How do Human Users Teach a Continual Learning Robot in Repeated Interactions?

Ali Ayub[1*], Jainish Mehta[1], Zachary De Francesco[1], Patrick Holthaus[2], Kerstin Dautenhahn[1] and Chrystopher L. Nehaniv[1]

***Abstract*— Continual learning (CL) has emerged as an important avenue of research in recent years, at the intersection of Machine Learning (ML) and Human-Robot Interaction (HRI), to allow robots to continually learn in their environments over long-term interactions with humans. Most research in continual learning, however, has been *robot-centered* to develop continual learning algorithms that can quickly learn new information on static datasets. In this paper, we take a *human-centered* approach to continual learning, to understand how humans teach continual learning robots over the long term and if there are variations in their teaching styles. We conducted an in-person study with 40 participants that interacted with a continual learning robot in 200 sessions. In this between-participant study, we used two different CL models deployed on a Fetch mobile manipulator robot. An extensive qualitative and quantitative analysis of the data collected in the study shows that there is significant variation among the teaching styles of individual users indicating the need for personalized adaptation to their distinct teaching styles. The results also show that although there is a difference in the teaching styles between expert and non-expert users, the style does not have an effect on the performance of the continual learning robot. Finally, our analysis shows that the constrained experimental setups that have been widely used to test most continual learning techniques are not adequate, as real users interact with and teach continual learning robots in a variety of ways.**

I. Introduction

We envision a future of general-purpose assistive robots that can help users with a variety of tasks in dynamic environments, such as homes, offices, shopping malls, etc. It would be necessary that such assistive robots are personalized to their users' needs and their environments [1]. However, over the long term, users' needs, preferences, and their environments will continue to change, which makes it impossible to pre-program the robot with all the tasks it might be required to perform. A solution to this problem is to allow people to continually teach their robots new tasks and changes in their environments on the fly, an approach known as continual learning (CL) [2], [3].

Continual learning has been extensively studied in recent years to allow robots to learn over long periods of time [3], [4]. As it is imperative for a robot to learn the objects in its environment, the majority of research on CL has focused on machine learning (ML) models for object recognition in recent years [4], [5]. Most of these techniques were tested on static object recognition datasets with a large number of training images for each object class. In real-world environments, however, robots will need to learn from individual interactions with their users who might be unwilling to provide a large number of training examples for each object.

In the past few years, robotics researchers developed CL techniques that can learn from only a few training examples per object, an approach known as Few-Shot Class Incremental Learning (FSCIL) [3], [6], [7]. Although FSCIL techniques produced promising results on real robots, they were only tested with systematically collected datasets by their experimenters. Overall, most research in continual learning has been *robot-centered*, to develop efficient CL algorithms that can learn from static datasets or interaction with robot experimenters. However, in the real world, robots will learn from real users who might be unfamiliar with robot programming and learning. Therefore, an equally important area of research in continual learning is *human-centered*, to understand how human users interact with and teach continual learning robots over the long term. To the best of our knowledge, we know of no other work on developing long-term user studies where human users teach modern CL models deployed on robots over multiple interactions.

In this paper, we have a human-centered focus to uncover the diversity and evolution of human teaching when interacting with a continual learning robot over repeated sessions. We developed a CL system that integrates a graphical user interface (GUI) with CL models of object learning deployed on the Fetch mobile manipulator robot [8]. We conducted a long-term between-participant study (N=40) where participants interacted with and taught everyday household objects to a Fetch robot that used two different CL models. We analyzed the data collected in the study to characterize various aspects of human teaching of a continual learning robot in an unconstrained manner. Our results highlight the variation in the teaching styles of different users, as well as the influence of the robot's performance on users' teaching styles over multiple sessions. Our results indicate that the constrained experimental setups traditionally used to test most CL models are inadequate, as real users teach continual learning robots in a variety of ways.

II. Related Work

In this section, we first present an overview of modern CL methods mostly tested without human users, and then

This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

[1]University of Waterloo, Waterloo, Ontario N2L 3G1, Canada
{*a9ayub, jm3mehta, zdefrancesco, cnehaniv, kdautenh} @uwaterloo.ca
[2]University of Hertfordshire, Hertfordshire AL10 9AB, England, UK
p.holthaus@herts.ac.uk

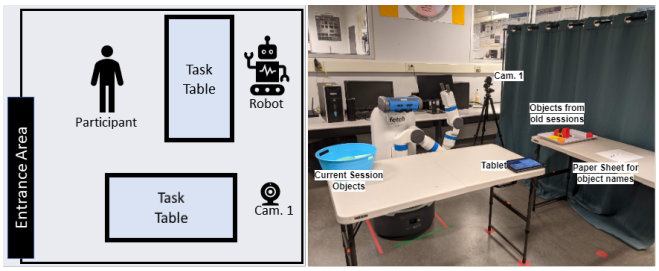


Fig. 1: (Left) Experimental layout for the CL setup with the participant and the robot. (Right) Corresponding real-world setup.

