Crowdsourcing the Perception of Machine Teaching

Jonggi Hong[1], **Kyungjun Lee**[1], **June Xu**[2], **Hernisa Kacorri**[3,1]
[1]Computer Science, [2]Electrical and Computer Engineering, and [3]Information Studies
University of Maryland, College Park, MD, USA
jhong12@umd.edu, kjlee@cs.umd.edu, junexu@terpmail.umd.edu, hernisa@umd.edu

ABSTRACT
Teachable interfaces can empower end-users to attune machine learning systems to their idiosyncratic characteristics and environment by explicitly providing pertinent training examples. While facilitating control, their effectiveness can be hindered by the lack of expertise or misconceptions. We investigate how users may conceptualize, experience, and reflect on their engagement in machine teaching by deploying a mobile teachable testbed in Amazon Mechanical Turk. Using a performance-based payment scheme, Mechanical Turkers ($N = 100$) are called to train, test, and re-train a *robust* recognition model in real-time with a few snapshots taken in their environment. We find that participants incorporate diversity in their examples drawing from parallels to how humans recognize objects independent of size, viewpoint, location, and illumination. Many of their misconceptions relate to consistency and model capabilities for reasoning. With limited variation and edge cases in testing, the majority of them do not change strategies on a second training attempt.

Author Keywords
teachable interfaces; interactive machine learning; object recognition; crowdsourcing; personalization

INTRODUCTION
As machine learning and artificial intelligence become more present in everyday applications, so do efforts to better capture, understand, and imagine this coexistence. Experts from diverse disciplines are working together and critically examining the impact of algorithmic decisions, their assumptions, and their biases [5, 7, 9, 14, 36]. Error-prone, computationally complex, and failing in ways unexpected by humans, such algorithms called early on for transparency, interpretability, accountability, and control [54, 56, 50, 18, 61]. More recently, these efforts have redoubled (surveyed in [1, 62]), fueled by funding and legal initiatives such as the DARPA Explainable Artificial Intelligence [24] and the European Union's General Data Protection Regulation [15], while feeding into future initiatives such as the Algorithmic Accountability Act [16].

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

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376428

Figure 1: Given an object category, MTurkers are called to choose three object instances and train a *robust* personal object recognizer using their mobile camera. Here we include examples from some of the participants' selected objects.

Machine teaching [51, 64] lies at the core of these efforts as it enables end-users and domain experts with no machine learning expertise to innovate and build AI-infused[1] systems. Beyond helping to democratize machine learning, it offers an opportunity for a deeper understanding of how people perceive and interact with such systems to inform the design of future interfaces and algorithms [3] – a perspective this paper shares.

Within this paradigm, teachable interfaces [48, 37] explore applications where users can explicitly train a model with their generated data and labels. While facilitating user control, the effectiveness of these applications can be hindered by the lack of expertise or misconceptions about machine learning. Though personalization is often the ultimate goal (*e.g.,* [34]), the interactive nature of these interfaces can help users in return to uncover basic machine learning concepts (*e.g.,* [27]).

In this paper, we examine how people conceptualize, experience and reflect on their engagement with machine teaching in the context of a supervised image classification task, a task where humans are extremely good compared to machines, especially when they possess prior knowledge of the image classes. Using a teachable interface for object recognition, we recruit participants ($N = 100$) through Amazon Mechanical Turk[2] to choose three objects in their environment and train a model to distinguish between them in real-time using the camera on their mobile phones, as shown in Figure 1.

[1]A term in Amershi *et al.*, 2019 [4] for "systems that have features harnessing AI capabilities that are directly exposed to the end user."
[2]https://www.mturk.com/

Why crowdsourcing. Beyond being utilized as a platform for obtaining labeled data quickly at low cost, crowdsourcing is also employed for behavioral and perception studies (*e.g.,* [26, 12, 52]) including those for understanding people's interactions with machine learning systems, surveyed in [59]. Allowing us to quickly recruit a large participant pool for this study, it also enables data collection outside a laboratory to obtain high variability and real-world illumination, backgrounds, and camera manipulations in the user's environment.

We build a web-based testbed for a mobile teachable object recognizer and ask participants to train and evaluate it on three objects of choice within an object category (Figure 1). Categories represent daily objects that span different characteristics such as size, shape, color, material, and function. Through a performance-based payment scheme [28], participants are called to iterate and reflect over their efforts with the goal of making their recognition models more *robust*. Serving as an oracle, they are tasked with delivering a teaching set to the recognition model to help it learn the classification task.

We conduct a contextualized quantitative analysis on the participants photos, their written responses, as well as their model performance. We find that diversity, important in machine learning, is deemed important by a majority of participants and incorporated in teaching strategies, drawing from parallels to how humans generalize across object size, viewpoint, location, and illumination [45]. Many misconceptions relate to consistency; few think that it is good to be consistent and teach with almost identical examples; others failed to be consistent on incorporating diversity across classes. While participants have good intuition on importance of discriminatory features in teaching but on evaluating their models, we observe susceptibility to missing edge cases. Last, we see that the majority of participants do not change strategies on a second attempt even though possess a reasonable intuition on what would be important. We see how our findings and insights can help better understand non-experts' interactions with machine teaching and guide the design of future teachable interfaces that can anticipate users misconceptions and assumptions.

RELATED WORK

We discuss prior work on machine teaching with a focus on teachable interfaces that most relate to our study. Prior work on behavioral studies using crowdsourcing is briefly mentioned to highlight elements that we draw from.

Machine Teaching

Machine teaching involves a teacher who knows the decision boundaries and designs an optimal training set for one or more students [64]. In this paper, the teacher is a human and the student is a classification model who is being trained to classify images of objects, as shown in Figure 2, though the inverse – machines teaching humans to classify images – is also an active area of research [31]. There is a rich literature on sequential machine teaching with humans as the teacher, *e.g.,* programming by demonstration for teaching robots to manipulate objects [58, 19]. However, in this review we focus on prior work that utilizes batch teaching, where examples are given as a set and their order does not matter.

Table 1: Related studies' characteristics juxtaposed with ours.

		[22]	[30]	[10]	[34]	[27]	[65]	This study
Setting	People	1,7,21	10	12	8	30	5	100
	Controlled	•	•	•	•	•	•	
	Real-world	•						•
People	Crowd							•
	Children					•	•	
	Disability			•	•			
Input	Sensing	•				•	•	
	Audio			•				
	Image				•			•
	Video	•	•					
Output	Recognition	•			•	•		•
	Detection			•			•	
	Control		•					
Analysis	Accuracy		•	•	•			•
	Behavior	•		•	•	•	•	•
	Feedback	•		•	•	•	•	•

Batch teaching is a very common paradigm for many real-world AI-infused systems, *e.g.,* using face recognition, fraud detection and speech recognition. This is typically done by experts in the field and end-users are hardly exposed to the underlying mechanisms that could help explain their limitations. Teachable interfaces[3] that fall under this machine teaching paradigm, have the potential to help in this direction as they can enable non-experts to uncover basic machine learning concepts (*e.g.,* [27]). Moreover, with advances in transfer learning [46, 55], they can spur innovation as end-users can re-purpose models trained on vast amounts of data for new but related tasks, *e.g.,* personalize assistive technologies [32].

