Robotics: Science and Systems 2023
Daegu, Republic of Korea, July 10-July 14, 2023

Investigating the Impact of Experience on a User's Ability to Perform Hierarchical Abstraction

Nina Moorman*
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
nmoorman3@gatech.edu

Nakul Gopalan*
School of Computing and Augmented Intelligence
Arizona State University
Tempe, USA
nakul.gopalan@asu.edu

Aman Singh
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
asingh609@gatech.edu

Erin Hedlund-Botti
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
erin.botti@gatech.edu

Mariah Schrum
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
mschrum3@gatech.edu

Chuxuan Yang
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
soyang@gatech.edu

Lakshmi Seelam
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
lseelam3@gatech.edu

Matthew C. Gombolay
School of Interactive Computing
Georgia Institute of Technology
Atlanta, USA
matthew.gombolay@cc.gatech.edu

***Abstract*—The field of Learning from Demonstration enables end-users, who are not robotics experts, to shape robot behavior. However, using human demonstrations to teach robots to solve long-horizon problems by leveraging the hierarchical structure of the task is still an unsolved problem. Prior work has yet to show that human users can provide sufficient demonstrations in novel domains without showing the demonstrators explicit teaching strategies for each domain. In this work, we investigate whether non-expert demonstrators can generalize robot teaching strategies to provide necessary and sufficient demonstrations to robots *zero-shot in novel domains*. We find that increasing participant experience with providing demonstrations improves their demonstration's degree of sub-task abstraction ($p < .001$), teaching efficiency ($p < .001$), and sub-task redundancy ($p < .05$) in novel domains, allowing generalization in robot teaching. Our findings demonstrate for the first time that non-expert demonstrators can transfer knowledge from a series of training experiences to novel domains without the need for explicit instruction, such that they can provide necessary and sufficient demonstrations when programming robots to complete task and motion planning problems.**

I. Introduction

Due to the diversity of end users and deployment settings, it is intractable for a robot to be pre-programmed to do any task in any environment One solution is to allow robots to learn new skills in situ, from end-users. Prior work in Learning from Demonstration (LfD) has investigated how to allow non-roboticist end-users to operate in the role of the robot teacher [8, 14, 44, 45, 47] in order to communicate personal preferences and leverage their domain knowledge [36, 42, 51]. However, many previous approaches require human demonstrators to learn how to teach the robot tasks in each domain, using videos or demonstrations from robot experts [3, 21, 39]. This approach does not scale up in enabling human users to teach a variety of tasks to a robot, as demonstrator training is domain-dependent, and an expert is still required to be in the training loop. In this work, we develop a series of demonstrator training tasks through which participants obtain knowledge about providing sufficient and necessary demonstrations that can generalize to novel domains.

Substantial emphasis in LfD has been placed on teaching robots single, short-horizon *skills*, such as picking up or making contact with an object [22, 28, 33, 37]. However, there is a lack of work enabling robots to learn long-horizon *tasks*, such as learning in-home assistive tasks or manufacturing process assembly operations, from human demonstrations. Such tasks can be considered multi-task problems. For example, setting a dinner table would require a robot to set multiple place settings, dependent on the number of guests, where each table setting consists of multiple objects that each require a different manipulation procedure. A demonstrator cannot be expected to provide demonstrations for each task specification of these multi-tasks, such as for each possible number of plate settings. Since long-horizon tasks require a robot to solve repetitive multi-task problems, demonstrators need to break up their demonstrations into shorter abstractions that a robot can reuse. Prior work has shown that demonstrators are capable of teaching abstractions when explicitly instructed on how to do so. Akgun et al., for instance, demonstrated

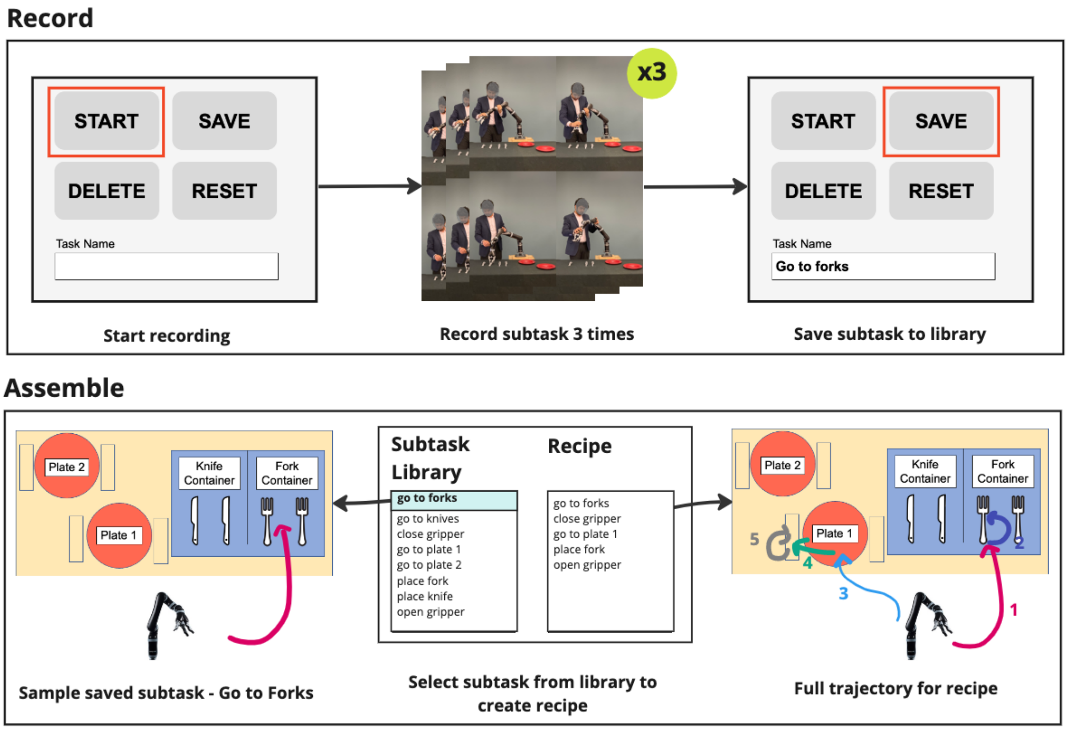


Fig. 1: Via the interface and kinesthetic teaching, participants record three demonstrations to save a sub-task. Saved sub-tasks are then available in the interface library. To execute a task, participants assemble a recipe from the set of recorded sub-tasks.

that users can teach Keyframe-based abstractions for robot movement [3]. Likewise, Mohseni-Kabir et al., demonstrated that, for hierarchical tasks, novel sub-tasks can be taught to a robot using combinations of previously taught sub-tasks [39]. However, if not prescribed how to teach the robot, prior work has found that participants struggle to provide demonstrations that exhibit abstractions sufficient for a hierarchical task [31]. Gopalan et al. similarly find that the majority of participants did not naturally teach the robot tasks using abstractions [21]. The authors compare various modes of demonstrator training and find that only videos where the experimenter demonstrates how to provide demonstrations using sub-task abstractions in the current domain enabled participants to use abstractions (in the same domain) that would be robust to novel task specifications [21]. In this work, we investigate whether, given enough practice in previous domains, participants can use sub-task abstractions to provide sufficient demonstrations in a novel domain without being told how to do so.

