Learning to Work in a Materials Recovery Facility: Can Humans and Machines Learn from Each Other?

HARRISON KYRIACOU, Worcester Polytechnic Institute (WPI), USA
ANAND RAMAKRISHNAN, WPI
JACOB WHITEHILL, WPI

Workplace learning often requires workers to learn new perceptual and motor skills. The future of work will increasingly feature human users who cooperate with machines, both to learn the tasks and to perform them. In this paper, we examine workplace learning in Materials Recovery Facilities (MRFs), i.e., recycling plants, where workers separate waste items on conveyer belts before they are formed into bales and reprocessed. Using a simulated MRF, we explored the benefit of machine learning assistants (MLAs) that help workers, and help train them, to sort objects efficiently by providing automated perceptual guidance. In a randomized experiment ($n = 140$), we found: (1) A low-accuracy MLA is worse than no MLA at all, both in terms of task performance and learning. (2) A perfect MLA led to the best task performance, but was no better in helping users to learn than having no MLA at all. (3) Users tend to follow the MLA's judgments too often, even when they were incorrect. Finally, (4) we devised a novel learning analytics algorithm to assess the worker's accuracy, with the goal of obtaining additional training labels that can be used for fine-tuning the machine. A simulation study illustrates how even noisy labels can increase the machine's accuracy.

CCS Concepts: • **Applied computing** → **Computer-assisted instruction**; **Interactive learning environments**; • **Human-centered computing** → Empirical studies in HCI.

Additional Key Words and Phrases: perceptual learning, automated feedback, object detection, MRF

ACM Reference Format:
Harrison Kyriacou, Anand Ramakrishnan, and Jacob Whitehill. 2021. Learning to Work in a Materials Recovery Facility: Can Humans and Machines Learn from Each Other?. 1, 1 (January 2021), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As AI continues to grow and transform workplaces, human workers will need to adapt and learn how to be productive in their new environments [7]. Workplace learning consists in acquiring not just cognitive skills, but also *perceptual* and *motor* skills [17] of how to see and manipulate physical objects. In many workplaces [19], machines and robots cannot do the job by themselves, but instead must work with human workers cooperatively to accomplish their goals. While robots can perform repetitive tasks without tiring, their

Authors' addresses: Harrison Kyriacou, Worcester Polytechnic Institute (WPI), USA, hkyriacou@wpi.edu; Anand Ramakrishnan, WPI, aramakrishnan@wpi.edu; Jacob Whitehill, WPI, jrwhitehill@wpi.edu.

perceptual skills can still lag behind humans' in many domains, especially in non-stationary environments, where the task is continually evolving, and deeper semantic understanding is required.

One particular way in which humans and AI can collaborate is when the machine provides humans with automated guidance about how to perceive, but leaves the final decision to the human worker about how to act. Since automated perception is so often driven by machine learning, we call this model a Machine Learning Assistant (MLA). The MLA could, for example, highlight important objects on a computer screen or augmented reality in wearable computing devices [5]. The idea could provide bidirectional benefits: The machine can help new workers to learn the task, give enough guidance to reduce workers' fatigue, and catch occasional mistakes. On the other hand, experienced human workers can give the machine expert supervision about the objects and how to deal with them, potentially improving the machine's accuracy over time.

Human-AI teaming has the potential to improve productivity and learning, but there are many unresolved questions. Automated guidance and feedback have the potential to reduce the amount of information that humans must visually scan and thus reduce their cognitive load; however, if the machine is inaccurate, then its feedback might be distracting. Automated assistance might help humans *perform* the task, but they might become dependent on it and never *learn* to complete the task themselves. Machine learning has been beneficial for training many new AI systems for automated perception in challenging domains (e.g., diagnosis of skin cancer from radiographs [6]), but even the best systems sometimes make mistakes; if the human user trusts the system too much, it might learn the skill incorrectly and perform disastrously.

