How to Give Imperfect Automated Guidance to Learners: A Case-Study in Workplace Learning

Jacob Whitehill and Amitai Erfanian

Worcester Polytechnic Institute, Worcester, MA, USA
{jrwhitehill,aberfanian}@wpi.edu

Abstract. In a workplace learning scenario in which workers in a simulated Material Recovery Facility learn to recognize and manipulate objects on conveyer belts, we studied how imperfect guidance from a machine learning (ML) assistant may impact learners' experience and behaviors. Specifically, in a randomized experiment ($n = 181$ participants from Amazon MTurk) we varied the assistant's False Positive (FP) and False Negative (FN) rates in detecting non-recyclable objects and assessed the impact on learners' performance, learning, and trust. We also explored a soft highlighting [8] condition, whereby the assistant provides fine-grained information about how confident it is. We found evidence that the FP/FN trade-off can impact learners' performance when working cooperatively with the assistant, and that the soft highlighting condition may generate less trust from learners compared to the other conditions. There was tentative evidence that workers' behaviors were impacted by the FP/FN trade-off of their assigned experimental condition even after the ML assistant was removed. Finally, in a follow-up study ($n = 27$) we found evidence that learners modulate their behaviors based on the fine-grained confidence values conveyed by the assistant.

Keywords: perceptual learning · simulation-based learning · AI-based feedback · material recovery facility

1 Introduction

As the field of machine learning (ML) continues to grow and proliferate into daily life, the number and diversity of subjects in which AI systems for education (AIEd) can provide learners with automated help will expand as well. Whereas classical AI methods such as expert systems were instrumental in creating intelligent tutoring systems (ITS) in highly structured domains such as computer programming and high school mathematics in the 1980s-90s [10], ML has opened new possibilities to provide scaffolding, guidance, and feedback in more flexible and open-ended domains such as medicine [14], sign language [4], teacher training [1, 17, 15], waste management [11], and many more. It also has the potential to augment standard ITS with new sensors that can better estimate students' emotions and thereby provide a more tailored learning experience [7].

In AIEd systems based on classical AI techniques, the automated feedback strategies are either manually programmed or inferred systematically from the rules of the subject matter (e.g., laws of algebra). In contrast, ML provides more flexibility because it can infer the correctness of a student's answer, and suggest helpful learning strategies, by harnessing swaths of both real and simulated data from previous learners – all without the need for manually crafted heuristics. ML, as the backbone of modern AIEd systems, has yielded powerful new feedback mechanisms in intelligent learning platforms, e.g., automated testing methods for novice programmers building interactive computer programs [13], and classroom observation systems that can give teachers feedback about the quality of their teaching based on audio [15] or video [1] of classroom interactions. The reach of AIEd can thereby expand from traditional classroom-oriented subject matter into more diverse fields, including the space of *workplace learning* where human workers may acquire not just cognitive but also perceptual and motor skills [16] to perform their jobs. This can manifest, for example, in an ITS or perhaps a collaborative "assistant" that provides visual cues about which types of objects in an airport screener are dangerous [2], or by suggesting a Python implementation to a function whose specification was entered by the user [3].

ML in AIEd – useful but imperfect: Along with the great potential of ML to expand the impact of AIEd come new challenges. One of the most severe is that, since the system's behavior is learned statistically so as to generalize to new scenarios (e.g., to new students with answers similar, but not identical, to those in the training set), its feedback is no longer guaranteed to be correct – sometimes the AIEd system may make mistakes. Concrete instances of these mistakes include telling a student that their solution to a computer programming task is incorrect when in fact it is correct; suggesting to them a solution path to a math problem that is either wrong or unnecessarily complicated; failing to recognize the hand gesture of a student learning sign language; etc. Mistakes on the part of the AIEd system may directly inhibit students' learning by misguiding and confusing them. Due to bidirectional influence of *learning* and *trust* between a student and their teacher [12], such mistakes can also indirectly and negatively impact learning if the student loses trust in the AI's ability to help them. Given an ML-based AIEd system's fallibility, it is important to consider how to set the learners' expectations of and structure their interactions with the system so as to optimize their learning.

