When Teaching A Robot, People Employ Different Feedback Strategies: Some Are More Effective Than Others

Nicholas C. Georgiou[1] **(nicholas.georgiou@yale.edu), Shuangge Wang**[1]**, Joel Banks**[1]**, Kate Candon**[1]**, Dražen Brščić**[2]**, Brian Scassellati**[1]

[1]Department of Computer Science, Yale University, *New Haven, CT, USA*

[2]Department of Social Informatics, Kyoto University, *Kyoto, Japan*

Abstract

To investigate the effects of human feedback strategies on machine learning (ML), we collected data from participants ($N = 36$) as they evaluated a robot with numeric feedback during a card game. We found that participants employed different partial credit feedback strategies for robot failures during the task (i.e., participants varied in how they scored the same robot failure actions). We then used the feedback from each participant to generate extrapolated feedback strategies. In simulations, we found that training a supervised ML model with these different extrapolated feedback strategies influenced how well the model was able to learn the task. Models trained with labels from some reasonable strategies significantly outperformed models trained with labels from other reasonable strategies. Participants' familiarity with ML, artificial intelligence, and the task did not significantly affect how well their extrapolated feedback strategy trained the model. These findings have implications for transferring learning algorithms into the real world.

Keywords: human teaching; machine learning; learning from human feedback; human-robot interaction.

Introduction

Numerical evaluations that assess the quality of an action or behavior have been utilized as a user-friendly modality through which people can teach a machine learner (Thomaz, Breazeal, et al., 2006; Knox & Stone, 2009; Chernova & Thomaz, 2014). When users are tasked with providing numeric feedback for evaluation, they must determine a feedback strategy, i.e., *what* and *how* different factors weigh into the score that they provide each time. Feedback strategies will likely vary between teachers, as people have different expectations, experiences, prior knowledge, and mental models related to teaching.

Nonetheless, as long as a teacher is providing consistent feedback that is aligned with the learning goal, what we refer to as a *reasonable* feedback strategy, one would expect that such a strategy would be sufficient to teach the learner the task. However, the downstream effects of different reasonable feedback strategies that people use are not obvious. It is unknown how different strategies will impact how effectively and efficiently a learning algorithm learns a task. This question is particularly important to study as machines are put into situations where they will be expected to learn directly from human feedback from different teachers.

To explore these ideas, we conducted an in-person user study in which participants were tasked with providing feedback to a robot through numerical evaluations. The experimental setup can be seen in Fig. 1. We first explored the feedback different participants provided throughout the task. Using the collected data, we fit a linear regression to each participant's collection of feedback to highlight the different feedback strategies that people employed. We considered these linear regressions to be *extrapolated feedback strategies* because they can provide an estimate for how participants would score unseen actions. To then explore the downstream effects of different reasonable strategies, we ran simulations in which we analyzed how effective these different extrapolated strategies were at training a supervised machine learning algorithm. Lastly, we discuss what our results show about the different feedback strategies that humans used to teach a robot a particular task, as well as potential real-world implications for our findings.

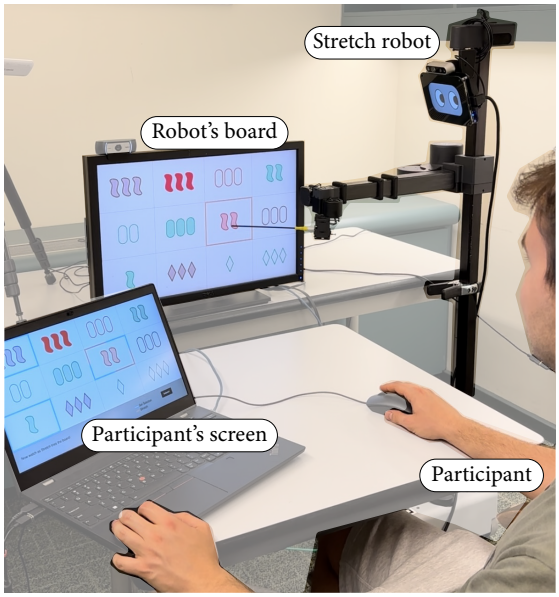

Figure 1: Experimental setup.