We look into prior work employing teachable interfaces, a term perhaps not originally used by the authors. Here, we focus on a subset of interactive machine learning literature, where users are called to generate all the training and testing examples for a personalized model. Table 1 presents representative examples of prior studies from 2011-2019 on gesture recognition for musicians [22], sign language [30] and educational applications [27], personalized sound detectors for people who are deaf/Deaf or hard-of-hearing [10], personal object recognizers for blind people [34], and physical activity classifiers for young athletes [65]. In contrast to this work, prior studies tend to have smaller participant pools and are typically conducted in a controlled setting, where the researchers are present. Partially this could be due to the user characteristics of interest; people with disabilities [10, 34], children [27], and students [65]. Another reason could be challenges in remote data collection as it would require a working prototype [10, 34] or specialized devices from the users [27, 65]. Our teachable object recognition testbed, utilizing built in camera in a mobile phone, and existing crowdsourcing platforms allow us to reach a larger participant pool that can be further scaled.

As shown in Table 1, the input modality for the teaching set was more often based on sensing [22, 27, 65] and videos [22, 30] with one example for sound [10] and photos [34]. For

[3] A term coined by Patel and Roy (1998) [48], where "the user is a willing participant in the adaptation process and actively provides feedback to the machine to guide its learning."

the last two, participants could not assess the quality of their teaching examples – participants who were deaf/Deaf or hard-of-hearing could not hear the sounds they recorded [10] and blind participants could not see the photos they took [34]. In this paper, we choose images as the input modality for the teaching set. This allows us to tap into a large user group of non-experts that can simply use their mobile phones to take the photos in a real-world setting. More so, by choosing an object classification task, an accessible task to many where they can serve as the oracle, we are given the opportunity to explore how humans teach a high-dimensional decision boundary to machines by feeding them only with few instances. More importantly, this modality allows us to visually inspect the teaching set for common patterns in users' behavior.

Similar to most of the prior work in Table 1, our analysis is based on observed behaviors and participant feedback. Leveraging prior work in neuroscience, we examine how non-experts' teaching strategies draw parallels in machine robustness to human robustness, where object recognition involves generalization across size, location, viewpoint and illumination [45]. While prior work did not include such a fine-grained analysis of the participants' input, it provided insights and anecdotal evidence that guided the design of our study such as the need for iterations [22, 65, 27], which may vary not only across participants but also due to the underlying algorithm and task [21]. For comparison purposes and time sake, we opted to keep the number of iterations constant at two. Similar to our study, the number of classes were limited (2-5) with an exception of 15 [34], where there were no iterations.

Crowdsourcing and Online Behavioral Studies

Despite the potential risks in data validity [47, 42], advantages such as subject anonymity, prescreening, diversity, efficiency, and low cost have made crowdsourcing platforms attractive for user studies both in social [41, 49] and cognitive [53] science with a focus on behavioral and perception studies (*e.g.,* [26, 12, 52]). A building block for the machine learning community, crowdsourcing has been utilized to generate data and annotations, validate existing systems, incorporate feedback from humans, and observe how people interact with machine learning models, surveyed in [59]. We build on this prior work adopting a performance-based payment scheme [28] to incentivize participants while ensuring a rate of $15/hour [25]. Perhaps the closest work to our study is that of Yang *et al*. [63], where online interviews with non-experts (N=98) were used to elicit how people with no machine learning expertise perceive machine learning processes. While the survey did not include hands-on interactions with a teachable interfaces, the findings stressed the need for future work on helping people build better learning algorithms, further motivating our work.

TESTBED: TEACHABLE OBJECT RECOGNIZER

To explore how non-experts conceptualize, experience and reflect on their engagement with machine teaching, we build a web-based teachable object recognizer for mobile phones. Participants can train, test, and re-train it to distinguish between three objects of their choice. In this case, a test corresponds to a 'direct' evaluation [22], where participants take photos of their objects in real-time and observe the model's behavior. To help us better contextualize our observations, participants also provide background information and feedback[4].

Human vs. machine	T=machine, S=machine	T=machine, S=human	**T=human, S=machine**	T=human, S=human
One vs. many	**One student**		Many student	
Batch vs. sequential	**Batch learning**		Sequential learning	
Teaching signal	Synthetic / constructive teaching	Hybrid teaching	**Pool-based teaching**	
Model-based vs. model-free	Model-based teaching	Graybox teaching	**Model-free teaching**	
Student awareness	The student anticipates teaching		**The student does not anticipate teaching**	
Angelic vs. adversarial	**Angelic teaching**		Adversarial teaching	
Theoretical vs. empirical	Theoretical teaching	Hybrid	**Empirical teaching**	

Figure 2: Characterization of our testbed in the machine teaching problem space [64], where T stands for teacher and S for student. A human T employs a pool-based, model-free, angelic, empirical teaching. The testbed has a single recognition model S learning in batch mode, unaware that is being taught, while considering T as a friend (no adversarial examples).

Our machine teaching problem. As shown in Figure 2, we adopt Zhu *et al*. [64] machine teaching problem space to characterize the teachable interface in our testbed as a system where human is the teacher and machine is the student. The teacher provides, in batch mode, a finite pool of examples consisting of labeled photos of objects as the teaching signal. The teacher takes a model free approach, treating the student as a blackbox, though we anticipate that humans may already have some assumptions on how the black box works or should work. The student, employing a convolutional neural network, does not anticipate teaching, *i.e.,* assuming training examples are independent and identically distributed and that there are no errors. More so, the teacher is considered a friend, *i.e.,* no adversarial training. Last, we assume that the teacher uses heuristic teaching methods to improve the performance of the student, the object recognition model in our case. We aim to better understand these heuristic methods, factors they may relate to, as well as assumptions that people may have.

Model. For each user, our testbed creates a new convolutional neural network using the Google Inception V3 [57] pre-trained on ImageNet [17]. Everytime the user provides a teaching set, the last layer of the pre-trained model gets replaced with a new softmax layer and re-trained with the user's images with 500 steps and a gradient descent learning rate of 10^{-2}. Models are trained on our 8 GPU server in real-time asynchronously; the app continues to run and ask users for open-ended feedback while the training continues in the back. The web interface communicates with the server using the Flask API [6].

Interface. As shown in Figure 3, initially the testbed asks for background information, technology experience, and familiarity with machine learning. Then, it provides five object category options: bottle, cereal, drink, snack, and spice, with three sample icons for each category indicative of the preferred shape. Categories are inspired by prior work on personal object recognizers [34] and are engineered to elicit objects that

[4]Questions and prompts can be found in the supplementary material.

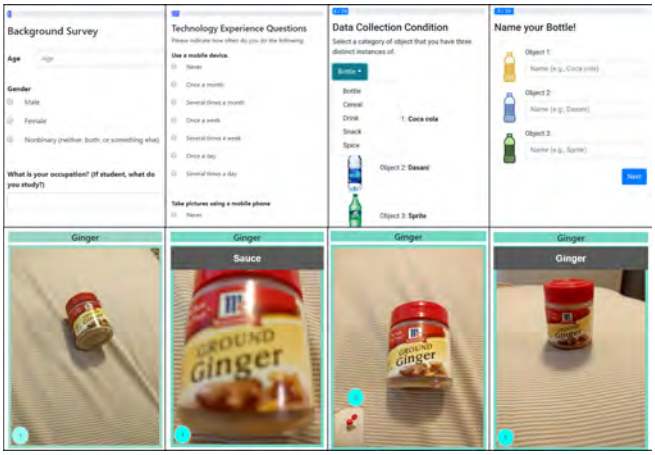

Figure 3: Testbed screenshots: questionnaires, category selection, object labeling, and camera view in training and testing.

are present in daily life but differ in size, shape, color, material, and function. Participants can choose to train only on one of the categories. To avoid object shape or size from being a factor in any observed inconsistencies between the classes, they are asked to use objects (a total of three) that fall within the same category; three, the smallest number for multiclass classification and previously used in teachable interfaces for non-experts [37], minimizes challenges in finding different object instances within a category in a real-world environment as well as the task completion time (already 40 mins long). After labeling their objects, participants are guided through five interactions with the machine learning model (the student)[5]:

Preliminary test (TS0): Participants are asked to take photos of their objects to see if the existing non-personalized model can recognize them. The instruction reads: *"Take a photo of an object (name at the top) by tapping on the camera screen. The existing model will try to predict it."* Given an object label displayed at the top, one takes a photo of the corresponding object and sees the recognition result (a label displayed for 3 seconds). This repeats 15 times (5 times per object in a random order). As expected, during this interaction recognition results will not match participant's labels as the generic model is based on Google's Inception V3 and is not yet personalized. There is a dual motivation behind this interaction. First, it helps familiarize with the interface, which simulates the native camera app. Second, it helps collect evaluation examples unbiased from one's teaching experience that is to follow.