How to present feedback on binary decisions: When developing ML-based AIEd systems, the question arises of (1) whether, on some binary decision, the machine should tend to err with more false positives (FP) or more false negatives (FN). Binary decisions are ubiquitous in intelligent learning platforms, e.g., judging whether a student's hand-written solution in a mathematics ITS is correct/correct, showing novice teachers a moment in a classroom observation video when a teacher seems to speak angrily (or not) to their student, or flagging an object as dangerous/non-dangerous in an augmented reality display to train airport security officers. False negatives (misses) can result in missed learning opportunities and overly optimistic self-assessments of learning, whereas false positives can confuse the student and damage trust in the AI. Related to the optimum FP/FN trade-off is the question of (2) whether and how to provide to the learner information about how *confident* the machine is in its own judgment; might this information help the learner to interpret the feedback more judiciously and preserve trust?

Case study on material recovery facilities: In this paper, we investigate these two questions within a workplace learning case-study on material recovery facilities (MRFs), i.e., recycling plants. MRFs sort objects by their materials so that they can be bailed and reprocessed, and they form an important part of waste management. MRF workers assist in this process by manually picking out items that were incorrectly sorted by machines, as well as removing objects that are dangerous. MRF work is arduous, with long shifts (often 10+ hours) and physically demanding conditions, and can be dangerous (e.g., due to syringes and other sharp objects on the belts). Because of high employee turnover, new MRF workers must frequently be trained. Due to continually new kinds of manufactured products that arrive at MRFs, workers must learn to recognize and physically manipulate objects with different materials and appearance. Helping new MRF workers to correctly recognize objects could thus boost the efficiency of the MRF, improve waste management, and improve safety for workers.

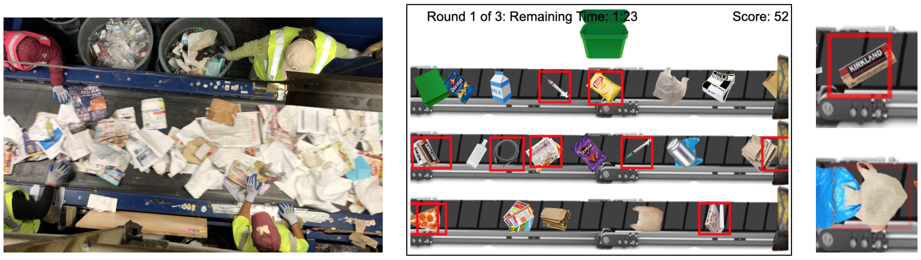

Fig. 1. Left: A material recovery facility (MRF) in which workers sort through items on conveyer belts. **Center**: The simulated MRF [11] used in our experiment on worker training with automated guidance. **Right**: Examples of soft highlighting [8] by the machine learning assistant to express its confidence to the learner about whether each highlighted object should be removed from the conveyer belts.

Research contributions: The central research question that we examine is: **How does the automated guidance provided by a machine learning assistant for an object detection task (recyclable vs. non-recyclable objects on a MRF conveyer belt) affect learners' performance, learning, trust, and behavior?** Aligned with our goal of examining *workplace learning*, we recruit our research participants ($n = 181$) from Amazon Mechanical Turk, a marketplace for online work. By varying the system's detection threshold (higher FP/lower FN; lower FP/higher FN; or a threshold-free "soft highlighting" [8] approach) while keeping its overall discriminability constant, we can explore how the learners in this task may respond differently to the information they receive in terms of their willingness to *complete* the task, their ability to *perform* the task when guidance is available, and the degree to which they *learn* to do the task independently. Further, we explore how much the learners' *trust* the system. In a follow-up study ($n = 27$), we also explored, for just the soft highlighting condition, whether participants use the confidence information conveyed by the highlight intensity to modulate their decisions about which objects to move. Our paper is, to our knowledge, one of the first to explore how imperfect guidance from an AIEd system can affect learners' experiences.

2 Related Work

Imperfect AI & Trust: The past few years have seen growing interest in the ML, human-computer interaction, as well as the AIEd communities, in how humans trust AI and how they can work together cooperatively. Kocielnik et al. [9] investigated how the FP/FN trade-off affected users' perceptions of the machine's accuracy as well as their acceptance of a ML-based tool that detects scheduling requests from free-text emails. They found that users had more favorable impressions of the system when the detection threshold was adjusted to give a lower FN rate in exchange for a higher FP rate. Hsu et al. [5] trained an automatic auto-grader of students' short-answer responses about computer science topics; they then investigated students' perceptions of the accuracy and fairness of the system. Further, they explored how students' (mis-)understanding of how the system worked internally could affect the answers that students constructed and submitted. Finally, in a research study toward increasing learners' trust in AI-based feedback systems, Hossain et al. [4] studied how experts perceive the feedback generated from an automatic explainable hand-gesture feedback system for learners of American Sign Language.

