Continual Learning through Human-Robot Interaction - Human Perceptions of a Continual Learning Robot in Repeated Interactions

Ali Ayub[1*], Zachary De Francesco[1], Patrick Holthaus[2], Chrystopher L. Nehaniv[3,1,2] and Kerstin Dautenhahn[1,2]

[1*]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, N2L 3G1, Ontario, Canada.
[2]School of Physics, Engineering and Computer Science, University of Hertfordshire, Hertfordshire, AL10 9AB, England, UK.
[3]Department of Systems Design Engineering, University of Waterloo, Waterloo, N2L 3G1, Ontario, Canada.

*Corresponding author(s). E-mail(s): a9ayub@uwaterloo.ca;
Contributing authors: zdefrancesco@uwaterloo.ca;
p.holthaus@herts.ac.uk; chrystopher.nehaniv@uwaterloo.ca;
kerstin.dautenhahn@uwaterloo.ca;

Abstract

For long-term deployment in dynamic real-world environments, assistive robots must continue to learn and adapt to their environments. Researchers have developed various computational models for continual learning (CL) that can allow robots to continually learn from limited training data, and avoid forgetting previous knowledge. While these CL models can mitigate forgetting on static, systematically collected datasets, it is unclear how human users might perceive a robot that continually learns over multiple interactions with them. In this paper, we developed a system that integrates CL models for object recognition with a Fetch mobile manipulator robot and allows human participants to directly teach and test the robot over multiple sessions. We conducted an in-person study with 60 participants that interacted with

our system in 300 sessions (5 sessions per participant). We conducted a between-subject study with three different CL models to understand human perceptions of continual learning robots over multiple sessions. Our results suggest that participants' perceptions of trust, competence, and usability of a continual learning robot significantly decrease over multiple sessions if the robot forgets previously learned objects. However, the perceived task load on participants for teaching and testing the robot remains the same over multiple sessions even if the robot forgets previously learned objects. Our results also indicate that state-of-the-art CL models might perform unreliably when applied on robots interacting with human participants. Further, continual learning robots are not perceived as very trustworthy or competent by human participants, regardless of the underlying continual learning model or the session number.

Keywords: continual learning, perceptions of robots, robot learning from human teachers, long-term human-robot interaction

1 Introduction

Assistive robots are becoming an integral part of our society in a variety of roles, such as caregivers, cleaning robots, or home assistants [1–6]. However, for robots to be able to assist people in daily environments over a long period of time, they must adapt to the changing needs of their users and their environments. As it would be impossible to pre-program all the tasks a robot needs to perform and all the items a robot might encounter in a person's daily environment, robots will need to continually learn interactively on the fly from their users who are likely unfamiliar with robotics and machine learning (ML).

To operate in daily environments, a general task for a robot is to learn and understand the objects in its environment [7–10]. Such a task is central to a variety of different service tasks such as fetching and carrying objects, cooking and meal preparation, doing the dishes and the laundry, etc. Various machine learning models have been developed in the last decade for achieving remarkable performance on object recognition tasks [11, 12]. However, one of the main challenges faced by robots using ML models to continually learn objects is *catastrophic forgetting* [13, 14]. Catastrophic forgetting[1] occurs when a continual learning (CL) agent forgets the previously learned knowledge when learning new information [15]. This, however, is in contrast with human memory that might gracefully forget detailed experiences but keeps abstract knowledge consolidated in long-term memory [16, 17]. One approach to avoid forgetting is to store the data of the previously learned tasks in memory and retrain the CL agent on the previously stored data plus the new information. However, this can lead to computational processing and memory storage issues

[1]Note that the term catastrophic forgetting is mainly used in Machine Learning literature to describe the phenomenon of an ML model forgetting most past knowledge on static datasets. However, when interacting with real users, the perception of forgetting might be far from "catastrophic".

for real-world robots with real-time constraints, limited onboard memory, and computational resources. In recent years, different research directions (some inspired by neuroscience [18, 19]) have been taken in the field of continual machine learning to mitigate the catastrophic forgetting problem without storing and relearning the complete dataset of the previous tasks [20–26]. While state-of-the-art (SOTA) CL models alleviate catastrophic forgetting, they still suffer from some forgetting when learning over a large number of repeated sessions [15, 18, 27–29].

Another challenge faced by continual learning robots is that their users might not provide a sufficiently large number of data (examples) to train an ML model. In the past few years, robotics researchers developed CL models that can learn continually from only a few training examples per object, while also mitigating catastrophic forgetting [30–32]. This problem is known as Few-Shot Incremental Learning (FSIL) [30, 31, 33, 34]. Although FSIL approaches have produced promising results on systematically collected "non-social" datasets by the experimenters, it is unknown how these systems might perform when learning from human participants. Further, it is also unknown how people might perceive robots that continually learn through interaction with their users.

For the long-term deployment of robots in human environments, it is critical that we understand how humans might perceive such robots, as these robots will interact with and operate around humans [4, 35–37]. Trust is one of the essential components for people's relationships with autonomous robots [38–41]. Most prior research showed that people lose trust in autonomous robots that might be imperfect. Research also showed that people's trust in robots could strengthen over time if they are involved in teaching the robot [42], however, this research was conducted with a simulated robot that did not actually learn from user instructions and used pre-programmed behavior. Other important factors that influence people's relationships with robots over the long term are perceptions of social attributes and usability of robots [4, 43–46]. Finally, for autonomous robots that learn through human teaching, it is also imperative to understand human perceptions of task load for interacting with and teaching the robot [47, 48]. Most of the prior research on analyzing human perceptions of robots has been conducted in a single interaction scenario, using hand-crafted, heuristic approaches. To the best of our knowledge, we know of no other work on testing CL or FSIL models deployed on robots that directly learn from human users on the fly over multiple interactions, which is the focus of this paper.

Here, we consider a system for socially guided continual learning (SGCL) and conduct an in-person user study to explore how people perceive a robot that continually learns common household objects over multiple interactions. We developed a system that integrates a graphical user interface (GUI) on an Android tablet with a CL model deployed on the Fetch mobile manipulator robot [49]. In this system, we focused solely on the continual learning of objects and avoided adding any extra social behaviours to the robot that

might affect human perceptions of the robot. We performed a long-term between-participant user study (N=60) where participants interacted with a fully autonomous Fetch mobile manipulator robot that used three different CL models: one that suffers from forgetting on static "non-social" datasets, another state-of-the-art (SOTA) approach for FSIL that mitigates forgetting, and the upper bound approach that stores and retrains all of the previous data when learning new information. We conducted 300 interactive sessions with 60 participants, where each participant taught 25 household objects to the robot in 5 sessions with 5 objects per session. We used four questionnaires in the study to answer the following research questions:

RQ1 How do human perceptions of trust, social attributes, task load, and usability evolve when interacting with a continual learning robot over multiple sessions?