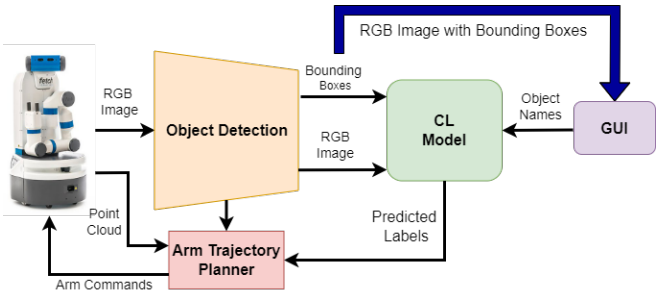


Fig. 2: Our complete CL system. Processed RGB images from the robot's camera are sent to the GUI for transparency and also passed on to the CL Model. The user sends object names to the CL model either for training the CL model or finding an object. The arm trajectory planner takes point cloud data, processed RGB data, and predicted object labels from the CL model as input and sends the arm trajectory for the Fetch robot to point to the object.

introduce current approaches to robot teaching, highlighting the need for a human-centered approach at the intersection of CL and human-robot interaction (HRI).

A. Continual Learning

The goal of CL models is to continuously adapt and learn new information over time while preserving past knowledge. Most research in the CL literature has focused on class-incremental learning (CIL) in which a machine learning model learns from labeled training data of different classes in each increment and is then tested on all the classes it has learned so far [4]. One of the main problems faced by class-incremental learning models is *catastrophic forgetting*, in which the model completely forgets the previously learned classes when learning new classes in an increment [9]. Various research directions have been pursued in the past to tackle the catastrophic forgetting problem, such as replay-based techniques that store and replay data of the old classes when learning new classes [4], [10], regularization techniques [11], [12], and generative replay based techniques that generate old data using stored class statistics [13], [14]. These techniques, however, are not suitable for learning from human users who might be unwilling to provide hundreds or thousands of images per object class.

In the past couple of years, researchers also developed class-incremental learning models that can learn from only a few labeled examples per class, a direction known as few-shot class incremental learning (FSCIL) [15]. However, CIL and FSCIL approaches were either tested on static datasets, or on data captured by a robot while interacting with experimenters in systematically controlled setups [2], [3], [15]. To the best of our knowledge, all of the FSCIL approaches were robot-centered and none of these approaches were tested with actual participants (users).

B. Human-Robot Teaching

Human-centered research for robot learning through HRI has been limited. A few user studies have been conducted in the past with simulated and real robots to understand the characteristics of human teaching. Most of these studies were conducted in Wizard of Oz setups where the robot did not learn from human teaching [16], [17]. Some research has been conducted on interactive reinforcement learning through HRI for learning manipulation tasks through physical human corrections, learning kitchen-related tasks in simulation, or learning natural language description of images from humans [18]–[20]. However, most of these studies were designed to test the performance of the reinforcement learning models or understand the perceptions of users towards these models and were not focused on understanding patterns of human teaching. Furthermore, these studies were only tested in a single interaction with users. However, for continual learning robots, it is imperative to design multi-session studies to understand how human teaching of continual learning robots evolves over the long term. In contrast to prior work, to the best of our knowledge, we conducted the first long-term user study at the intersection of continual machine learning and HRI, to understand patterns of human teaching with a continual learning robot over multiple interactions.

III. Method

We investigated human teaching patterns when interacting with a continual learning robot to teach an object recognition task. The subsections below describe our CL system and the method for our long-term study.

A. Continual Learning System

In this experiment, in each session, the user taught the robot household objects in a table-top environment and then tested the robot to find and point to the requested object on the table. Figure 1 shows the table-top experimental setup for this study. The simplicity of the setup and the task makes it clear what the user should do to teach the robot different objects, and what the robot should do to find the learned objects during the testing phase.

For this setup, we developed a CL system for the object recognition task, which integrates CL models with a Fetch mobile manipulator robot [8], as well as a graphical user interface (GUI) for interactive and transparent learning from human users. Figure 2 shows our system for the object recognition task. In this system, the user interacts with the robot through the GUI on an Android tablet (Figure 3). The user provides labels of new objects placed in front of the robot through the GUI and saves the images of objects processed through the object detection module in the robot's