In this paper we explore the potential for human-AI teaming and how humans and machine learning assistants can learn from each other. We examine the particular setting of a Materials Recovery Facility (MRF), i.e., recycling plant where human workers separate different kinds of trash before they are baled and reprocessed (see Fig. 1 (left)). MRFs form an important part of the waste management cycle. Sorting trash at MRFs is physically demanding, with shifts lasting 10+ hours and considerable back-breaking leaning over the conveyer belt, and can be dangerous given the kinds of non-recyclable items that they must handle. (During a recent visit by our research group to a MRF near our university, we learned that non-recyclable items that arrive at MRFs can include syringes, bowling balls, and deer carcasses.) While robots for MRFs are starting to emerge, it is unlikely that the sorting task can be completely automated in the foreseeable future due to the complex perception and object manipulation that must be performed.

In our study, we investigate the following **research questions**: How does the availability, as well as the accuracy, of a machine learning assistant (MLA) affect workers' performance on perceptual (object detection) tasks? How does it affect their learning of the task? Do users trust the more accurate MLAs more than the less accurate ones? How can the machine infer the worker's accuracy, and how can it use the human worker's actions (whether or not to remove an object from the conveyer belt) to improve its own accuracy?

2 RELATED WORK

We are unaware of any prior work that examines quite the same research questions as we do (see above). However, there are several related fields:

Perceptual learning and motor learning: A vast literature rooted in psychology and cognitive science studies how humans learn to recognize and distinguish subtle visual and auditory signals. Practical applications include identifying pathologies from radiographs, comparing fingerprints, and recognizing voice tones. Motor learning, which is studied by the sports medicine and kinesiology communities, is concerned

Fig. 1. **Left**: Conveyer belts for trash sorting at a real Materials Recovery Facility (MRF). **Right**: The Simulated MRF that we developed and deployed for our experiments. Some objects are highlighted with a box that is generated by a Machine Learning Assistant (MLA) to help the worker to recognize objects that need to be removed from the conveyer belts.

with helping humans to acquire or re-develop motor skills such as walking or throwing a ball. Some of the literature from these fields examines the role of *feedback*, i.e., providing information to the student on the correctness of their response (for the perceptual task) [18] or performance (for motor tasks) [1]. Other works explore the utility of *guidance* mechanisms, including the highlighting of the most salient information in the image to help novices adopt the best practices of perceptual experts [11, 16], or subliminal visual cues [5] in augmented-reality contexts.

Learning from robots: Over the past 15 years, interactive teaching robots have been created to help students to learn various skills [14, 20–22]. However, we are unaware of any prior work that examined how students' learning depends on the accuracy of the robot's feedback or guidance.

Intelligent tutoring systems: The MLA that we deployed inside the simulated MRF can be seen as an intelligent agent that continuously evaluates the student's performance and provides guidance to help the student learn. This is similar to the goals of intelligent tutoring systems (ITS). While ITS have most commonly been used to teach cognitive skills such as mathematics, computer programming, and reading [9, 12, 23], examples also exist of ITS to help students develop perceptual expertise, e.g., microscopic diagnosis of inflammatory skin diseases [3]. One important difference is that our MLA has only imperfect knowledge of the answers themselves, i.e., which objects on the conveyer belt are recyclable. This makes the knowledge tracing [2] problem more difficult. We explore this in more detail in Sec. 6.

Item response theory and consensus algorithms: One of the tasks performed by our MLA is to estimate whether the student's perceptual skill has surpassed that of the MLA itself. Estimating students' abilities based on their answers to test questions is the central problem tackled by item response theory (IRT). In contrast to standard IRT, in our scenario, the answers to the test items themselves are unknown since the MLA itself only has an imperfect classifier to estimate each object's label. Over the past decades, many *consensus algorithms* have been devised for this exact purpose. For example, the seminal work by Dawid & Skene [4] illustrates how the students' accuracies and the items' labels can be inferred simultaneously. Their algorithm provides an iterative solution based on Expectation-Maximization. In contrast, we derive an analytical solution by harnessing the specific constraints of our problem (see Sec. 6).