Soft highlighting to convey the machine's confidence: In research within the intersection of data visualization, ML, and human-computer interaction, Kneusel et al. [8] explored whether human workers benefit more from a ML-based object detector on visual perception tasks, such as examining satellite imagery for specific objects, when they have access to the *confidence* of the machine's predictions. They devised a "soft highlighting" mechanism, whereby the machine's detections were colored with an intensity proportional to the confidence of the predictions, and found that human workers performed better on the task with "soft" highlights than with binary (hard) object predictions.

MRF training system: Our work builds on a prior workplace learning study by Kyriacou et al. [11], who created a MRF simulator to compare different training strategies for human workers who are learning to sort different objects on the conveyer belts. Similarly to our work, they examined how different accuracy characteristics of a machine learning assistant can affect learners' performance. In contrast to their work, we hold the overall discriminability of the assistant constant and manipulate the FP/FN trade-off in isolation. Moreover, we introduce a new soft-highlighting condition and explore how the learners may benefit from and mimic the behaviors exhibited by the assistant.

3 Experiment I

We conducted a randomized experiment in which participants learned to recognize and manipulate garbage objects of different types in a simulated MRF. The simulator we used was based on that of [11] but extended to support a soft highlighting condition (described above). During the training rounds, the participants received automated guidance from a ML assistant about whether each object is recyclable. This can potentially help them to learn to recognize the object types more quickly and thereby perform better on the task.

3.1 Participants

The participants ($n = 181$) in our study were adults (≥18 years) on Amazon Mechanical Turk who had earned a "Masters" qualification.

3.2 Experimental Conditions

At the outset of the experiment, each participant was randomly assigned to one of three different conditions: (1) FP=0/FN=0.11, i.e., the machine learning assistant will correctly highlight 89% of the non-recyclable objects and miss 11%; it will never falsely flag a recyclable object as non-recyclable; (2) FP=0.11/FN=0; the assistant never misses a non-recyclable object but occasionally falsely flags recyclable object as non-recyclable; and (3) Soft highlighting: every single object on the conveyer belt is highlighted, but the intensity of the highlight corresponds to the confidence of the machine's prediction. Importantly, the overall *discriminability* of the object detector used for automated guidance was held constant across all three conditions. The particular FP/FN values were chosen based on previous work [11] with this simulator, which suggested that the discriminability of the ML assistant needs to very high in order to be useful.

3.3 MRF Simulator

The simulator contains three conveyer belts, each of which has a different speed and moves a never-ending stream of objects from left to right. In the task, the goal is to remove only and all the non-recyclable items (syringes, broken glass, coat-hangers, batteries, etc.) from the conveyer belts into a trash-can (a green bucket shown at the top of the screen).

Machine learning assistant: During the training rounds (see Section 3.4), the learner is aided by a simulated machine learning assistant that detects (with imperfect accuracy) whether each object is recyclable or not. Across all three experimental conditions, the discriminability of the assistant is held constant at 0.945. Specifically, we quantified *discriminability* as the widely used Area Under the Receiver Operating Characteristics Curve (AUROC) metric. The AUROC is equivalent to the probability, in a 2-alternative forced-choice task, that the detector can correctly recognize the non-recyclable object from a random pair of objects (one recyclable, one not). In the two experimental conditions corresponding to FP=0/FN=0.11 and FP=0.11/FN=0, a random number generator determines whether each object is highlighted, depending on the FP and FN rates and on whether the object is recyclable. In the third condition corresponding to soft highlighting, no threshold is used, and instead the intensity of the red color is proportional to a real-valued confidence score C for that object in the interval [0,1]. In particular, if the object is recyclable, then C is drawn from a Beta distribution, i.e., $C \sim \mathrm{Beta}(2, 4.675)$; if it is non-recyclable, $C \sim \mathrm{Beta}(4.675, 2)$. Under this generative process, non-recyclable objects tend to have higher C values and thus are highlighted with a brighter red box. Moreover, the parameters for the probability distributions were chosen so that the discriminability of the assistant is exactly 0.945 (i.e., $P(C_a > C_b) = 0.945$ where C_a and C_b are the confidence values of a random non-recyclable and recyclable object, respectively).