Instead of training demonstrators from scratch in each domain they encounter, we want the knowledge the demonstrator learns about providing demonstrations in one domain to be transferable to novel domains. We develop a three-hour training procedure with five domains and corresponding unpersonalized expert videos. This procedure enables demonstrators to transfer acquired knowledge about providing sufficient and necessary demonstrations – gained through trial, error, and expert feedback – to novel domains where no expert feedback is available. In this work, we study the impact of domain experience on users' ability to provide demonstrations when teaching a robot. Here, we define domain experience as the number of training domains for which (i) the participant has taught the robot a task via demonstration and (ii) the demonstrator has subsequently watched a training video showing a robotics expert providing the optimal teaching strategy. We employ a standard way to formulate multi-modal[1], multi-task problems in robotics [19], namely Task and Motion Planning (TAMP), for our sub-task representations.

To the best of our knowledge, our findings show for the first time that humans can transfer knowledge from a few training experiences to provide sufficient TAMP demonstrations in a novel domain. In this work, we contribute the following.

1) We design a novel user study and set of training domains to investigate whether participants improve their ability to teach the robot via demonstrations over time, without the use of a curriculum.
2) Our results demonstrate that participants are able to generalize knowledge about task abstraction ($p < .001$), teaching efficiency ($p < .001$), and redundancy ($p < .05$) zero shot to novel domains.
3) We additionally find that prior teaching experience impacts sub-task count ($p = .011$), that sub-task count impacts perceived workload ($p = .005$) as well as robot likeability ($p = .007$), and that participant agreeableness impacts teaching duration ($p = .006$).

[1]For a definition of mode, see Section II

II. PRELIMINARIES

In this section, we define terms pertaining to our work. We include a separate related works section (Section V) preceding our discussion to contextualize our findings.

Multi-task problems – In multi-task problems, the objects that the robot interacts with remain the same. However, the number of objects, their locations, or the order in which the robot interacts with the objects changes between sub-tasks [12]. This is often accomplished by leveraging similarities between the sub-tasks [50].

Multi-modal tasks – A mode is a sub-manifold of robot motion within which the robot's contact specification, with respect to different objects in the world, remains constant [4, 5, 25, 26]. A multi-modal task is one where the robot transitions between at least two modes to solve a task. For example, to pick up a block, the robot first is restricted to a mode where all its motion is confined to a sub-manifold within which the robot's gripper is not in contact with any object. After picking up the block, the mode of the robot is the sub-manifold within which it is in continuous contact with the block. Similarly, a *Long Horizon Task* is a task where the robot needs to perform multiple mode switches to solve the task. Thus, per our definition, multi-modal tasks have at least one mode switch (≥ 1), and long horizon tasks have several (> 2) mode switches. In our work, the robot is solving multi-modal tasks.

Task and Motion Planning (TAMP) – Robotics problems require an interplay between symbolic and continuous domains. For example, to pass medicines to a patient, the robot needs to make a high-level symbolic plan to know which boxes of medicines to pick up and pass to a patient. This plan and its corresponding state are symbolic and discrete over the type and quantities of medicine box objects required. However, to pick up a box the robot needs to create a continuous motion plan without collisions such that the box is in the robot's hand. This motion planning problem occurs over the continuous state of the robot's joints. Such problems, that exhibit an interplay between symbolic and continuous plans, are TAMP problems. We choose to define our domains as TAMP problems, as they require an interplay between symbolic goal states and continuous motion from the robot.

Sub-task based abstraction – Transitions between the symbolic states of a TAMP problem are called sub-tasks[2] [19]. For example, when a robot moves to pick up a cup, the state of the world transitions symbolically, such that the cup is in the robot's hand. In TAMP formulations, the sub-tasks are described by preconditions (`pre`) and effects (`eff`), as well as constraints (`con`) that must hold for all continuous actions for the duration of time the action is being taken. We provide a sample mathematical TAMP formulation for the sub-tasks of the medicine dispensing domain in the Appendix.

Sufficient sub-tasks – In our domains, a sub-task is deemed *sufficient* if the sub-task changes the symbolic state of the world and results in at most one mode change. For a sub-task to change the symbolic state of the world, the change must go beyond a negligible change in the robot's pose. Moreover, limiting the sub-task to at most one mode change ensures that the robot can change its interaction with only one object within the sub-task. Such design of sub-tasks ensures that a sub-task transition affects only a small set of symbolic state variables at a time. These sub-tasks can then be sequenced by a task planner to reach a larger set of the symbolic state space, allowing maximal generalizability in the tasks that can be solved within the domain.

Redundant sub-tasks – A sub-task is deemed redundant if its goal can be met by another sufficient sub-task or a combination of sufficient sub-tasks previously taught. Sub-task redundancy is defined with respect to a given set of demonstrations being taught.

Necessary sub-tasks – Similarly, a sub-task is deemed necessary if its goal can *not* be met by another sufficient sub-task or a combination of sufficient sub-tasks previously taught. Sub-task necessity is defined with respect to a given set of demonstrations being taught. A sub-task is deemed necessary if it is a not a redundant sub-task.

Domain experience – We define domain experience, a metric for demonstrator training, as the number of domains experienced thus far in the user study. Note that for each domain, this experience entails participants first providing demonstrations to the robot, then observing the optimal teaching sub-task breakdown in the form of a video.

[2]Sub-tasks are referred to as actions in [19]; however, we refer to these actions as sub-tasks to prevent confusion between low-level robot actions and TAMP level actions.