3 SIMULATED MATERIALS RECOVERY FACILITY (SMRF)

To investigate how human workers learn to sort recyclable from non-recyclable items in a MRF, and how their task performance and learning curves might depend on the availability and accuracy of a Machine Learning Assistant (MLA) that highlights recyclable items to make them easier to spot, we created a Simulated Materials Recovery Facility (SMRF) game[1]. In the SMRF (see Fig. 1 (right)), there are three conveyor belts operating at different speeds that carry different kinds of objects – some recyclable (e.g., paper, bottles, bags) and some not (e.g., food scraps, broken glass, and some dangerous items such as syringes). There are different specific instances of each kind of object (e.g., various kinds of food scraps). Objects of different types are randomly placed on the conveyer belt and move from left to right. Each object may be rotated to various degrees. Moreover, objects may occlude each other, and sometimes it is necessary to move items (by clicking and dragging them) slightly on the belt to get a clear view of what is underneath.

Like a real MRF, the user's job is to remove from the conveyor belts the non-recyclable items and put them into trash bins (the green bin at the top of the figure). The game is divided into three rounds (each 3 minutes long), and within each round, the user receives a score whose value is updated and displayed continuously throughout gameplay. If the user correctly moves all the non-recyclable objects and none of the recyclable objects into the trash bin, they receive 100 points (perfect score). For every non-recyclable item that they either neglect to move into the trash and every recyclable item they erroneously move into the trash, they lose 1 point (down to a minimum score of 0). Auditory feedback in the form of distinct sounds for correct and incorrect actions is also given throughout the game. Because of the perceptual difficulty in recognizing the different kinds of objects, and because of the game's fast pace, the task is quite challenging and requires considerable concentration.

3.1 Machine Learning Assistants

We created Machine Learning Assistants (MLAs) to help the user both to learn the task and to perform it. The MLA's job is to highlight for the user (see the red rectangles in Fig. 1 (right)) the objects that the MLA believes are non-recyclable and thus should be moved into the trash bin. It is still up to the user, however, to move these items into the trash. The MLA thus provides perceptual assistance but still requires the human users to do the work. Algorithmically, the MLAs are object detectors, based on the Faster R-CNN design [15], that are trained to identify non-recyclable objects (syringes, garden hoses, batteries, cables, soiled boxes, food, and shattered glass) in the work area. We trained different MLAs with different accuracies: Low (L), High (H), and Perfect (P). See Table 1 for their associated True Positive Rate, False Positive Rate, Area Under the ROC Curve, and a number of Extra False Alarms (false positive detections that occurred outside the conveyer belts) per frame. Within each game round, the user is given (depending on their randomly assigned experimental condition) one of these three detectors, or possibly no detector at all.

Implementation: Faster R-CNN is a state-of-the-art neural network architecture for creating object detectors. Since our goal was to create a game that could be played in a normal web browser (such as on Amazon Mechanical Turk), and since running Faster R-CNN model would be too slow in a browser, we pursued a hybrid strategy: In the actual SMRF with which we conducted experiments on human subjects, the object detector was *simulated*, given the ground-truth positions of all the objects (which are, of course,

[1]The game can be played on Amazon S3: https://conveyerbeltgame.s3.amazonaws.com/instructions.html

MLA	AUC	TPR	FPR	# Extra FA
Perfect (P)	1	1	0	0
High (H)	0.945	0.99	0.1	3.19
Low (L)	0.65	0.55	0.25	3.19
None (N)	–	–	–	–

Table 1. Accuracy characteristics of the different Machine Learning Assistants (MLAs).

known by the simulator), as well as the accuracy characteristics of a *real* Faster R-CNN detector that we trained and evaluated offline. The accuracy characteristics include the probabilities of true positive and false positive detections, as well as the positional noise in localizing each detected object precisely. Our hybrid strategy produced realistic noisy predictions and jitter that occurs with real-world object detectors, and it also allowed us to generate MLAs with varying accuracy profiles (P, H, L, or N).

Training: For the convolutional blocks, we used a Resnet50-V2. We trained the detectors from scratch using sampled (at 1Hz) screenshots and associated labels (objects' positions and types) from the SMRF. We used 23050 images for training, and 11532 images for validation. We trained the model using Adam (lr=10e-3) and early stopping with a patience of 10 epochs based on the validation loss. Finally, we computed statistics of the true positive rate, false positive rate, and localization error.