Train 1 (TR1): Participants are asked to train the object recognizer with the following instructions: *"Train our object recognizer to identify robustly your objects anywhere, anytime, for anyone. We will randomly choose one of your objects and ask you to take 30 photos of it. You will be paid $2 extra if your examples pass our robustness test."* Here, we hint that model robustness means to be able to recognize an object anywhere, anytime, for anyone. Motivated by Ho *et al.* [28] performance-based payment scheme, we also create the impression of a 'secret' test distinguishing examples best for robustness, though on our end this is merely a naive quality examination (*e.g.,* photos of objects in a screen rather than in the real-world). As shown in Figure 3, given an object label displayed at the top, participants take 30 sequential photos. This repeats 3 times (1 time per object in a random order). Thus, the first teaching set comprises 90 photos (30 per object).

Test 1 (TS1): Similar to TS0, participants are asked to *"Test the trained object recognizer again to see how robust it is."* Here, recognition labels match participants' labels except in cases of misclassification, where an object is misrecognized as one of the other two. Again, no confidence scores are shown.

Train 2 (TR2): Participants are given an opportunity to re-train their model from scratch with the following instructions: *"You told us what you would do differently, now show us! On the next screen, take 30 more pictures of the requested object. You will be paid $3 extra if this training does better than the previous one in our robustness test."*

Test 2 (TS2): As in TS1, users can test the re-trained model. The instruction given to the participant was *"The object recognizer is trained again. Test the trained object recognizer."*

Eliciting Feedback. The testbed includes the following open-ended questions: *"What did you think was important to consider when training the object recognizer?"* after TR1; *"If you were to retrain the system to make it more robust, what would you do differently?"* after TS1; *"How did you position the object in the image?"*, *"How did you decide the distance of the camera from the object?"*, and *"How did you decide which side of the object is visible in the image?"* at the end.

CROWDSOURCING STUDY

We deploy our testbed in Amazon Mechanical Turk (IRB #1255427-1) and investigate how non-experts crowdworkers teach a machine a high-dimensional decision boundary such as a fine-grained image classification with a few examples only.

Participants

We recruited 143 participants over 10 days. However, data from 43 were excluded from the analysis – 7 helped in piloting, 1 used the same object for all classes, 3 took photos of objects in display screens, 2 took photos with no objects. The other 30 had technical problems by attempting the task simultaneously

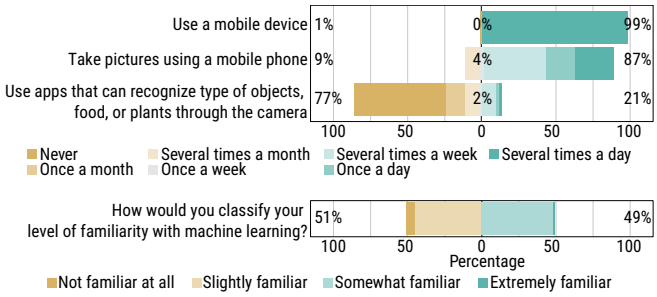

Figure 4: Participants' technology experience and familiarity with machine learning mostly ranging from slightly (have heard of it but don't know what it does) to somewhat familiar (I have a broad understanding of what it is and what it does).

[5]All instructions can be found in the supplementary material.

with our system failing to distribute them across the 8 GPUs, losing data from 12 and interrupting the task for other 18; all were compensated and the bug was fixed. The 100 participants who were included in the dataset ranged from 20 to 60 in age (μ=32.6, σ=8.3); 49 were male, 50 female, and 1 non-binary with 90 reporting being right handed. No one reported a visual or motor impairment. As shown in Figure 4, the majority of participants are frequent users of mobile devices taking photos with them weekly, though many of them don't use any applications for recognizing objects, food, or plants. When asked about familiarity with machine learning, 6 reported never having heard of it, 45 had heard of it but didn't know what it does, 48 had a broad understanding of what it is and what it does, and only one reported having extensive knowledge.

Procedure

With the goal of attracting non-experts in machine learning, we opted for a HIT description that minimizes technical terms: *"You will be asked to take photos of everyday products such as soda cans, cereal boxes, and spices to teach your phone to automatically recognize them. To see how well the object recognition works you will test it by giving a single photo at the time."* A warning message was displayed if participants attempted to start the study from a device other than a mobile phone. Only one participation was allowed.

Through piloting, we estimated that a study session could be successfully completed within 30-40 minutes. Adopting a $15/hour compensation rate [25] all participants received a total of $10 once all the data collection was completed. To incentivize participants, we used a performance-based payment scheme [28], where this amount was split as $5 flat participation, $2 bonus for passing *"our robustness test"* in the first attempt to train, and $3 bonus for achieving a better performance in *"our robustness test"* the second time around. Given that objects differ across participants it was not possible to have an ideal *'secret robustness test'*; bonus was decided merely on a quality check. While the testbed's connection is persistent and one could do other tasks in between, we observe that participants took on average 35.57 minutes (14.21-79.86, σ=12.85) to complete the study, very close to our estimates.

ANALYSIS OF BEHAVIOR

We explore how participants conceptualize, experience and reflect on their engagement with machine teaching by looking at the photos they took for the teaching and testing sets as well as changes in their behavior when repeating the process. Observations are contextualized with participants' responses.

Visual Attributes in Photos. We collected a total of 22,500 photos from 100 participants across all training and testing interactions. To uncover patterns in participants' teaching strategies, photos were coded using thematic coding [11]. Two researchers independently created initial codebooks of visual attributes in photos across four dimensions, *i.e.,* size, location, viewpoint, and illumination; prior work on visual object understanding [45] indicates that our ability to recognize objects generalizes across these dimensions. We want to see how participants draw parallels from their understanding of robustness in these dimensions to enable machines to do the same.

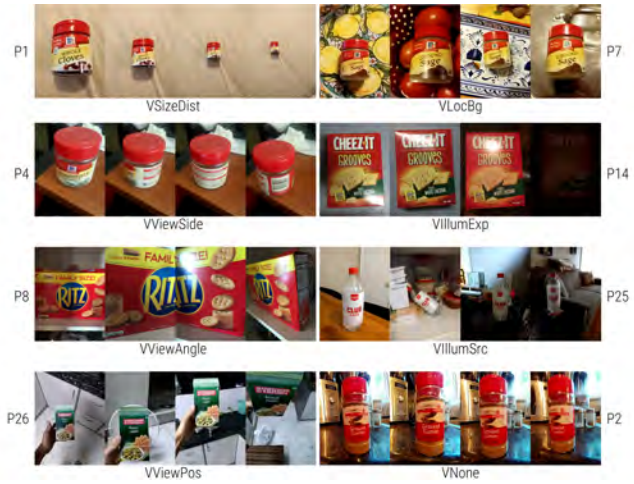

Figure 5: Examples of variation attributes in teaching sets.

