2024 33rd IEEE International Conference on
Robot and Human Interactive Communication (RO-MAN)
August 26-30, 2024. Pasadena, California, USA.

2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN) | 979-8-3503-7502-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/RO-MAN60168.2024.10731359

How Much Progress Did I Make? An Unexplored Human Feedback Signal for Teaching Robots

Hang Yu[1], Qidi Fang[1], Shijie Fang[1], Reuben M. Aronson[1], and Elaine Schaertl Short[1]

***Abstract*— Enhancing the expressiveness of human teaching is vital for both improving robots' learning from humans and the human-teaching-robot experience. In this work, we characterize and test a little-used teaching signal: *progress*, designed to represent the completion percentage of a task. We conducted two online studies with 76 crowd-sourced participants and one public space study with 40 non-expert participants to validate the capability of this progress signal. We find that progress indicates whether the task is successfully performed, reflects the degree of task completion, identifies unproductive but harmless behaviors, and is likely to be more consistent across participants. Furthermore, our results show that giving progress does not require extra workload and time. An additional contribution of our work is a dataset of 40 non-expert demonstrations from the public space study through an ice cream topping-adding task, which we observe to be multi-policy and sub-optimal, with sub-optimality not only from teleoperation errors but also from exploratory actions and attempts. The dataset is available at `https://github.com/TeachingwithProgress/Non-Expert_Demonstrations`.**

I. Introduction

Robots have already firmly become part of our daily lives, making it crucial to learn from users, especially non-expert users. Learning from Demonstration (LfD) enables robots to learn new skills by observing expert policies [1], [2] while Learning from Human Feedback (LfHF) allows robots to adapt to human preferences or correct wrong behaviors by learning or shaping a policy [3], [4], [5]. More recent work has further shown that using human feedback and demonstrations together can make learning even more effective by reducing the data needs for human feedback [6] and loosening the requirements of demonstrations to be near-optimal [7]. However, while interest in learning fully or partially from humans is high, there is relatively little research on what the most effective forms of human feedback are, especially in *combination* with human demonstrations.

Human feedback and human demonstrations can be complementary due to the difference in human knowledge they carry. Demonstrations carry relatively dense and global information including policies and goals, and tend to be less accurate [8]. Human feedback carries relatively sparse and local information such as the correctness or a rating of a robot's action, and giving high-quality feedback can be much easier than giving high-quality demonstrations [9]. Perfect demonstrations are hard to obtain while purely learning from human feedback requires many human labels, often obtained at significant time and expense. To address these challenges and improve the quality of learning, human demonstrations can be combined with human feedback: demonstrations can be used to train an initial policy to improve the sample efficiency of feedback [6], and feedback can be used to refine the policy learned from demonstrations [8].

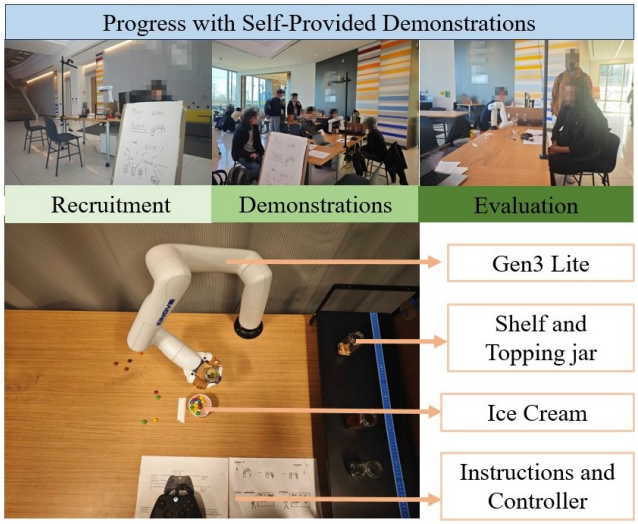

Fig. 1: Public space study with an ice cream topping-adding task to collect demonstrations and progress from non-experts.

In previous work using human feedback alongside demonstrations, the forms of human feedback used are directly adapted from LfHF [6], [8], which might not be effective for evaluating a demonstration. The quality of a demonstration or a partial demonstration is typically assessed by comparing it to another [8], [7], [1]. This approach of comparing or ranking demonstrations is extremely hard for naive users, especially when the trajectory is only a partial demonstration [10]. It also neglects the objective quality of a demonstration or a partial demonstration itself: a pair of demonstrations might both be good or both bad, making preferences difficult to provide. This is especially true for non-expert demonstrations: non-expert demonstrations can be noisy, multi-policy, and yet still succeed (we show this in subsection V-C). While feedback like binary evaluation and scalar feedback is capable of assessing the quality of a demonstration, comparative information is unavailable for binary feedback and is unreliable for scalar feedback [11].

In this work, we characterize a novel type of human feedback for robot learning: *progress*, which is used to capture the completion of a task. We show that progress can indicate the extent of task completion, determine if a task is completed, and be robust to unproductive behaviors.

[1] Tufts University, Medford, MA 02144, USA

Furthermore, when compared to scalar feedback, progress is more consistent when demonstrations are noisy and does not need extra workload and time. We show the capability of progress through two online studies and one in-person study. For the online studies, we recruit 76 crowd-sourced workers to provide scalar feedback and progress with pre-recorded expert demonstrations over three simple tasks and one long-horizon task. For the in-person study, we recruit 40 passersby to provide demonstrations in an ice cream topping-adding task, shown in Figure 1. The demonstrators then provide progress and scalar feedback to their own demonstrations.