4 EXPERIMENT

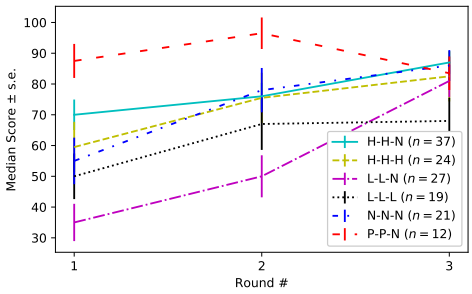
Given the SMRF and MLAs, we conducted an experiment during August 2020 among $n = 140$ participants on Amazon Mechanical Turk. Each participant was randomly assigned to one of 6 experimental conditions – H-H-N, H-H-H, L-L-N, L-L-L, N-N-N, and P-P-N – that indicate the MLA's accuracy at each game round. The number of participants per condition is shown in Table 2[2]. Completing the experiment takes about 12 minutes in total. Each participant received a base payment of $1 for completing the task. To incentivize good performance, participants earned an extra $0.50 per 100 points scored in the game as a bonus payment.

The experiment begins with an introduction that explains the game goal, the Machine Learning Assistant, and the scoring system. Next, the game rules and controls are explained, including the descriptions of the different kinds of objects, and how to move non-recyclable objects from the conveyor belts to the trash bin. The user receives visual examples and an English description of each kind of non-recyclable object.

The game consists of three rounds. Before each round, the user must press a "Start" button. The software automatically records each user's game score, the accuracy of their assigned MLA for each round, and the history of objects that they dragged. After all three rounds, the user is given a validated survey [8] to ask them about their level of trust in the MLA they received. The survey consists of 12 Likert items (1-7 scale) with questions such as, "The system is dependable", "I can trust the system", etc.

Dependent measures: We assessed *performance* as the score (0-100) during the third (and final) round of the game, regardless of whether an MLA was present during that round. We defined *learning* as the ability to perform the task *without* the MLA and thus assessed it as the score obtained during the third round, but only for conditions in which no MLA was present during that round (i.e., H-H-N, L-L-N, P-P-N, and N-N-N)[3]. Finally, we assessed *trust* (12-84) using the validated survey instrument [8] mentioned above.

[2]Given 4 agent types (L, H, P, N) and 3 game phases, there could theoretically be $4^3 = 64$ conditions, but this is far too many to be practical. We instead chose 6 conditions that we believed would be most revealing.
[3]One could alternatively define learning as the difference in final vs. first round scores; however, a fair comparison would require an extra "round 0" with No detector, which would have lengthened the experiment.



		Median Game Score ± s.e.		
Condition	n	**Round 1**	**Round 2**	**Round 3**
H-H-N	37	70.0 ± 4.95	76.0 ± 4.36	87.0 ± 3.68
H-H-H	24	59.5 ± 8.20	75.5 ± 8.15	82.5 ± 8.17
L-L-N	27	35.0 ± 6.06	50.0 ± 6.83	81.0 ± 5.25
L-L-L	19	50.0 ± 7.41	67.0 ± 8.42	68.0 ± 7.56
N-N-N	21	55.0 ± 7.56	78.0 ± 7.21	86.0 ± 5.02
P-P-N	12	87.5 ± 5.53	96.5 ± 5.12	83.5 ± 5.45

Fig. 2. Results of the experiment conducted on Amazon Mechanical Turk: Median game scores and their standard error estimates, separated by round, for each experimental condition.

	Median # Dragged Objects ± s.e.		
Condition	**Round 1**	**Round 2**	**Round 3**
H-H-N	63.0 ± 4.95	67.0 ± 4.36	53.0 ± 3.68
H-H-H	66.0 ± 8.20	75.0 ± 8.15	75.5 ± 8.17
L-L-N	54.0 ± 6.06	57.0 ± 6.83	54.0 ± 5.25
L-L-L	57.0 ± 7.41	68.0 ± 8.42	71.0 ± 7.56
N-N-N	54.0 ± 7.56	58.0 ± 7.21	50.0 ± 5.02
P-P-N	73.0 ± 5.12	77.5 ± 4.74	42.5 ± 5.04

Table 2. Median numbers of dragged objects and their standard error estimates.