III. METHODS

In this section, we describe our study design, research questions, metrics, domains, and experimental procedure.

Study Design

We conducted a 1×4 within-subjects experiment with twenty-eight participants, seven per ordering condition (see Appendix for the domain ordering of each condition). Participants experience five domains in this study, a practice domain that all participants experience first, and the four ordered domains. We control for the ordering of the remaining four domains using a Latin square, ensuring that the participant count per condition is balanced. The independent variable in this study is the number of domains encountered thus far.

The robot employed in this study is the JACO arm (Gen2, three fingers for a total of seven degrees of freedom) [11] attached to a hand-crafted base located next to the experiment's table, as seen in Figure 1. We additionally designed a user interface that allows users to record and save sub-tasks they demonstrate to the robot; interface design decisions can be found in the Appendix. Participants can then use the interface to combine different sub-tasks to accomplish a task. We name the group of sub-tasks assigned to a particular task a *recipe*. We require the participant to record three demonstrations for each sub-task in order to capture variability in the way the participant moves the robot for robustness to noise.

Research Questions

RQ1*: What is the impact of domain experience on the quality of demonstrations?* We investigate whether participants can perform zero-shot transfer to novel domains of any acquired knowledge as measured by sub-task abstraction score, teaching efficiency, and sub-task redundancy.

RQ2*: What is the effect of demonstration abstraction on participants' perceived workload?* We hypothesize that higher abstraction scores will reduce the repetitiveness of participant demonstrations, thereby reducing perceived workload.

RQ3*: Do participant demographics impact the quality of demonstrations?* We investigate whether participant demographics, such as prior robotics experience and prior teaching experience, impact the quality of their demonstrations. We posit that participants with robotics or teaching experience will teach the robot, via demonstration, more effectively and efficiently.

RQ4*: Does domain type impact the quality of demonstrations?* We hypothesize that the domain type will impact the sub-task count and redundancy, abstraction score, and teaching duration.

Metrics

The objective metrics we collect in our user study are as follows. These metrics are collected per domain, for each participant.

- **Abstraction Scores:** We employ the abstraction scoring method validated in [21]. In this scoring method, one point is allotted for each instance of a sufficient sub-task employed to accomplish a task, and one point is awarded for sufficient sub-tasks that could be constructed by composing other sufficient sub-tasks. The latter ensures that finer-grain abstractions are scored higher than coarser-grain abstractions, within reason (no points are awarded for gratuitously low-level abstractions such as *move left*). The abstraction scoring rubrics employed in each domain can be found in the Appendix. This metric allows us to evaluate the sufficiency of demonstration sub-tasks.
- **Redundancy Score:** We count the number of redundant sub-tasks taught, i.e., sub-tasks whose function can be fulfilled by another existing sub-task or a combination of existing sub-tasks. This metric allows us to evaluate the necessity, independently from the sufficiency, of demonstration sub-tasks.
- **Sub-task Count:** We count the total number of sub-tasks taught to the robot in each domain, that are employed to accomplish a task.
- **Teaching Duration:** We measure the total time the participant taught the robot, including time using the interface and time spent providing the kinesthetic demonstrations.

The subjective metrics in our user study are as follows. The details of hand-crafted surveys, Cronbach's alpha, and qualitative results and quotes from interview questions are in the Appendix.

Pre-study Questionnaire:

- **Demographic Information:** We collect participants' age, gender, education, and race/ethnicity.
- **Personality** We employ the Big Five Personality survey [20], consisting of fifty questions rated on a seven-point scale (Very Strongly Disagree=1 to Very Strongly Agree=7).
- **Prior Robotics Experience** We obtain participants' prior robotics experience through a hand-crafted single-item question rated on a scale from 0 to 10+ years.
- **Prior Teaching Experience** We obtain participants' prior teaching experience through a hand-crafted, 5-question survey rated on a five-point scale (Strongly Disagree=1 to Strongly Agree=5).
- **Negative Attitude towards Robotics** We employ the Negative Attitudes Towards Robotics (NARS) Scale [49], composed of 14-questions rated on a seven-point scale (Strongly Disagree=1 to Strongly Agree=7). We report results on the three sub-scales: negative situations, negative social influence, and negative emotions.

Post-domain Questionnaire:

- **Teaching Strategy** After completing each domain, we ask participants to "please explain your strategy and thought process when teaching the robot in this domain."

Post-study Questionnaire:

- **Workload** We use the NASA Task Load Index (NASA TLX) [23] to obtain perceived workload.
- **Impression of Agent** We use the Perceived Intelligence and Likeability sub-scales of the Godspeed Questionnaire Series, rated on a 5-point scale [7].
- **Post-Interview** We ask participants five post-interview questions. The question list and qualitative results can be found in the Appendix.

Domains

We employ five domains in this study, each comprised of three tasks that the participant must teach the robot, as seen in Figure 2. The set of tasks in each domain was designed to be repetitive and time-consuming to encourage participants to use sub-task abstractions in order to avoid recording repetitive sub-tasks. Additionally, we chose these domains because they are representative of common household chores that humans could reasonably be asked to teach a robot: setting the table, packing lunch, gardening, and dispensing medication. Each task has a distinct objective with differing numbers of objects and goals but requires similar types of abstractions from the participant.

Furthermore, when teaching a robot a task in a residential or "unsanitized" setting, there will likely be objects in the environment that are unrelated to the task being taught. We thus employ distractor items in this study, which are present but not relevant to the list of tasks the participant must teach the robot in the domain, to realistically represent such settings.

Participants were given unlimited time to record sub-task demonstrations and build recipes using these recorded sub-tasks in the interface. The optimal sub-task list for each task in

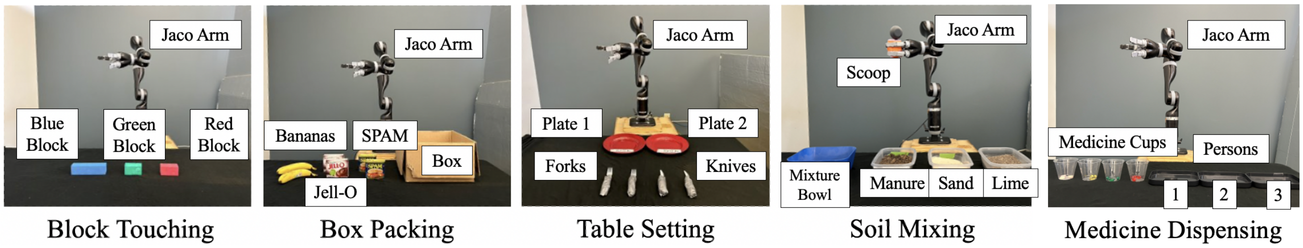


Fig. 2: This figure depicts the five domains in which participants taught the robot.

each domain can be found in the Appendix. We now describe the domains in this study (Fig. 2).