The main contribution of this work is to demonstrate that *progress* has information beyond rating and ranking, and has great potential in interactive learning. Our validation studies covered a wide range of scenarios and involved 116 participants in total. Our results also showed that non-expert demonstrations are multi-policy and sub-optimal, but sub-optimal in a meaningful way. Finally, we released a dataset online with 40 non-expert labeled demonstrations from a public space study, which may better-reflect the types of demonstrations that can be expected from real-world deployments than typical expert or in-lab demonstration datasets.

II. Background

By leveraging human knowledge, interactive machine learning allows learning agents to adapt to the needs of individual users and improve sample efficiency [9]. Human knowledge can take a variety of forms, such as semantic representation [12], numerical feedback [5], eye gaze [13], gestures [14], facial expressions [3], and demonstrations [15]. In this work, we focus on human feedback and human demonstrations. Separately, each of these approaches has significant limitations: on the one hand, inferring a policy from human feedback requires a large number of interactions [5], [16], [17], and on the other, error-free demonstrations are rare and expensive to obtain in the real world [15]. To address this, many approaches in Interactive Machine Learning seek to combine human feedback and human demonstration to compensate for the limitations of each.

Learning from Human Feedback Learning from Human Feedback has emerged as a promising technology for robots or machine-learning agents to learn from humans via interactions [18]. LfHF, in general, refers to methods that have three components: feedback collection, policy or reward shaping, and policy optimization. Human feedback can be in a variety of forms, such as verbal [19], [20], numerical [4], [5], [21], and implicit [3], [22]. Three representative works of learning from explicit human feedback are Policy Shaping [4], TAMER [5], and Preference-based policy learning [23]. Human feedback has also been applied to modern Large Language Models (LLMs) [24] to further improve the performance of trained models by having humans rate the outputs with binary critiques. Despite a large body of work that has been done, human feedback is mostly used as reward signals.

Learning from Human Demonstrations In robotics, Learning from Demonstrations (LfD) is a method that facilitates robots to learn new skills by imitating humans [25]. The use of LfD offers several advantages, including eliminating the need for expert programming [26], high data efficiency [27], safety for learning [28], and guaranteed performance [29]. The research interest in teaching robots via demonstrations has steadily advanced. LfD methods are capable of producing optimal behaviors with clean demonstrations and sufficient error-free demonstrations [30]. However, due to the optimal assumption on the demonstrations, LfD methods like generative adversarial imitation learning [31] or behavior cloning [32] failed to acquire optimal policy for many robot tasks since even human experts would make mistakes while providing demonstrations [33]. Weighting [34], [33], [2] or ranking [1], [35] demonstrations are considered to be robust methods of learning from noisy demonstrations. However, learning from multiple users and learning from imperfect demonstrations are still challenging problems [15]. One previous work used a technique they refer to as "reward sketching"; in practice the annotators were instructed to provide progress [36]. While their work demonstrates the potential of progress for guiding learning, it did not closely investigate the properties of progress or take full advantage of it as a teaching signal, instead using large numbers of these "reward sketching" annotations as a loose approximation for a dense reward function.

Using Human Demonstrations and Human Feedback Recent work has demonstrated that combining human feedback and human demonstrations could overcome many disadvantages of using one of them solely, including safety [37], sample efficiency [6], and accuracy [8]. Specifically, work from [6] used demonstrations to train an initial model for efficiently collecting preferences from users. Work from [8] built on [6] and used a model-based method to reduce data from humans. Using human rankings, work from [7] has achieved super-human demonstration performance. Although prior work has achieved great success in consolidating human feedback with human demonstrations, the source of human feedback and human demonstrations are likely from experts or pre-trained agents [38], and the human feedback they used could be more informative. Our work differs from prior work by focusing on non-experts. We conducted our studies with crowd-sourced workers and random passersby. We showed that progress is informative and consistent when human demonstrations are multi-policy and non-optimal.

III. Progress

Our goal is to improve learning from human feedback with a new teaching signal: progress. In this section, we first define progress, and then introduce our hypothesis.

A. Progress

We hypothesize that *progress* provides complementary information to demonstrations beyond rewards. We characterize *progress* as the accumulative task completion rate over a task based on the current observation, ranging from totally incomplete to complete fully. Our intuition is that: *A teaching signal would be more robust to sub-optimal demonstrations and more consistent among users if human teachers could*

have objective references while providing the teaching signal. For progress specifically, users can use start states and finish states as references. In this work, we use the *progress* signal as a range from 0 to 100. A *progress* value of 0 indicates that the task has not yet begun, while a value of 100 signifies task completion. In between, we expect that for any given task t, current state s, any action a_i and a_j, and any previous state s_i and s_j:

$$prog_t(s_i, a_i, s) = prog_t(s_j, a_j, s) \forall i, j \quad (1)$$

Ideally, progress is independent of the path taken to the state, irrespective of the sequence of preceding states. However, human feedback is known to be noisy and can only be considered consistent if we view it as an approximate value [11]. Thus, we collect progress by presenting users with a trajectory instead of a single state to increase reliability.

B. Hypotheses

We expect that people naturally estimate task completion in their daily lives, so progress should be not hard to give. We also expect that participants would use progress to describe the completion degree of a task. We hypothesize:

H1. Giving progress does not require extra workload and extra time compared to giving scalar feedback.

H2. Progress describes the completion rate of a task.

H3. Progress could correctly indicate if the task is complete even if the robot has made mistakes.

H4. Progress is more consistent than scalar feedback when demonstrations are non-optimal.

IV. Progress with Non-Self-Provided Demonstrations

We first crowd-sourced users to provide progress with pre-recorded demonstrations for our online studies. This allows us to explore the effectiveness of utilizing progress alone, and examine its applicability across a range of intricate scenarios. We conducted two online studies: Online study I uses three simple tasks to evaluate the workload of giving progress, while online study II involves a long-horizon task comprising six sub-tasks and five scenarios to assess the utility of progress.