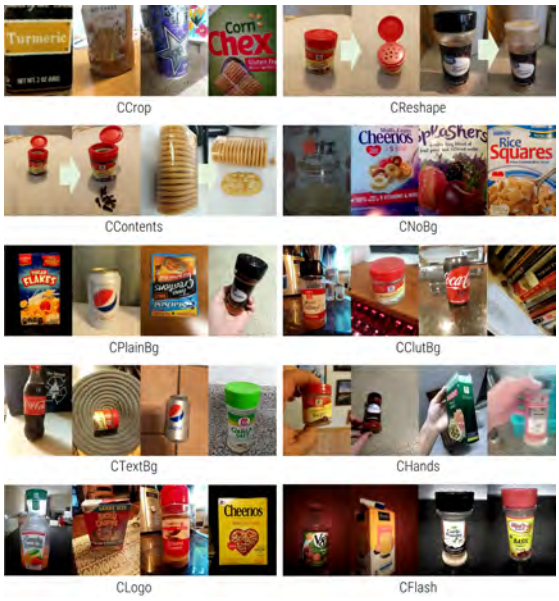

Figure 6: Sample photos considered by the count attributes.

Researchers discussed disagreements to produce a final codebook, shown in Tables 2–4 with examples in Figures 5 and 6. There are two types of attributes: binary and count. Binary attributes capture presence of variation or inconsistency within a teaching or testing set of photos. If a participant varied photos for an object along an attribute such as distance (*VSizeDist*) or background (*VLocBg*), the corresponding attribute is 1; otherwise 0. Similarly, variation inconsistency across the three objects is captured through binary attributes, named *ISize, ILoc, IView, IIllum*. Count attributes indicate the number of photos within a set with a certain characteristic such as presence of participant's hand (*CHands*) and use of flashlight (*CFlash*) or a quality issue such as dark (*QDim*) and blurry (*QBlurry*) photos. There was substantial agreement (Cohen's kappa=0.80).

Subjective Feedback. Participants' responses to the open-ended questions were also analyzed with a thematic coding ap-

Table 2: Variation attributes, true if a variation is present for at least one object.

Variation	Definition
VSizeDist	True if camera **distance**, ratio of object height to frame, differs for two or more photos using [0, 0.25), [0.25, 0.5), [0.5, 1.0), and [1.0, ∞) bins.
VLocBg	True if the **background** differs for two or more photos, *i.e.,* different locations or perspectives of a space.
VViewSide	True if the **side** of objects differs for two or more photos.
VViewAngle	True if the **angle** between the camera and the object with the same side of object differs for two or more photos.
VViewPos	True if the **position** of the object in the camera frame, center, top left, top right, bottom left, or bottom right, differs for two or more photos.
VIllumExp	True if the **exposure** to light differs for two or more photos taken at the same location.
VIllumSrc	True if the **source** of light differs for two or more photos because they were taken at different locations.

Table 3: Inconsistency attributes, true if there is an inconsistency in variation across the three objects.

Count	Definition
ISize	True if the camera **distance** varies in the photos for one or two objects but not all three.
ILoc	True if the **background** varies in the photos for one or two objects but not all three.
IView	True if size, angle, or position capturing **viewpoint** varies in the training photos for one or two objects but not all three.
IIllum	True if light exposure or source capturing **illumination** varies in the training photos for one or two objects but not all three.

Table 4: Count attributes, number of photos with a given characteristic including those looking at quality issues.

Count	Definition
CCrop	Number of photos where the object is **cropped**, *i.e.,* object is close to the camera, out of frame, or obscured by another object.
CReshape	Number of photos where the object was **reshaped** (*e.g.,* opening a lid of a package).
CContents	Number of photos where the **contents** inside a package was taken out of the container or the inside of the package is visible.
CNoBg	Number of photos where were the background is **not visible** because the photos are filled with the object completely.
CPlainBg	Number of photos where the background includes two or fewer colors with **no or very simple textures**.
CClutBg	Number of photos where the background is **cluttered** with objects other than the object of interest.
CTextBg	Number of photos where the background includes a wall, floor, or furniture with **texture**.
CHands	Number of photos where the participant's **hand(s)** is visible in the photo.
CLogo	Number of photos where the side with the **logo** (or label) of the object was visible in the photos.
CFlash	Number of photos where the brightness varies in different parts of the photo like using **flashlight**.
QSmall	Number of photos where the object is too **small** (height of the object < 25% of the height of the photo).
QDim	Number of photos where the brightness of the photo is too **dark** to recognize texture or edge of the object.
QBlurry	Number of photos where the object of interest is **blurry**.
QIrrelevant	Number of photos where the photo includes only **irrelevant** objects without the object of interest.

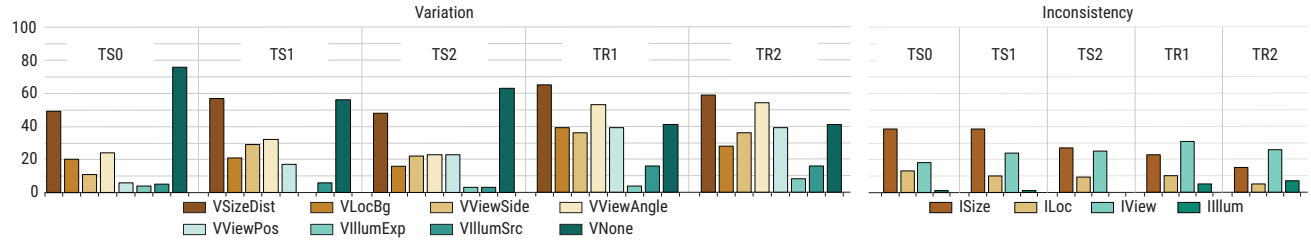

Figure 7: Number of participants per variation and inconsistency attribute across all five interactions with the model: preliminary test (TS0), train 1 (TR1), test 1(TS1), train 2 (TR2) and test 2 (TS2). The graphs on the left indicate how participants incorporate diversity in their photos in terms of object size, viewpoint, location, and illumination when they train and debug their models.

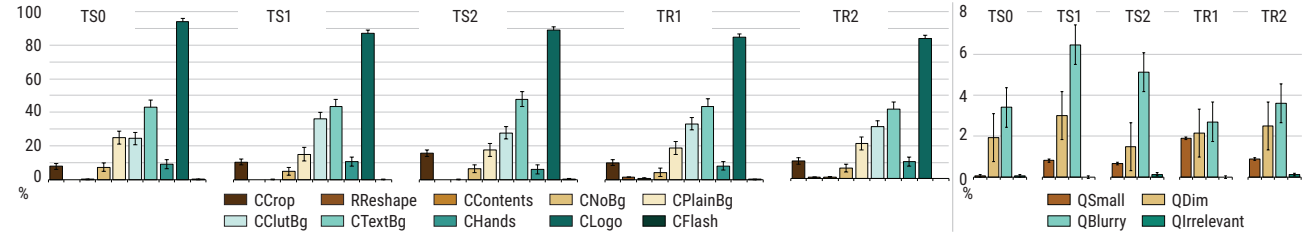

Figure 8: Percentage of photos per participant given a count attribute, with standard error as error bars. Participants took photos mostly with the logo on it and many of them against a textured or cluttered background. Often the objects were cropped in the camera frame and sometimes participants' hands were included in the photos. Surprisingly, few participants opened the object and trained the model on their content as well. The most common quality issues were blurry and dim photos though not that prevalent.

proach [11]. The same two researchers who coded the photos, created initial codebooks and merged them through discussions resolving disagreements. Responses were coded independently with a substantial agreement (Cohen's kappa=0.73).