Block Touching: A blue, green, and red block are placed in front of the robot. Participants are asked to teach the robot to touch the blocks in a particular order using the robot gripper.

Box Packing: Two plastic bananas, two Jell-O boxes, and two Spam cans, along with a cardboard box are laid out in front of the robot. Participants are asked to teach the robot to pack (pick up and place) a combination of these food items into the cardboard box.

Table Setting: Two forks, two knives, and two plates are placed in front of the robot. Participants are asked to teach the robot to set the table by picking up the utensils and placing them in designated locations around the two plate settings.

Soil Mixing: A bucket of manure, a bucket of sand, and a bucket of lime are placed in front of the robot, along with a mixing bowl into which scoops of each of these materials are to be poured. Participants are asked to teach the robot to create different soil mixtures for different plants. In this domain, the scoop is placed in the robot's gripper by the experimenter.

Medicine Dispensing: Four kinds of medicine (red pill cup, green pill cup, yellow pill cup, and TUMS pill cup) along with three trays labeled persons 1, 2, and 3 are placed in front of the robot. Participants are asked to dispense the proper medication to each person by picking and placing medicine cups into the appropriate person's tray.

Procedure

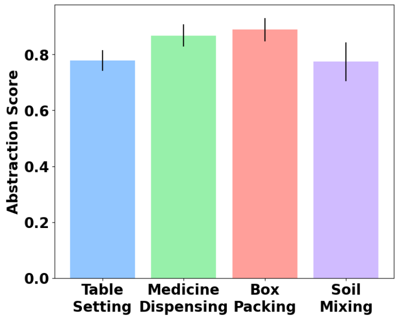
This study was approved by our university's Institutional Review Board (IRB), protocol #H22450. We recruited all participants through advertisements on campus. The study took three hours, and participants were compensated with a $50 Amazon gift card, given the long duration of the study. The procedure of the study is as follows.

Participants first take the pre-study questionnaire, comprised of surveys to collect demographic information, personality measures, prior robotics experience, prior teaching experience, and negative attitude towards robots. After the pre-study questionnaire, participants start the training portion of the study. To begin, they observe the introduction video. The introduction video[3] introduces the study, the robot, and the interface used to teach the robot sub-tasks. The video then consists of a conceptual description of how to optimally teach the robot to make an omelet. The optimal sub-tasks described for this example included (1) going to the egg carton, (2) picking up an egg, (3) going to the pan, and (4) breaking an egg into the pan. This portion of the video motivates breaking up the task into sub-tasks that can be called many times, to generalize to an omelet of any quantity of eggs. It also suggests recording the "go to the egg container" sub-task separately from the "pick up the egg" sub-task, allowing the robot to generalize going to an egg carton whose location has been moved. Finally, this video communicates that the sub-tasks can be called from generalized starting positions so multiple sub-tasks can be chained together without going back to a home position first. We note that this initial omelet domain is experienced entirely virtually, and we do not show the participant how it would be taught on the physical robot.

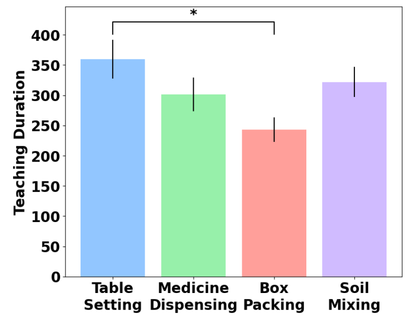
Next, participants teach the robot to complete the three block touching tasks in the demo (i.e., practice) domain. After this demo domain, the participant explains their teaching strategy, recorded via a voice recording. We note that after this demo domain, we do not show the participant a video of the optimal way to teach the robot. This block-touching domain serves to familiarize the participant with moving the robot and using the interface.

Then, for the testing portion of the study, participants teach the robot how to accomplish tasks in four different domains: box packing, table setting, soil mixing, and medicine dispensing. Each participant experiences one ordering condition, which defines the order in which the domains are encountered. All participants experience each of these domains (within subjects). The four domain ordering conditions are listed in the Appendix. For each of these four domains, participants are introduced to the domain verbally, then asked to teach the robot how to do three tasks in that domain using the interface, as seen in Figure 1. To teach the sub-tasks, participants provide kinesthetic demonstrations in which participants physically manipulate the robot. After teaching the robot, the participants answer the post-domain interview question and then observe a video showing the optimal way of teaching the robot in that domain (communicating the proper sub-task breakdown) prior to experiencing the next novel domain. The optimal teaching strategy video for each domain was designed to communicate how to optimally teach the robot, listing the optimal sub-tasks for the domain, along with how to teach and record those sub-tasks on the robot using the interface. Next, the videos show how to use the sub-tasks to build the recipe for one task in

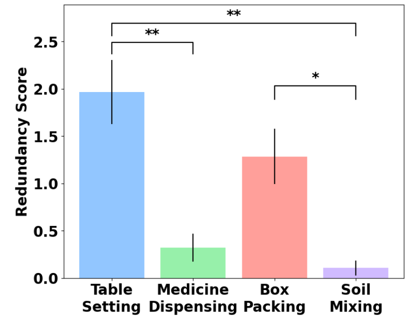
[3]The videos employed in this study can be found at https://sites.google.com/view/moormanetal-rss2023.



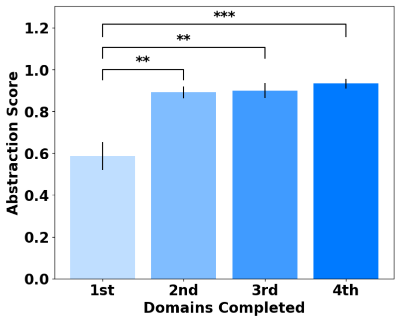
(a) Abstraction score does not significantly differ across domains.