A. Online study setups

First, we validated the workload of providing progress and verified that progress contains unique information relative to scalar feedback in a range of 0 to 100. For each study, we recruited two groups of participants from an online platform. One group of participants was only asked to give progress and the other group of participants was only asked to give scalar feedback. Participants in two groups watched the same demonstrations in the same order.

1) Online study I: We recruited two groups of 20 participants from Amazon Mechanical Turk to provide progress annotations to a robot performing three related tasks. The three tasks we used are reaching, pouring, and spinning, shown in Figure 2, which are sub-tasks of the task we used in the online study II. For each task, participants gave 10 progress or 10 scalar feedback annotations. After giving all progress or scalar feedback, a NASA TLX [39] questionnaire was given to measure the workload. All demonstrations are perfect demonstrations except the robot made one mistake during the spinning task at step 5.

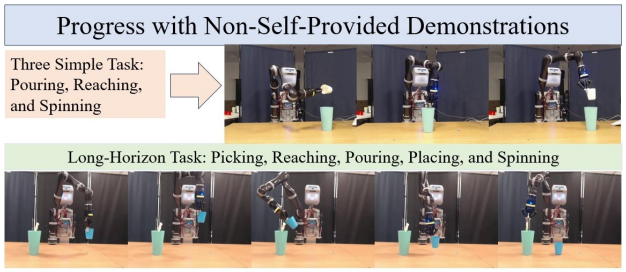

Fig. 2: Online Study Setups. Three simple tasks for comparing the workload of giving progress, and a long-horizon task for comparing the applicability of progress.

2) Online study II: In study II, we used a long-horizon task with five representative types of variation in task performance to show the capability of progress. The task we use is a tea steeping task (also shown in Figure 2), which is a task combining picking, reaching, pouring, and spinning. We recruited another two groups of 18 participants from Prolific, a crowdsourcing platform for scientific research, to give scalar feedback or progress. We chose five demonstrations each representing a distinct scenario: Perfect (everything is performed flawlessly), Imperfect (the cup is dropped onto the table rather than being placed carefully), Unaware (failed to pick the stir, and spun without a stir), Corrected (failed to pick the stir but went back to pick it after a few spinning), and Failure (the cup was dropped at the beginning). The results and further illustrations of five cases are shown in Figure 4. For scenarios other than Corrected, participants were asked to give 15 progress or scalar feedback. In Corrected, 20 progress or scalar feedback were collected.

B. Quantitative analysis

To analyze the data, we used t-tests [40] and Bayesian statistics with the schemes present in [41]. For t-test results, we use Shapiro-Wilk tests to determine if the data is from a normal distribution. If the data is from a normal distribution, we use a standard independent samples t-test. Otherwise, we apply Kruskal-Wallis H Tests and Wilcoxon Rank-Sum Tests to our results. For Bayesian statistics, a Bayes Factor (BF) is used. We interpret BF lower than 3 as "no evidence" for the alternative hypothesis, between 3 to 10 as "moderate evidence", and 30 or above as "strong evidence".

C. Giving progress is not time-consuming and not hard

We used the NASA TLX form to measure the workload, and we recorded the average task completion time. The results are shown in Figure 3. We did not find any significant difference between giving scalar feedback and progress in all the dimensions of the NASA TLX results using t-tests and BF ($p > 0.45$ and $0.3 < BF < 1$ for all dimensions).

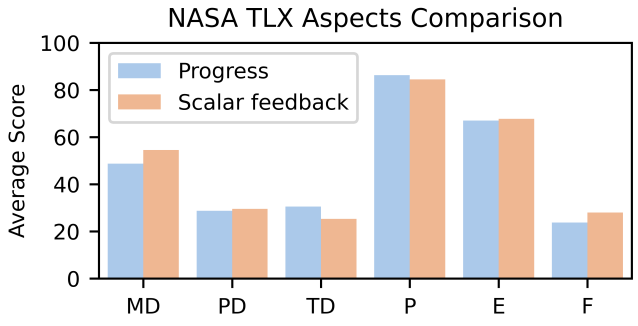

Fig. 3: Online Study One Results. The workload of giving progress has no difference from giving scalar feedback.

		Average	IQR	Comparison to Perfect
Perfect	Progress	91.5	2.75	
	Scalar F	95.1	0.75	
Imperfect	Progress	89.0	15.0	$p = 0.71$, $BF = 0.34$
	Scalar F	62.7	43.75	$p < 0.01$,$BF > 1000$
Unaware	Progress	61.3	21.25	$p < 0.01$, $BF = 98.75$
	Scalar F	23.1	50.0	$p < 0.01$, $BF > 1000$
Corrected	Progress	94.1	8.25	$p = 0.62$, $BF = 0.33$
	Scalar F	82.0	23.75	$p < 0.05$, $BF = 2.01$
Failure	Progress	23.4	17.5	$p < 0.01$, $BF > 1000$
	Scalar F	91.0	10.0	$p = 0.30$, $BF = 0.46$

TABLE I: Average progress and scalar feedback at the last step. Progress shows the ability to indicate task completion even if the demonstration is not perfect.

The average task completion time for evaluating the demonstrations using progress is 22 minutes 3 seconds, and the average task completion time for using scalar feedback is 23 minutes 10 seconds. The average completion time is lower for progress by 5.06%, but we did not find any significant difference using t-test and BF ($p = 0.442$, and $BF = 0.444$). **H1** is supported.

