of teaching a task to an agent over doing the task alone using instructions. The second experiment examines the extent to which these teachable agents are trusted to handle the tasks, if they were taught by the workers. In this section, we discuss the key implications of our

findings and recommend some guidelines for the design of teachable agents.

Conversation as an information gathering mechanism. Results from experiment 1 suggests that conversations can be used as a viable information gathering mechanism in interactive systems. This is specifically relevant in situations where the agent is the interface to an underlying machine learning algorithm. An important benefit of using conversational interface is that clarifications can be obtained naturally during the interaction process. For instance, when the agent in our experiments did not understand something, it would simply ask the human-teachers to re-type, or even click something relevant in the chat interface. Strategically guided conversations can not only make the interaction more natural for humans, but also impact the factors that directly relate to the performance of the underlying machine learning algorithm such as limited training samples, imbalanced classes, and biased data.

Explicit nature of teaching and learning. Teaching and learning process in interactive systems should be explicit such that humans who teach are aware of the fact that their teaching will impact agent's performance. Likewise, machine learners should also convey that they are learning from the interaction. Both these points were observed after some participants mentioned that they were unsure if their agent was learning. Note that this explicitness is not usually seen in traditional machine learning systems like recommendation engines or ad platforms that silently gather the data from humans to inform relevant contents or advertisements. For agent-driven interactive machine learning systems, human teachers should be able to control when they want to casually interact with their agents and when they want the agent to make progress in learning; likewise, the agent should demonstrate their learning in a timely fashion in order to continually engage with the human teacher. This can be operationalized by presenting distinguishable audio or visual cues to illustrate appropriate interaction states of the agent when it is listening, thinking (retrieving information from knowledge source), or learning (acquiring new knowledge through interactions). This additional granularity is a subtle but important distinction from common interaction states exhibited by most of the existing conversational systems. Effectively conveying these system states can also contribute towards trust-building factors in conversational interactions.

Treating humans as teachers rather than mere annotators. The success of agents depend critically on their ability to efficiently learn from non-expert humans. Although active learning allows machine learners to improve their performance by actively querying humans and obtain labels, humans are not always willing to be simple oracles unless monetary benefits are involved. This is evident from the results of Experiment 1 where participants significantly preferred teaching condition over self-classification, both in terms of enjoyment and usefulness of the task. Therefore it is important to embrace humans as teachers, and provide sufficient teaching guidance to optimize machine learning.

While our experimental results highlight the possibility of integrating teachable conversational agents for crowdsourcing, it is necessary to discuss several tradeoffs in adopting them for general application domains. First, it is worth acknowledging that teaching an agent with the expectation to share the workload is a form of personalized interaction. Producing agents that can approximate a particular skillset by directly learning from human teachers is a beneficial technique for short term interactions. However, it is important to realize that learning is a lifelong process and humans also improve over time. Therefore, if teachable-agents are only used for sharing human workload but the teaching-learning process stops after few short-term interactions, then soon the agent's expertise on the task will no longer approximate the right amount of expertise originally acquired from the

human-teacher. Thus, without continuous learning the agent may not meet the right amount of expectations in extended interactions, which may hurt the dynamic of trust between human-AI collaboration. Moreover, letting individuals teach an agent also implies addressing challenges arising from individual biases and problems related to data cogency. Thus, the use of such agents should be highly restricted to individual use-cases under specific application domains such as task delegation for teachable skills.

8 CONCLUSION

In this work, we described the idea of leveraging conversational interactions as an interface between humans and an interactive machine learning system. Through two experiments, we investigate (1) the effectiveness of learning by teaching paradigm within the context of crowdsourcing studies, and examine whether teaching a task is more beneficial for crowdworkers than doing the same task for themselves; (2) whether teaching a task to an agent can make people delegate similar tasks to the agent where monetary compensation is involved. In conclusion, this paper aims to take one step in the direction of studying the effectiveness and perception of teachable conversational agents. Understanding the breakdowns across these facets will be important for building teachable agents that can reliably learn, be adopted, and benefit human teachers through the conversational interaction.