Scoring: In the pre-test, post-test, and each training round, the user's score starts at 100 and decreases by 1 whenever (a) the user misses a non-recyclable object and it moves off the screen to the right; or (b) the user incorrectly moves a recyclable item from the belt into the trash-can. Hence, the maximum score in each round is 100, and it can decrease to 0 (or even below) if the participant makes many mistakes. The training and testing rounds were essentially the same, except that (a) the pre-test was shorter (1 min) than the other rounds (3 min), and (b) the ML assistant was available only during the training rounds.

Good performance in the task requires quick and accurate visual recognition of the objects and manipulation of the objects (using the mouse) from the belt to the trash-can. There is also some strategy that can be beneficial to task performance, e.g., prioritizing one belt over another due to the different speeds, moving objects slightly so as to reduce occlusion, etc. Participants received $1.50 for completing the task; to incentivize good performance, they could also earn a reward that increased linearly with their scores up to a maximum of $2.50.

3.4 Procedures

The experiment consisted of (1) a study overview; (2) task instructions; (3) pre-test (1 min), during which no help from the ML assistant was provided; (4) two rounds of training (3 min each), during which the machine learning assistant provided automated guidance; (4) post-test (3 min), during which no help from the machine learning assistant was provided; and (5) questionnaire about trust in the system. All in all, the study takes about 12-15 min to complete.

3.5 Task instructions

The instructions to the learners describe and show examples of the objects that they should move from the belts into the green trash bucket. In addition, the instructions explain that the machine learning assistant automatically flags certain objects as likely non-recyclable using a red rectangle. To set the users' expectations and promote effective usage of the assistant, the instructions explain: "*The object detector is imperfect. You may find it useful to you as you* *play the game as a way of focusing your attention on the correct objects, but you should not rely on it completely. In particular, sometimes the detector will miss non-recyclable objects and therefore not flag it. It may also make 'false alarms' and highlight a recyclable object.*"

3.6 Trust Questionnaire

To assess participants' trust in the AI system after completing the task, we asked them to complete a validated questionaire on trust in technology [6], which consists of 12 Likert items (e.g., strongly disagree to strongly agree with: "the system is deceptive") about the users' impressions of the following attributes: *deceptive*, *underhanded*, *suspicious*, *wary*, *harm*, *confident*, *security*, *integrity*, *dependable*, *reliable*, *trust*, and *familiar* (some of these items are negatively scored). The maximum score (most trusting) is 84 and minimum score is 12.

3.7 Measures

The simulator automatically records the participant's score after every pre-test, post-test, and training round. The scores on the pre-test tend to be higher than in the other phases of the experiment since the pre-test is shorter (only 1 min), and hence the participants have less time to make mistakes. The simulator also records the responses to the trust questionnaire.

3.8 Results

A total of $n = 181$ total participants completed the task; in the FP=0/FN=0.11, FP=0.11/FN=0, and soft-highlighting conditions, their numbers were 56, 60, and 65, respectively. These numbers are not stat. sig. different ($\chi^2(2) = 0.67403$, $p = 0.7139$) from a uniform distribution, and hence we do not conclude that any condition caused participants to drop out more often than another.

Fig. 2 (**left**) shows the mean score (along with its standard error) across the three rounds separately for each experimental condition. The difference in mean pre-test scores across condition was not statistically signficant ($F(2) = 0.906, p = 0.406$). As a general trend, participants in the FP=0/FN=0.11 condition tended to receive higher scores, followed by those in the soft highlighting condition, and then the FP=0.11/FN=0 condition.

Fig. 2 (**right**) shows the mean trust score (and s.e.) for each condition.

Task Performance with Automated Guidance We used a linear mixed-effect model (repeated-measures design with subject id as the random effect and pre-test score as a covariate) to analyze the scores during the training rounds 1 & 2 when workers had access to the assistant. We found that: (1) scores were stat. sig. higher when participants received guidance from the assistant compared to on the post-test, when they did not receive guidance ($t(360) = 3.325$, $p < 0.001$); and (2) participants in the FP=0.11/FN=0 condition received