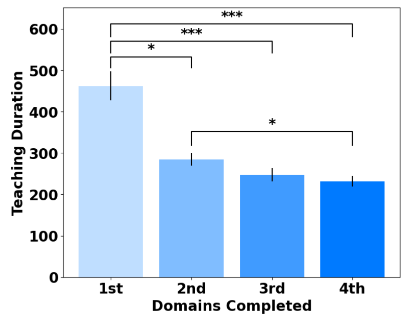
(b) Teaching duration differs across the different domains.



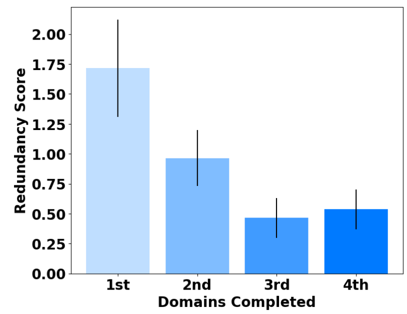
(c) Redundancy score differs across the different domains.



(d) Abstraction score increases with more domain experience.



(e) Teaching duration decreases with more domain experience.



(f) Redundancy score decreases overall with more domain experience.

Fig. 3: We depict results with respect to domain type (top row) and domain experience (bottom row).

the domain.

Between each domain, the experimenter reset the environment, placing the proper domain's items on the table. After experiencing all four domains, participants take the post-survey questionnaire, comprised of surveys to collect perceived workload and impressions of the robotic agent. Finally, participants answer the post-interview questions.

For each demonstration saved via the interface, we record the robot trajectories along with a third-person perspective video of the participant moving the robot, collected using a Kinect camera. While participants record their sub-tasks, the experimenter takes detailed notes on participant behavior, recording which sub-tasks they record. These notes, along with the interface's saved recipes (i.e., an ordered list of sub-tasks applied to each task in each domain), were used to obtain the abstraction score and redundancy score for each domain. Three coders scored participant abstraction and redundancy scores, resulting in an intra-class correlation coefficient of 0.998 for abstraction scores and 0.755 for redundancy scores.

IV. Results

We conducted our study with 28 participants (39.26% female, mean age = 22.89, standard deviation = 1.63). Before running statistical tests, we first checked that our data met parametric assumptions via Shapiro-Wilk's test and Levene's test. Due to our statistical models not passing tests for normality, we employ non-parametric tests throughout our analysis. We employ Bonferroni correction when applying multiple tests for the same hypothesis to reduce the risk of Type I errors [46]. To test RQ1 and RQ4 we employ the Friedman rank sum test, where we report χ^2(degree of freedom) and p-value. For follow-up pairwise comparisons, we employ the Nemenyi Wilcoxon-Wilcox all-pairs test, for which we report the p-value. To test RQ2 and RQ3 we employ Spearman's rank correlation test, where we report ρ and p-value.

Research Question 1

We first study the impact of domain experience on the quality of demonstrations. This hypothesis investigates whether participants can perform zero-shot transfer of knowledge regarding sub-task abstraction, teaching efficiency, and sub-task redundancy to novel domains.

We note that the block touching domain was the demo task, intended to familiarize the participant with the robot and the interface, to isolate the effect of learning in the actual test rounds. As the participants do not observe the optimal demonstration after this demo task, we do not include the block-touching domain in our domain experience.

Abstraction Score – Through a Friedman test, we find a main effect of the participant's domain experience on the participant's domain abstraction score ($\chi^2(3) = 28.056, p < .001$). We conduct pairwise comparisons using a Nemenyi Wilcoxon-Wilcox all-pairs test, visualized in Figure 3d, and find significance between the first and second domain ($p = .006$), the first and third domain ($p = .001$), and the first and fourth domain experienced ($p < .001$).

We first observe in Figure 3d that abstraction scores improve between the first and second domains. This finding points to participants' ability to transfer knowledge about sub-task abstraction zero-shot to a novel domain. We further observe that abstraction scores improve between the first domain and all subsequent domains. This finding supports our hypothesis that participants improve the level of abstraction of their demonstrations as they gain domain experience.

Our results show that abstraction scores, on average, are monotonically increasing. While the statistically significant improvement in abstraction score occurs after the first domain, the results show a positive trend in subsequent rounds. The diminishing but positive improvement is consistent with prior work finding that human task performance improves logarithmically with practice [43].

Teaching Duration – We find significance with respect to teaching duration and domain experience ($\chi^2(3) = 41.796, p < .001$). We find the significant pairs (Figure 3e) to be between the first and second domain ($p = .014$), the first and third domain ($p < .001$), and the first and fourth domain ($p < .001$), as well as between the second and fourth domain ($p = .041$). This finding indicates that participants provide demonstrations more efficiently over time.

Sub-task Redundancy– We find a main effect with respect to learning experience and sub-task redundancy ($\chi^2(3) = 8.018, p = .046$), but find no pairwise significance, (Figure 3f). This finding suggests that there may be a trend between domain experience and sub-task redundancy, but more data are needed.

Research Question 2

We investigate the effect of demonstration sub-task abstraction on participants' perceived workload.

Sub-task Count – We perform a Spearman's correlation test and find significance between sub-task count and perceived workload ($\rho = -.519, p = .005$). These findings imply that sub-task count is negatively correlated with perceived workload. High sub-task count means breaking up the task into many smaller sub-tasks, each of which can be reused to avoid redundant demonstrations. One possible explanation of this finding is that fewer sub-tasks for a task indicate more repetitive demonstrations.

Research Question 3

We investigate whether participant demographics impact the quality of demonstrations.

Teaching Experience – We find significance between prior teaching experience and sub-task count ($\rho = -.473, p = .011$). This finding is evidence that increased prior teaching experience is negatively correlated with sub-task count. This gained understanding of the impact of prior teaching experience on sub-task count could be used to improve the existing curriculum designed to teach demonstrators how to provide sufficient demonstrations.

Likeability – We find significance between sub-task count and robot likeability ($\rho = -.501, p = .007$). This finding is evidence that increased robot likeability is negatively correlated with sub-task count.

Agreeableness – Next, we find significance between teaching duration and the agreeableness sub-scale of the Big Five Personality survey ($\rho = .503, p = .006$). This finding is evidence that participant agreeableness is negatively correlated with the efficiency with which they provide demonstrations, namely that more agreeable participants utilize more time to provide demonstrations.

Negative Social Influence – Finally, we find significance between teaching duration and the negative social influence sub-scale of the Negative Attitude towards Robotics survey ($\rho = .577, p = .001$). This finding is evidence that higher teaching duration is correlated to perceptions of negative robot social influence, i.e., participants that are warier of robots take more time to provide demonstrations.