5 RESULTS

The median[4] game score for each condition with associated standard error estimates is shown separately by round in Fig. 2. Table 2 shows the median number of dragged objects (i.e., objects dragged from the conveyer belt into the trash bin, regardless of whether they were recyclable) by round and condition.

5.1 Does a MLA help users to perform the task?

We compared game scores in rounds 1 and 2 of participants who received the High-accuracy detector (conditions H-H-H, H-H-N), the Low-accuracy detector (L-L-N, L-L-L), the Perfect-accuracy detector (P-P-N), and No detector at all (N-N-N). Using a non-parametric repeated measures test for longitudinal data [13], we found that there was a significant effect of experimental condition ($F = 4.23, df = 4.38, p < 0.01$) and round ($F = 27.85, df = 1.85, p < 0.001$) on scores. Then, Kruskal-Wallis H-tests revealed that, in each of Rounds 1 and 2, the Perfect-accuracy detector helped users to obtain a higher score compared to the High-accuracy (round 1: $H = 5.58, p = 0.018$; round 2: $H = 9.22, p = 0.002$) or Low-accuracy detectors (round 1: $H = 20.44, p < 0.001$, round 2: $H = 18.04, p < 0.001$), and also compared to no detector at all (round 1: $H = 11.36, p < 0.001$, round 2: $H = 10.01, p = 0.002$). The Low-accuracy detector actually *hurt* performance compared to having no detector at all (round 1: $H = 5.44, p = 0.02$, round 2: $H = 5.96, p = 0.01$). The High-accuracy detector improved performance in round 1 ($H = 5.21, p = 0.02$), but not in round 2 ($H = 0.24, p = 0.63$), compared to no detector at all. The results suggest that a MLA must have very high accuracy in order to boost humans' task performance. We explore a possible explanation in Sec. 5.3.

[4]Game scores were not normally distributed, and hence we used non-parametric statistical tests.

5.2 Does a MLA help users to learn?

For this analysis, we considered only those participants who received No MLA during round 3 and thus had to complete the task independently: We compared game scores in round 3 of participants who, during rounds 1 and 2, received the High-accuracy detector (H-H-N), the Low-accuracy detector (L-L-N), the Perfect-accuracy detector (P-P-N), and No detector at all (N-N-N). The only statistically significant difference ($H = 5.70, p = 0.017$) was that receiving High-accuracy detectors (H-H-N) led to higher performance compared to receiving the Low-accuracy detector (L-L-N). Interestingly, even though the Perfect-accuracy detector led to the highest performance during rounds 1 and 2 (Sec. 5.1), the scores of users in that condition (P-P-N) suffered when they no longer had the MLA. In fact, the median round-3 score (87.0) of H-H-N participants was slightly (though not stat. sig.) higher compared to the P-P-N condition (83.5). In summary, the MLAs did not provide a better training experience to the human workers compared to not giving any guidance at all (N). We did find evidence, however, of how a low-accuracy MLA can impair users' learning.

An interesting distinction is whether workers are learning *conceptually* which objects should be recycled, versus *mechanically* how to drag them. The fact that the round 2 scores were higher than round 1 scores (Wilcoxon signed-rank test of score differences, $p < 0.001$) in the P-P-N condition – in which the worker could always accept the MLA's suggestions without needing to decide whether they were correct – is consistent with the hypothesis that the learning was at least partially mechanical.

5.3 What happens when the MLA is removed?

We examined the number of dragged objects during round 3 of participants who received an MLA during rounds 1 and 2 but not round 3 (conditions H-H-N, L-L-N, and P-P-N). Using a non-parametric repeated measures test, we found a significant effect of experimental condition ($F = 6.49, df = 4.54, p < 0.001$) and round ($F = 118.22, df = 1.79, p < 0.001$) on number of dragged objects. For participants in H-H-N or L-L-N, median scores during round 3 were higher than the corresponding median scores during rounds 1 and 2. On the other hand, the median number of dragged objects *decreased* in round 3 compared to rounds 1 and 2 (see Table 2). This suggests that workers were not just slowed down by possible visual distraction from the imperfect MLAs; rather, they were actively dragging the *wrong* objects. It examplifies how automated assistance can hurt the worker in terms of performance and learning. In contrast, for the P-P-N participants, both the number of dragged objects and the game scores tended to decrease from round 2 to round 3.