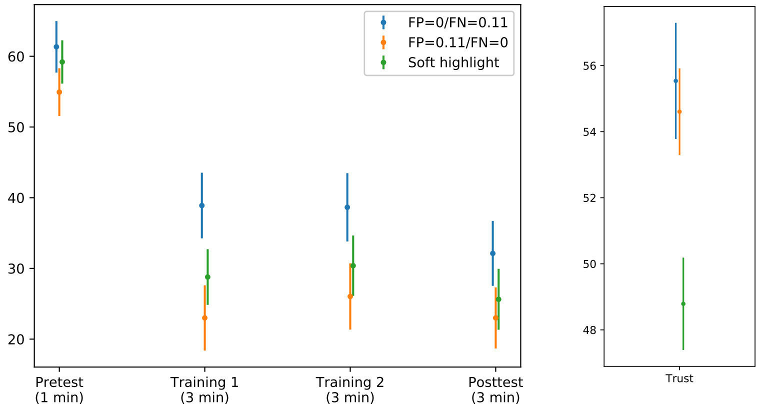

Fig. 2. Left: Mean scores (and error bars for standard errors, shifted slightly for readability) for the different phases of the study, split by condition. Note that, as the pre-test is only 1 min and participants have less time to make mistakes, the scores tend to be higher. **Right**: Mean (and s.e.) trust questionnaire results, split by condition.

stat. sig. lower scores compared to those in the other conditions ($t(178) = -1.992$, $p = 0.048$). The soft-highlighting condition was also associated with a lower score, but the effect was not stat. sig. ($t(178) = -1.341$, $p = 0.1816$).

In terms of participants' actions, we found a stat. sig. difference in the number of non-recyclable objects moved into the trash-can ($F(2) = 6.782$, $p = 0.001$), but not in the number of recyclable objects ($F(2) = 1.301$, $p = 0.275$), that depends on the experimental condition. In particular, participants in the FP=0.11/FN=0 condition moved more non-recyclable objects (an average of 67.3 combined over rounds 2 & 3) compared to those in the FP=0/FN=0.11 (average of 59.9) or soft highlighting (58.1) conditions.

Learning to Perform the Task Without Guidance We used ANOVA (with pre-test score as a covariate) to assess whether participants' scores on the post-test differed with their experimental conditions. The difference was not stat. sig. ($F(2) = 0.62$, $p = 0.539$). In terms of participants' behaviors, we found a borderline effect in the number of trashed non-recyclable objects ($F(2) = 2.734$, $p = 0.0677$): the average number of such objects was higher for participants in the FP=0.11/FN=0 condition (32.4 objects) and soft highlighting (31.7) conditions compared to those in the FP=0/FN=0.11 (27.6) condition.

Trust With an ANOVA, we found that the trust score calculated from the questionnaire responses was stat. sig. related to the experimental condition $F(2) = 4.173$, $p = 0.0177$). In particular, the soft-highlighting condition was stat. sig. negatively associated with trust ($t = -2.778$, $p = 0.006$).

3.9 Discussion

Workers benefited from the guidance of the ML assistant, as evidenced by their higher scores during training rounds compared to the post-test. The fact that, during the training rounds, the mean scores in the FP=0/FN=0.11 condition were less than 82 (which, based on how the simulator was constructed, is the expected value if the participants perfectly followed the assistant's guidance) could indicate that participants were reluctant to accept all the assistant's advice, or that doing so was too difficult (possibly due to the fast pace of the simulation). Also, the fact that mean scores in the FP=0.11/FN=0 condition during training rounds were greater than 0 (which, due to the higher number of recyclable compared to non-recyclable objects, was the expected value for perfectly following the assistant in this condition), but were lower during the post-test than during training rounds, suggests that participants in this condition did use the MLA, but they did not follow its guidance blindly. In our experiment, a higher False Positive rate of the assistant was associated with lower performance. This contrasts with the result of [9], thus suggesting that the optimal FP/FN trade-off is likely task-dependent based on the cost of each kind of mistake.

There was no stat. sig. impact of condition on *learning*. However, there was a borderline effect of condition on the number of non-recyclable objects that were trashed during the post-test. This result is encouraging, especially given the short study duration, and suggests that the training effect of the MLA may linger in users' behaviors even after the MLA has been removed.

The soft highlighting condition was advantageous neither in terms of task performance during training rounds nor in terms of learning. This contrasts with [8], who found that participants could perceive more effectively when provided with soft highlights rather than hard detections. Learners indicated that they trusted the soft highlighting condition the least among the three assistant types.