RQ2 Is there a difference in participants' perceptions of trust, social attributes, task load, and usability of a continual learning robot for different continual learning models?

The remainder of the paper is organized as follows: Section 2 reviews related work including background on continual learning and prior continual learning approaches, prior work on robot learning from human teachers, and research on perceptions of robots. Section 3 explains our unique socially guided continual learning setup and the continual learning models used in the study. Section 4 describes the hypotheses guided by the RQs, participants recruited for the study, experimental setup and the procedure, and the measures used to evaluate the data collected from the study. Section 5 describes the detailed results of the study, followed by Section 6 to discuss the implications of the results. Finally, Section 7 concludes the article, followed by Section 8 that discusses the limitations of the study and directions for future research.

2 Related Work

In this section, we first present an overview of continual learning and ML methods for continual learning that are mostly tested without human users, followed by methods for robot learning that are designed to learn from human users, albeit in a single interaction. We then describe research on evaluating human perceptions of robots over single and multiple interactions.

Continual Learning. The standard continual learning (CL) problem for an object recognition task is defined as: Suppose a CL model $\mathcal{M}$ gets a stream of labeled training datasets $D^1, D^2, ...$ over multiple increments, where $D^t = \{x_i^t, y_i^t\}_{i=1}^{|D^t|}$ is the dataset in the tth increment, x_i^t is the $i$the data point in D^t with label y_i^t. L^t is the set of object classes in the tth training dataset, where $L^j \cap L^k = \varnothing, \forall j \neq k$. During the testing phase, if the model $\mathcal{M}$ is given the increment label when predicting the class label of a data point, this setup is known as task-incremental learning [18, 24, 27]. In contrast, for the class-incremental learning (CIL) setup, the model $\mathcal{M}$ is tested in increment t on data points belonging to any of the previous classes $(L^1, ..., L^t)$ without access

to the increment label [15, 50–52]. CIL is a more realistic continual learning setup, as robot users might not be willing to (or even remember) the increment label when asking the robot to predict the class label of an object. Therefore, we mainly review CIL approaches in this paper.

Class-Incremental Learning. Various research directions have been taken in the past to develop CIL models that can mitigate the catastrophic forgetting problem [15, 30, 50, 51, 53]. Most existing class-incremental learning (CIL) methods avoid catastrophic forgetting by storing a portion of the training samples from previous classes and retraining the model on a mixture of the stored data and new data [15, 50, 51, 54]. However, this approach does not scale as additional data exhausts memory capacity limiting performance in real-world applications. To avoid this problem, some CL approaches use regularization techniques [18, 27]. Although these approaches solve the memory storage issues, their performance is significantly inferior to approaches that store old class data. Another set of approaches uses generative replay to avoid storing raw data and generate old data using stored class statistics [19, 28, 29, 55, 56]. Generative approaches, however, do not scale well to learning over longer sequences and their performance deteriorates drastically. One of the major concerns for all CIL approaches is that they perform poorly when learning from limited training data [30, 31]. Therefore, they are not suitable for learning from human users who might be unwilling to provide hundreds or thousands of images per object class.

Few-Shot Incremental Learning. In the past couple of years, CL researchers developed class-incremental learning models that continually learn from a small number of training examples per class. This setup is known as few-shot incremental learning (FSIL) [30, 31]. All of these approaches train a CL model on a large number of object classes (called base classes) with a large dataset in the first increment to learn a good representation of the data. In the next increments, the model utilizes the representation learned in the first increment to learn new classes with only a few training images per class [30, 31, 33, 34, 57, 58]. These approaches, however, were only tested on static, simple datasets (e.g. MNIST [59]) and not on real robots that might not have perfect data available. Although a few CL approaches [7–9] have been tested with robots in recent years, most of them have been tested on only a small number of object classes (usually 10 or fewer). Further, none of these FSIL approaches were tested with real participants, and the data was captured by the experimenters in systematically controlled setups. Non-expert users, however, might not be aware of the underlying CL models, therefore they might provide imperfect data to the robot during teaching (or testing). It is also unclear how human participants might perceive CL systems on robots, and if they consider such systems to be feasible and easy to use.

Robot Learning with Human Teachers. A few studies have been conducted in the past with human participants to teach robots different manipulation tasks [60] or object classes [10, 61]. For example, Bobu et al. [60] developed a reinforcement learning technique for a manipulator robot that can perform

simple manipulation tasks with human assistance. Thomaz et al. [10] developed an object learning system that allowed a robot to learn object names and simple affordances from interactions with human participants. Human participants taught 6 simple objects to a social robot, which used a support vector machine (SVM) based method for learning these objects. Thomaz et al. showed that there were significant performance differences when machine learning models learned from human teachers rather than using systematically collected object datasets. Although these studies developed and tested ML techniques with human teachers, they were only tested in a single interaction with the users. Note that a single teaching interaction with the robot might not be a correct indicator of human perceptions of a continual learning robot, as user perceptions might change when teaching and testing the robot for the same task over multiple interactions (see Section 5 for results). In addition, single-session research with robots can suffer from the novelty effect which is well known in HRI research [62–64].

Perceptions of Robots. For the long-term deployment of robots in human environments, it is critical that we understand how humans might perceive such robots [4, 35–37]. Humans ascribe social traits and meaning to any agent in motion [65], but controlled experiments can help understand humans' perceptions about different aspects of robots [66–68]. For example, different studies have been conducted in the past that evaluate humans' trust in social robots exhibiting erroneous behaviors [38, 39, 69, 70]. Most of these studies indicated that the robot's performance on different tasks might affect humans' trust in the robot. Other studies analyzed human perceptions of social robots in terms of competence and warmth for a variety of tasks [35, 43]. Studies have also been conducted for domestic and industrial robots to understand their perceived usability [4, 71–73]. Although a few of these studies have been conducted to understand long-term interaction (maximum three sessions) with robots [35, 70], these studies used hand-crafted, heuristic approaches, and not any modern CL approaches. Unlike these prior research, a unique aspect of our study is that the robot directly learns new objects through human teaching and then uses the learned knowledge to autonomously find objects.

Recently a few studies analyzed human perceptions of robots learning directly from human demonstrations. For example, Schrum et al. [74] tested how human perceptions changed when the robot provided feedback related to human teaching. This approach, however, was only tested in a single session and was not used to test a continual learning approach. Liu et al. [75] presented a framework for learning online from human users but did not test their approach with real participants.

In contrast to previous work, to the best of our knowledge, we conducted the first user study at the intersection of continual learning and HRI, to understand human perceptions of a robot regarding various aspects (trust, competence, system usability), when the robot continually learns from human users over five sessions.