Overall, our results show that enabling crowdworkers to teach an agent can elicit more information from them during a task. While we do see a performance improvement between pre- and post-interaction tasks, we do not explore the actual learning outcomes of these human-teachers after the interaction. An interesting area to explore for follow up work can specifically focus on long and short-term memory changes across longitudinal studies. More experiments can be conducted to understand the dynamics of trust on teachable agents in the presence of embodied agents. In future work, it will be interesting to explore different modalities of the interaction with teachable agents beyond the textual conversational interaction. Possible extensions should also consider the use of voice-based agents or embodied agents like physical robots to validate these results in different settings.

REFERENCES

- [1] Ernest Adams. 2014. Fundamentals of game design. Pearson Education.
- [2] Alan Agresti. 2003. Categorical data analysis. Vol. 482. John Wiley & Sons.
- [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] Frank J Balbach and Thomas Zeugmann. 2009. Recent developments in algorithmic teaching. In International Conference on Language and Automata Theory and Applications. Springer, 1–18.
- [5] John A Bargh and Yaacov Schul. 1980. On the cognitive benefits of teaching. *Journal of Educational Psychology* 72, 5 (1980), 593.
- [6] Gautam Biswas, Krittaya Leelawong, Daniel Schwartz, Nancy Vye, and The Teachable Agents Group at Vanderbilt. 2005. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence* 19, 3-4 (2005), 363–392.
- [7] Sean Brophy, Gautam Biswas, Thomas Katzlberger, John Bransford, and Daniel Schwartz. 1999. Teachable agents: Combining insights from learning theory and computer science. In *Artificial intelligence in education*, Vol. 50. Citeseer, 21–28.
- [8] 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.
- [9] Maya Cakmak and Andrea L Thomaz. 2010. Optimality of human teachers for robot learners. In 2010 IEEE 9th International Conference on Development and Learning. IEEE, 64–69.
- [10] Maya Cakmak and Andrea L Thomaz. 2014. Eliciting good teaching from humans for machine learners. Artificial Intelligence 217 (2014), 198–215.

Proc. ACM Hum.-Comput. Interact., Vol. 6, No. CSCW2, Article 331. Publication date: November 2022.

- [11] Davide Calvaresi, Yazan Mualla, Amro Najjar, Stéphane Galland, and Michael Schumacher. 2019. Explainable Multi-Agent Systems through Blockchain Technology. In Proceedings of the 1st International Workshop on eXplanable TRansparent Autonomous Agents and Multi-Agent Systems (EXTRAAMAS 2019).
- [12] Justine Cassell. 2000. More than just another pretty face: Embodied conversational interface agents. Commun. ACM 43, 4 (2000), 70–78.
- [13] Kimmy Wa Chan, Stella Yiyan Li, Jian Ni, and John JianJun Zhu. 2021. What feedback matters? The role of experience in motivating crowdsourcing innovation. *Production and Operations Management* 30, 1 (2021), 103–126.
- [14] Catherine C Chase, Doris B Chin, Marily A Oppezzo, and Daniel L Schwartz. 2009. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology* 18, 4 (2009), 334–352.
- [15] Veena Chattaraman, Wi-Suk Kwon, Juan E Gilbert, and Yishuang Li. 2014. Virtual shopping agents. Journal of Research in Interactive Marketing (2014).
- [16] Nalin Chhibber. 2019. Towards the Learning, Perception, and Effectiveness of Teachable Conversational Agents. Master's thesis. University of Waterloo.
- [17] Nalin Chhibber and Edith Law. 2019. Using conversational agents to support learning by teaching. arXiv preprint arXiv:1909.13443 (2019).
- [18] Nalin Chhibber and Edith Law. 2021. Towards Teachable Conversational Agents. arXiv preprint arXiv:2102.10387 (2021).
- [19] Michelene TH Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian LaVancher. 1994. Eliciting self-explanations improves understanding. *Cognitive science* 18, 3 (1994), 439–477.
- [20] Leigh Clark, Abdulmalik Ofemile, Svenja Adolphs, and Tom Rodden. 2016. A multimodal approach to assessing user experiences with agent helpers. ACM Transactions on Interactive Intelligent Systems (TiiS) 6, 4 (2016), 29.
- [21] Peter A Cohen, James A Kulik, and Chen-Lin C Kulik. 1982. Educational outcomes of tutoring: A meta-analysis of findings. American educational research journal 19, 2 (1982), 237–248.
- [22] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services. ACM, 43.
- [23] C Brad Crisp and Sirkka L Jarvenpaa. 2013. Swift trust in global virtual teams: Trusting beliefs and normative actions. Journal of Personnel Psychology 12, 1 (2013), 45.
- [24] Edward L Deci and Richard M Ryan. 2012. Self-determination theory. (2012).
- [25] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a learning science for complex crowdsourcing tasks. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 2623–2634.
- [26] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In Proceedings of the ACM 2012 conference on computer supported cooperative work. 1013–1022.
- [27] Rochelle E Evans and Philip Kortum. 2010. The impact of voice characteristics on user response in an interactive voice response system. *Interacting with Computers* 22, 6 (2010), 606–614.
- [28] 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.
- [29] Pedro Fialho, Luísa Coheur, Sérgio Curto, Pedro Cláudio, Ângela Costa, Alberto Abad, Hugo Meinedo, and Isabel Trancoso. 2013. Meet EDGAR, a tutoring agent at MONSERRATE. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 61–66.
- [30] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. 2015. Training workers for improving performance in crowdsourcing microtasks. In Design for Teaching and Learning in a Networked World. Springer, 100–114.
- [31] Mahtab Ghazizadeh, John D Lee, and Linda Ng Boyle. 2012. Extending the Technology Acceptance Model to assess automation. Cognition, Technology & Work 14, 1 (2012), 39–49.
- [32] Sally A Goldman and Michael J Kearns. 1995. On the complexity of teaching. J. Comput. System Sci. 50, 1 (1995), 20-31.
- [33] Sandy JJ Gould, Anna L Cox, and Duncan P Brumby. 2016. Diminished control in crowdsourcing: An investigation of crowdworker multitasking behavior. ACM Transactions on Computer-Human Interaction (TOCHI) 23, 3 (2016), 1–29.
- [34] Andrew Guillory and Jeff A Bilmes. 2011. Simultaneous learning and covering with adversarial noise. (2011).
- [35] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. 2015. Incentivizing high quality crowdwork. In Proceedings of the 24th International Conference on World Wide Web. 419–429.
- [36] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.
- [37] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. When children teach a robot to write: An autonomous teachable humanoid which uses simulated handwriting. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction. ACM, 83–90.
- [38] Matthias Jerusalem and Ralf Schwarzer. 1979. The general self-efficacy scale (GSE).[Updated 2006 Oct 7].