Results

What Are Non-experts' Teaching and Debugging Strategies? We explore how variation[6], inconsistency, and other attributes manifest on participants' image sets when they are first called to train the object recognizer on objects of their choice.

Incorporating diversity in teaching. Diversity plays an important role in machine learning [23]. When incorporated in the teaching set, it ensures that examples can provide more discriminatory information to help the model learn. By looking at participants' photos (results in Figure 7) and by reading their responses, we find that the majority of the participants share this intuition, but not all. In detail, 23 participants (age 21–60, μ=37.57, σ=9.87) did not include any kind of variation in their TR1 teaching set – 3 of them reported never having heard of machine learning, 12 had heard of it but did not know what it does, and 8 had a broad understanding of what it is and what it does. Immediately after training, when asked about what they considered important, 5 participants referred to the need for consistency, which in this context contradicts the way machines and people learn. For instance, P6 said *"I figured I needed to be consistent when I took the picture so they looked similar."* and P30 *"Keeping the pictures the same."* Others, who did not consider this type of consistency, mentioned that it is important to have a good quality photo where the object is well framed (4) with visible labels (8) and images that are clear (6) with ample light (2). Without even having tested their model, P2 said: *"Getting different angles and perspectives so the trainer could recognize it more easily"* – a contradiction to their initial teaching set that had no variation. We observed that in TR2, P2 reflected on this observation and varied both the object size and viewpoint. Only two other participants from this group did so as well, P5 and P18. They said having the "name and color in" is important in TR1 but also varied the camera distance (P5) and angle (P18) in TR2.

However, the majority of participants ($N = 77$) diversified examples in their first attempt. They varied either size ($N = 65$) or viewpoint ($N = 63$), with some considering location ($N = 39$) and illumination ($N = 19$). Light exposure was least diverse ($N = 4$). Looking at responses on important considerations for training, many participants ($N = 52$) mentioned these strategies[7] and reflected on the need for diversity with concrete terms such as *"different", "various", "all", "many", "multiple", "every", "variety", and "difference"* combined with *"angles", "views", "sides", "facets", "background", "lighting", "distance", and "positioning"*. These terms correspond to the four dimensions of our coding scheme informed from prior work on visual object understanding [45], highlighting that humans' strategies for machine teaching parallel their own abilities. However, only 11 participants (age: $\mu = 34, \sigma = 8.71$) incorporated diversity in their teaching set across all four dimensions – 3 reported having heard of machine learning with no further understanding, and 8 had a broad understanding of what it is and what it does.

Being fair and consistent between classes. Model consistency across classes is a desirable trait in machine learning with many social implications for fairness, whose definition is still being debated in the community (*e.g.,* [44, 43]). There is anecdotal evidence on non-experts learning to balance class proportions in the training set over multiple iterations [22, 65]. By keeping the number of training examples constant, we look into their behavior across other potential disparate treatments. Given that many participants considered diversity important for good performance, we explore how fair[8] (*i.e.,* consistent) they are in incorporating diversity across their three objects, with results shown in Figure 7. Beyond the 23 participants who did not introduce any variation for any object, we find that there were 30 other participants that were consistent. This is promising, especially since this included participants from all levels of familiarity with machine learning: not familiar at all ($N = 1$), slightly familiar ($N = 11$), somewhat familiar ($N = 17$), and the only participant in our study that reported being extremely familiar ($N = 1$). While none of these participants explicitly mentioned consistency as important, we find that more than half of them ($N = 16$) continued doing so in their second attempt at training, in TR2. For the remaining 47 participants, their inconsistencies were found in variations related to all four dimensions: object size ($N = 21$), viewpoint ($N = 31$), location ($N = 10$), and illumination ($N = 5$).

Deciding what to show in the teaching set. We analyze the fine-grained count attributes in teaching and training sets (Figure 8) to uncover common teaching patterns across participants. Khan *et al*. [35] observed that one of the most prominent teaching strategies for a binary classification task among non-experts, called the *extreme* strategy, is consistent with the "curriculum learning" principle [8, 40], where participants start with the most extreme examples and continue with those closer to the decision boundary[9]. While our batch teaching task does not allow for a similar sequential analysis, we find that almost all participants ($N = 98$) included the logo (or label) of objects in their teaching sets; on average 84.9% ($SD = 25.0$) of any participants' images included logos. This indicates that participants understand that logos and labels tend to include the most discriminatory features, which serve as the most extreme examples. Then, through variation they add less discriminative viewpoints that are closer to the decision boundary. Indeed, 18 participants explicitly mentioned logos or labels being important in training. For instance, P36 said *"... trying to have a constant label view"* and P46 *"... a clear shot of the front of the package with minimal background*

[6] A preliminary analysis of this appears in a work-in-progress [29].

[7] All questions, instructions, and prompts prior to training were carefully edited not to prime participants towards our coding attributes.

[8] In this work classes are object instances that fall within the same category and consequently share similarities such as shape, size, and material in the context of the decision making task of incorporating variation. Thus, we consider "individual fairness" [20], where "similar individuals should be treated similarly", and explore whether object instances within a category are being treated the same by a participant when introducing variation in the training photos.

[9] In the Khan *et al.* [35] study participants did not generate the examples but they ordered them as most representative of the two classes and chose to teach one by one using all of them or a subset.

interference." When looking deeper at these responses though, we find that many of the participants assumed that the machine would read the text. For example, P28 said *"It [the model] recognizing the different cereals by name"* and P44 *"Getting a clear shot where the writing and the size are clear."*

In terms of the background, we find that the majority were textured ($N = 66$) or cluttered ($N = 62$), while many used plain ($N = 48$) and a few none at all ($N = 11$) – the latter two are preferred since very few varied the object location. We observe that 26 participants included their hands in the photos. The presence of hands has been leveraged to better distinguish objects by modeling the contextual relationship between grasp types and object attributes [13] or to estimate the object of interest in a clutter environment [38, 39]. However, given this study's fine-grained task, the grasp is expected to be similar across object of the same category. Thus, the presence of the hand doesn't really help, especially if it is not applied consistently across classes. More surprisingly, we observe that 8 participants reshaped their objects, *e.g.,* opened the lid, and 4 decided to train on the content of the object as well, *e.g.,* cinnamon powder. When asked what is important for training, one of these participants, P76, said: "Getting lots of different angles and different ways the spice could be portrayed." In general, there were not many photos with quality issues. Participants took clear photos in most cases and many of them mentioned the importance of image quality in their responses, but some ($N = 36$) mistakenly took a few blurry photos. Also, objects sometimes appeared too small ($N = 17$) and occasionally the light was dim ($N = 9$).

Debugging and including edge cases in testing. When asked to evaluate their model in TS1, many participants ($N = 30$) did not diversify their images at all – 2 of them reported never having heard of machine learning, 17 had heard of it but didn't know what it does, and 11 had a broad understanding of what it is and what it does. This means that they did not check whether the recognizer is robust. We also find that compared to training, fewer participants diversify their testing set across object size ($N = 57$), viewpoint ($N = 49$), location ($N = 21$) and illumination ($N = 6$). This could be explained by many factors such as: smaller number of photos in testing (15) compared to training (90); difficulty in conceptualizing robustness; assumptions about machine's generalizing capabilities; not anticipating future uses of the model under different circumstances; or simply minimizing efforts for this HIT. Logos were still included by the majority of the participants ($N = 98$) and the same number of participants ($N = 11$) took photos that did not include any background, keeping their testing data consistent with their training examples. Similar to what Zimmermann *et al*. [65] observed, participants "enacted [testing] practices wherein their models appeared to have high reliability but questionable validity." We also find that participants took fewer photos with plain background ($W = 756$, $Z = 2.17$, $p = .030$, $r = 0.15$), and objects that were too small ($W = 126.5$, $Z = 2.61$, $p = .011$, $r = 0.18$) using a Wilcoxon signed rank test. None of the interesting object reshaping, or content images present in training, carried over to testing; a similar behavior to Kacorri *et al*. [34], with "exaggerated" variation in training unobserved in testing.