5.4 Do users trust the MLA?

There was strong agreement between the accuracy of the MLA (Low, High, or Perfect) and participants' trust scores as computed from their responses to the trust survey: The median trust scores for the L, H, and P detectors were 47.5, 56.0, and 65.0, respectively. Although users trusted the Low-accuracy detector less than the other two MLAs, they also dragged more objects when the MLA was available compared to when it was not, suggesting they still trust the MLAs too much and follow their automated suggestions too willingly.

6 LEARNING FROM THE HUMAN USER

Here we explore how the human worker can help the machine to improve its own accuracy. From the machine's perspective, there is an ongoing stream of objects that go by on the conveyer belt. The machine

classifier can classify each object, and it knows its own average accuracy α (defined below), but it does not know the ground-truth label of any object. The human user provides implicit feedback by moving objects (or not) from the conveyer belt into the trash can, but their accuracy β is not known a priori by the machine. Over relatively long sequences, β can change due to learning, fatigue, and other factors; however, we assume it is constant over relatively short sequences. (For an example of how to model its evolution over time, see the paper of Khajah et al. [10].) Under these assumptions, what can the machine infer about β? This is the computational task tackled by item response theory (IRT), for the particular setting where the accuracy of one "student" (the machine) is known and of the other (the human user) is unknown.

6.1 Algorithm: Inferring the Human User's Accuracy Without Ground-Truth Labels

Let $\alpha \in (0.5, 1]$ be the probability that the machine classifies an object correctly (note that we assume the machine's accuracy is better than chance). We assume that α is known, e.g., via prior estimation on a validation set. Let $\beta \in [0, 1]$ be the accuracy of the human user, which is unknown a priori. For a set of n objects with binary labels with equal probability of each class, let n_{agr} be the number of objects on which the machine's judgment agrees with the human's, and let $f = \frac{n_{\text{agr}}}{n}$.

On a single trial, the probability a that the machine's and the human's judgments agree is

$$a = \alpha\beta + (1 - \alpha)(1 - \beta) \tag{1}$$

i.e., they either are both correct, or they are both incorrect. Over n trials, the log-likelihood, as a function of β, of agreeing on n_{agr} objects and disagreeing on $(n - n_{\text{agr}})$ is then

$$\mathcal{L}(\beta) = \log \left[a^{n_{\text{agr}}} (1 - a)^{n - n_{\text{agr}}} \right] = n_{\text{agr}} \log a + (n - n_{\text{agr}}) \log(1 - a)$$

To find the maximum-likelihood estimator $\widehat{\beta}$, we find where $\frac{\partial \mathcal{L}}{\partial \beta}$ vanishes, i.e., $\left(\frac{n_{\text{agr}}}{a} - \frac{n - n_{\text{agr}}}{1 - a} \right) \frac{\partial a}{\partial \beta} = 0$. After substituting for a (with Eq. 1) and doing some algebraic manipulation, this yields the solution $\widehat{\beta} = \frac{f + \alpha - 1}{2\alpha - 1}$ but only if this value is inside the interval $[0, 1]$. If $\frac{\partial \mathcal{L}}{\partial \beta}$ is never 0 within this interval, then since $\mathcal{L}$ is a smooth function and its derivative is thus continuous everywhere, $\mathcal{L}$ must be strictly monotonic. In that case, the solution must be at one of the interval's endpoints, i.e., $\widehat{\beta} \in \{0, 1\}$.