D. *Progress is used to signify the completion rate of a task*

The results of online study II are shown in Figure 4. Each line represents the average progress or scalar feedback of 18 participants under a certain scenario. While scalar feedback and progress can both describe the quality of a demonstration, progress and scalar feedback focus on different information.

Progress describes the degree to what extent a task has been completed. Progress was low at the beginning, increased as the task was completed more, stayed the same or became lower while the robot was not acting towards finishing the task (see Failure before spinning, Unaware after missing the stirrer, and Corrected before re-picking the stirrer). A higher value of progress indicates more successful task completion. For instance, in the Unaware scenario, the progress at the final step is greater than in the Failure scenario but lower than in the Perfect scenario, due to 5 out of 6 sub-tasks being completed in the Unaware, as opposed to 1 out of 6 in the Failure. We calculate Pearson corrections between the progress function from participants and an accumulative progress function created by true labels (+1 for completing part of the task, 0 for not completing anything, and -1 for backtracking) for each scenario. The correlations are 0.983 for Perfect scenario ($p < 0.001$), 0.982 for Imperfect scenario ($p < 0.001$), 0.974 for Unaware scenario ($p < 0.001$), 0.977 for Corrected scenario ($p < 0.001$), and 0.895 for Failure scenario ($p < 0.001$). **H2** is supported.

Scalar feedback reflects the quality of a trajectory, and only a single trajectory. The average values of scalar feedback were generally high if there were no imperfections, and dramatically changed if there were any mistakes had been made, no matter if the mistakes would affect the robot completing the task (step 10, dropping cup, in Imperfect scenario) or if the robot did better in previous sub-tasks (step 12 to step 15 in Failure).

E. *Progress indicates if a task is complete*

We showed average progress and average scalar feedback at the last step for all participants and statistical analysis results between each scenario and the Perfect scenario in Table I. We found that progress could correctly indicate if the task is complete, while the indication of task completion was not captured by scalar feedback. The task has been completed in three scenarios, Perfect, Imperfect, and Corrected. The average progress of three completed scenarios at the last step is about 90, and there was no evidence showing that there is any difference between the other two and the Perfect case ($p = 0.714, BF = 0.34$ for Imperfect, and $p = 0.620, BF = 0.33$ for Corrected). For the Unaware and Failure scenarios, there is strong evidence indicating that the task was not completed ($avg = 61.3, p < 0.001, BF = 98.75$ for Unaware, and $avg = 23.4, p < 0.001, BF > 1000$ for Failure). **H3** is supported.

F. *Progress is robust and more consistent to sub-optimality*

In the Imperfect scenario, when the robot dropped the cup onto the table, the progress remained at a similar level and did not affect the eventual progress. The conclusion holds the same in the Corrected scenario. The progress only changed slightly when the robot missed the stirrer and started spinning unproductively, and progress at the end is similar to Perfect. Scalar feedback, on the other hand, changed dramatically in all these cases. Participants used scalar feedback to indicate if a demonstration was clean and good, but the possibility that the errors were "harmless explorations" or "fixable mistakes" is not captured by scalar feedback. Moreover, for every scenario other than Perfect, progress has a lower average standard deviation in each scenario compared to scalar feedback. The average standard deviation for Imperfect: 16.1 for progress and 22.5 for scalar feedback ($p = 0.018$, $BF = 3.19$), Unaware: 16.8 for progress and 21.1 for scalar feedback ($p = 0.060$, $BF = 1.41$), Corrected 16.1 for progress and 20.3 for scalar feedback ($p = 0.058$, $BF = 1.39$), and Failure: 15.6 for progress and 28.3 for scalar feedback ($p < 0.001$, $BF > 1000$). **H4** is supported.