Related Work

Evaluative feedback – values provided by a teacher to assess a learner's behaviors and actions – can be used as a reward in interactive reinforcement learning or as labeled training data in interactive ML (Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos, Bobes-Bascarán, & Fernández-Leal, 2022). To

be able to effectively incorporate evaluative feedback into ML, understanding how human teachers provide feedback is critical (Thomaz & Breazeal, 2008). Exploring how human teaching differs from how machines expect to learn has led to the development of algorithms that improve robot learning, such as TAMER (Knox & Stone, 2008), COACH (MacGlashan et al., 2017), and REPaiR (Kessler Faulkner, Schaertl Short, & Thomaz, 2020). However, these works focus on the incorporation of human reward into the ML loop (i.e., differences in human-provided and environment-provided reward), rather than on the effects of differences in feedback strategies between human teachers.

Recent work in human-robot interaction (HRI) has looked into how people's feedback changes over time (Candon et al., 2024), how robot competency affects feedback quality (Wang, Wang, Goncharova, Scassellati, & Fitzgerald, 2025), how binary feedback differs from scalar feedback between users (Yu, Aronson, Allen, & Short, 2023), and how feedback can be broken into multiple dimensions (Huang, Aronson, & Short, 2024). Despite these advancements, insights into how people's underlying feedback strategies differ and how these strategies affect learning are understudied.

Prior work has studied how and when people employ different training strategies, defined by how much they focus on allocating penalty vs. reward (Loftin et al., 2016, 2014). Other work focuses on modeling humans' preference-based feedback (Lee, Smith, Dragan, & Abbeel, 2021) and evaluating how synthetic and human labelers differ when training preference-based reinforcement learning models (Metcalf, Sarabia, Fedzechkina, & Theobald, 2024). Much of the work in this area studies how to learn varying preferences among people (i.e., different end goal). Instead, we are studying how to learn the same task (i.e., same end goal) via different approaches, as well as effects of these approaches on learning.

Based on the gaps identified in prior work, our guiding research questions were:

RQ1: When given the same teaching objective with the same end goal, do the feedback strategies that people employ to teach a robot vary?

RQ2: Do the feedback strategies extrapolated from different human teachers impact how effectively an ML algorithm can learn to complete the task?

Methodology - Data Collection

To investigate our research questions, we asked participants to provide feedback to a robot performing a task. Importantly, we required participants to provide numeric evaluations, but did not instruct them on the strategy they should use to determine their scores. The robot did not learn during the interaction, but the data was used to train models in simulation (see Simulations sections). The following subsections describe the details of our data collection methodology.

Teaching Task - Set

We chose Set, a card selection game, as our teaching task. The main objective of Set is to select three cards that meet certain requirements, classifying them as a *set*, out of a board of 12 cards. Each card has four feature dimensions: number, shape, color, fill. Each feature dimension can have one of three distinct values (see Fig. 2 for examples). To be considered a set, for each of the four dimensions, the three cards must either have the same value or three different values.

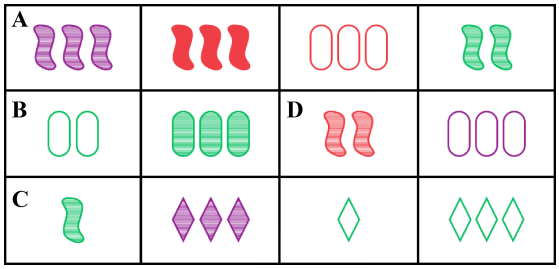

Figure 2: Set board example. Cards A, B, and C are not a set. Despite each having a different number, they break the requirements of a set for three dimensions (shape: 2 squiggle, 1 oval; color: 2 green, 1 purple; fill: 2 striped, 1 outlined). The action of selecting cards A, B, and C can be represented as $[c = 2, f = 1]$: two cards are from the solution, one of the dimensions (number) meets the set requirements. Cards A, C, and D are a set because they have the same shape and fill and have different colors and numbers. Cards A, C, and D can be represented as $[c = 3, f = 4]$.

For our version of Set, we use three row by four column boards, where each board includes only one set, that is, only one solution. We represent the board space as a $3 \times 4 \times 3^4$ dimensional space. There are $\binom{12}{3} = 220$ three-card selections for a Set board with twelve cards, with only one of these selections constituting a success (the selection of the set).

Set is very controllable — it is straightforward to verify solutions and to generate unique states that satisfy various constraints. The game can be explained quickly, yet is cognitively challenging. It is simple enough that there are a limited number of likely feedback strategies, but complex enough that there is not one strategy more obvious than others. Set was carefully selected to help us explore the potential effects of different strategies on learning.