Do Teaching Strategies Evolve Through Iteration?

Prior work indicates that the interactive nature of teachable interfaces can help users uncover machine learning concepts [27]. We ask participants whether they would do something differently were they to retrain the model for a second time and offer a bonus if they could make it even more robust.

Updating teaching strategies to improve performance. "Is this information a signal or noise" was one of the most common debug strategies by experts [63]. We investigate whether participants employ a similar approach by comparing TR2 to TR1 in terms of the variation, inconsistency, and other image characteristics, which serve as information signals for the model. Using a McNemar test for binary and Wilcoxon signed rank test for count attributes, we find the only significant difference is variation of location as observed by changes in the photo background (VLocBg). More participants diversified the background in their teaching set on the first attempt than the second ($\chi^2(1, N = 100) = 4.35, p = .037, \phi = 0.21$, the odds ratio is 11.86). As in Zimmermann *et al*. [65], we suspect that participants were trying to maximize performance by increasing consistency between their training and testing data, even though in our prompts we had defined robustness as ability to recognize the objects anywhere, anytime, for anyone. No other significant differences were observed, though this could be partially explained by limitations in the binary nature of our variation and inconsistency attributes failing to capture changes in magnitude. We shed light into other possible explanations by looking at participant's responses.

When asked about what they would do differently if they were to retrain, some ($N = 22$) said *"nothing"*, *"wouldn't do it differently"*, and *"would not change anything"*. Few said they had nothing to change because they were satisfied with the performance in TS1 ($N = 6$). For instance, P23 said *"Nothing it seems very robust after the learning phase."* This was not a surprise given that in TS1 participants did not opt for a thorough evaluation, as discussed above. *"Having no idea what to change"* was also mentioned by some ($N = 19$) reflected by terms such as *"not sure"*, *"unsure"*, *"I can't think of anything"*, *"have no idea"*, or *"don't know"*. Indeed, we find that the models of these 22 participants perform well on their own test data with an average F_1 score of 0.981 ($SD = 0.048$)[10] and significantly better than the rest of the participants ($U = 1472, Z = 5.22, p < .001, r = 0.52$); a trend that carries over to the second attempt.

Few participants wanted to change elements of the teaching process such as improving the testbed ($N = 3$), taking photos faster ($N = 1$), adding more classes ($N = 2$), or adding more samples ($N = 6$). Yang *et al*. [63] characterized the latter as "most non-experts' only strategy to improve a model's performance." Others focused on improving the quality of their teaching set such as better focus ($N = 5$), more light ($N = 2$), show labels ($N = 2$), better framing with a certain distance ($N = 1$), and centering ($N = 1$). Few participants ($N = 2$) explicitly mentioned the importance of the background, with P83 saying *"I would try to change the color of the background to ensure that it knows what the actual object is. I think it was*

[10] Only recognition labels are available in testing and no scores.

confused by the curry because of the black stove background which may look like the black cap of the cumin." Surprisingly, one participant (P85) pointed to discriminatory limitations of their objects uncovering challenges in fine-grained classification by stating *"Change objects to not look so similar."*

Last, some participants ($N = 22$) explicitly indicate that adding more variation in their training set is something they would do. For instance, P14: *"I would take a wider variety of angles"* and P21: *"Take picture from many different locations lighting and positions."* Only one, P36 mentioned doing so in testing, *"Test different sizes"*. When examining what they actually did in their second attempt at training, we find differing approaches: some indeed started incorporating new variations ($N = 13$), some perhaps changed the magnitude as variations were present in both first and second attempt ($N = 5$), and others ($N = 4$) did not make those changes. While variation for these 22 participants was mostly limited to the 4 dimensions (size, viewpoint, location, and illumination), few other participants ($N = 5$) indicated that they would also include different forms of the same object, *e.g.,* different containers, perhaps difficult within this study.

ANALYSIS OF PERFORMANCE

We report the performance of the models that the participants train by looking at the predicted labels during the first and second round of testing using the F_1 score measure (F-score).

Relating observed behavior to performance. Participants achieved on average a 0.75 ($SD = 0.38$) F-score in their first attempt to train the model. Using a multiple linear regression, we explore how attributes capturing their behavior in teaching and testing may relate to the relative performance of their models. While this performance is far from an ideal controlled robustness[11], it can provide some context for the observations above such as participants' behavior in the second attempt. We use a square root transform of the F-score[12] as the dependent variable. As independent variables, we use variation, inconsistency, and count attributes in TR1 and TS1 and their interaction. For model selection, we use stepwise variable selection based on Akaike information criterion (AIC) [2] with results shown in Table 5. We find that only 28% of the variability in recognition performance is accounted by this model, as indicated by the adjusted R-squared metric. While this is modest, it is not surprising, as there are many factors that can contribute to the performance of an image classification algorithm. For instance, performance can vary based on object similarities, a common challenge in fine-grained classification; a similarity that is not directly captured by our attributes.

In training, we find that variation in light exposure (VIllumExp) relates positively with the F-score, though very few participants included this type of diversity in their teaching set. We also see that the number of images where the object is taken against a plain background (CPlainBg) has a negative relationship with model performance. Though counter-intuitive, we suspect that lack of diversity in the background might have contributed to a model that does not generalize well, *e.g.,* when tested. This seems to be supported by the negative relationship of the number of cluttered background images during testing.

Attempt	Variable	Estimate	Std. Error	t value
TR1	(Intercept)	0.939	0.048	19.79***
	VIllumExp	0.167	0.063	2.64**
	VIllumSrc	-0.076	0.049	-1.55
	CCrop	0.000	0.002	0.12
	CPlainBg	-0.002	0.001	-2.50*
	CTextBg	-0.001	0.001	-1.55
TS1	VSizeDist	-0.068	0.037	-1.81.
	VViewSide	0.108	0.038	2.83**
	VViewPos	-0.089	0.045	-1.97.
	CCrop	0.048	0.012	4.04***
	CClutBg	-0.007	0.003	-2.14*
	QBlurry	-0.016	0.009	-1.74.
TR1*TS1	CCrop	-0.001	0.000	-3.16**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.157 on 87 degrees of freedom
Multiple R-squared: 0.3681, Adjusted R-squared: 0.2809
F-statistic: 4.223 on 12 and 87 DF, p-value: 3.195e-05

Table 5: Modeling recognition performance based on attributes capturing variation, inconsistency, and other characteristics.

In testing, we find that variation in object size (VViewSide) relates positively with the F-score. We also see that the number of images where objects appear to be cropped (CCrop) has a positive relationship with model performance. A plausible explanation could be that these attributes capture participants' behavior of zooming in on the object's most discriminative features, thus helping the model to distinguish objects. However, when considered as an interaction between training and testing (TR1*TS1-CCrop), this attribute appears to be negatively related to the model performance perhaps pointing to the sensitivity for consistency between the two – if you crop objects in one case, then it helps to do so in the other as well.