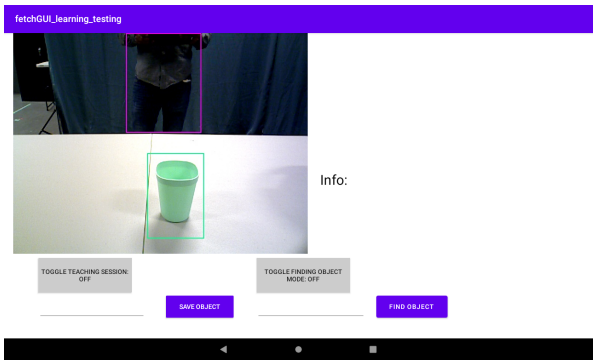


Fig. 3: The graphical user interface (GUI) used to interact with the robot. The RGB camera output with bounding boxes is on the top left. The buttons at the bottom can be used to teach objects to the robot and ask it to find objects in the testing phase. The top right of the GUI shows information sent by the robot to the user.

memory. The robot then uses the saved object images in each session to train the CL model. After teaching, the user can test the robot by asking it to find objects on the table through the GUI. The robot passes the pre-processed images to the CL model to get the predicted object labels. If the object requested by the user is found, the robot finds the 3D location of the object on the table and points to the object using its arm.

1) Continual Learning Models: We consider two CL models in this study. For the first model, we consider a naïve finetuning (FT) approach [4] in which a convolutional neural network (CNN) [21] is trained on the image data of the object classes in each increment (i.e. in an interactive session with the user). The model does not train on any of the objects learned in the previous increments (sessions) and therefore it forgets the previously learned objects. This model can serve as a baseline for forgetting in continual learning [4], [10].

For the second model, we consider a state-of-the-art CL approach specifically designed for FSCIL in robotics applications [3]. This approach, termed centroid-based concept learning (CBCL), mitigates forgetting by creating separate clusters for different object classes. CBCL stores cluster centroids of object classes in memory and uses these centroids to make predictions about labels of new objects. More details about these models can be found in [3], [4]. Note that all of these models were only tested on systematically collected object datasets in prior work, and have never been tested in real-time with human participants.

B. Participants

We recruited 40 participants (19 female (F); 21 male (M), all students) from the University of Waterloo, between the ages of 18 and 37 years ($M = 23.48$, $SD = 4.49$). 20 participants (ages: $M = 24.15$, $SD = 4.21$, 10 F, 10 M) were randomly assigned to *FT* condition, and the other 20 (ages: $M = 22.78$, $SD = 4.68$, 9 F, 11 M) were randomly assigned to *CBCL* condition. The participants had diverse backgrounds in terms of their majors, but most of them (65%) were engineering and computer science students. Based on their self-assessments in a pre-experiment survey, 40% of the participants reported that they were familiar with robot programming, 55% reported that they had previously interacted with a robot, 5% were familiar with the Fetch robot, and 10% had previously participated in an HRI study. For the remainder of the paper, we will call participants with prior robot programming experience 'experts' and the rest of the participants 'non-experts'. All procedures were approved by the University of Waterloo Human Research Ethics Board.

C. Research Questions

We analyze the data collected in our study to answer the following research questions and test the associated hypotheses:

RQ1 How do different human users label objects when teaching a continual learning robot over multiple sessions?

H1.1 Labelling strategies for objects vary among different users.

RQ2 Does the continual learning robot's performance affect the way users teach over multiple sessions?

H2.1 Classification performance of the robot affects the teaching style of the participants over multiple sessions.

H2.2 Users teach a robot that forgets previous objects differently than a robot that remembers previous objects.

RQ3 Do users change the way they teach the continual learning robot over multiple sessions?

H3.1 Teaching styles of users change over multiple sessions regardless of the CL model.

RQ4 Is there a difference in teaching style and robot performance for expert and non-expert users?

H4.1 Continual learning robots taught by expert users perform better than the ones taught by non-expert users.

H4.2 There is a difference between the teaching styles of expert and non-expert users.

D. Procedure

We conducted five repeat sessions (each lasting ∼20-30 minutes) with each participant in a robotics laboratory. All sessions were video recorded. We also stored the image data of the objects taught and tested by the participants. Each participant was randomly assigned to one of the two experimental conditions using one of the two CL models, CBCL and FT. Before their first session, each participant was asked to complete a consent form and a pre-experiment survey online. After completing the consent form and the pre-experiment survey, the experimenter greeted the participant and gave a brief oral introduction to the experiment. The participant then interacted with the robot in a demo session to understand how to teach and test the robot. In the demo phase, the robot did not learn any objects.

During the demo phase, the experimenter explained to the participant how to start a teaching session using the GUI, teach an object to the robot, and test the robot to find the object. The participant then tried teaching a demo object (this object was not used later) to the robot. The participant then

Fig. 4: The twenty-five objects used in our study.

tested the robot to find the demo object on the table using the GUI. After the demo phase (∼5 minutes), the experimenter gave a paper sheet, which served as a memory aid, to the participant to write down the names of the objects of the current session. In this way the participants could remember the object names when they needed the robot to find these objects in the next sessions. The experimenter then took the tablet from the participant and loaded the program for the actual session on the tablet. The experimenter handed the tablet back to the participant and placed five objects to be taught in the session on one side of the table. The experimenter then mentioned to the participant that they can start their session and start teaching the five objects.