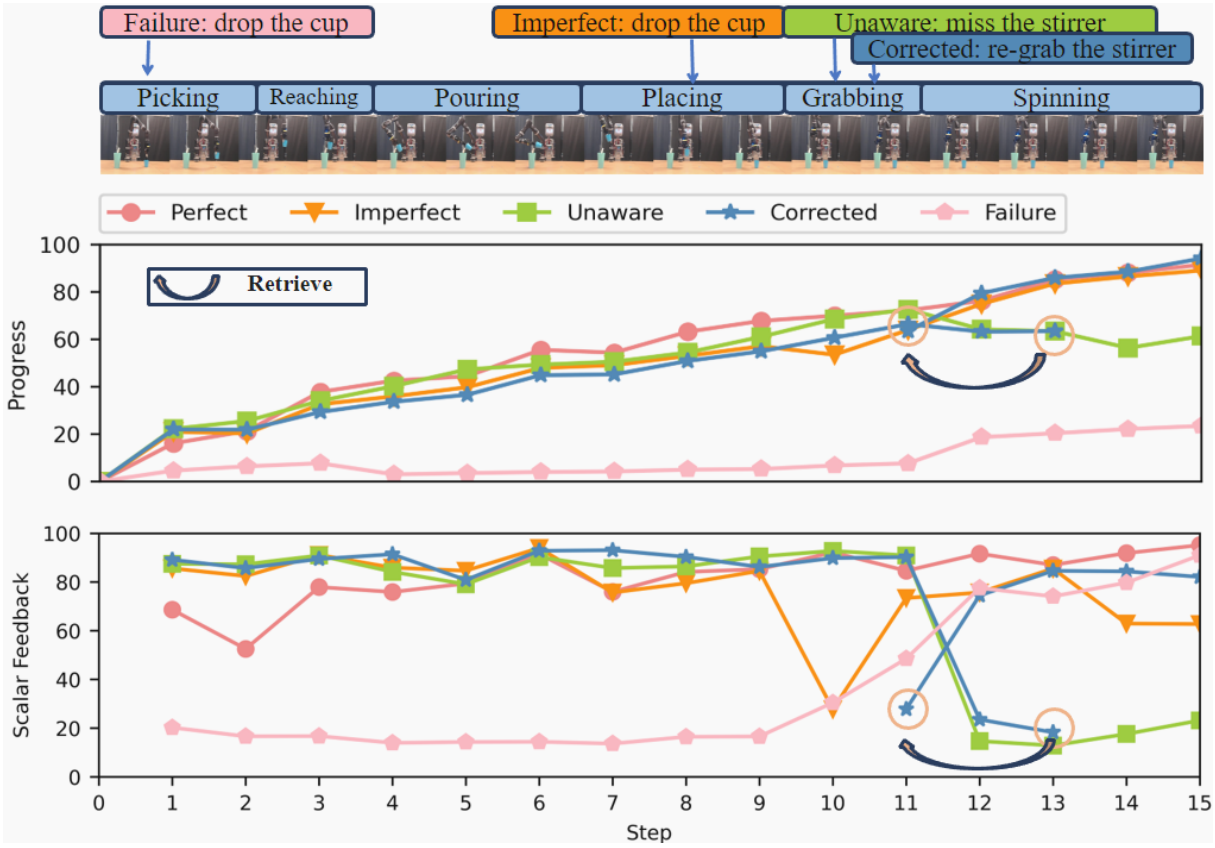

Fig. 4: Online Study II Results. Progress and scalar feedback carry different information. Scalar feedback reflects the optimality of a trajectory, while progress reflects the degree towards completing the task. Progress is more consistent than scalar feedback when the demonstration is non-optimal across participants. Progress is also capable of indicating if the task is completed successfully even if the robot has made a minor mistake or made a faulty mistake but fixed it.

V. Progress with Self-Provided Demonstrations from Non-Experts

While progress can be effective in being used along with expert demonstrations, using progress to annotate non-expert demonstrations may still be challenging. We expected most hypotheses will hold. We conducted a public space study to validate the applicability of progress with self-provided demonstrations from non-experts. We recruited 40 participants to provide demonstrations, progress, and scalar feedback in an ice cream topping-adding task. Participants were recruited from the atrium of a university building and the overall participation time was about 15 minutes. This work is approved by the Institutional Review Board and all data collected was anonymous. We released all data we collected from our public space study as a dataset, along with the example scripts that read the data from files. The dataset is available at: `https://github.com/TeachingwithProgress/Non-Expert_Demonstrations`

A. Experiment Setup

The study was settled in the lobby of a university building. Each participant was asked to first give one demonstration and then watch a replay of the demonstration. The task we asked participants to demonstrate is an ice cream topping-adding task. The goal for participants is to pick up a topping from a shelf and pour the topping into an ice cream via teleoperating a robot arm. The shelf is located on the right side of the workspace, and there are four toppings available which are located at four locations. The participants controlled the arm by using an Xbox controller. The arm was a Kinova Gen 3 Lite arm with six DoF. The setup and the workspace are shown in Figure 1. We recorded the demonstrations in 5 HZ. During the replay, the arm would stop every 10% of the frames, and we would ask participants for one progress and one scalar feedback for the replayed partial trajectory.

B. Experiment Procedure

We recruited participants by asking people who walked by our setup. Of the 40 participants, 22 participants were male, 14 participants were female, and 4 participants preferred not to say. If the participants agreed to join the study, we first asked them if they were familiar with robots, and excluded them if they said yes. We then asked them to fill out a consent form. Then we introduced them to the ice cream topping adding task, and how to use an Xbox controller to teleoperate the arm. Participants had up to 3 minutes to practice the

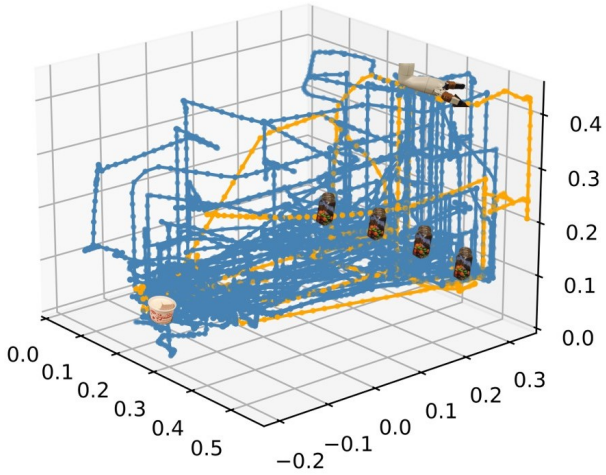

Fig. 5: 3D visualization of 40 non-expert demonstration trajectories. The positions of objects are marked out with images. The blue trajectories are successful demonstrations and the orange trajectories are faulty demonstrations. Most of the demonstrations succeed, while the policies are diverse.

task before giving a demonstration. Each participant only had one chance to give a demonstration and could not retry. After giving a demonstration, the experimenter introduced participants to the replay evaluation task. The experimenter would introduce progress and scalar feedback in detail to reduce the difference in understanding of the signals among participants. During the replay, the arm would stop ten times. For each time, the experimenter would briefly explain progress and scalar feedback again, and ask participants for two signals in a random order.