Based on our personal experience playing Set and on a pilot study, we hypothesized that there were two main components that people would focus on when determining what feedback score to provide to the robot: 1) the number of cards, c, from the board solution included in the three-card selection, and 2) the number of feature dimensions in the selection, f, that fulfill the Set rules criteria. A three-card selection with a high c value indicates that the selection is close to the specific solution for a current board, whereas a high f value indicates that a selection is close to fulfilling the criteria for feature dimensions with respect to the rules of the game. Scoring proportionally to c and/or f are all reasonable strategies, and it is not obvious which strategy people would use, nor which is more effective in providing feedback for machine learning.

Importantly, in Set, c and f can vary independently. A

three-card selection can include cards from the board solution (high c), but have no feature dimensions achieving the rules of Set (low f), or vice-versa. Alternatively, a three-card selection could be high in both c and f (e.g., two cards from solution, third card breaks Set rules on only one feature dimension), or could be low in both c and f (e.g., three cards not in solution and breaking rules on all four dimensions).

Therefore, we have $c \in \mathbf{C} = \{0,1,2,3\}$ and $f \in \mathbf{F} = \{0,1,2,3,4\}$. By definition, $c = 3$ if and only if $f = 4$, because if the three cards from the solution are selected, then they meet the Set requirements on all four dimensions. Thus, $c = 3$ and $f = 4$ is the success action for the robot. There are $|\{\mathbf{C}\backslash\{3\}\} \times \{\mathbf{F}\backslash\{4\}\}| = 12$ failure variations of three-card selections. From this point forward, a three-card selection from the robot will be represented in terms of c and f. See Fig. 2 and Fig. 3 for examples.

Participants

We recruited 45 participants for data collection. Nine participants were excluded due to technical issues. The final 36 participants reported their age ($M = 25.83$ years, $SD = 4.65$ years) and gender (15 male, 21 female). 27 participants were undergraduate or graduate students. When self-reporting familiarity with ML, 8 participants reported that they knew it well, 4 knew a fair amount, 14 knew a little, and 10 had heard of it. For familiarity with AI, 6 reported that they knew it well, 7 knew a fair amount, 17 knew a little, and 6 had heard of it. For prior experience with Set, 13 reported that they had played before, 22 had not, and 1 was unsure.

All participants provided consent for the study and audio-visual recording. Participants received $15 as compensation. The study was reviewed and approved by an Institutional Review Board. The study took approximately one hour.

Robot

We used a Hello Robot Stretch 2 (Kemp, Edsinger, Clever, & Matulevich, 2022) in our Set task. Stretch is a lightweight mobile robot with a manipulator. We replaced the default end effector gripper with a simple pointer to suit our task.

To make Stretch appear more social, we added a small HDMI monitor to its head that displayed two eyes, based on the Shutter robot (Thompson, Narcomey, Lew, & Vázquez, 2024), as seen in Fig. 1. With the help of this monitor, the Stretch performed anthropomorphic behaviors, such as blinking, nodding as a greeting or as a response to the participant's feedback, peering towards the participant while selecting a card, expressing happiness or sadness, and turning/tilting its head to and from the participant and Set board.

Procedure

In our data collection study, participants provided numeric feedback to the Stretch robot as it was playing Set. Participants provided a score to the robot after each three-card selection on a board. All data collection sessions were conducted by the same experimenter, who used a script for consistency. The procedure consisted of five phases.

Pre-Interaction Phase: Participants completed a consent form and demographics survey outside of the study room.

Tutorial Phase: The participant was brought into the study room by the experimenter and completed a tutorial explaining the rules of finding a set on a laptop in the room. After the participant completed the tutorial, the experimenter explained how to play Set. For each round, participants were shown a 3×4 card board on the laptop screen (participant's screen in Fig. 1), which contained only one solution. When participants selected cards on the laptop screen via mouse clicks, they were highlighted in blue. After they submitted their three-card selection, the system confirmed whether or not the participant found the set. If the participant did not find the set within one minute, the solution was revealed to the participant. After the participant found the set or their time expired, they rated the perceived difficulty of the board on a scale from 1 (very easy) to 10 (very hard).