The experimenter then went to a secluded area and the participant taught and tested objects to the robot. At the end of the session, the experimenter came out of the secluded area and asked the participant to finish a post-experiment survey. The participant then scheduled their next session. In the next four sessions, the same procedure was repeated, except for replacing the objects to be taught between sessions. Figure 4 shows the 25 objects used in our study. Participants were also told that they can bring a maximum of two objects per session of their own choice in sessions 3-5 to teach to the robot. If participants brought their own objects, we replaced some of the objects from our set (Figure 4) with participants' objects (total objects taught over 5 sessions was still 25). Participants did not go through a demo interaction in the next four sessions. At the end of the last session, the experimenter asked the participant to have a short interview to answer some questions describing their experience with the robot. This interview was audio recorded. Analyses of the post-experiment survey and audio interview are not reported since they go beyond the scope of this paper, and will be reported in future publications. Examples of the teaching and testing phases are shown in the supplementary video.

E. Measures

We used both qualitative and quantitative measures to analyze the data for the two conditions. We analyzed the object names given by the participants to different objects using the image data stored for objects during teaching sessions. We report the variety and frequency of labels used by the participants for each object. We also coded the video recordings to calculate the frequency of teaching by the participants in all 5 sessions, and if they re-taught any objects to the robot in case the robot was not able to correctly find them on the table.

Object	No. of Different Labels	Most Common Label
Green Cup	10	Cup (59%)
Honey	13	Honey (46.5%)
Bowl	10	Bowl (65%)
Glue	6	Glue (76%)
Spoon	6	Spoon (81%)
Apple	3	Apple (90%)
Banana	3	Banana (90%)
Red Cup	14	Red Cup (25%)
Blue Marker	11	Marker (58%)
Orange	5	Orange (77%)
Mug	7	Mug (72%)
Fork	6	Fork (76%)
Sharpie	8	Sharpie (48%)
Plate	10	Plate (61%)
Stapler	6	Stapler (86%)
Book	4	Book (86%)
Red Marker	4	Red Marker (31%)
Blue Pen	7	Pen (60%)
Pepsi	7	Pepsi (54%)
White Bottle	8	Water Bottle (62%)
Coca Cola	8	Coke (36%)
Milk	7	Milk (77%)
Phone	5	Phone (68%)
7Up	12	7Up (44%)
Water Bottle	8	Water (47%)

TABLE I: The number of different labels given by the participants to all 25 objects in the study together with the most common label for each object with the percentage of participants that chose this label. Objects are ordered from top to bottom as they were taught in 5 sessions with 5 objects per session. Note that the first column shows some reference names for the objects to be able to identify them individually in the paper.

We also analyzed the performance of two CL approaches. Classification accuracy per session (increment) has been commonly used in the CL literature [4], [6] for quantifying the performance of CL models for object recognition tasks. Therefore, for each session, during the testing phase, we recorded the total number of objects tested by the participant and the total number of objects that were correctly found by the robot. Using this data, we calculated the accuracy $\mathcal{A}$ of the robot in each session as:

$$\mathcal{A} = \frac{number\ of\ objects\ correctly\ found}{number\ of\ objects\ tested} \quad (1)$$

We use the accuracy of the models to determine the teaching quality of the participants in each condition and over multiple sessions. Further, using the image data stored for the objects, we calculated the average number of times each object was taught by the participants in each session to determine the effort spent by the participants in teaching the robot. Finally, we analyze how the above-mentioned variables are affected by the sessions, choice of the CL model, and previous robot programming experience of the participants.

IV. Results

In this section we present the results of our analysis in terms of different labeling strategies and teaching styles of the participants. We also report the effect of participants' teaching styles on the robot's performance and vice versa.

	Accuracy	No. images	Teaching phases	Reteaching
Session number	**0.002 ****	0.819	0.975	0.639
CL model	**<0.0001 ******	0.181	0.644	0.4
Programming experience	0.328	**<0.0001 ******	**0.096 .**	0.772
Session number : CL model	**<0.001 *****	0.95	0.963	0.862
Session number : Programming experience	0.624	0.681	0.865	0.679
CL model : Programming experience	0.417	**<0.0001 ******	**0.071 .**	0.147
Session number : CL model : Programming experience	0.991	0.734	0.988	0.335

TABLE II: Results (p values) of the three-way ANOVA using session number, continual learning model, and previous programming experience as independent variables. Columns for the accuracy of the models, number of images per object, number of teaching phases, and reteaching misclassified objects show p values for the dependent variables. Significance levels ($. := p < 0.1; * := p < .05; ** := p < 0.01; *** := p < 0.001; **** := p < 0.0001$).