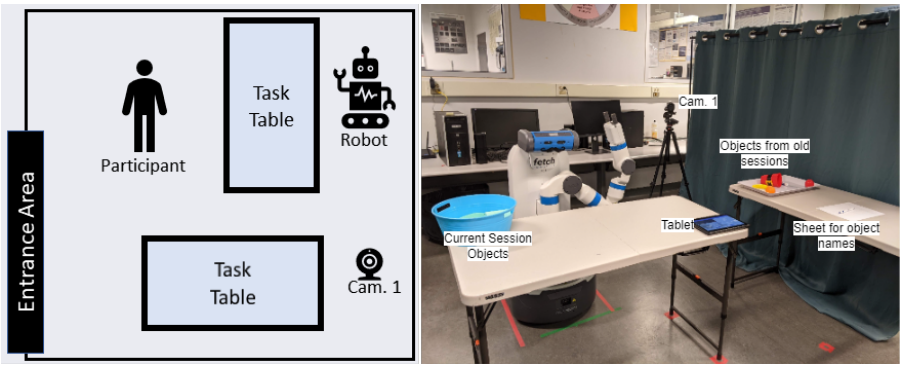


Fig. 1: (Left) Experimental layout for the SGCL setup with the participant and the robot. (Right) Corresponding real-world setup.

3 Socially Guided Continual Learning

We studied human perceptions of a continual learning robot in the context of an object recognition task. In this setup, the robot learns household objects from the user (in multiple sessions) on a table-top environment, and then finds and points to the requested object on the table after learning them from the user. 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 consider a socially guided continual learning (SGCL) system for the object recognition task, which integrates continual learning (CL) models with the robot for interactive and transparent learning from human users. Figure 2 shows the SGCL system for the object recognition task. In this system, in each session (or increment) t the user interacts with the robot through a graphical user interface (GUI) to teach the robot L_t number of objects. The robot captures images of the L_t objects and pre-processes them, getting the labels of the processed object images from the user to generate a dataset $D^t = \{x_i^t, y_i^t\}_{i=1}^{|D^t|}$, where x_i^t is the ith image in the dataset with the class label y_i^t. The CL model $\mathcal{M}$ then trains on the dataset D^t. Note that unlike static CL setups (such as FSIL [30]), the number of objects per object class in a session is not fixed as it is dependent on the number of times the user teaches an object to the robot. Further, there can be an overlap in the object classes taught in different sessions depending on how the user labels the objects i.e. $L^j \cap L^k \neq \varnothing$, for any $j \neq k$. For example, the user can name two different cups in different sessions with different names, such as "green cup" and "red cup", or they can name both of the cups as "cup". Such labelling differences among users are generally considered a problem (labelling bias) when developing general-purpose ML models trained offline on large datasets. However,

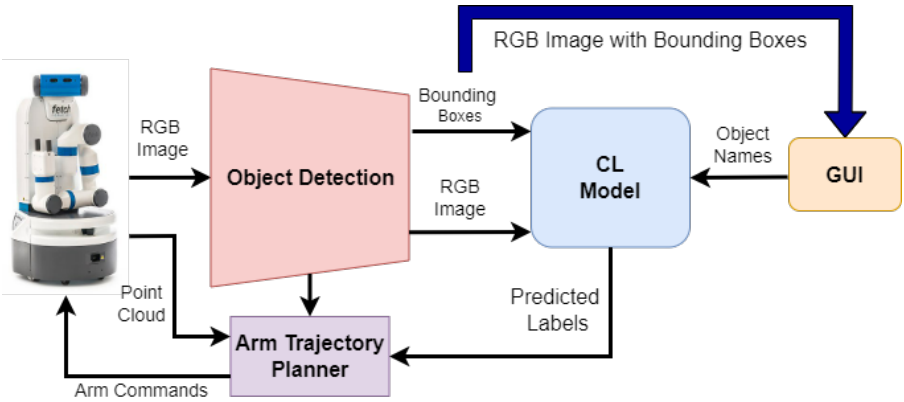


Fig. 2: Our complete SGCL system. Processed RGB images from robot's camera are sent to the GUI for transparency and also passed on to the CL Model. The user sends object names to the CL model either for training the CL model or finding an object. The arm trajectory planner takes point cloud data, processed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

our study focuses on personalized autonomous robots that will learn continually from their individual users about objects in their unique environments. Therefore, we did not constrain the users to teach the robot in the exact same way.

In the testing phase, the robot receives the request from the user through the GUI to find an object. The robot passes the pre-processed images to the CL model to get the predicted object labels. If the object is found, the robot finds the 3D location of the object on the table and points to the object using its arm. Note that the user has flexibility in terms of the total number of objects to be tested in an increment, as well as which objects to test (old or new objects). Therefore, unlike static CL setups, the test set of objects is not fixed in each session. Due to this flexibility in SGCL, results for CL models were quite different from the results on static datasets (see section 5 for details). It was important for us to introduce this flexibility for the user, to make the setup more similar to real-world scenarios.

3.1 Continual Learning Models

The main goal of our study is to do an in-depth analysis of how users perceive CL models over repeated, long-term interactions. To do such an analysis, it is important to choose a meaningful baseline. The naive finetuning (FT) approach [15] has been used extensively in CL literature as a baseline on static datasets. Therefore, we chose to test FT as our study's baseline model. The FT approach uses a convolutional neural network (CNN) [11] that 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. More details on this model can be found in [9, 15]. Please see Section A for further details on the choice of FT as a baseline for our study.

For the second model, we consider a SOTA CL approach specifically designed for FSIL in robotics applications [30]. This approach, termed centroid-based concept learning (CBCL), uses a CNN pre-trained on the ImageNet dataset [76] as a feature extractor for object images. In each increment, t, the model receives a small number of images for some object classes and extracts feature vectors for the object images using the pre-trained feature extractor. CBCL then clusters the feature vectors of all the object classes in the increment and generates a set of centroids $C^y = \{c_1^y, ..., c_{n_y}^y\}$ for each object class separately, where n_y is the total number of centroids for class y. CBCL avoids forgetting by generating separate centroids for each class in a new increment t, without changing the centroids of the previously learned classes. For the classification of a new object, CBCL finds the distance of the feature vector of the test object from the centroids of all the object classes. CBCL then uses a weighted voting scheme to find the most common class among the closest centroids to the test feature vector. The most common class is predicted as the object class for the test feature vector. More details about CBCL can be found in [30]. CBCL has been shown to produce promising results when learning from systematically collected object datasets by experts (researchers). However, it was never trained or tested in real-time with human participants.