Robot Introduction Phase: The experimenter explained to the participant that their goal was to help the robot learn Set by providing feedback as it tried to solve each board. Participants were told to provide feedback after each three-card selection made by the robot. They were asked to provide a score on a scale from 1 to 10, with one being very poor and 10 being excellent, via the laptop keyboard. They were also told that they should feel free to speak with Stretch throughout the study. They were told that the robot would not adjust its behavior based on their feedback during the session, but that the feedback would be used to help the robot learn in the future. After the interaction protocol was explained, the participant pressed a button on the laptop to wake up the robot. The robot opened its eyes and turned its head to the experimenter. The experimenter greeted the robot by saying "Hello Stretch". In response, the robot displayed happy eyes and nodded its head. The robot then turned its head to the participant, displayed happy eyes and nodded its head. The experimenter told the participant to press a button to begin the feedback interaction and left the room.

Robot Interaction Phase: The participant and Stretch then went through eight different Set boards. For each board, the participant first submitted a card selection, followed by a difficulty score. The correct solution was then highlighted in blue on the participant's screen. Next, Stretch selected three cards on its own board, which was identical to the participant's, but did not show the true solution. The robot's board was displayed on the monitor in front of the participant and to the side of Stretch (see Fig. 1). The robot's arm was used to select cards on the board. When pointing to a card, the robot's head faced the board, the robot's arm moved to the desired card, the robot's face rotated and its eyes peered towards the participant, and its wrist rotated such that its pointer appeared to touch the card. At the end of this motion, the representative card on both the robot's and participant's boards were highlighted in orange. After Stretch selected three cards, it turned back to the user and waited for a ring or buzzer noise that indicated whether its selection was correct or incorrect. The robot

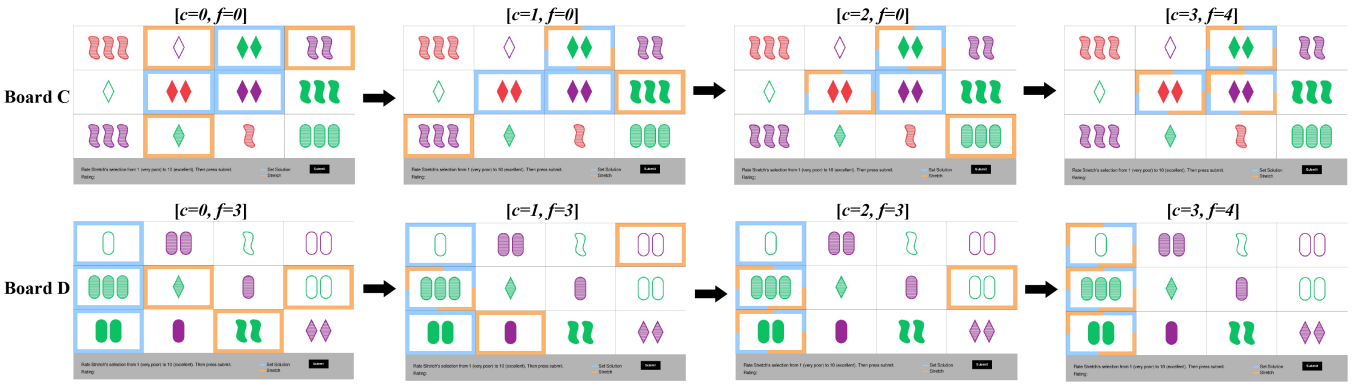

Figure 3: Robot action sequences for Boards C and D. Sequences progress left to right. The cards in the board solution (i.e., the set) are outlined in blue. For each action, the robot selects three cards, which are outlined in orange, or in both orange and blue if the card is also in the solution. Each robot action is represented by c, how many cards from the solution it contains, and f, how many of the feature dimensions satisfy the Set rules. Each action's $[c, f]$ is shown above the board. Board C shows the robot getting closer to the board solution in terms of c, but consistently doing poorly in terms of f. Board D shows the robot also getting closer to the board solution in terms of c, but consistently doing well in terms of f.

reacted with happy or sad eyes depending on the noise. The cards that the robot selected were predefined (see Robot Action Sequences section). The participant provided a feedback score to Stretch for each three-card selection that the robot made. The robot nodded to the participant to acknowledge the receipt of the feedback. Stretch attempted the board until it finally found the set. Then, the participant moved on to the next board. Stretch waited for the participant to complete their turn, and then it attempted the new board that the participant had just attempted. After all boards were completed, Stretch closed its eyes and put its head down, indicating that it had returned to sleep.