Given the machine's accuracy α and the fraction of trials on which the machine agrees with the human's judgments, $\widehat{\beta}$ can thus be easily computed. How high must the fraction of agreement f be so that the human's estimated accuracy is higher than that of the machine, i.e., $\widehat{\beta} > \alpha$? If $\frac{f + \alpha - 1}{2\alpha - 1} \in [0, 1]$, then since $\alpha > 0.5$ (and thus $2\alpha - 1 > 0$), we have: $\widehat{\beta} > \alpha \iff f > 1 - 2\alpha + 2\alpha^2$. If, on the other hand, $\frac{f + \alpha - 1}{2\alpha - 1} > 1$, then $f > \alpha$, which is subsumed by the second inequality above.

Suppose the machine estimates that $\widehat{\beta} > \alpha$. In that case, one sensible strategy is to trust the human's labels more than the machine's. This can be useful in two ways: (1) In a setting where a robot not only perceives but moves objects into the trash based on input from the human user, following the human's judgments over the machine's will increase the team's overall accuracy. (2) In a setting where the MLA perceives but does not act, the human's labels can be used as noisy feedback for fine-tuning.

6.2 Experiment: Fine-Tune with Noisy Human Labels

As an illustration of the second point above, we conducted a machine learning experiment to estimate how the machine's perceptual accuracy α might improve by fine-tuning on the human's noisy labels. In particular,

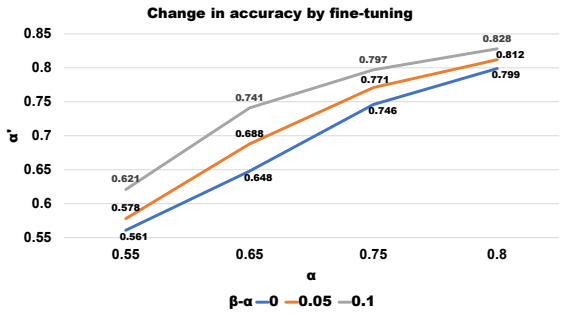


Fig. 3. Accuracy gain (α' vs. α) of the object detector using labels from a worker with accuracy β.

we varied the baseline accuracy α of the machine as well as the accuracy $\beta > \alpha$ of the human user via a simulation. The human labels were accepted by the machine as ground-truth, despite not being entirely accurate. We then assessed to what extent these labels improved the machine's accuracy via fine-tuning.

Procedures: We trained object detectors, using the same procedure as in Sec. 3.1 combined with early stopping[5] on the *test* set, to achieve 4 different accuracies – 0.55, 0.65, 0.75, and 0.80 – quantified as the percent of correctly classified objects when the prior probability of each class (recyclable, non-recyclable) is equal. We then obtained simulated noisy human labels on 5000 images with the accuracy difference $(\beta - \alpha) \in \{0.00, 0.05, 0.1\}$, where 0.00 was a control. We then fine-tuned our trained detector model on the obtained labeled images but with a reduced learning rate (10e-5). For each combination of α and $(\beta - \alpha)$, we fine-tuned and then computed the machine's new accuracy α' after fine-tuning.

Results are given in Fig. 3. When the accuracy of the human worker exceeded that of the machine (i.e., $(\beta - \alpha) \in \{0.05, 0.1\}$), we consistently obtained improvements in α' compared to α. As expected, this trend is not observed for the control condition $\beta - \alpha = 0.00$, where we actually see a slight decrease in performance. These results provide an example of how MLAs can learn from human workers in human-AI teams.

7 CONCLUSIONS

We explored how human workers in a simulated materials recovery facility (MRF) might benefit, in terms of task performance and learning, from a machine learning assistant (MLA) that provides guidance as to which objects are non-recyclable. Results from a randomized experiment ($n = 140$) suggest: (1) The accuracy of the MLA significant affects both learning and performance, but there was no clear benefit of even very high-accuracy MLAs compared to having no MLA at all. (2) The users reported greater trust in the more accurate MLAs, but they still followed the guidance of the less-accurate detectors more often than they should have, and their performance suffered as a consequence. Finally, (3) we derived a novel algorithm to infer the human worker's accuracy based on the MLA's imperfect knowledge of the label of each object (recyclable or not). A machine learning experiment, in which we demonstrated how obtaining labels from trained human users and using them for fine-tuning can improve the machine's own accuracy, provided an example of how humans and AI can learn from each other. Overall, our results underline the difficulty in designing worker training experiences, based on automated guidance from an MLA, that actually help them

[5]As an alternative to early stopping, we also tried training on a smaller training set until convergence; it gave similar results.

to learn to perform the task independently. Future work will explore how to modulate the accuracy of the MLA across multiple rounds, as well as devise new kinds of visual guidance, so as to teach workers more effectively. **Acknowledgement**: We gratefully acknowledge NSF grants #1928506 and #1822830.