Improving performance the second time around. As shown in the previous analysis, we observe few changes in participants' teaching strategies in the second training as captured by our attributes – though some participants said they would do things differently. We find that this is also reflected when comparing the performance of their second model to the first. On average, participants achieved a 0.746 ($SD = 0.38$) F-score the first time and a 0.749 ($SD = 0.28$) the second with no significant change ($W = 80.5, Z = -0.16, p = .871$). However, participants who indicated they would do nothing to improve their model after the first attempt ($N = 22$), seem to achieve significantly higher performance than the rest ($U = 1472, Z = 5.22, p < .001, r = 0.52$) and this is a consistent trend across both attempts ($U = 1459.5, Z = 5.12, p < .001, r = 0.51$). Looking at these relative low F-scores for such a simple 3-way classification task, it is surprisingly that the second group of participants did not further improve their performance even though they expressed reasonable strategies. Perhaps the incentives were not strong enough and they had a higher threshold for errors, or there was not enough time and iterations to try things out. It could simply be that their object instances were too similar. Indeed, the majority (N=38) of the participants in this group had chosen spices.

[11] Such a neutral test is unrealistic in our study since participants choose different objects in different environments.

[12] Transformation is used to meet the normality assumption.

DISCUSSION

We see how our results, some being new insights, others strengthening prior empirical and anecdotal evidence, can help better understand non-experts' interactions with machine teaching and guide the design of future teachable interfaces. We highlight some of them with the following suggestions:

Account for teaching strategies: Our observations suggest that non-experts mainly tend to teach with clear representative examples and sometimes incorporate examples that are closer to the decision boundary through variation, which draws from parallels to how humans generalize for similar recognition tasks. In the case of object recognition, these were object size, viewpoint, location, and illumination [45]; though all four were considered only by a few. Our analysis also suggest that beyond class imbalance [22, 65], there can be other disparate treatments such as inconsistency in the way variation is incorporated across classes.

Anticipate misconceptions: A prevalent misconception relates to consistency. While it is true that consistency between training and testing data will result in better performance, assuming they both represent real-life examples, some thought that being consistent entails teaching with multiple identical examples with no variation whatsoever. Other misconceptions relate to the capabilities of the machine for reasoning. For example, participants would train with visually disparate examples from both the container and its content separately. Others would assume that the models were able to infer the text.

Help users craft evaluation examples: Our observations indicate that testing examples tend to be less diverse or not at all. Thus, it is no surprise to see many people wanting to change nothing, being satisfied with the performance, or not knowing what to do. Even those who did change their behavior when training for a second time, it was to not vary the background rather than making their model more generalizable. Help may look different based on the goal of the teachable interface. If it is personalization (*e.g.,* [33]), then it could mean guiding the user to generate examples that are more representative of future use cases [22]. However, if it is an application intended to uncover machine learning concepts (*e.g.,* [27]) perhaps promoting more model-breaking examples [60] would be more appropriate; though in the context of a teachable interface this could lead to users training the model with less authentic data to simply improve its performance [65].

This work has several limitations listed below:

Task: We explore machine teaching in a narrow context, that of a supervised 3-way image classification task. This allows us to dive deep in our analysis using a fine-grained scheme when coding participants examples informed from prior work on visual object understanding. However, it also limits the generalizability of our findings. We attempt to overcome this by connecting our results with that of prior work when possible. Three, the smallest number for multiclass classification, was selected to minimize challenges in finding different object instances within a category in a real-world environment as well as the task completion time (already 40 minutes long).

Study: While teachable object recognizers are real-world applications [38], they are typically intended for blind users. Thus, the sighted participants may lack motivation in this study. We attempt to compensate for this lack of incentives with a performance-based payment scheme [28] creating the impression that we have a 'secret' test to distinguish models that are more 'robust'; though on our end this is merely a naive quality examination. By doing so, combined with the fact that the testbed shows only the predicted labels but no confidence scores in testing, we might have limited participants' criteria for model evaluation [22] to just correctess.

Analysis: Through crowdsourcing we were able to quickly recruit a large participant pool and collect data outside a lab in the users' environment. However, this limited our control over the object instances that participants could use as well as the opportunity to create our own evaluation set for comparing the performance of the models against the same data.
To allow some time before testing for the photos to be received on our server and the models to be trained on our GPUs, participants were asked to review their training photos and select 10 out of 30, 5 out of 10, and 1 out of 5. We are still analyzing these data while considering more fine-grained variation and inconsistency attributes.

CONCLUSION AND FUTURE WORK

We have presented a crowdsourcing study, where MTurkers choose three objects in their environment and iteratively train a model to distinguish between them in real-time using the camera on their mobile phones. By doing so, we were able to explore, with a large participant pool ($N = 100$), an instance of a machine teaching problem with a task where many non-experts can serve as the oracle. Our findings and insights can contribute to the ongoing discussion on how non-experts conceptualize, experience, and reflect on their engagement with machine teaching. To allow for study replicability and future comparisons, we have provided a detailed description of our testbed, its framing within the machine teaching problem space from Zhu *et al*. [64], and the list of questions and prompts used in the study.

Our results are based on a fine-grained analysis of the participants' examples contextualized by their responses, background, and model performance. We discuss how they can guide the design of future teachable interfaces to anticipate users tendencies, misconceptions, and assumptions. Given our research group's interest in teachable interfaces for accessibility [33], our next step will be to explore whether these insights and data from sighted participants could be leveraged for the design of effective teachable object recognizers for blind users. Our rationale is that insights from this study can perhaps enable us to decouple non-experts misconceptions from challenges in camera manipulations among blind users [38].

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful comments on an earlier draft of this paper. This work is supported by NSF (#1816380). Kyungjun Lee is supported by NIDILRR (#90REGE0008).

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article Paper 582, 18 pages. DOI: http://dx.doi.org/10.1145/3173574.3174156

[2] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* 19, 6 (December 1974), 716–723. DOI: http://dx.doi.org/10.1109/TAC.1974.1100705

[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 (Dec. 2014), 105–120. DOI: http://dx.doi.org/10.1609/aimag.v35i4.2513

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 3, 13 pages. DOI: http://dx.doi.org/10.1145/3290605.3300233

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[6] Flask API. 2010. Browsable Web APIs for Flask. (2010). https://www.flaskapi.org

[7] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Cal. L. Rev.* 104 (2016), 671.

[8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*. Association for Computing Machinery, New York, NY, USA, 41–48. DOI: http://dx.doi.org/10.1145/1553374.1553380

[9] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data. *Information, Communication & Society* 15, 5 (2012), 662–679. DOI: http://dx.doi.org/10.1080/1369118X.2012.678878

[10] Danielle Bragg, Nicholas Huynh, and Richard E. Ladner. 2016. A Personalizable Mobile Sound Detector App Design for Deaf and Hard-of-Hearing Users. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '16)*. Association for Computing Machinery, New York, NY, USA, 3–13. DOI: http://dx.doi.org/10.1145/2982142.2982171

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. DOI: http://dx.doi.org/10.1191/1478088706qp063oa

[12] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. 2011. Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6, 1 (23 8 2011), 3–5. DOI:http://dx.doi.org/10.1177/1745691610393980

[13] Minjie Cai, Kris Kitani, and Yoichi Sato. 2018. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. (2018).

[14] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now 2017 Report. *AI Now Institute at New York University* (2017). https://ainowinstitute.org/AI_Now_2017_Report.pdf

[15] European Commision. 2016. European Union General Data Protection Regulation (GDPR). (2016). Retrieved September 1, 2018 from https://www.darpa.mil/program/explainable-artificial-intelligence

[16] US Congress. 2019. S.1108 - Algorithmic Accountability Act of 2019. (2019). https://www.congress.gov/116/bills/s1108/BILLS-116s1108is.pdf

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255. DOI: http://dx.doi.org/10.1109/CVPR.2009.5206848

[18] Nicholas Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. (2014). DOI: http://dx.doi.org/https://doi.org/10.7916/D8ZK5TW2

[19] Rüdiger Dillmann. 2004. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems* 47, 2 (2004), 109 – 116. DOI: http://dx.doi.org/https://doi.org/10.1016/j.robot.2004.03.005 Robot Learning from Demonstration.