Post-Interaction Phase: The experimenter used semi-structured questions to ask about the participant's approach to determine feedback scores for the robot. Then, the participant completed survey questions about their interaction.

Robot Action Sequences

For each Set board, there was a predefined sequence of three-card selections for the robot, which we will call a *robot action sequence*. Except for the first tutorial board, the robot showed different types of improvements before finally arriving at the solution. The types of improvements were driven by c and f such that at least one of c or f increased throughout the sequence. Each participant saw the same eight boards, each with a fixed robot action sequence (see Table 1). After Board 1, Boards A-G were presented in randomized order. See Fig. 3 for examples.

Board	$[c, f]$	Robot Action Sequence $[c, f]$
1	$[0, 0]$	[0,0] → [0,0] → [0,0] → [3,4]
A	$[0, \uparrow]$	[0,0] → [0,1] → [0,2] → [0,3] → [3,4]
B	$[2, \uparrow]$	[2,0] → [2,1] → [2,2] → [2,3] → [3,4]
C	$[\uparrow, 0]$	[0,0] → [1,0] → [2,0] → [3,4]
D	$[\uparrow, 3]$	[0,3] → [1,3] → [2,3] → [3,4]
E	$[\uparrow, \uparrow]$	[0,1] → [1,2] → [2,3] → [3,4]
F	$[\uparrow, \downarrow]$	[0,3] → [1,2] → [2,1] → [3,4]
G	$[\downarrow, \uparrow]$	[2,1] → [1,2] → [0,3] → [3,4]

Table 1: Robot Action Sequences. The second column indicates the type of improvement exhibited by the robot during the sequence. A number indicates that the c or f value stayed consistent until the set was found, an up arrow indicates that the value increased, and a down arrow indicates that the value decreased. The third column shows the detailed sequence of robot actions, represented in the $[c, f]$ space.

Results - Data Collection

In this section, we present results on the numeric feedback provided by participants during their interaction with Stretch.

Feedback Score Statistics

Overall, participants provided a mean score of 5.53 (SD = 3.18) for the robot's actions. Participants used most of the unique numbers ($M = 7.33, SD = 1.55$, min = 4 unique numbers, max = 10 unique numbers) on the feedback scale. On successes, participants gave the robot perfect scores most of the time ($M = 9.87$, $SD = 0.58$). On failures, participants provided a variety of scores with a mean of $M = 4.20$ (SD = 2.36). The distribution of participants' feedback scores, broken down by success and failure, are shown in Fig. 4.

Difficulty Ratings

Participants provided an average difficulty score of 5.92 ($SD = 2.56$) across all boards. Table 2 shows the difficulty ratings and feedback scores, broken down by how long it took participants to complete a board (or not). Pearson's correlation showed a strong positive correlation between board completion time (incomplete = 60s) and difficulty rating ($r = 0.72, p < 0.001$), but no correlation between board completion time and feedback score provided to the

Figure 4: Distribution of participants' feedback scores.

Duration	Boards (N)	Difficulty Rating	Robot Actions (N)	Feedback Score
$< 30s$	71	3.07 ± 1.53	302	5.43 ± 3.31
$\geq 30s, < 60s$	44	5.27 ± 1.69	194	5.59 ± 3.12
incomplete	173	7.35 ± 2.00	728	5.55 ± 3.15

Table 2: Mean $\pm$ SD of self-reported difficulty ratings and feedback scores provided to the robot, grouped by how long it took each participant to find the solution on each board.

robot ($r = 0.03, p = 0.37$) or between difficulty rating and feedback score ($r = 0.01, p = 0.65$).

Participant Feedback Strategies

To examine the feedback strategies that participants used to determine feedback scores, we fit a linear regression for each participant that predicted the feedback score as a dependent variable, with c and f as independent variables (β_c and β_f are the learned coefficients). These participant models approximated feedback scores well with an average $R^2 = 0.84$ ($SD = 0.09$, min = 0.61, max = 0.98). Every participant model was statistically significant ($p < 0.001$), suggesting that these regressions reasonably align with what participants considered when determining their scores. Each participant's β_c and β_f can be seen in Fig. 5. This plot demonstrates that participants weighed factors differently when determining what score to provide as feedback to the robot. Additionally, Fig. 5 illustrates that there is no apparent pattern with respect to self-reported familiarity with ML. There was also no observed pattern in plots with familiarity with AI and Set.