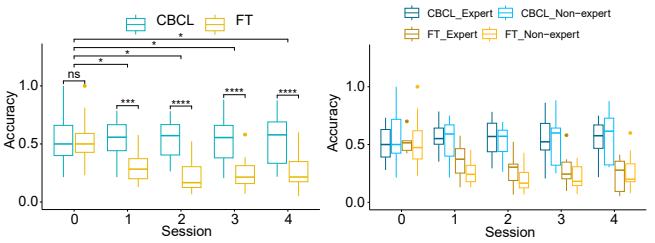
A. Object Labeling by Human Teachers

Table I shows the number of different labels given to the 25 objects by 40 participants in the study. To identify each object we add a generic name for each object in the table. For example, for the plastic apple used in our study, we identify it as an apple in the table. Overall, there was a significant variation in the labeling of objects by the participants, ranging from 3 (for Apple) to 14 (for Red Cup) different labels for the objects. Among such labels, some were quite simple and generic, such as *Honey*, *Bowl*, *Milk*, etc. whereas some were quite specific, such as *Almost Empty Yellow Honey Jar* , *Light Green Flat Bowl*, *Empty Milk Carton*, etc. We also report the most common label given to each object and the percentage of participants that chose that label. The consensus among the participants for labeling the objects varied from 25% for *Red Cup* to 90% for *Apple*.

We also noticed some unique labeling strategies by the participants. Some of the participants labeled different objects in different sessions using the same label. For example, multiple participants gave the label *Cup* to *Green Cup* in Session 1, *Red Cup* in Session 2, and *Mug* in Session 3. In total, 10 out of 40 participants (25%) gave the same label to at least two different objects. Further, some participants also gave multiple labels to the same objects. For example, one participant labeled *Milk* as both *Milk Box* and *Milk Pouch*. Overall, there were 7 out of 40 participants (17.5%) that gave more than one label to at least one object. Finally, we noticed that some participants gave labels that did not match the objects. For example, one participant named glue *Insert Stick Joke Here*, another participant named bowl *Plate*, and another participant named stapler *The Better Robot*.

B. Participants' Teaching Styles and Robot Performance

We performed a three-way ANOVA with three independent variables: the two conditions (CBCL, FT), session number, and previous robot programming experience of the participants. The ANOVA was performed to understand the effect of the three independent variables on the teaching style of the participants and the robot's performance in the testing phase. The dependent variables were the classification accuracy of the robot in the testing phases for 200 sessions, the average number of images per object shown by the participants in each session, the number of teaching phases started by the participants in each session, and the number of times participants retaught misclassified objects in each session.



(a) Accuracy of models by session.

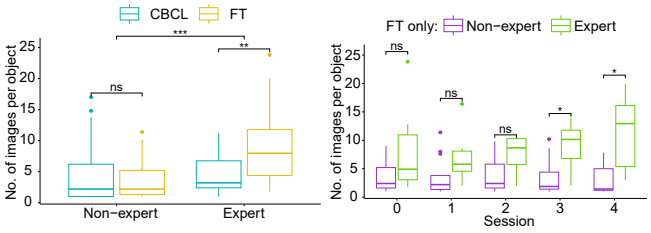
(b) Accuracy of models for experts and non-experts.

Fig. 5: Boxplots for *accuracy* for two conditions. Significance levels ($* := p < .05; *** := p < 0.001; **** := p < 0.0001$) are indicated on bars between columns.

Table II represents the p values and significance levels for the ANOVA. For classification accuracy, we see a significant effect based on the session number and the choice of the CL model (CBCL or FT condition), and the interaction between the session number and the CL model. For the number of images taught per object, we noticed a significant effect based on the previous programming experience of the participants, and the interaction between the CL model and the programming experience. For the number of teaching phases per session, we only saw a borderline effect by the programming experience and interaction between the CL model and the programming experience. Finally, reteaching of misclassified objects was not significantly affected by any of the independent variables.

For significant ANOVAs, we performed the post hoc Tukey HSD test. However, the data for sub-groups for some dependent variables were not normally distributed, therefore, we also performed the Wilcoxon rank sum test [22] with false discovery rate correction [23] for pairwise comparisons between sub-groups for each dependent variable.

1) Model Accuracy: As the data for classification accuracy was normally distributed, we performed the post hoc Tukey HSD test for significant ANOVAs. Figure 5 shows the average classification accuracy of the continual learning robot over five sessions. As displayed in Figure 5a, the accuracy is significantly affected by the choice of the CL model. For the first session, both models have similar



(a) Number of images per object for experts and non-experts, and two conditions.

(b) Number of images per object by session for experts and non-experts in FT condition.

Fig. 6: Boxplots for *number of images per object*. Significance levels ($* := p < .05; *** := p < 0.001; **** := p < 0.0001$) are indicated on bars between columns.