Finally, for the third model, we consider the batch learning (called joint training (JT) in this paper) approach that stores the image data of the object classes from previous increments and retrains using stored data when learning new objects. JT has been used in CL literature as an upper bound for continual learning on static datasets. JT trains a CNN model on a combined image dataset of the new and old object classes in each increment, and therefore its training time continues to increase with each increment. More details about this model can be found in [9, 15]. For our study, similar to [9], we used a CNN pre-trained on the ImageNet dataset instead of using a CNN with random weights for FT and JT, to enable learning from a few training examples. In this paper, we integrate FT, CBCL, and JT in a fully autonomous system that allows users to experience these different ML models in real time through the Fetch mobile manipulator robot [49].

4 Method

To answer the two research questions (RQ1,2 in Section 1), we tested different hypotheses, related to those research questions, in a repeated measures study where users interacted over five sessions with the system (Section 3).

4.1 Hypotheses

The following hypotheses are guided by previous research that was discussed in Section 2: Prior HRI research showed that users' perceptions of trust, usability,

and social attributes are correlated with the performance of the robot, whereas the perceptions of task load are correlated to the time and effort spent in interacting with the robot. Further, prior CL research showed that CL models can forget previous knowledge over time, and thus their performance decreases. However, there is a difference in the rate of forgetting for different CL models.

Note, H$n.m$ is the mth hypothesis related to the research question n, e.g. H1.3 is the third hypothesis to answer RQ1.

Trust

H1.1 Users' perceptions of trust decrease in the robot over multiple sessions regardless of the CL model.

H2.1 A robot that forgets is perceived as less trustworthy than a robot that remembers most previously learned objects.

H2.2 A robot that retrains on all previous objects is perceived as more trustworthy than a robot that does not retrain on all previous objects.

Social Attributes

H1.2 Users' perception of the social attributes of the robot decreases over multiple sessions regardless of the CL model.

H2.3 The social attributes of a robot that forgets are perceived to be worse than those of a robot that remembers most previous objects.

H2.4 The social attributes of a robot that retrains on all previous objects are perceived to be better than those of a robot that does not retrain on all previous objects.

Task Load

H1.3 Users' perception of the task load for teaching the robot increases over multiple sessions regardless of the CL model.

H2.5 The task load for teaching and testing a robot that forgets is less than a robot that remembers most previous objects.

H2.6 The task load for teaching and testing a robot that retrains on all previous objects is more than a robot that does not retrain on all previous objects.

Usability

H1.4 Users' perceptions of the usability of the robot decrease over multiple sessions, regardless of the CL model.

H2.7 Users perceive a robot that remembers most previous objects to be more useful and easier to use than a robot that forgets.

H2.8 Users perceive a robot that retrains on all previous objects to be more useful and easier to use than a robot that does not retrain on all previous objects.

4.2 Fetch Mobile Manipulator Robot

Manipulator robots with an RGB-D camera are well-suited for recognizing and manipulating objects. In our setup (Figure 1), we use the Fetch mobile manipulator robot [49]. Fetch consists of a mobile base and a 7 DOF arm. The robot also contains an RGB camera, a depth sensor, and a Lidar sensor. These sensors can be used for 3D perception, SLAM mapping, and obstacle

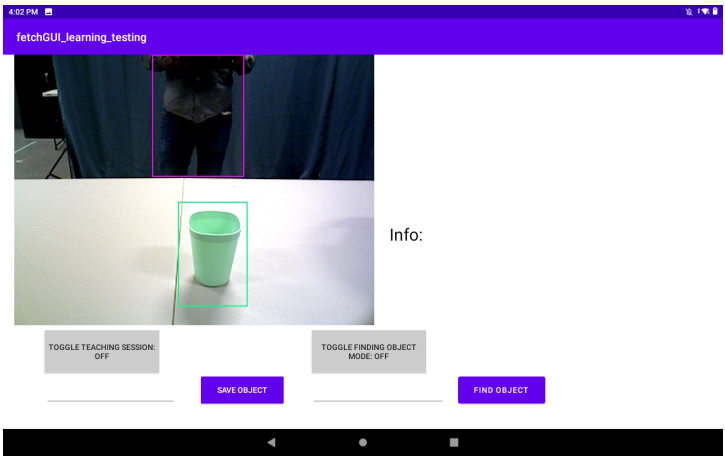


Fig. 3: The graphical user interface (GUI) used to interact with the robot. The RGB camera output with bounding boxes is on the top left. The buttons at the bottom can be used to teach objects to the robot and ask it to find objects in the testing phase. The top right of the GUI shows information sent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

detection in the robot's environment. In our setup, we do not ask the robot to manipulate objects or move its base, allowing us to solely focus on continual learning which is principally about learning and recognizing objects. We mainly use the RGB-D camera to recognize objects and the 7 DOF arm to point to objects. We use ROS packages available with the Fetch robot for moving the torso, and the arm of the robot. We did a safety analysis of the robot (approved by our University's ethics review board), and also adopted several mitigating strategies. Therefore, the robot was considered safe to be used with human participants in our study.

As there can be multiple objects on the table in front of the robot's camera, we process the RGB images further by passing them through a generic object detector [77]. The object detector finds regions in the image that are likely to contain objects (Figure 3). The detected regions are filtered using non-max suppression [78] to remove any overlaps. We also filter out the detected objects that are not on the table (background objects, participant interacting with the robot as seen in Figure 3) using the depth perception of the objects. The resulting regions are cropped into separate images for objects detected on the table and then forwarded to the CL model.

4.3 Graphical User Interface

For users to be able to interact and teach the robot different objects in an open-ended manner, we created a simple graphical user interface and deployed it on an Android tablet. Figure 3 shows a screenshot of the GUI. The top left side of the GUI shows the pre-processed camera output of the robot which contains bounding boxes for detected objects. The camera output was used as

a transparency device so that the participants could clearly understand what the robot was seeing on the table. On the bottom left of the GUI, there is a toggle button that can be used to start a teaching session with the robot. Once the button is pressed, it turns green indicating that the system was in the teaching phase. After starting the teaching phase, participants can type the name (class label) of the objects in the space below the toggle button. Participants can save an image of the object using the save button next to the empty space. The bottom right of the GUI contains another toggle button that can be used by the participants to start the testing phase. The button turns green once pressed. During the testing phase, participants can type the name (class label) of the object to be found on the table in the space below the testing toggle button. Participants can then press the Find Object button next to the empty space to ask the robot to find the requested object on the table. Finally, the top right section of the GUI shows the messages communicated by the robot to the user during the session. The robot also spoke these messages using a text-to-speech module available in ROS.

Note that we did not use an NLP (natural language processing) based interaction system because designing an open-ended NLP system for teaching objects is a challenging problem [79], and it is prone to additional errors during the speech-to-text and natural language understanding (NLU) phases. These errors could distract participants from the main research goal of the study (continual learning of household objects). Our goal was to study participants' perceptions of the continual learning system, and not to evaluate the communicative capabilities of the robot.