C. Non-Expert Demonstrations Are Multi-Policy and Noisy

We collected 40 demonstrations from 40 participants. All trajectories are visualized in Figure 5. We also visualized the locations of the objects used in the experiment. The blue trajectories are successful demonstrations and the orange trajectories are failed demonstrations. We found that non-expert demonstrations are noisy and contain a variety of policies, even though 34 out of 40 are ultimately successful. This suggests that policies can be both successful and sub-optimal, which supports our intuition: assessing the quality of demonstrations by comparing is likely to lose information if there are many "good enough" policies. We also observed that the noise in the demonstrations is not only teleoperating errors but also explorations. For instance, we observed that 14 participants slightly shook the topping jar to test if the gripper firmly held the jar when picking the jar, and 17 participants poured a few toppings out first to see if the jar was right above the ice cream when pouring the toppings. This highlights the importance of detecting unproductive behaviors, and suggests that **noisy demonstrations from partially trained agents or perfect demonstrations with injected noise are inappropriate approximations of noisy human demonstrations**: human demonstrations are "noisy" in a meaningful way. For the six failed demonstrations,

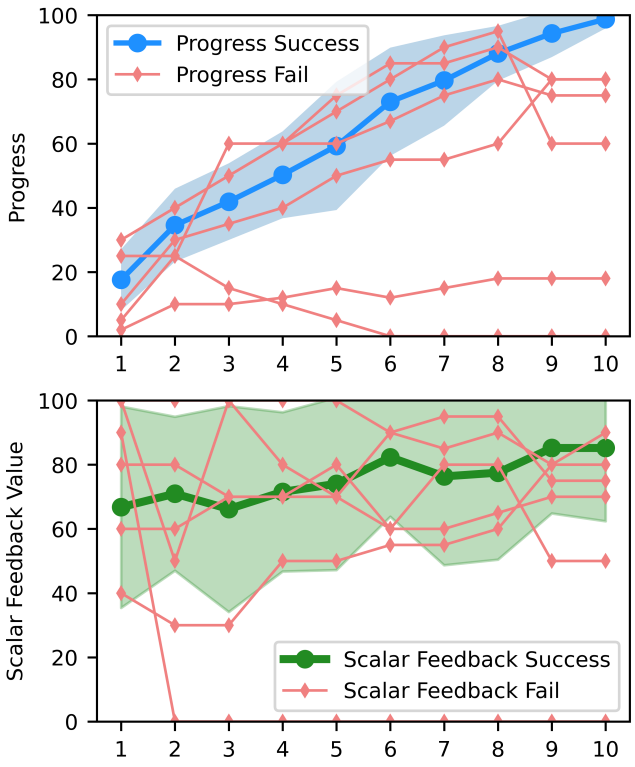

Fig. 6: Progress and scalar feedback over 32 Demonstrations. Progress increases as the task approaches completion, indicates task success, and is consistent among participants.

three of them were because the topping jar was accidentally dropped while reaching the ice cream, and two of them were because the gripper did not successfully pick up the topping jar. The most common faulty cases are similar to the Failure scenario and the Unaware scenario we used in our online study, which confirmed the validity of the design of our online study.

D. Progress indicates task completions and is more consistent across participants than scalar feedback

We collected progress and scalar feedback from 34 participants (two participants' progress and scalar feedback were excluded since all progress and scalar feedback they provided were 100). The results are shown in Figure 6. We plot the average progress and scalar feedback for all successful demonstrations, and progress and scalar individually for all failure demonstrations. We are not able to determine if progress from participants correctly describes the task completion rate since we do not have ground truth labels, but average progress for successful demonstrations did start from low and increased as the demonstrations were reproduced which is a strong indication that progress correlates with task completion rates.

We successfully identified all failed demonstrations by only looking at progress at the last step. We use 90 as the divide value, which is the average progress at the last step for successful demonstrations in our online study. If progress at

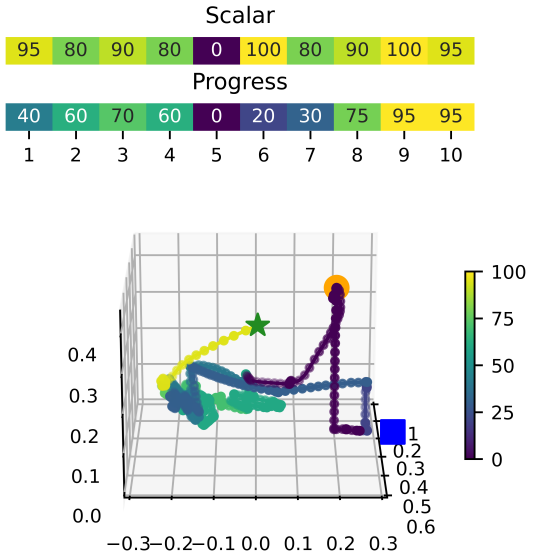

Fig. 7: 3D trajectory of one participant's demonstration. The trajectories are colored with progress, and both scalar feedback and progress signals are indicated in a timeline above the figure. The brown circle represents the start point. The green star represents the endpoint. The blue square is the position of the topping jar. The participant reset the arm to the start position after demonstrating a few steps to adjust the grip point.

the last step is less than 90, the task is incomplete, otherwise the task is completed. For all failed demonstrations, the end progress is less than 90 and the average progress at the end for successful demonstrations is 98.8. **H3** is supported. Moreover, progress is significantly more consistent across participants even though their demonstrations are multiple policies and in different quality ($avg\_std_{progress} = 11.6$, $avg\_std_{scalarfeedback} = 25.5$, $p < 0.001$, $BF > 1000$). **H4** is supported.