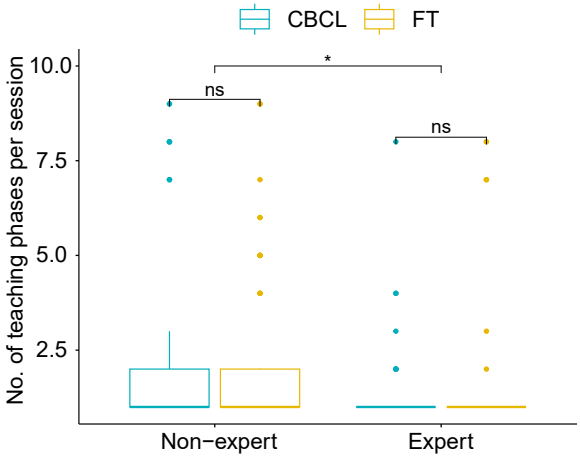


Fig. 7: Boxplot for *number of teaching phases per session* for experts and non-experts in two conditions. Significance levels ($* := p < .05$) are indicated on bars between columns.

accuracy ($\mu = 0.53$, $\sigma = 0.19$ for CBCL; $\mu = 0.52$, $\sigma = 0.18$ for FT). For the next four sessions, there is a statistically significant difference between the two models: when comparing CBCL ($\mu = 0.54$, $\sigma = 0.16$) to FT ($\mu = 0.29$, $\sigma = 0.13$) with $p = 0.0002$ for session 2, comparing CBCL ($\mu = 0.54$, $\sigma = 0.15$) to FT ($\mu = 0.22$, $\sigma = 0.13$) with $p < 0.0001$ for session 3, comparing CBCL ($\mu = 0.53$, $\sigma = 0.20$) to FT ($\mu = 0.24$, $\sigma = 0.13$) with $p < 0.0001$ for session 4, and comparing CBCL ($\mu = 0.55$, $\sigma = 0.19$) to FT ($\mu = 0.26$, $\sigma = 0.14$) with $p < 0.0001$ for session 5. Further, when considering the two models separately, significant differences are seen between the first and the subsequent sessions for FT only.

As evident from the ANOVA, there was no statistically significant difference in classification accuracy for expert and non-expert users (based on their previous programming experience). Results in Figure 5b correlate with the ANOVA.

2) Number of Images per Object: We performed the post hoc Tukey HSD test for the significant ANOVAs for the number of images as the dependent variable. Figure 6a details the difference between the two CL models and expert and non-expert participants in terms of the number of images taught per object. There is a statistically significant difference between CBCL and FT for experts only, with ($\mu = 4.48$, $\sigma = 2.92$) for CBCL and ($\mu = 8.76$, $\sigma = 5.75$) for FT with $p < 0.0001$. Further, we also notice a statistically significant difference between experts and non-experts irrespective of the CL model with ($\mu = 3.93$, $\sigma = 3.71$) for non-experts and ($\mu = 6.09$, $\sigma = 4.67$) for experts with $p = 0.0001$. However, this difference seems to stem from participants in the FT condition only.

To further investigate the experts and non-experts in the FT condition, we performed a Wilcoxon rank sum test [22] between experts and non-experts in the FT condition over five sessions. As displayed in Figure 6b, there is a statistically significant difference between experts and non-experts for sessions 4 and 5 only i.e. when comparing experts ($\mu = 9.05$, $\sigma = 4.41$) to non-experts ($\mu = 3.68$, $\sigma = 3.25$) with $p = 0.035$, $W = 10.5$ in session 4, and comparing experts ($\mu = 11.49$, $\sigma = 7.03$) to non-experts ($\mu = 3.14$, $\sigma = 2.49$) with $p = 0.035$, $W = 10.0$ in session 5.

3) Number of Teaching Phases per Session: As the ANOVA for the number of teaching phases was not significant, we did not perform a post hoc Tukey HSD test. However, we performed a Wilcoxon rank sum test for the borderline values in ANOVA. Most of the data values were 1, indicating that most participants started only a single teaching phase in most sessions. Figure 7 shows the number of teaching phases per session for the two conditions and for experts and non-experts. There was no statistically significant difference between the two conditions, however, as displayed in the figure, there was a statistically significant difference between experts ($\mu = 1.69$, $\sigma = 1.78$) and non-experts ($\mu = 2.23$, $\sigma = 2.28$) with $p = 0.047, W = 5427$.

Further, 20 out of 40 participants had at least one session where they started more than one teaching phase with the robot. Overall, 50 out of 200 sessions had more than one teaching phase ranging from 2 to 9 teaching phases in a single session.