Research Question 4

We now investigate whether domain type impacts the quality of demonstrations.

Teaching Duration – Through a Friedman rank sum test, we find significance in the teaching duration among domains ($\chi^2(3) = 8.656, p = .034$). We find one significant pair between table setting and box packing domains ($p = .036$). We plot domain type against teaching time, as seen in Figure 3b.

Sub-task Redundancy – Through a Friedman rank sum test, we find significance in the redundancy score among domains ($\chi^2(3) = 33.836, p < .001$). We find the significant pairs to be between table setting and medicine dispensing ($p = .006$), table setting and soil mixing ($p < .001$), and box packing and soil mixing ($p = .023$) (Figure 3c).

Sub-task Count – Through a Friedman rank sum test, we find significance in the unique sub-task count among domains ($\chi(3)^2 = 45.87, p < .001$). A Nemenyi-Wilcoxon-Wilcox all-pairs test yields significant pairs for table setting and box packing ($p < .001$), table setting and medicine dispensing ($p = .006$), and table setting and soil mixing ($p < .001$).

Abstraction Score – Finally, we note that we find no significance between the abstraction score and domain, as seen in Figure 3a.

V. Related Works

In this section, we discuss relevant prior work in robot learning from human demonstrators and hierarchical task representations to further contextualize and motivate our results prior to the discussion.

Learning from Human Demonstrators

The field of LfD explores how human demonstrations can be used to teach robots new skills [14, 44, 8, 41, 13, 27]. LfD enables the agent to learn from a small set of examples, i.e., demonstrations provided by a teacher rather than learning from lengthy exploration, i.e., experience collected in an environment [6]. The mode of demonstration collection depends on whether the LfD algorithm aims to model the human feedback

[32], latent reward function [1, 18, 52], or unknown robot policy directly [27]. Additional LfD design decisions include accounting for prior robotics experience, demonstrator sub-optimality, and demonstration heterogeneity [48, 47, 13, 9, 40]. **In this paper, we evaluate how well non-experts can teach robots kinesthetically without explicitly being taught by domain experts.**

Much prior work in LfD has focused on enabling robots to perform short-horizon skills [28, 33, 22, 37]. There has been a lack of approaches using LfD to train robots to perform long-horizon tasks and multi-tasks, which would require the robot to accomplish a series of shorter sub-tasks. LfD approaches to multi-task learning are often expensive since demonstrators need to be taught how to provide demonstrations for each of the tasks required [16]. We show that demonstrators can generalize knowledge about providing sufficient and necessary demonstrations to novel tasks. **These findings suggest that demonstrators do not need to be taught how to teach in each domain explicitly, making long-horizon, multi-task LfD more tractable.**

Hierarchical Task Representations

Prior work has investigated how the hierarchical nature of a task can facilitate long-horizon task completion [2, 35, 38]. Breaking up a long-horizon task hierarchically and abstracting the task components into repurposable sub-tasks reduces planning depth and allows for faster planning [24, 29]. This representation of the task affords the agent the ability to adapt to novel environments that share features with the distribution of environments previously experienced [17].

Currently, LfD demonstration collection requires multiple demonstrations for each possible configuration of a multi-task setup. One way to make this process more tractable is for demonstrators to break up the task into a series of shorter abstractions that the robot could reuse for multiple configurations within the same domain. Prior work has shown that users can teach robots using task abstractions [3, 39]. Akgun et al. found that demonstrators can teach abstractions for robot end-effector movement [3], and Mohseni-Kabir et al. found that participants could teach the robot novel sub-tasks using previously taught sub-tasks as building blocks [39]. However, in most prior work establishing a human demonstrator's ability to provide usable demonstrations that contain abstractions, the participants are shown precisely how to teach the robot. They are then asked to reproduce the method of robot teaching that was prescribed. Cakmak et al. compare written and video demonstrator instruction, and find that trial and error plays a large role in the learning process [10]; we note that these demonstrators learn and are evaluated on the same task. Teaching a robot using abstractions without this guidance is not intuitive to non-experts [31]. Gopalan et al. find that the majority of participants are unable to provide sufficiently abstracted demonstrations naturally, and find that, even when told to employ abstractions, demonstrators struggle to provide demonstrations robust to minor changes in the task specification, such as item multiplicity or item location [21]. In this work, we investigate whether demonstrators' ability to provide sufficient sub-task abstractions improves over time, as they practice providing demonstrations in multiple different domains. **Our findings support participants' ability to learn to provide sufficient sub-task abstractions in novel domains, with enough practice.**

There have been algorithmic approaches to learning tasks from user demonstrations without requiring the demonstrations to specify task abstractions [34, 30, 15]. However, these approaches require the collection of demonstration datasets that would not be scalable for multi-task settings. In order to investigate the scalability of training non-expert demonstrators, we additionally investigate whether participants can generalize knowledge about providing demonstrations zero-shot, to novel domains. **Our findings support that, rather than training demonstrators in each domain they encounter, experimenters could train demonstrators in a handful of training domains, for demonstrators to generalize this training to novel domains.**

VI. Discussion

In this work, we studied the effect of domain experience and participant demographics on the quality of LfD demonstrations. Previous works have studied short horizon skills [28, 33, 22, 37] and teaching robots tasks using experts [21, 39, 3, 14, 44, 8, 41, 13, 27]. In our work we allow users to refine their teaching strategies through demonstrator training. We then show that they can effectively teach the robot in new domains where they do not receive expert guidance. In this section, we give perspective over our results. We go over prominent results relating to RQ1, RQ2, and RQ3.

Impact of domain experience on demonstration sufficiency, necessity, and efficiency (RQ1).

We find that participant abstraction score is positively impacted by the number of domains experienced ($p < .001$), meaning that over time participants provide demonstrations that manifest higher levels of abstraction. We further find that teaching duration is negatively impacted by the number of domains experienced ($p < .001$). This indicates that over time participants take less time to provide demonstrations.

These findings suggest that participants can generalize knowledge gained about providing demonstrations efficiently, using more sub-task abstraction, from previously experienced domains to a novel domain. **These findings indicate that demonstrators can be trained to efficiently provide sufficient demonstrations to new domains, zero-shot.**

Impact of prior teaching experience on sub-task count (RQ3).