4.4 Participants

We recruited 63 participants (35 female (F); 28 male (M), all students) from the University of Waterloo, between the ages of 18 and 37 years ($M = 23.12$, $SD = 4.04$). Out of the 63 participants, three dropped out before completing the study. Out of the 60 participants, all three conditions were randomly assigned 20 participants each (ages: $M = 22.7$, $SD = 4.58$, 9 F, 11 M for *CBCL* condition, ages: $M = 24.53$, $SD = 4.03$, 10 F, 10 M for *FT* condition, ages: $M = 22.2$, $SD = 2.96$, 15 F, 5 M for *JT* condition). Based on their self-assessments in a pre-experiment survey, 35% 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 8% had previously participated in an HRI study. For the rest of the paper, we 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.

4.5 Procedure

We conducted five repeat sessions (each ∼20-30 minutes) with each participant in a robotics laboratory. All sessions were video recorded. Each participant

was randomly assigned to one of the three experimental conditions using one of the three CL models (CBCL, finetuning, and joint training). Before their first session, each participant was asked to complete a consent form and a pre-experiment survey online (using Qualtrics [80]). After completing the consent form and the pre-experiment survey, the experimenter greeted the participant and gave a brief oral introduction to the experiment. Participants were told to consider Fetch as their personal household robot and they can teach the robot household objects over five days such that the robot learns 25 objects over time. Next, the participant was directed to the study area of the lab and was handed an Android tablet with the GUI (see Figure 3) loaded. The participant was told that they would first run a demo session with the robot to understand how to teach and test the robot. The experimenter also mentioned that during the demo phase, the robot would not be learning any of the objects shown in front of the camera.

The experimenter explained that each session with the robot will consist of a teaching phase and a testing (i.e. finding an object) phase. The experimenter then used a blue cup as a demo object (this object was not used later) and placed it on the table. The experimenter then asked the participant to stand in the designated area in front of the table and start a teaching session by pressing the "Toggle Teaching Session" button on the GUI. Once the button was pressed, it turned green, and the robot sent a message on the tablet stating, "Entered teaching mode. You can now start teaching me objects." The robot spoke the same message through its speakers. The experimenter then asked the participant to type the name of the object in the text box below the toggle button. The participant was also told that they can name the object whatever they like. After the participant named the object, they were told that they can save the object by pressing the "Save Object" button next to the text box. Once the "Save Object" button was pressed, the robot stated, "[OBJECT NAME TYPED IN THE TEXT BOX] has been saved". The experimenter then mentioned to the participant that they can save each object as many times as they want by placing the object at different places on the table at various angles. They further mentioned that a similar procedure can be used to teach the other four objects in the session. The participant was then told that once they are finished saving all five objects in a session, they can press the toggle button again to end the teaching session. Once the button was pressed, it turned grey, and the robot stated, "I am learning the objects, please wait." The robot then stated, "Left teaching mode". The experimenter explained to the participant that the robot would learn the objects shown by them in the session and then communicate to them when it had finished learning and left the teaching mode.

The experimenter then explained the testing phase to the participant. The experimenter asked the participant to press the "Toggle Finding Object Mode" button to start the testing phase of the session. Once the button was pressed, it turned green and the robot stated, "Entered finding mode". The experimenter then placed two other objects on the table alongside the demo object (a total of

three objects on the table). The experimenter then mentioned that during the testing phase, the participant can place one or up to three (a suggestion, not a requirement) objects on the table. The experimenter further mentioned that after placing the objects, the participant can type the name of the object to be found by the robot in the text box below the toggle button. The experimenter then asked the participant to type "cup" in the text box to ask the robot to find this object on the table. Once they finished typing, the experimenter asked the participant to press the "Find Object" button. After the participant pressed this button, the robot stated, "I will point to the cup now. Please make sure that you are at a safe distance from me." The robot then moved its torso and arm to point to the cup on the table, and after finishing it stated, "I am done". The experimenter then mentioned that the participant could ask the robot to find the objects taught in the current session and in the previous sessions by placing them on the table and using the "Finding Object Mode". The experimenter then asked the participant to press the "Toggle Finding Object Mode" button again to leave the testing phase. Once the button was pressed, it turned grey and the robot stated, "Left finding mode".

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 taught in the current session. The paper sheet was kept by the experimenter and handed to the participants at the start of each session. This way the participants could remember the object names when they needed the robot to find these objects in the next sessions[2]. 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 started teaching the five objects to the robot. Once the participant finished teaching, they moved to the testing phase. During the testing phase, they asked the robot to find the objects taught in the current session and the previous sessions. The participant could switch back to the teaching phase and reteach objects misclassified by the robot, as many times as they desired. After the final testing phase was finished, the experimenter came out of the secluded area and stated, "Thank you for coming today. We have a few questions about your experience today. Could you please answer them on this tablet?" The experimenter gave a different tablet to the participant to answer questionnaires in Qualtrics [80] format. After finishing the questionnaire, the experimenter thanked the participant. The participant then scheduled their next session.

[2]Note that in real-world situations participants will not need to write down the names on a sheet as they will be interacting with the objects multiple times, but in our setup, it was not possible to interact with the objects long enough to remember their names from one session to the next. Sessions were typically a few days apart. On average, the time elapsed between the first and the last session was 25.3 days, with a maximum of 52 days. Further, the average time between any two consecutive sessions was 6.2 days, with a maximum of 29 days.

In the next four sessions (each ∼20-30 minutes), the same procedure was repeated, except for changing the objects to be taught in each session. Figure A1 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 A1) with participants' objects (the total number of objects taught over 5 sessions remained 25). Further, 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. Participants were remunerated $30 CAD if they participated in all five sessions. Otherwise, if they did not complete all five sessions, they were remunerated $6 CAD/session. Analysis of the audio data collected during the interviews is beyond the scope of this paper and will be reported in future work.

Examples of the teaching and testing phases are shown in the supplementary video. Our code is available at https://github.com/aliayub7/cl_hri.

4.6 Measures

To verify the hypotheses and thus evaluate the different learning models, we applied a range of quantitative measures in both experimental conditions.

Subjective Measures. After each trial, we asked participants to fill in the following questionnaire scales as subjective measurements aimed to test the hypotheses. We measured people's trust in the robot using the cognition-based trust subscale of Madsen's *Human-Computer Trust (HCT)* questionnaire [81] to address **H1.1**, **H2.1** and **H2.2**. The scale contains six individual questions that can be rated on a 5-point Likert scale, ranging from "Strongly disagree" to "Strongly agree". We further used the *Robot Social Attributes Scale (RoSAS)* [82] to measure how people rate the robot's social attributes to be able to accept or reject **H1.2**, **H2.3** and **H2.4**. The scale asks participants how closely they associate 18 attributes with the robot, using a Likert scale ranging from 1 to 7. A combination of these items forms three principal factors "warmth", "competence", and "discomfort". Additionally, we used the *Nasa-Task Load Index (NASA-TLX)* [83] to estimate participants' mental workload to gain insights about **H1.3**, **H2.5**, and **H2.6**. *TLX* is comprised of six questions that participants rate on a 21-point scale, ranging from "Very low" to "Very high", resulting in a single factor. Finally, we estimated an overall usability score using the *System Usability Scale (SUS)* [84] to address **H1.4**, **H2.7**, and **H2.8**. This scale is presented in ten questions on a 5-point Likert scale, ranging from "1 - Strongly disagree" to "5 - Strongly agree", to form a single factor. Note that we were only interested in long-term changes in the robot's social perception and hence we only employed RoSAS in the first and last session, while the other three questionnaires were presented in all five repeat sessions to allow for observing changes in between sessions.