- [39] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71.
- [40] Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D Pantasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. 2020. Curiosity Notebook: A Platform for Learning by Teaching Conversational Agents. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts. 1–9.
- [41] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A Terry, and Krzysztof Z Gajos. 2016. Curiosity killed the cat, but makes crowdwork better. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. 4098–4110.
- [42] Ludovic Le Bigot, Patrice Terrier, Virginie Amiel, Gérard Poulain, Eric Jamet, and Jean-François Rouet. 2007. Effect of modality on collaboration with a dialogue system. *International Journal of Human-Computer Studies* 65, 12 (2007), 983–991.
- [43] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. 2119–2128.
- [44] Jae-Eun Lim, Joonhwan Lee, and Dongwhan Kim. 2021. The effects of feedback and goal on the quality of crowdsourcing tasks. International Journal of Human–Computer Interaction (2021), 1–13.
- [45] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2018. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal* of Librarianship and Information Science (2018), 0961000618759414.
- [46] Giuseppe Lugano. 2017. Virtual assistants and self-driving cars. In 2017 15th International Conference on ITS Telecommunications (ITST). IEEE, 1–5.
- [47] 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. ACM, 5286–5297.
- [48] James N MacGregor. 1988. The effects of order on learning classifications by example: heuristics for finding the optimal order. Artificial Intelligence 34, 3 (1988), 361–370.
- [49] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In 11th australasian conference on information systems, Vol. 53. Citeseer, 6–8.
- [50] Dominic W Massaro, Michael M Cohen, Sharon Daniel, and Ronald A Cole. 1999. Developing and evaluating conversational agents. In *Human performance and ergonomics*. Elsevier, 173–194.
- [51] H David Mathias. 1997. A model of interactive teaching. journal of computer and system sciences 54, 3 (1997), 487-501.
- [52] Masaki Matsubara, Ria Mae Borromeo, Sihem Amer-Yahia, and Atsuyuki Morishima. 2021. Task Assignment Strategies for Crowd Worker Ability Improvement. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–20.
- [53] Panagiotis Mavridis, Owen Huang, Sihang Qiu, Ujwal Gadiraju, and Alessandro Bozzon. 2019. Chatterbox: Conversational interfaces for microtask crowdsourcing. In Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization. 243–251.
- [54] Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of management journal 38, 1 (1995), 24–59.
- [55] Robert R Morris, Mira Dontcheva, and Elizabeth M Gerber. 2012. Priming for better performance in microtask crowdsourcing environments. *IEEE Internet Computing* 16, 5 (2012), 13–19.
- [56] Ευάγγελος Μουρελάτος. 2019. Quality of work in online labor markets: an empirical study in paid crowdsourcing environments. Ph.D. Dissertation. http://hdl.handle.net/10889/13366
- [57] Krista R Muis, Cynthia Psaradellis, Marianne Chevrier, Ivana Di Leo, and Susanne P Lajoie. 2016. Learning by preparing to teach: Fostering self-regulatory processes and achievement during complex mathematics problem solving. *Journal* of Educational Psychology 108, 4 (2016), 474.
- [58] John F Nestojko, Dung C Bui, Nate Kornell, and Elizabeth Ligon Bjork. 2014. Expecting to teach enhances learning and organization of knowledge in free recall of text passages. *Memory & Cognition* 42, 7 (2014), 1038–1048.
- [59] Sharon Oviatt, Colin Swindells, and Alex Arthur. 2008. Implicit user-adaptive system engagement in speech and pen interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 969–978.
- [60] Aannemarie Sullivan Palinscar and Ann L Brown. 1984. Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and instruction* 1, 2 (1984), 117–175.
- [61] Neil Patel, Sheetal Agarwal, Nitendra Rajput, Amit Nanavati, Paresh Dave, and Tapan S Parikh. 2009. A comparative study of speech and dialed input voice interfaces in rural India. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 51–54.
- [62] Paul A Pavlou. 2003. Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International journal of electronic commerce* 7, 3 (2003), 101–134.