4 Experiment II

Soft highlighting gives the learner information about the machine's confidence in its guidance. Does the learner's likelihood of picking an object and moving it to the trash-can increase when the object is highlighted with higher confidence? To explore this, we conducted a follow-up experiment ($n = 27$ participants on Mechanical Turk) consisting of just the soft highlighting condition. We assessed whether the probability distribution $P(\text{pick} \mid c, \text{isRecyclable})$ is equal to $P(\text{pick} \mid \text{isRecyclable})$, where pick indicates whether or not the participant picked the object; C is the confidence score of the object; and isRecyclable is either recyclable or non-recyclable. If the distributions are equal (for both values of isRecyclable), then by definition, the participants' decisions to pick or not pick the objects is independent of the confidence scores. From Bayes' rule, we obtain:

$$P(\text{pick} \mid c, \text{isRecyclable}) = \frac{P(c \mid \text{pick}, \text{isRecyclable})P(\text{pick} \mid \text{isRecyclable})}{P(c \mid \text{isRecyclable})}$$

Hence, for either value of isRecyclable, we can evaluate the hypothesis that the decision to pick/not pick is independent of c by testing whether $\frac{P(c \mid \text{pick,isRecyclable})}{P(c \mid \text{isRecyclable})}$ equals 1 for all values of C. The numerator can be estimated from the log data of participants' actions in the MRF simulator about the confidence scores of the objects they picked and whether they were recyclable or not; the denominator is given by the Beta distributions as described in Section 3.3.

4.1 Results

To estimate $P(c \mid \text{pick}, \text{isRecyclable})$ for isRecyclable $=$ false and pick $=$ true, we computed a histogram over the confidence scores of the picked objects using a bin width of 0.1 over the interval [0,1]. We were primarily interested in whether there was a population-level association between picking behaviors and the confidence scores, and hence we computed a histogram over the events of all $n = 27$ participants, summed over both training rounds, separately for the recyclable and non-recyclable objects. We then computed the corresponding histogram based on the generative process of the confidence scores themselves using the appropriate Beta distribution (see Section 3.3). A χ^2-test showed a stat. sig. difference between the distributions, both for isRecyclable $=$ true ($\chi^2(9) = 28.30$, $p < 0.001$) and for isRecyclable $=$ false ($\chi^2(9) = 38.33$, $p < 0.001$). Also, Fig. 3 shows the probability ratio $\frac{P(c \mid \text{pick,isRecyclable})}{P(c \mid \text{isRecyclable})}$ both for the recyclable (**left**) and non-recyclable (**right**) cases. Despite some outliers that are partly due to small numbers of observations in the outer-most histogram bins, participants' likelihood of picking an object tends to increase with larger C.

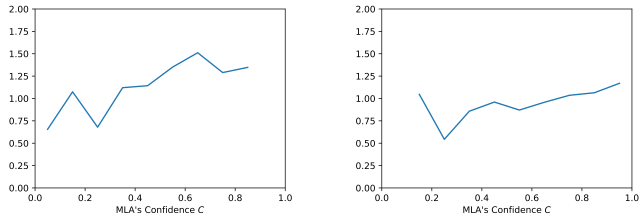

Fig. 3. Left: Relative probability increase of picking up and disposing of a *recyclable* object (**left**) or a *non-recyclable* object (**right**), given the assistant's confidence C.

4.2 Discussion

We found stat. sig. support for the hypothesis that learners do take into account the assistant's fine-grained confidence, as conveyed by the soft highlights, when deciding whether or not to pick or an object. This effect was observed for both the recyclable and the non-recyclable objects, indicating that the influence of the confidence scores goes beyond simple distinctions in the object's recyclability. From the perspective of the designer of an AIEd system to provide better guidance and feedback to the user, this result is encouraging as it shows an example of when the confidence information is perceived by and acted on by the learner.

5 Conclusions

We conducted two experiments ($n = 181$, $n = 27$) on a workplace learning task in which participants learned to recognize and manipulate objects in a simulated MRF. We varied the FP/FN trade-off of the ML assistant that automatically, but imperfectly, highlighted objects that were deemed to be non-recyclable and thus should be moved to the trash-can. Moreover, we explored whether and how soft highlighting, rather than binary thresholding, of the guidance may impact participants' performance, learning, and trust. Our results indicate that (1) the FP/FN trade-off can affect learners' cooperative task performance with the assistant; (2) there was tentative evidence that the assigned experimental condition affected learners' behaviors – in terms of how many non-recyclable objects were trashed – even after the assistant was removed; and (3) soft highlighting impacted learners' trust as well as their behaviors in the task.