4) Reteaching after Misclassification: The ANOVA for the dependent variable reteaching after misclassification was not significant and there were no borderline values. Therefore, we did not perform a post hoc Tukey HSD test. Overall, we noticed that 18 out of 40 participants retaught at least one object after it was misclassified by the robot during the testing phase. In total there were 46 out of 200 sessions in which participants retaught misclassified objects with a maximum of 7 reteaching of misclassified objects.

Note that the above statistic only counts the reteaching of misclassified objects from the current session only i.e. if an object taught in the previous sessions was misclassified and retaught in a session it is not covered in the above statistic. Overall, there were only 11 sessions when participants retaught at least one object from the previous sessions, with a maximum number of 4 old objects taught in a session. In terms of the number of participants, only 6 out of 40 participants retaught objects from previous sessions in subsequent sessions.

V. Discussion

Results from the qualitative and quantitative analyses of the data collected in our study allow us to validate the hypotheses in Section III-C and answer the research questions.

For object labeling, we noticed significant variations in the labeling strategies of different participants. None of the 25 objects used in the study had a single consistent label across all 40 participants, even for simple objects, such as *Apple*. Further, some participants also labeled different objects with the same label, and some participants gave multiple labels to the same object. These strategies significantly affected the performance of the continual learning robot as depicted by the high standard deviation in classification accuracy of the two CL models (Figure 5a). As a consequence, we can accept **H1.1**. These results also indicate the need for developing personalized robots that adapt to their users' labeling strategies and learn, and understand, their environment such that both the user and the robot can effectively communicate about the entities in the environment.

The classification accuracy of the continual learning robot was significantly affected by the choice of the CL model which was expected as the FT model forgets previous objects over the five sessions. However, classification accuracy was not affected by the previous robot programming experience of the participants. This result was surprising as it indicates that even expert users who have previous programming experience might not be familiar with continual learning over the long term. Therefore, the teaching effectiveness of both expert and non-expert users might be similar for a continual learning robot. Consequently, we have to reject **H4.1**.

We quantified participants' teaching styles by calculating the number of images taught per object, the number of teaching phases started in each session, the number of times objects were retaught after being misclassified by the robot, and the number of times objects from previous sessions were taught by the participants. For the number of images per object, we did not find a statistically significant difference regarding the choice of the CL model or the session number in ANOVA. However, the previous robot programming experience of the participants did have a significant effect on the number of images per object. Particularly, this difference occurred because of a significantly high number of images shown by expert users in the FT condition. Upon further investigation, this difference occurred because expert users showed a significantly larger number of objects than non-expert users during later sessions in the FT condition. These results show that based on their previous experiences expert users might try to compensate for the degraded performance of the robot in later sessions by teaching more images per object. Note, that this might still not affect the robot's classification performance, as users might not be familiar with continual learning.

In terms of the number of teaching phases per session, there was no statistically significant effect of the choice of the CL model or the session number. However, we did see a significant difference between expert and non-expert users. Particularly, we noticed that non-expert users started more teaching phases with the robot in each session. Note that in the demo session participants were shown only a single teaching phase. Therefore, this result indicates that non-expert users might teach continual learning robots differently than the experimenter, i.e., not entirely following their instructions. Further, taking into account the number of images taught per object, this result indicates that non-expert users might teach comparatively few objects to the robot. However, based on the robot's performance they might re-teach the same objects again.

For reteaching objects based on misclassification by the robot, we did not see any significant effect of the session number, choice of the CL model, or the previous programming experience of the participants. However, we did notice that almost half of the participants retaught objects if they were misclassified by the robot. This result indicates that, unlike static datasets, the continual learning robots might get more data for the objects if the robot misclassifies them in a session. Finally, we also noticed that almost half (45%) of the participants also retaught some of the objects from previous sessions to the robot. Note that in the study instructions, and during the demo phase, participants were not told that they cannot re-teach old objects, therefore many of the participants retaught objects from previous sessions if the robot misclassified them during the testing phases. These results further demonstrate the difference between constrained CL test setups and testing in the real world with real users. Particularly, unlike constrained CL setups, users will reteach objects that they previously taught the robot if the robot does not classify them correctly. Finally, these results show that most users in the study were motivated to improve the performance of the robot, even though they were not given any specific incentive to do so. This is quite promising, as it indicates that users might be motivated to improve the performance of their personal robots over long-term interactions. Based on the above results **H2.1** can be accepted partially as we noticed that almost half of the participants retaught objects to the robot based on the robot's classification performance. **H2.2** can also be accepted partially as we noticed a difference in the number of images per object between CBCL and FT for expert users. Furthermore, **H3.1** has to be rejected as we noticed a change in the number of images per object for expert users in the FT condition only. Finally, we can accept **H4.2** partially, as we did notice a difference in the number of images per object, and the number of teaching phases for expert and non-expert users. However, there was no difference in terms of reteaching old objects between expert and non-expert users.