E. Progress allows the awareness of backtracking

As mentioned in subsection V-C, participants would do explorations while demonstrating difficult parts of the task. We also observed that some participants went back a few steps to adjust the grip pose or the pouring position. For instance, one participant failed to pour the toppings into the ice cream because the grasp point was not optimal. Then the participant decided to reset the arm to the initial position, demonstrated the entire task from the beginning again, and succeeded. We plot the trajectory for that demonstration along with progress and scalar feedback, shown in Figure 7. As indicated by progress, the arm was reset to the initial position between step 4 and step 5, and the task was successfully completed afterward.

VI. Discussion

In this work, we investigated and closely defined an under-explored teaching signal, *progress*, and conducted three different studies to show the usefulness of progress across a variety of scenarios. We showed that giving progress is not hard and progress carries information beyond scalar feedback and other teaching signals in prior work. We expect that progress will be effective in many applications other than just using along with demonstrations.

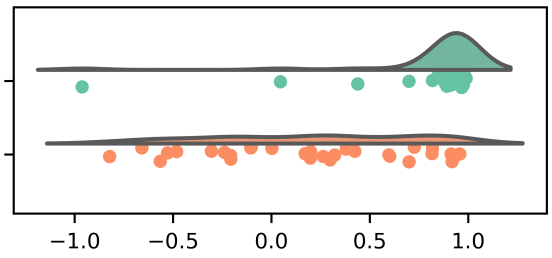

Fig. 8: Correlations between progress and rewards (top), and correlation between scalar feedback and rewards (bottom) over 32 participants. The correlations between progress and rewards have an average of 0.83.

Prevent Reward Hacking Progress could be a powerful signal to indicate reward hacking. Reward hacking is a phenomenon in which a learning agent learns to achieve high rewards by performing unintended actions instead of finishing the task. For instance, a cleaning robot gets a +1 reward every time it cleans a room. The robot, instead of cleaning one room and going to the next room, repetitively ejects dirt in a room, then cleans it, and thus achieves high rewards. Using progress, we can easily identify that the robot is not advancing towards task completion.

Inverse Reinforcement Learning We observed a similarity between progress functions and reward functions and expect that a progress function might be a representation of a reward function. We trained a reward function using demonstrations we collected in the public space study using Adversarial Inverse Reinforcement Learning [42], and calculated the rewards for each demonstration. We then calculated the Pearson correlation between the rewards of the demonstration and the progress, as well as the correlation between the rewards and the scalar feedback from each participant in Figure 8. We found that progress is strongly correlated with the learned reward function ($avg_r = 0.83$) and the average correlation is significantly higher ($p < 0.001$, $BF > 1000$) than scalar feedback ($avg_r = 0.19$).

Data Filtering and Ranking Progress could also be used to rank demonstrations. For example, a demonstration with a progress of 60 at the end should be ranked lower than a demonstration with a progress of 90. Moreover, progress can be applied as a data filter especially when the demonstrations are sub-optimal.

A key area for future work is to compare the model performance between the model trained using progress and the model trained using other types of human feedback. Training a reliable model from limited non-expert demonstrations and non-expert annotations is challenging but future work could expand our data with more demonstrations along with more types of human teaching signals such as preference.

VII. Conclusion

In conclusion, we defined *progress* in detail and found that *progress* could be used to describe completion degrees of a task, indicate if a task is complete, and be more consistent across users, without requiring extra workload or time compared to giving scalar feedback. We collected 40 non-expert demonstrations along with progress and scalar feedback, and released them as a dataset. We found that non-expert demonstrations are multi-policy and mostly successful, while noisy in a meaningful way. Our work suggests that progress is information-rich and is worth more attention to develop new methods to effectively leverage the novel information from progress.

ACKNOWLEDGMENT

The work described here was supported in part by the US National Science Foundation (IIS-2132887).

References

[1] D. S. Brown, W. Goo, and S. Niekum, "Better-than-demonstrator imitation learning via automatically-ranked demonstrations," in *Conf. on robot learning*. PMLR, 2020, pp. 330–359.

[2] L. Chen, R. Paleja, and M. Gombolay, "Learning from suboptimal demonstration via self-supervised reward regression," in *Conf. on robot learning*. PMLR, 2021, pp. 1262–1277.

[3] Y. Cui, Q. Zhang, B. Knox, A. Allievi, P. Stone, and S. Niekum, "The empathic framework for task learning from implicit human feedback," in *Conf. on Robot Learning*. PMLR, 2021, pp. 604–626.

[4] T. Cederborg, I. Grover, C. L. Isbell Jr, and A. L. Thomaz, "Policy shaping with human teachers." in *IJCAI*, 2015, pp. 3366–3372.

[5] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The tamer framework," in *Proc. of the fifth Int. Conf. on Knowledge capture*, 2009, pp. 9–16.

[6] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in atari," *Advances in neural information processing systems*, vol. 31, 2018.

[7] D. S. Brown, W. Goo, and S. Niekum, "Ranking-based reward extrapolation without rankings," *arXiv preprint arXiv:1907.03976*, 2019.

[8] M. Palan, N. C. Landolfi, G. Shevchuk, and D. Sadigh, "Learning reward functions by integrating human demonstrations and preferences," *arXiv preprint arXiv:1906.08928*, 2019.

[9] C. Arzate Cruz and T. Igarashi, "A survey on interactive reinforcement learning: Design principles and open challenges," in *Proc. of the 2020 ACM designing interactive systems Conf.*, 2020, pp. 1195–1209.

[10] C. Laidlaw and S. Russell, "Uncertain decisions facilitate better preference learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 070–15 083, 2021.

[11] H. Yu, R. M. Aronson, K. H. Allen, and E. S. Short, "From "thumbs up" to "10 out of 10": Reconsidering scalar feedback in interactive reinforcement learning," in *2023 IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 4121–4128.