REFERENCES

[1] Suzete Chiviacowsky and Gabriele Wulf. 2007. Feedback after good trials enhances learning. *Research quarterly for exercise and sport* 78, 2 (2007), 40–47.

[2] Albert T Corbett and John R Anderson. 1994. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction* 4, 4 (1994), 253–278.

[3] Rebecca S Crowley and Olga Medvedeva. 2006. An intelligent tutoring system for visual classification problem solving. *Artificial intelligence in medicine* 36, 1 (2006), 85–117.

[4] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C* 28, 1 (1979), 20–28.

[5] Richard W DeVaul, Alex Pentland, and Vicka R Corey. 2003. The memory glasses: subliminal vs. overt memory support with imperfect information. In *IEEE Symposium on Wearable Computers*.

[6] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 546, 7660 (2017), 686–686.

[7] National Science Foundation. [n.d.]. NSFs 10 Big Ideas.

[8] 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.

[9] Shamya Karumbaiah, Rafael Lizarralde, Danielle Allessio, Beverly Woolf, Ivon Arroyo, and Naomi Wixon. 2017. Addressing Student Behavior and Affect with Empathy and Growth Mindset. *International Educational Data Mining Society* (2017).

[10] Mohammad Khajah, Rowan Wing, Robert Lindsey, and Michael Mozer. 2014. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *EDM*.

[11] Ronald T Kneusel and Michael C Mozer. 2017. Improving human-machine cooperative visual search with soft highlighting. *ACM Transactions on Applied Perception (TAP)* 15, 1 (2017), 1–21.

[12] James A Kulik and JD Fletcher. 2016. Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research* 86, 1 (2016), 42–78.

[13] Kimihiro Noguchi, Yulia R Gel, Edgar Brunner, and Frank Konietschke. 2012. nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* 50, 12 (2012).

[14] Aditi Ramachandran, Sarah Strohkorb Sebo, and Brian Scassellati. 2019. Personalized robot tutoring using the assistive tutor POMDP (AT-POMDP). In *AAAI*, Vol. 33.

[15] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[16] Brett Roads, Michael C Mozer, and Thomas A Busey. 2016. Using highlighting to train attentional expertise. *PloS one* 11, 1 (2016), e0146266.

[17] Richard A Schmidt and Craig A Wrisberg. 2008. *Motor learning and performance: A situation-based learning approach*. Human kinetics.

[18] Aaron R Seitz. 2020. Perceptual Expertise: How Is It Achieved? *Current Biology* 30, 15 (2020).

[19] Jilles Smids, Sven Nyholm, and Hannah Berkers. 2020. Robots in the Workplace: a Threat toor Opportunity forMeaningful Work? *Philosophy & Technology* 33, 3 (2020), 503–522.

[20] Samuel Spaulding and Cynthia Breazeal. 2017. Learning behavior policies for interactive educational play. In *Models, Algorithms, and HRI Workshop*.

[21] Fumihide Tanaka, Aaron Cicourel, and Javier R Movellan. 2007. Socialization between toddlers and robots at an early childhood education center. *PNAS* 104, 46 (2007).

[22] Alessia Vignolo, Henry Powell, Luke McEllin, Francesco Rea, Alessandra Sciutti, and John Michael. 2019. An adaptive robot teacher boosts a human partners learning performance in joint action. In *IEEE International Conference on Robot and Human Interactive Communication*. 1–7.

[23] Zhihong Xu, Kausalai Wijekumar, Gilbert Ramirez, Xueyan Hu, and Robin Irey. 2019. The effectiveness of intelligent tutoring systems on K-12 students' reading comprehension: A meta-analysis. *British Journal of Educational Technology* 50, 6 (2019), 3119–3137.