We find that prior teaching experience is negatively correlated with sub-task count ($p = .011$), indicating that participants with more teaching experience record fewer sub-tasks. We note that we don't find significance between prior teaching experience and abstraction score or redundancy score. **This finding indicates that increasing teaching experience will**

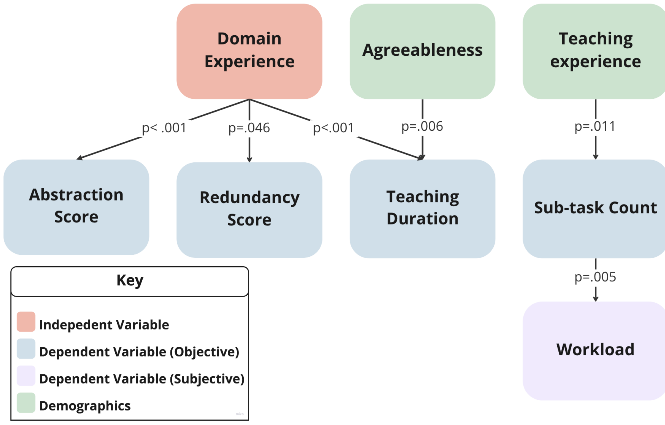


Fig. 4: Depicted is a summary of our significant results.

increase sub-task efficiency, though not at the expense of sub-task sufficiency or necessity.

Since general teaching experience does not appear to translate to demonstration quality, our findings highlight the need for a way to teach demonstrators how to provide sufficient and necessary sub-tasks. Our results show that we contribute a scalable and generalizable method for training LfD demonstrators, by exposing demonstrators to multiple domains in which they practice and observe the optimal teaching method.

Impact of sub-task count on perceived workload (RQ2).

We find that participant workload is negatively correlated with their sub-task count ($p = .005$). This indicates that a lower sub-task count correlated with a higher perceived workload. We hypothesize that this is due to the lengthier process of demonstrating and recording under-abstracted sub-tasks. **In addition to abstractions being useful for robust robot learning, this finding suggests that participants find correct abstractions less effort to teach, as observed via lower perceived workload.**

Impact of robot likeability, participant agreeableness, and negative attitudes on demonstrations (RQ3).

We find that robot likeability is negatively correlated with sub-task count ($p = .007$). **This suggests that people rated the robot as more likeable when the teaching was less involved.**

On the other hand, we find that participant agreeableness is positively correlated with teaching duration ($p = .006$). This finding suggests that demonstrators with higher agreeableness take longer when providing demonstrations, though not at the expense of sub-task count, abstraction score, or redundancy. **This finding indicates that more agreeable demonstrators take their time when recording demonstrations.** We posit this is because these participants either wanted to please the experimenter or because they wanted to be thorough in order to be helpful.

Participants that perceived robots as more socially negative additionally took longer to teach the robot ($p = .001$). **Participants that are warier of robots take more time to provide demonstrations**, therefore we posit that addressing negative robot perceptions will reduce the time people take to teach robots.

Limitations and Future Work

A limitation of our work is that our participants are composed primarily of college students. We additionally report limitations in our study design. Firstly, in this work participants do not observe the learned robot behavior (resulting from their demonstrations). This absence of observing the subsequent consequences on the environment of their demonstrations may have reduced participant urgency and desire to improve. Secondly, we note the possible impact of the experimenter expectancy effect on participant behavior, since the experimenter took notes on the participant's demonstrations throughout the study.

As our abstraction scoring method is taken from prior work which validates it on a real robot system [21], we do not validate the scoring method again in this study. In future work, we propose to further validate our abstraction score, redundancy score, and original subjective surveys. We additionally propose to explore demonstrators' ability to perform the zero-shot transfer, with regards to providing useful demonstrations, to novel domains when providing demonstrations that operate under different sub-task specifications.

VII. Conclusion

Learning from demonstration enables non-expert end-users to be involved in robot learning. However, providing usable demonstrations is not intuitive to most demonstrators. Demonstrators have to be trained in order to provide demonstrations that would be usable, and this training is often domain-specific. Instead of teaching demonstrators how to provide sufficient demonstrations in all possible domains, we propose to teach demonstrators such that what they learn can generalize to novel domains. In this work, we study the impact of experience providing demonstrations across multiple domains on the quality of demonstrations for LfD. We find that as participants gain domain experience they are able to generalize knowledge about sub-task abstraction ($p < .001$), teaching efficiency ($p < .001$), and sub-task redundancy ($p < .05$) zero shot, to novel task domains. We show that with a few hours of training, we can teach human demonstrators to provide sufficient, necessary, and efficient demonstrations in novel domains.

Acknowledgments

This work was supported by a grant from the National Science Foundation (IIS-2112633) and a gift from Konica Minolta, Inc.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] Bernardo Aceituno and Alberto Rodriguez. A hierarchical framework for long horizon planning of object-contact trajectories. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 189–196. IEEE, 2022.

[3] Baris Akgun, Maya Cakmak, Karl Jiang, and Andrea L Thomaz. Keyframe-based learning from demonstration. *International Journal of Social Robotics*, 4(4):343–355, 2012.

[4] Rachid Alami, Thierry Siméon, and Jean-Paul Laumond. A geometrical approach to planning manipulation tasks. the case of discrete placements and grasps. In *International Symposium on Robotics Research*, 1991.

[5] Rachid Alami, Jean-Paul Laumond, and Thierry Siméon. Two manipulation planning algorithms. In *Proceedings of the Workshop on Algorithmic Foundations of Robotics*, 1995.

[6] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57 (5):469–483, 2009.

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81, 2009.

[8] Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. Robot programming by demonstration. In *Springer handbook of robotics*, pages 1371–1394. Springer, 2008.

[9] Daniel S. Brown, Wonjoon Goo, and Scott Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 330–359. PMLR, 30 Oct–01 Nov 2020. URL https://proceedings.mlr.press/v100/brown20a.html.

[10] M. Cakmak and L. Takayama. Teaching people how to teach robots: The effect of instructional materials and dialog design. *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 431–438, 2014.

[11] Alexandre Campeau-Lecours, Véronique Maheu, Sébastien Lepage, Hugo Lamontagne, Simon Latour, Laurie Paquet, and Neil Hardie. Jaco assistive robotic device: Empowering people with disabilities through innovative algorithms. In *Rehabilitation Engineering and Assistive Technology Society of North America (RESNA) Annual Conference*, 2016.