Proc. ACM Hum.-Comput. Interact., Vol. 6, No. CSCW2, Article 331. Publication date: November 2022.

- [63] Reid Porter, James Theiler, and Don Hush. 2013. Interactive machine learning in data exploitation. Computing in Science & Engineering 15, 5 (2013), 12–20.
- [64] Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. Academy of management review 23, 3 (1998), 393–404.
- [65] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.
- [66] Stefan Schaffer, Robert Schleicher, and Sebastian Möller. 2015. Modeling input modality choice in mobile graphical and speech interfaces. *International Journal of Human-Computer Studies* 75 (2015), 21–34.
- [67] F David Schoorman, Roger C Mayer, and James H Davis. 2007. An integrative model of organizational trust: Past, present, and future.
- [68] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. Hey Alexa, What's Up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 857–868.
- [69] John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. 1980. Speech act theory and pragmatics. Vol. 10. Springer.
- [70] Burr Settles. 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.
- [71] 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).
- [72] David Traum, Priti Aggarwal, Ron Artstein, Susan Foutz, Jillian Gerten, Athanasios Katsamanis, Anton Leuski, Dan Noren, and William Swartout. 2012. Ada and Grace: Direct interaction with museum visitors. In *International conference* on intelligent virtual agents. Springer, 245–251.
- [73] Weiquan Wang, Lingyun Qiu, Dongmin Kim, and Izak Benbasat. 2016. Effects of rational and social appeals of online recommendation agents on cognition-and affect-based trust. *Decision Support Systems* 86 (2016), 48–60.
- [74] Noreen M Webb. 1983. Predicting learning from student interaction: Defining the interaction variables. Educational psychologist 18, 1 (1983), 33–41.
- [75] Alex C Williams, Gloria Mark, Kristy Milland, Edward Lank, and Edith Law. 2019. The perpetual work life of crowdworkers: How tooling practices increase fragmentation in crowdwork. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.
- [76] Teng Ye, Sangseok You, and Lionel Robert Jr. 2017. When does more money work? Examining the role of perceived fairness in pay on the performance quality of crowdworkers. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
- [77] Ming Yin, Yiling Chen, and Yu-An Sun. 2013. The effects of performance-contingent financial incentives in online labor markets. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 27.
- [78] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In Advances in neural information processing systems. 649–657.