Objective Measures. We also used an objective measure to analyze the performance of the three CL approaches and how it correlates with user perceptions of trust, social attributes, and usability of the continual learning robot. Classification accuracy per session (increment) has been commonly used in the continual learning literature [15, 30, 31] 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:

$$\mathcal{A} = \frac{total\ number\ of\ object\ correctly\ found\ in\ the\ session}{total\ number\ of\ objects\ tested\ in\ the\ session} \quad (1)$$

We also report the average number of times each object was taught by the participants in the three conditions to determine the task load for teaching the robot.

5 Results

Visual inspection using quantile-quantile plots as well as applying Shapiro-Wilk tests for normality [85] suggest that none of the scores obtained using questionnaire scales are normally distributed, requiring non-parametric tests to test for potential differences. Consequently, we evaluated all questionnaire scales using a Wilcoxon rank sum test [86] comparing the scores between the two models and five sessions, respectively. We also applied Bonferroni correction [87] with the Wilcoxon rank sum test to avoid false positives in our multiple statistical hypotheses testing. The analysis of each model's accuracy is discussed below, followed by the discussion of subjective measures from questionnaires, as described in Section 4.6. For the remainder of the paper, we term the finetuning model as *FT* (theoretically suffers from forgetting), the few-shot incremental learning (FSIL) model CBCL (SOTA model designed for FSIL to mitigate forgetting), as *CBCL*, and the batch learning model as *JT* (theoretical upper bound for continual learning that retrains on the data of previous sessions).

5.1 Classification Accuracy

Figure 4 shows the classification accuracy of the three models averaged over all the participants per model. In the first session, the classification accuracy of both CBCL and FT is similar ($\mu =\sim 50\%$), whereas the classification accuracy of JT is higher ($\mu =\sim 60\%$). However, for the second session, FT's accuracy significantly decreased ($\mu =\sim 30\%$), and it further decreased in the next three sessions ($\mu =\sim 25\%$). CBCL's accuracy remained similar ($\mu =\sim 50\%$) in all five sessions. JT's accuracy stayed consistent in the first two sessions ($\mu =\sim 60\%$), however, it significantly decreased in the third session ($\mu =\sim 45\%$) and stayed consistent for the next three sessions. Huge variations were seen in classification accuracy for all three models in all five sessions. This variation

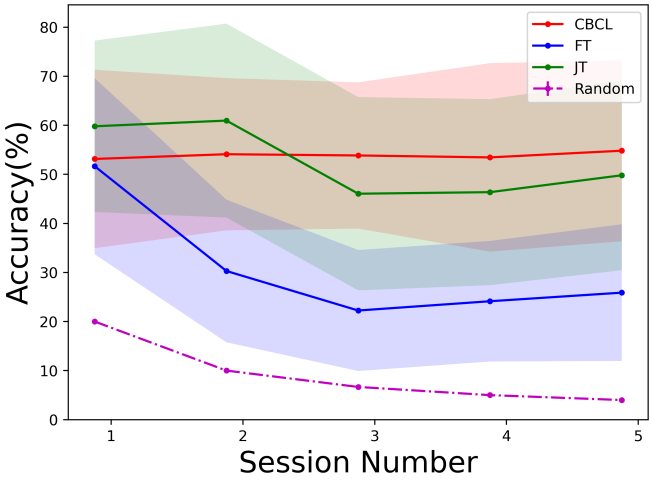


Fig. 4: Average classification accuracy of the three CL models over 5 sessions. The dotted line at the bottom represents the random chance of predicting a correct object in each session. The shaded areas represent the standard deviation for CL models.

was because of the differences in the classification accuracy of the models for different participants.

For the *FT* condition we noticed that in later sessions many participants tested the robot on more new objects than old objects[3] which caused the accuracy of FT for those participants to be comparable to CBCL's accuracy (notice the high standard deviation in Figure 4). Also, some users gave the same name to objects in different sessions. For example, some users named "green cup" in Session 1, "red cup" in Session 2, and "black mug" in Session 3, as "cup". In such cases, FT was able to remember the previous instances of "cup" and thus the test accuracy for FT was higher.

Finally, we also calculated the average number of times each object was shown to the robot (number of training images per object) in the two conditions. Participants in all three conditions showed a similar number of images per object ($\mu = 4.43$, $\sigma = 3.47$ for CBCL, $\mu = 5.19$, $\sigma = 4.20$ for FT, and $\mu = 4.82$, $\sigma = 3.77$ for JT), with the highest average number of images per object in the *FT* condition.

5.2 Cognition based trust

Scores for *cognition-based trust* on *HCT* are calculated as mean values of six individual items with a minimum value of 1 and a maximum value of 5, resulting in an overall value of $\mu = 2.48, \sigma = 0.79$. Figure 5 details how this score

[3]Participants were not told they had to test the robot on old objects. Instead, they had flexibility regarding which objects they wanted to test in each session.