Further indicative of the idea that different people used different teaching strategies, not all participants' β_c and β_f were found to be significant ($p < 0.05$) in the linear regression predicting their feedback scores. More specifically, seven participants only valued β_f, nine only valued β_c, and 20 valued both. These different strategies were consistent with how participants explained their reasoning for choosing what score to provide for feedback in their post-interaction interviews.

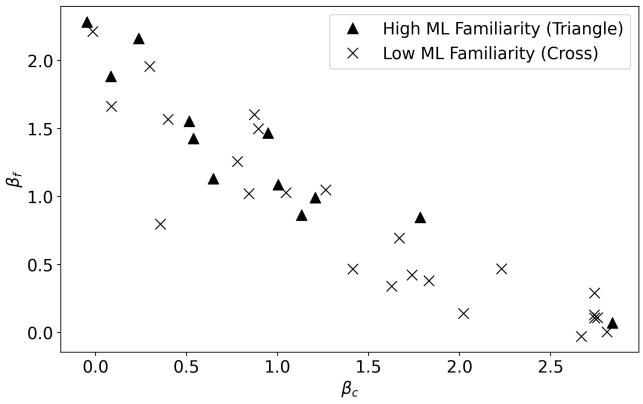

Figure 5: Participants' linear regression coefficients, labeled by self-reported familiarity with ML. High familiarity = knew it well or a fair amount; low familiarity = knew it a little or had heard of it.

Methodology - Simulations

We wanted to test how different strategies would affect an ML algorithm's ability to learn the task from raw input features. The linear regressions from the previous section use handcrafted input features, c and f. However, we wanted to train models to learn the task without this domain knowledge. As the input dimensionality increases, the amount of data required to train ML algorithms increases (Koutroumbas & Theodoridis, 2008), so it becomes intractable to collect sufficient data from real humans in an in-person user study. Therefore, we used the linear regressions as *extrapolated feedback strategies* (one per participant) to automatically predict feedback scores for given board-action pairs. We used these predicted feedback scores to train models to learn the task directly from a Set board.

For each extrapolated feedback strategy, we trained an individual multilayer perception (MLP[1]) to learn the task (i.e., picking the solution out of a given board). We used a reduced space of the Set board by having 8 cards to make data curation more tractable. We sampled a training dataset of 10 million boards, each containing only one *valid* set with random cards and robot action. To evaluate the model's accuracy on a board, we searched the action space for the action that generated the highest output using the MLP, and checked if that action matched the true solution.

Results - Simulations

In this section, we present the results from MLPs that were trained using different extrapolated feedback strategies.

[1]The state and action representations are converted into a one-hot encoding and concatenated before being fed into the MLP. The MLP outputs a scalar feedback score, which is evaluated using mean squared error (MSE) loss. The MLP consists of four hidden layers with 512, 256, 128, and 64 neurons, respectively, and employs ReLU activation (Fukushima, 1969). The model is trained for 15 epochs with a batch size of 512, and optimization is performed using the Adam optimizer with a learning rate of 0.001 (Kingma & Ba, 2014).