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. DOI: http://dx.doi.org/10.1145/2090236.2090255

[21] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. DOI: http://dx.doi.org/10.1145/604045.604056

[22] 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 (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 147–156. DOI: http://dx.doi.org/10.1145/1978942.1978965

[23] Z. Gong, P. Zhong, and W. Hu. 2019. Diversity in Machine Learning. *IEEE Access* 7 (2019), 64323–64350. DOI:http://dx.doi.org/10.1109/ACCESS.2019.2917620

[24] David Gunning. 2017. Explainable Artificial Intelligence (XAI). (2017). Retrieved September 1, 2018 from https://www.darpa.mil/program/explainable-artificial-intelligence

[25] Kotaro Hara, Abigail Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey P. Bigham. 2018. A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 449, 14 pages. DOI: http://dx.doi.org/10.1145/3173574.3174023

[26] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 203–212. DOI:http://dx.doi.org/10.1145/1753326.1753357

[27] Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 415, 11 pages. DOI:http://dx.doi.org/10.1145/3290605.3300645

[28] 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 (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 419–429. DOI: http://dx.doi.org/10.1145/2736277.2741102

[29] Jonggi Hong, Kyungjun Lee, June Xu, and Hernisa Kacorri. 2019. Exploring Machine Teaching for Object Recognition with the Crowd. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper LBW0279, 6 pages. DOI: http://dx.doi.org/10.1145/3290607.3312873

[30] I. I. Itauma, H. Kivrak, and H. Kose. 2012. Gesture imitation using machine learning techniques. In *2012 20th Signal Processing and Communications Applications Conference (SIU)*. 1–4. DOI: http://dx.doi.org/10.1109/SIU.2012.6204822

[31] E. Johns, O. M. Aodha, and G. J. Brostow. 2015. Becoming the expert - interactive multi-class machine teaching. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2616–2624. DOI: http://dx.doi.org/10.1109/CVPR.2015.7298877

[32] Hernisa Kacorri. 2017a. Teachable Machines for Accessibility. *SIGACCESS Access. Comput.* 119 (Nov. 2017), 10–18. DOI: http://dx.doi.org/10.1145/3167902.3167904

[33] Hernisa Kacorri. 2017b. Teachable machines for accessibility. *ACM SIGACCESS Accessibility and Computing* 119 (2017), 10–18.

[34] Hernisa Kacorri, Kris M. Kitani, Jeffrey P. Bigham, and Chieko Asakawa. 2017. People with Visual Impairment Training Personal Object Recognizers: Feasibility and Challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5839–5849. DOI: http://dx.doi.org/10.1145/3025453.3025899

[35] Faisal Khan, Bilge Mutlu, and Jerry Zhu. 2011. How Do Humans Teach: On Curriculum Learning and Teaching Dimension. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1449–1457. https://tinyurl.com/vfuaxvp

[36] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 411.

[37] Google Creative Lab. 2017. Teachable machine. (2017). https://experiments.withgoogle.com/teachable-machine

[38] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. 2019. Revisiting Blind Photography in the Context of Teachable Object Recognizers. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '19)*. Association for Computing Machinery, New York, NY, USA, 83–95. DOI: http://dx.doi.org/10.1145/3308561.3353799

[39] Kyungjun Lee and Hernisa Kacorri. 2019. Hands Holding Clues for Object Recognition in Teachable Machines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 336, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300566

[40] Y. J. Lee and K. Grauman. 2011. Learning the easy things first: Self-paced visual category discovery. In *CVPR 2011*. 1721–1728. DOI: http://dx.doi.org/10.1109/CVPR.2011.5995523

[41] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods* 49, 2 (4 2017), 433–442. DOI:http://dx.doi.org/10.3758/s13428-016-0727-z Leib Litman and Jonathan Robinson share first authorship of this article.

[42] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44, 1 (2012), 1–23. DOI: http://dx.doi.org/10.3758/s13428-011-0124-6

[43] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. (2019).

[44] Arvind Narayanan. 2018. FAT* tutorial: 21 fairness definitions and their politics. *New York, NY, USA* (2018).

[45] Thomas J Palmeri and Isabel Gauthier. 2004. Visual object understanding. *Nature Reviews Neuroscience* 5, 4 (2004), 291. DOI:http://dx.doi.org/10.1038/nrn1364

[46] S. J. Pan and Q. Yang. 2010. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering* 22, 10 (Oct 2010), 1345–1359. DOI: http://dx.doi.org/10.1109/TKDE.2009.191

[47] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. (2010). https://repub.eur.nl/pub/31983

[48] Rupal Patel and Deb Roy. 1998. Teachable interfaces for individuals with dysarthric speech and severe physical disabilities. In *Proceedings of the AAAI Workshop on Integrating Artificial Intelligence and Assistive Technology*. Citeseer, 40–47.

[49] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153 – 163. DOI: http://dx.doi.org/10.1016/j.jesp.2017.01.006

[50] Ben Shneiderman and Pattie Maes. 1997. Direct Manipulation vs. Interface Agents. *Interactions* 4, 6 (Nov. 1997), 42–61. DOI: http://dx.doi.org/10.1145/267505.267514

[51] Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *CoRR* abs/1707.06742 (2017). http://arxiv.org/abs/1707.06742

[52] Daniel J. Simons and Christopher F. Chabris. 2012. Common (Mis)Beliefs about Memory: A Replication and Comparison of Telephone and Mechanical Turk Survey Methods. *PLOS ONE* 7, 12 (12 2012), 1–5. DOI: http://dx.doi.org/10.1371/journal.pone.0051876

[53] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. 2017. Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences* 2017 (Jan. 2017), 1–13. DOI:http://dx.doi.org/10.1016/j.tics.2017.06.007

[54] Lucy A Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

[55] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 403–412.

[56] William R. Swartout. 1983. XPLAIN: a system for creating and explaining expert consulting programs. *Artificial Intelligence* 21, 3 (1983), 285 – 325. DOI: http://dx.doi.org/10.1016/S0004-3702(83)80014-9

[57] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2818–2826. DOI: http://dx.doi.org/10.1109/CVPR.2016.308

[58] Andrea L. Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6 (2008), 716 – 737. DOI: http://dx.doi.org/10.1016/j.artint.2007.09.009

[59] Jennifer Wortman Vaughan. 2018. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. *Journal of Machine Learning Research* 18, 193 (2018), 1–46. http://jmlr.org/papers/v18/17-234.html

[60] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019. Trick Me If You Can: Human-in-the-loop Generation of Adversarial Question Answering Examples. *Transactions of the Association for Computational Linguistics* 7, 0 (2019), 387–401. https: //transacl.org/ojs/index.php/tacl/article/view/1711

[61] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 601, 15 pages. DOI: http://dx.doi.org/10.1145/3290605.3300831

[62] Daniel S Weld and Gagan Bansal. 2018. Intelligible Artificial Intelligence. *arXiv preprint arXiv:1803.04263* (2018).

[63] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. 2018. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 573–584. DOI: http://dx.doi.org/10.1145/3196709.3196729

[64] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N. Rafferty. 2018. An Overview of Machine Teaching. *CoRR* abs/1801.05927 (2018). http://arxiv.org/abs/1801.05927

[65] Abigail Zimmermann-Niefield, Makenna Turner, Bridget Murphy, Shaun K. Kane, and R. Benjamin Shapiro. 2019. Youth Learning Machine Learning through Building Models of Athletic Moves. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children (IDC '19)*. Association for Computing Machinery, New York, NY, USA, 121–132. DOI: http://dx.doi.org/10.1145/3311927.3323139