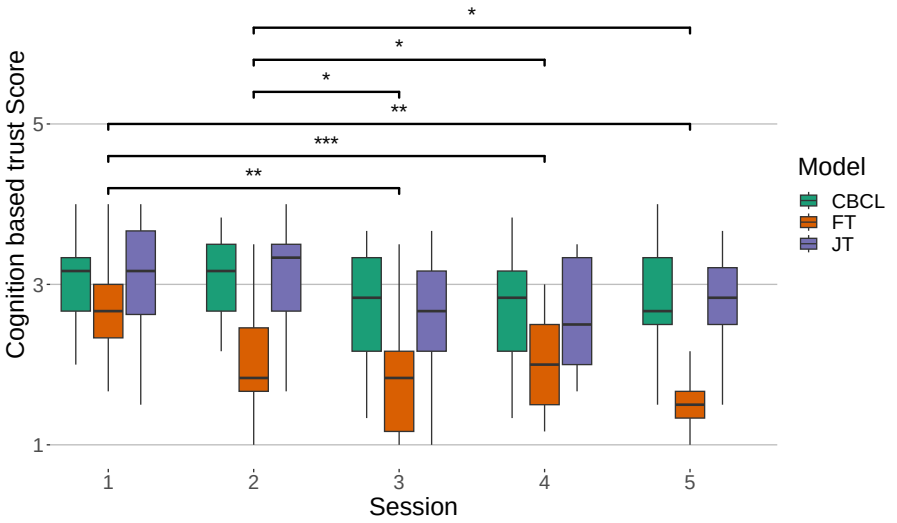


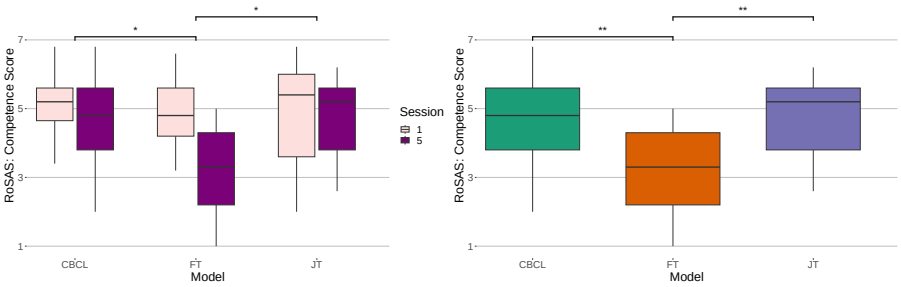
Fig. 5: Boxplots for *cognition based trust* scores on the *HCT* scale, ranging from 1 to 5. Significance levels ($* := p < .05$; $** := p < 0.01$; $*** := p < 0.001$) are indicated on bars between the columns.

differs between the subsequent experimental sessions. In particular, as displayed in Figure 5, trust decreases significantly only in the *FT* condition when comparing the first session with any of the subsequent sessions, i.e. when comparing session 1 ($\mu = 2.52, \sigma = 0.96$) to session 2 ($\mu = 1.92, \sigma = 0.79$; $p = 0.02584, W = 117$), to session 3 ($\mu = 1.66, \sigma = 0.81$; $p = 0.00329, W = 97$), to session 4 ($\mu = 1.75, \sigma = 0.88$; $p = 0.01015, W = 97.5$), and when comparing to session 5 ($\mu = 1.51, \sigma = 0.66$; $p = 0.00025, W = 69.5$). When considering the *CBCL* condition, a statistically significant difference is seen only between session 1 ($\mu = 3.07, \sigma = 0.53$) and session 4 ($\mu = 2.63, \sigma = 0.67$; $p = 0.03689, W = 137.5$), whereas no significant differences in scores can be observed between any of the sessions for the *JT* condition.

Moreover, *cognition based trust* scores are significantly different between the *CBCL* condition ($\mu = 2.86, \sigma = 0.65$) and *FT* condition ($\mu = 1.88, \sigma = 0.89$) with ($p < 0.0001, W = 1930.5$), and between the *JT* condition ($\mu = 2.67, \sigma = 0.87$) and *FT* condition with ($p < 0.0001, W = 7344$), when looking at all sessions combined, and consistently across all five sessions (see Table C2 for details).

5.3 Robot social attributes

Overall scores on *RoSAS* are calculated by averaging across individual items that belong to one of the subscales, ranging from 1 to 7. Resulting attribute scores are *warmth*: $\mu = 2.82, \sigma = 1.36$, *competence*: $\mu = 4.54, \sigma = 1.35$, and *discomfort*: $\mu = 1.93, \sigma = 0.89$. No significant differences with regard



(a) RoSAS competence scores by model in all sessions

(b) RoSAS competence scores by model in session 5 only

Fig. 6: Boxplots for *competence* scores on the *RoSAS* scale, ranging from 1 to 7. Significance levels ($* := p < .05$; $** := p < 0.01$; $*** := p < 0.001$) are indicated on bars between columns.

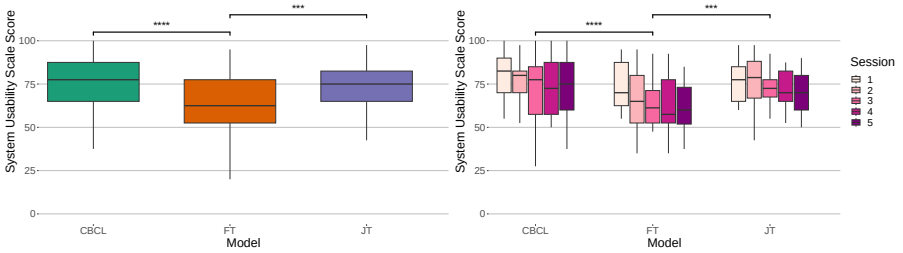
to *warmth* or *discomfort* can be observed when comparing experimental conditions or experimental sessions (Table C5). *Competence* scores (Figure 6a), are significantly different between the *CBCL* condition ($\mu = 4.77, \sigma = 1.29$) and the *FT* condition ($\mu = 4.01, \sigma = 1.47$) with $p = 0.0132, W = 604.5$, and between the *JT* condition ($\mu = 4.8, \sigma = 1.30$) and the *FT* condition with $p = 0.0155, W = 1127$ (Table C5). The difference in the results between conditions is mainly driven by the ratings in the last session, whereas scores in the first session were not statistically distinguishable (Table C5). In contrast, scores after the last session differ significantly ($p = 0.0065, W = 105.5$) between the *CBCL* condition ($\mu = 4.49, \sigma = 1.41$) and the *FT* condition ($\mu = 3.27, \sigma = 1.17$), and between the *JT* condition ($\mu = 4.63, \sigma = 1.19$) and the *FT* condition with $p = 0.0013, W = 333.5$. Accordingly, scores significantly drop between sessions only in the *FT* condition (more details in Table C5).

5.4 Task load index

Scores for the NASA task load index (TLX) are calculated as average values of six individual items (21-point scale, which is then translated into a score that ranges from 0 to 100). For all three models combined, the task load index remains at $\mu = 26.93, \sigma = 12.19$. No significant differences were seen between the three conditions overall (*CBCL* condition: $\mu = 26.4, \sigma = 13.0$, *FT* condition: $\mu = 28.3, \sigma = 13.3$, *JT* condition: $\mu = 26.2, \sigma = 10.3$) or between any of the five sessions (see Table C4 for details).