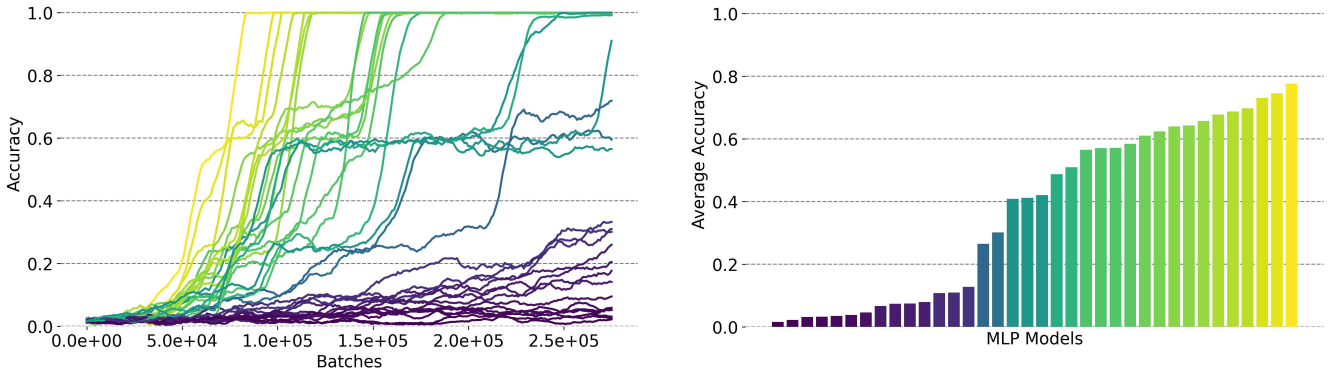

Figure 6: Smoothed training accuracy over time (left) and time-averaged accuracy (right) of each MLP model, trained using a different extrapolated feedback strategy, labeled by the same color in both figures.

There were very large differences in performance, depending on which extrapolated feedback strategy was used to train the MLP. The average post-training accuracy on 1000 unseen boards for the trained MLPs was 62.9% ($SD = 41\%$, max = 100%, min = 3.2%). The average accuracy throughout training (where accuracy was reported every 1000 batches) also greatly varied between extrapolated feedback strategies for training ($M = 37.3\%$, $SD = 27\%$, max = 77%, min = 1.6%) The smoothed training accuracies (window size=10 batches) reported every 1000 batches can be seen in Fig. 6, and bar graphs showing the average accuracy of each MLP can be seen in Fig. 6.

Familiarity with ML, AI, and Set

To explore if the participants' self-reported familiarities with ML, AI, or the task (i.e., Set) had a significant impact on how well their extrapolated feedback strategies trained the MLP, we mapped each MLP to the participant whose extrapolated feedback strategy was used to train it. Then, we grouped the MLPs' post-training accuracy results by the participant's familiarity (or not) with ML[2], with AI[2], or with prior experience (or not) playing Set. We performed Mann-Whitney U tests[3]. We did not find significant differences when comparing the accuracy of the MLPs separated by familiarity with ML ($p = 0.13, U = 99.5$), with AI ($p = 0.25, U = 114.5$), or prior Set experience ($p = 0.27, U = 111.0$).

Discussion

In this work, we investigated how people provided numeric feedback to evaluate a robot attempting a task. Importantly, all participants provided partial credit to the robot (i.e., they did not provide the same value for every failure action of the robot). This is important because machine learning algorithms prefer granular feedback to learn more effectively and efficiently. Additionally, the ways in which participants provided this partial credit differed among participants – they valued different factors when they provided a score for the robot's failure actions. In our simulations, the drastic differences in learning performance of some models compared to others suggest that not all partial-credit feedback strategies were effective, even if they are seemingly reasonable and consistent.

We found that the teacher's perceived difficulty of the task exhibited low correlation with the feedback score they provided. This suggests that their own performance on a given board did not significantly influence their feedback. It could be interesting to explore if (and how much) people's difficulty perceptions affect their feedback in other machine learning tasks, as well.

We realize that the specifics of Set do not necessarily generalize to other tasks. However, there are many complex, real-world tasks that allow for different interpretations of what a reasonable feedback strategy may be. When people provide evaluative feedback, the strategies they use to determine the score can vary. Set provided us with a simple, controlled way to examine this. Our study highlights the following phenomenon: different teachers may employ consistent, reasonable feedback strategies that have substantially different effects on how well an ML algorithm learns.

This phenomenon has several real-world implications. Most importantly, when a specific learning algorithm is chosen to learn a task from numeric evaluations, the strategy employed by the teacher is critical. It should not be assumed that just because a teacher's feedback is consistent and seemingly coherent, the strategy will result in effective learning. We found that even if a participant was familiar with ML, AI, or the task, this prior knowledge did not translate into extrapolated feedback strategies that were more effective at teaching the MLP than those with less self-reported familiarity. Developers should not assume that people with a better understanding of the task or the technology will necessarily be more effective teachers.

[2]To be considered familiar, a participant had to self-report that they knew it well or they knew it a fair amount.

[3]We used this test because the Shapiro-Wilk Test of Normality showed the groups' accuracies to be non-normally distributed.

Acknowledgments

This work was supported by the National Science Foundation (NSF) awards IIS-2106690 and IIS-1955653, and Office of Naval Research (ONR) award N00014-24-1-2124. Dražen Brščić was supported by JST Moonshot Grant JPMJMS2011, and JSPS Kakenhi Grants JP24H00722 and JP23K11271. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or ONR.