[12] Rich Caruana. *Multitask learning*. Springer, 1998.

[13] Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conference on robot learning*, pages 1262–1277. PMLR, 2021.

[14] Sonia Chernova and Andrea L Thomaz. *Robot Learning from Human Demonstration*. Springer, 2014.

[15] Luis C. Cobo, Charles L. Isbell, and Andrea L. Thomaz. Automatic task decomposition and state abstraction from demonstration. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '12, page 483–490, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0981738117.

[16] Tesca Fitzgerald and Andrea L Thomaz. Skill demonstration transfer for learning from demonstration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, pages 187–188, 2015.

[17] Tesca Fitzgerald, Ashok Goel, and Andrea Thomaz. Abstraction in data-sparse task transfer. *Artificial Intelligence*, 300:103551, 2021.

[18] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

[19] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual review of control, robotics, and autonomous systems*, 4:265–293, 2021.

[20] Lewis R Goldberg. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1): 26, 1992.

[21] Nakul Gopalan, Nina Moorman, Manisha Natarajan, and Matthew Gombolay. Negative result for learning from demonstration: Challenges for end-users teaching robots with task and motion planning abstractions. In *Proceedings of Robotics: Science and Systems.*, 2022.

[22] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. *CoRL*, 2022.

[23] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[24] Valentin N Hartmann, Ozgur S Oguz, Danny Driess, Marc Toussaint, and Achim Menges. Robust task and motion planning for long-horizon architectural construction planning. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6886–6893. IEEE, 2020.

[25] Kris Hauser and Jean-Claude Latombe. Multi-modal motion planning in non-expansive spaces. *The International Journal of Robotics Research*, 29(7), 2010.

[26] Kris Hauser and Victor Ng-Thow-Hing. Randomized multi-modal motion planning for a humanoid robot ma-

nipulation task. *The International Journal of Robotics Research*, 30(6), 2011.

[27] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

[28] Auke Jan Ijspeert, Jun Nakanishi, Heiko Hoffmann, Peter Pastor, and Stefan Schaal. Dynamical movement primitives: learning attractor models for motor behaviors. *Neural computation*, 25(2):328–373, 2013.

[29] Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical planning in the now. In *Workshops at the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[30] Wonchul Kim, Chungkeun Lee, and H. Jin Kim. Learning and generalization of dynamic movement primitives by hierarchical deep reinforcement learning from demonstration. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3117–3123, 2018. doi: 10.1109/IROS.2018.8594476.

[31] Moritz Knaust and Dorothea Koert. Guided robot skill learning: A user-study on learning probabilistic movement primitives with non-experts. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, pages 514–521. IEEE, 2021.

[32] W. B. Knox and P. Stone. Interactively shaping agents via human reinforcement: the tamer framework. In *K-CAP '09*, 2009.

[33] George Konidaris and Andrew Barto. Skill discovery in continuous reinforcement learning domains using skill chaining. *Advances in neural information processing systems*, 22:1015–1023, 2009.

[34] George Konidaris, Scott Kuindersma, Roderic Grupen, and Andrew Barto. Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, 31(3):360–375, 2012. doi: 10.1177/0278364911428653.

[35] Sanjay Krishnan, Animesh Garg, Richard Liaw, Lauren Miller, Florian T Pokorny, and Ken Goldberg. Hirl: Hierarchical inverse reinforcement learning for long-horizon tasks with delayed rewards. *arXiv preprint arXiv:1604.06508*, 2016.

[36] Clemente Lauretti, Francesca Cordella, Eugenio Guglielmelli, and Loredana Zollo. Learning by demonstration for planning activities of daily living in rehabilitation and assistive robotics. *IEEE Robotics and Automation Letters*, 2(3):1375–1382, 2017.

[37] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[38] Boyao Li, Jiayi Li, Tao Lu, Yinghao Cai, and Shuo Wang. Hierarchical learning from demonstrations for long-horizon tasks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4545–4551. IEEE, 2021.

[39] Anahita Mohseni-Kabir, Charles Rich, Sonia Chernova, Candace L Sidner, and Daniel Miller. Interactive hierarchical task learning from a single demonstration. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, pages 205–212, 2015.

[40] Vivek Myers, Erdem Biyik, Nima Anari, and Dorsa Sadigh. Learning multimodal rewards from rankings. In *Conference on Robot Learning*, 2021.

[41] Kai Ploeger, Michael Lutter, and Jan Peters. High acceleration reinforcement learning for real-world juggling with binary rewards. In *Conference on Robot Learning*, pages 642–653. PMLR, 2021.

[42] Harish Ravichandar, Athanasios S Polydoros, Sonia Chernova, and Aude Billard. Recent advances in robot learning from demonstration. *Annual review of control, robotics, and autonomous systems*, 3:297–330, 2020.

[43] Frank Ritter and Lael Schooler. The learning curve. In *International encyclopedia of the social and behavioral sciences.*, 2001.

[44] Stefan Schaal. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, 3(6):233–242, 1999.

[45] Mariah L Schrum, Erin Hedlund, and Matthew C Gombolay. Improving robot-centric learning from demonstration via personalized embeddings. *arXiv preprint arXiv:2110.03134*, 2021.

[46] Mariah L. Schrum, Muyleng Ghuy, Erin Hedlund-Botti, Manisha Natarajan, Michael J. Johnson, and Matthew C. Gombolay. Concerning trends in likert scale usage in human-robot interaction: Towards improving best practices. *J. Hum.-Robot Interact.*, nov 2022. doi: 10.1145/3572784. URL https://doi.org/10.1145/3572784.

[47] Mariah L Schrum, Erin Hedlund-Botti, and Matthew Gombolay. Reciprocal mind meld: Improving learning from demonstration via personalized, reciprocal teaching. In *6th Annual Conference on Robot Learning*, 2022.

[48] Mariah L Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C Gombolay. Mind meld: Personalized meta-learning for robot-centric imitation learning. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 157–165. IEEE, 2022.

[49] Dag Sverre Syrdal, Kerstin Dautenhahn, Kheng Lee Koay, and Michael L Walters. The negative attitudes towards robots scale and reactions to robot behaviour in a live human-robot interaction study. *Adaptive and emergent behaviour and complex systems*, 2009.

[50] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5(1):30–43, 2018.

[51] Zuyuan Zhu and Huosheng Hu. Robot learning from demonstration in robotic assembly: A survey. *Robotics*, 7(2):17, 2018.

[52] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.