VI. Conclusions

In this paper, we considered a human-centered approach to continual learning to understand how users interact with and teach continual learning robots over the long term. We designed a long-term between-participant HRI study with a continual learning robot using two different CL models and analyze the data to understand the different teaching

styles of participants, and how these styles are influenced by the performance of the robot over multiple sessions. Our results indicate that different users teach household objects to the continual learning robot in a variety of ways, which can also affect the classification performance of the CL models. Moreover, the results show that the classification performance of the robot can also influence the teaching style of the users, which is different from constrained CL test setups. The results also show that the previous programming experiences of the users can also significantly influence the way they interact with and teach the continual learning robot over multiple sessions. Finally, our results demonstrate the limitations of current CL test setups and CL models. Therefore, based on the results of this study, we recommend future CL models focus on adapting to the teaching style of their users, and that CL models should be tested in more realistic test setups.

VII. Limitations and Future Work

We conducted our study in an unconstrained setup, where participants could teach and test the robot flexibly. However, the study was conducted in a robotics lab and not in a realistic household environment. In future work, we plan to conduct a similar study in a smart home with the same robot to understand the influence of the household environment on the interactions and teaching styles of the users.

We conducted the user study with a mix of expert and non-expert users, however, they were all university students between the ages of 18 and 37 years. In future work, we plan to conduct this study with participants who might be less familiar with robots to understand the effectiveness of continual learning robots for assistive applications. Finally, the study was conducted with one particular robot and with two CL models. Expanding this work to other robots and CL models can help us understand the larger design space of continual learning robots and users' teaching patterns when interacting with these robots.

Despite these limitations, our user study took the first step toward a human-centered approach to continual learning by integrating machine learning-based CL models with HRI. We hope that our results can help ML and HRI researchers design CL models that can adapt to their users' teaching styles and test these models in realistic experimental setups where embodied agents interact with human users.

References

[1] J. Saunders, D. S. Syrdal, K. L. Koay, N. Burke, and K. Dautenhahn, ""Teach Me–Show Me"—End-user personalization of a smart home and companion robot," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 27–40, 2016.

[2] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand, "Online object and task learning via human robot interaction," in *2019 International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 2132–2138.

[3] A. Ayub and A. R. Wagner, "Tell me what this is: Few-shot incremental object learning by a robot," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[4] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental classifier and representation learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] T. L. Hayes, K. Kafle, R. Shrestha, M. Acharya, and C. Kanan, "Remind your neural network to prevent catastrophic forgetting," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 466–483.

[6] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-shot class-incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] C. Zhang, N. Song, G. Lin, Y. Zheng, P. Pan, and Y. Xu, "Few-shot incremental learning with continually evolved classifiers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 12 455–12 464.

[8] M. Wise, M. Ferguson, D. King, E. Diehr, and D. Dymesich, "Fetch and freight: Standard platforms for service robot applications," in *IJCAI, Workshop on Autonomous Mobile Service Robots*, 2016.

[9] R. M. French, "Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference," *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pp. 335–340, 2019.

[10] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 13, pp. 3521–3526, 2017.

[12] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935–2947, Dec 2018.

[13] A. Ayub and A. Wagner, "Eec: Learning to encode and regenerate images for continual learning," in *International Conference on Learning Representations*, 2021.

[14] O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi, "Learning to remember: A synaptic plasticity driven framework for continual learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019, pp. 11 321–11 329.

[15] X. Tao, X. Chang, X. Hong, X. Wei, and Y. Gong, "Topology-preserving class-incremental learning," in *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*. Springer-Verlag, 2020, p. 254–270.

[16] P. Ramaraj, C. L. Ortiz, and S. Mohan, "Unpacking human teachers' intentions for natural interactive task learning," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 1173–1180.

[17] T. Kaochar, R. T. Peralta, C. T. Morrison, I. R. Fasel, T. J. Walsh, and P. R. Cohen, "Towards understanding how humans teach robots," in *User Modeling, Adaption and Personalization: 19th International Conference, UMAP 2011, Girona, Spain, July 11-15, 2011. Proceedings 19*. Springer, 2011, pp. 347–352.

[18] A. Bajcsy, D. P. Losey, M. K. O'Malley, and A. D. Dragan, "Learning from physical human corrections, one feature at a time," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 141–149.

[19] E. Senft, P. Baxter, J. Kennedy, S. Lemaignan, and T. Belpaeme, "Supervised autonomy for online learning in human-robot interaction," *Pattern Recognition Letters*, vol. 99, pp. 77–86, 2017.

[20] R. Krishna, D. Lee, L. Fei-Fei, and M. S. Bernstein, "Socially situated artificial intelligence enables learning from human interaction," *Proceedings of the National Academy of Sciences*, vol. 119, no. 39, p. e2115730119, 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[22] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945. [Online]. Available: http://www.jstor.org/stable/3001968

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.