Limitations: The experiments we conducted were short (about 12-15 minutes). Moreover, the learning task we examined was about perception and motor control in a fast-paced environment. It is possible that different trends would be found for different learning tasks (e.g., cognitive rather than perceptual).

5.1 Future Research

Better framing of the AI guidance/feedback: Our results suggest that learners trust the extra confidence information provided by the soft highlighting condition less than a simpler binary guidance mechanism. In order to be both more trustworthy and effective, the learners might need more thorough framing, and possibly even some form of "pre-training", about how the system works, what its intentions are, and how they ought to make use of it.

Explaining the optimal FP/FN trade-off: Given the contrasting results of our study compared to [9], it would be interesting to change the scoring system so that different mistakes (FP versus FN) resulted in different penalties, and to investigate whether this affected the optimal FP/FN trade-off.

Acknowledgements This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL #2019805. The opinions expressed are those of the authors and do not represent views of the NSF. This research was also supported by NSF award #1928506 (FW-HTF-RL).

References

1. Ahuja, K., Kim, D., Xhakaj, F., Varga, V., Xie, A., Zhang, S., Townsend, J.E., Harrison, C., Ogan, A., Agarwal, Y.: Edusense: Practical classroom sensing at

scale. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **3**(3), 1–26 (2019)

2. Akcay, S., Breckon, T.: Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging. Pattern Recognition **122** (2022)
3. Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al.: Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)
4. Hossain, S., Kamzin, A., Amperayani, V.N.S.A., Paudyal, P., Banerjee, A., Gupta, S.K.: Engendering trust in automated feedback: A two step comparison of feedbacks in gesture based learning. In: International Conference on AIED (2021)
5. Hsu, S., Li, T.W., Zhang, Z., Fowler, M., Zilles, C., Karahalios, K.: Attitudes surrounding an imperfect ai autograder. In: CHI Conference on Human Factors in Computing Systems (2021)
6. Jian, J.Y., Bisantz, A.M., Drury, C.G.: Foundations for an empirically determined scale of trust in automated systems. Intl. journal of cognitive ergonomics **4** (2000)
7. Karumbaiah, S., Lizarralde, R., Allessio, D., Woolf, B., Arroyo, I., Wixon, N.: Addressing student behavior and affect with empathy and growth mindset. International Educational Data Mining Society (2017)
8. Kneusel, R.T., Mozer, M.C.: Improving human-machine cooperative visual search with soft highlighting. ACM Transactions on Applied Perception (TAP) **15** (2017)
9. Kocielnik, R., Amershi, S., Bennett, P.N.: Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In: CHI Conference on Human Factors in Computing Systems (2019)
10. Kulik, J.A., Fletcher, J.: Effectiveness of intelligent tutoring systems: a meta-analytic review. Review of educational research **86**(1), 42–78 (2016)
11. Kyriacou, H., Ramakrishnan, A., Whitehill, J.: Learning to work in a materials recovery facility: Can humans and machines learn from each other? In: Learning Analytics and Knowledge (LAK) Conference. pp. 456–461 (2021)
12. Landrum, A.R., Eaves Jr, B.S., Shafto, P.: Learning to trust and trusting to learn: A theoretical framework. Trends in Cognitive Sciences **19**(3), 109–111 (2015)
13. Nie, A., Brunskill, E., Piech, C.: Play to grade: Testing coding games as classifying markov decision process. Neural Information Processing Systems **34** (2021)
14. Randhawa, G.K., Jackson, M.: The role of artificial intelligence in learning and professional development for healthcare professionals. In: Healthcare management forum. vol. 33, pp. 19–24. SAGE Publications Sage CA: Los Angeles, CA (2020)
15. Schlotterbeck, D., Uribe, P., Jiménez, A., Araya, R., van der Molen Moris, J., Caballero, D.: Tarta: Teacher activity recognizer from transcriptions and audio. In: International Conference on AIED (2021)
16. Schmidt, R.A., Wrisberg, C.A.: Motor learning and performance: A situation-based learning approach. Human kinetics (2008)
17. Zylich, B., Whitehill, J.: Noise-robust key-phrase detectors for automated classroom feedback. In: Intl. Conf. on Acoustics, Speech and Signal Processing (2020)