References

Candon, K., Georgiou, N. C., Zhou, H., Richardson, S., Zhang, Q., Scassellati, B., & Vázquez, M. (2024). React: Two datasets for analyzing both human reactions and evaluative feedback to robots over time. In *Proceedings of the 2024 acm/ieee international conference on human-robot interaction* (p. 885–889). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3610977.3637480 doi: 10.1145/3610977.3637480

Chernova, S., & Thomaz, A. (2014, 04). Robot learning from human teachers. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *8*, 1-121. doi: 10.2200/S00568ED1V01Y201402AIM028

Fukushima, K. (1969). Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, *5*(4), 322–333.

Huang, J., Aronson, R. M., & Short, E. S. (2024). Modeling variation in human feedback with user inputs: An exploratory methodology. In *Proceedings of the 2024 acm/ieee international conference on human-robot interaction* (pp. 303–312).

Kemp, C. C., Edsinger, A., Clever, H. M., & Matulevich, B. (2022). The design of stretch: A compact, lightweight mobile manipulator for indoor human environments. In *2022 international conference on robotics and automation (icra)* (pp. 3150–3157).

Kessler Faulkner, T. A., Schaertl Short, E., & Thomaz, A. L. (2020). Interactive reinforcement learning with inaccurate feedback. In *2020 ieee international conference on robotics and automation (icra)* (p. 7498-7504). doi: 10.1109/ICRA40945.2020.9197219

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knox, W. B., & Stone, P. (2008). Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th ieee international conference on development and learning* (p. 292-297). doi: 10.1109/DEVLRN.2008.4640845

Knox, W. B., & Stone, P. (2009). Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on knowledge capture* (pp. 9–16).

Koutroumbas, K., & Theodoridis, S. (2008). *Pattern recognition*. Academic Press.

Lee, K., Smith, L., Dragan, A., & Abbeel, P. (2021). B-pref: Benchmarking preference-based reinforcement learning. *Neural Information Processing Systems (NeurIPS)*.

Loftin, R., MacGlashan, J., Peng, B., Taylor, M., Littman, M., Huang, J., & Roberts, D. (2014). A strategy-aware technique for learning behaviors from discrete human feedback. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 28).

Loftin, R., Peng, B., MacGlashan, J., Littman, M. L., Taylor, M. E., Huang, J., & Roberts, D. L. (2016). Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *Autonomous agents and multi-agent systems*, *30*, 30–59.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., ... Littman, M. L. (2017). Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th international conference on machine learning - volume 70* (p. 2285–2294). JMLR.org.

Metcalf, K., Sarabia, M., Fedzechkina, M., & Theobald, B.-J. (2024, Mar.). Can you rely on synthetic labellers in preference-based reinforcement learning? it's complicated. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(9), 10128-10136. doi: 10.1609/aaai.v38i9.28877

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, A. (2022, aug). Human-in-the-loop machine learning: a state of the art. *Artif. Intell. Rev.*, *56*(4), 3005–3054. Retrieved from https://doi.org/10.1007/s10462-022-10246-w doi: 10.1007/s10462-022-10246-w

Thomaz, A. L., & Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*, *172*(6), 716-737. doi: https://doi.org/10.1016/j.artint.2007.09.009

Thomaz, A. L., Breazeal, C., et al. (2006). Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai* (Vol. 6, pp. 1000–1005).

Thompson, S., Narcomey, A., Lew, A., & Vázquez, M. (2024). Shutter: A low-cost and flexible social robot platform for in-the-wild deployments. In *Companion of the 2024 acm/ieee international conference on human-robot interaction* (p. 94–96). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3610978.3641090 doi: 10.1145/3610978.3641090

Wang, S., Wang, A., Goncharova, S., Scassellati, B., & Fitzgerald, T. (2025). Effects of robot competency and motion legibility on human correction feedback. In *2025 20th acm/ieee international conference on human-robot interaction (hri)* (p. 789-799). doi: 10.1109/HRI61500.2025.10974241

Yu, H., Aronson, R. M., Allen, K. H., & Short, E. S. (2023). From "thumbs up" to "10 out of 10": Reconsidering scalar feedback in interactive reinforcement learn-

ing. In *2023 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 4121-4128). doi: 10.1109/IROS55552.2023.10342458