[12] I. Kostavelis and A. Gasteratos, "Learning spatially semantic representations for cognitive robot navigation," *Robotics and Autonomous Systems*, vol. 61, no. 12, pp. 1460–1475, 2013.

[13] A. Saran, E. S. Short, A. Thomaz, and S. Niekum, "Understanding teacher gaze patterns for robot learning," in *Conf. on Robot Learning*. PMLR, 2020, pp. 1247–1258.

[14] P. M. Yanik, J. Manganelli, J. Merino, A. L. Threatt, J. O. Brooks, K. E. Green, and I. D. Walker, "A gesture learning interface for simulated robot path shaping with a human teacher," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 1, pp. 41–54, 2013.

[15] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.

[16] R. Arakawa, S. Kobayashi, Y. Unno, Y. Tsuboi, and S.-i. Maeda, "Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback," *arXiv preprint arXiv:1810.11748*, 2018.

[17] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep tamer: Interactive agent shaping in high-dimensional state spaces," in *Proc. of the AAAI Conf. on artificial intelligence*, vol. 32, no. 1, 2018.

[18] S. Casper, X. Davies, C. Shi, T. K. Gilbert, J. Scheurer, J. Rando, R. Freedman, T. Korbak, D. Lindner, P. Freire, *et al.*, "Open problems and fundamental limitations of reinforcement learning from human feedback," *arXiv preprint arXiv:2307.15217*, 2023.

[19] P. Goyal, S. Niekum, and R. J. Mooney, "Using natural language for reward shaping in reinforcement learning," *arXiv preprint arXiv:1903.02020*, 2019.

[20] G. Kuhlmann, P. Stone, R. Mooney, and J. Shavlik, "Guiding a reinforcement learner with natural language advice: Initial results in robocup soccer," in *The AAAI-2004 workshop on supervisory control of learning and adaptive systems*. San Jose, CA, 2004.

[21] D. Arumugam, J. K. Lee, S. Saskin, and M. L. Littman, "Deep reinforcement learning from policy-dependent human feedback," *arXiv preprint arXiv:1902.04257*, 2019.

[22] D. Xu, M. Agarwal, E. Gupta, F. Fekri, and R. Sivakumar, "Accelerating reinforcement learning using eeg-based implicit human feedback," *Neurocomputing*, vol. 460, pp. 139–153, 2021.

[23] R. Akrour, M. Schoenauer, and M. Sebag, "Preference-based policy learning," in *Machine Learning and Knowledge Discovery in Databases: European Conf., ECML PKDD 2011, Athens, Greece, September 5-9, 2011. Proc., Part I 11*. Springer, 2011, pp. 12–27.

[24] OpenAI, :, J. Achiam, Adler, and et al., "GPT-4 Technical Report," *arXiv e-prints*, p. arXiv:2303.08774, Mar. 2023.

[25] S. Chernova and A. L. Thomaz, *Robot learning from human teachers*. Morgan & Claypool Publishers, 2014.

[26] Z. Zhu and H. Hu, "Robot learning from demonstration in robotic assembly: A survey," *Robotics*, vol. 7, no. 2, p. 17, 2018.

[27] H. Ravichandar, S. R. Ahmadzadeh, M. A. Rana, and S. Chernova, "Skill acquisition via automated multi-coordinate cost balancing," in *2019 Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7776–7782.

[28] J. Umlauft and S. Hirche, "Learning stable stochastic nonlinear dynamical systems," in *Int. Conf. on Machine Learning*. PMLR, 2017, pp. 3502–3510.

[29] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: learning attractor models for motor behaviors," *Neural computation*, vol. 25, no. 2, pp. 328–373, 2013.

[30] F. Sasaki, T. Yohira, and A. Kawaguchi, "Sample efficient imitation learning for continuous control," in *Int. Conf. on learning representations*, 2018.

[31] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[32] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, "Survey of imitation learning for robotic manipulation," *Int. Journal of Intelligent Robotics and Applications*, vol. 3, pp. 362–369, 2019.

[33] Y.-H. Wu, N. Charoenphakdee, H. Bao, V. Tangkaratt, and M. Sugiyama, "Imitation learning from imperfect demonstration," in *Int. Conf. on Machine Learning*. PMLR, 2019, pp. 6818–6827.

[34] F. Sasaki and R. Yamashina, "Behavioral cloning from noisy demonstrations," in *Int. Conf. on Learning Representations*, 2020.

[35] Y. Wang, C. Xu, B. Du, and H. Lee, "Learning to weight imperfect demonstrations," in *Int. Conf. on Machine Learning*. PMLR, 2021, pp. 10 961–10 970.

[36] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, *et al.*, "Scaling data-driven robotics with reward sketching and batch reinforcement learning," *arXiv preprint arXiv:1909.12200*, 2019.

[37] D. Brown, R. Coleman, R. Srinivasan, and S. Niekum, "Safe imitation learning via fast bayesian reward inference from preferences," in *Int. Conf. on Machine Learning*. PMLR, 2020, pp. 1165–1177.

[38] J. Huang, R. M. Aronson, and E. S. Short, "Modeling variation in human feedback with user inputs: An exploratory methodology," 2024.

[39] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proc. of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[40] R. S. Witte and J. S. Witte, *Statistics*. John Wiley & Sons, 2017.

[41] J. van Doorn, D. van den Bergh, U. Böhm, F. Dablander, K. Derks, T. Draws, A. Etz, N. J. Evans, Q. F. Gronau, J. M. Haaf, *et al.*, "The jasp guidelines for conducting and reporting a bayesian analysis," *Psychonomic Bulletin & Review*, vol. 28, pp. 813–826, 2021.

[42] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," 2018.