5.5 Usability

System usability scores are calculated as average values of ten individual questions (5-point scale, every second item inverted, which is then translated into a score that ranges from 0 to 100). The overall score is at $\mu = 70.41, \sigma = 16.29$, with significant differences between the *CBCL* condition ($\mu = 74.3, \sigma = 16.6$)



(a) Overall scores by model (b) Scores by session for three models

Fig. 7: Boxplots for *usability* scores on the *SUS* scale, ranging from 0 to 100. Significance levels ($* := p < .05$; $** := p < 0.01$; $*** := p < 0.001$) are indicated on bars between columns.

and the FT condition ($\mu = 63.7, \sigma = 18.8$) at $p < 0.001, W = 3441$, and between the JT condition ($\mu = 72.7, \sigma = 13.6$) and the FT condition at $p < 0.001, W = 6521$. In terms of individual sessions, the scores significantly differ between the *CBCL* condition and the FT condition in sessions 2, 4, and 5, and between the JT condition and the FT condition in sessions 3, 4, and 5 (see Table C3). No statistically significant difference is seen between the *CBCL* condition and the JT condition in any of the sessions. When investigating differences between the sessions, no significant change can be found in the *CBCL* condition and the JT condition, whereas in the FT condition, scores in session 1 ($\mu = 73.3, \sigma = 14.0$) are significantly higher than scores in session 3 ($\mu = 62.6, \sigma = 15.9$; $p = 0.039, W = 130.5$), session 3 ($\mu = 58.2, \sigma = 22.3$; $p = 0.023, W = 108$), and session 5 ($\mu = 59.0, \sigma = 17.4$; $p = 0.007, W = 107$). No statistically significant difference was seen between sessions 1 and 2.

6 Discussion

Results obtained in the repeated measures experiment with the interactive system allow us to validate the hypotheses introduced in Section 4 and conclusions to be drawn with regards to the research questions (**RQ1** How do human perceptions of trust, social attributes, task load, and usability evolve when interacting with a continual learning robot over multiple sessions? **RQ2** Is there a difference in participants' perceptions of trust, social attributes, task load, and usability of a continual learning robot for different continual learning models?). As a general observation, results seem to be influenced by the change over the course of repeated sessions, how participants interacted with the models, and if the model forgot previous objects.

In comparison to other studies [69], overall *cognition-based trust* is rated at mediocre levels only. Such a result is within our expectations for CBCL and FT as forgetting plays an important role and the cognitive function of the system is therefore not reliably identifiable by the user. However, for JT the result was

surprising as this model theoretically should not suffer from any forgetting. One reason for mediocre trust towards JT can be that JT was originally designed to learn from a large number of training samples, whereas participants only showed a few images per object (Section 5.1) to the robot. Further, over the five sessions, the robot learned an incrementally larger number of objects for both CBCL and JT conditions, however, unlike prior work [42], this did not have an accumulation effect on the perceptions of trust toward the robot. We believe that the imperfect nature of object teaching might have influenced the user's impression of the system because even the CBCL and JT approaches achieved only $\sim 45 - 60\%$ classification accuracy in all sessions. Considering the three conditions, trust towards the system is lower in the *FT* condition as opposed to the *CBCL* and *JT* conditions, where it remains on similar levels, except between sessions 1 and 3 for the *CBCL* condition where we saw a statistically significant drop in trust. This indicates that people, over time, lose trust in a model that forgets learned objects but they keep a similar amount of trust if it remembers previous objects. As a consequence, we can support **H1.1** (*Users' perceptions of trust decrease in the robot over multiple sessions regardless of the CL model*) but we only find support for **H2.1** (*A robot that forgets is perceived as less trustworthy than a robot that remembers most previously learned objects*) in the *FT* condition and for one session in the *CBCL* condition. Hence **H2.1** can only be supported partially by our data. This result is consistent with the experiment's objective measures since trust seems to correlate with the classification performance of both models. The classification accuracy for *FT* condition decreased because of forgetting and so did the trust. For *CBCL* condition, both the trust and the accuracy levels stayed similar, although accuracy remains $\sim$50% whilst *cognition based trust* decreases slightly during the course of the experiment with a significant decrease in session 3. For *JT* condition, accuracy slightly decreased in the final three sessions, and so did the trust in the robot. Finally, our data support **H2.2** (*A robot that retrains on all previous objects is perceived as more trustworthy than a robot that does not retrain on all previous objects*) partially, as there is a statistically significant difference in perceptions of trust between the *FT* condition and the other two conditions, but there is no significant difference observed between *CBCL* condition and *JT* condition. These results are promising indicating that users trust SOTA CL models, such as CBCL, that do not store and retrain on previous data similar to the theoretical upper bound JT.

For the robot's social attributes, warmth and discomfort scores stay reasonably low in all repeated sessions in a functional scenario with no extra social cues added to the robot, and where the interaction with the robot happens indirectly through the medium of a screen. The model choice (i.e. condition) also does not influence the experienced discomfort or warmth of the robot, making all equally good choices in terms of users' perceptions of these social attributes. This result was encouraging, showing that even a forgetful model caused little discomfort to the users. However, *competence* is perceived as significantly lower by the participants after interacting multiple times with a

robot that forgets (*FT* condition). In contrast, when using the CBCL and JT approaches, competence is rated similarly as in the first session. Therefore, all three hypotheses **H1.2** (*Users' perception of the social attributes of the robot decreases over multiple sessions regardless of the CL model*), **H2.3** (*The social attributes of a robot that forgets are perceived to be worse than those of a robot that remembers most previous objects*), and **H2.4** (*The social attributes of a robot that retrains on all previous objects are perceived to be better than those of a robot that does not retrain on all previous objects*) can be supported partially by our data. For **H2.3**, a significant difference was seen only for competence, but not for warmth and discomfort between the three conditions and over multiple sessions. Results for discomfort are interesting because they indicate that users feel little discomfort interacting with a continual learning robot even if the robot's performance decreases over multiple sessions. Thus, only the results for competence are supported by the objective measure (classification accuracy) of the experiment. Further, for **H2.4**, similar to the results for trust, there was no difference for all three social attributes between CBCL and the theoretical upper bound JT.

All three models had similarly low task load ratings, which is expected for *FT* condition as the model is simple and continues to forget previous objects. However, even for more complex models that mitigate forgetting, participants' workload did not increase. Neither the accuracy of the model nor any subsequent iterations affect the task load and hence **H2.5** (*The task load for teaching and testing a robot that forgets is less than a robot that remembers most previous objects*) cannot be supported by our data. Similarly, both **H1.3** (*Users' perception of the task load for teaching the robot increases over multiple sessions regardless of the CL model*) and **H2.6** (*The task load for teaching and testing a robot that retrains on all previous objects is more than a robot that does not retrain on all previous objects*) are not supported since we cannot find evidence that would support any difference between the conditions with regard to task load. There is no correlation between the task load and the model's performance and the model choice. However, task load seems to be linked with the total number of images shown per object, as participants for all three models showed only a few images per object. These results are quite promising as they indicate the feasibility of personalized continual learning robots that directly learn from their users. The results also suggest that researchers might need to focus more on the task (and task load) than the choice of the ML model alone when developing continual learning robots.

The experiment results suggest an effect of model choice on the system's *usability* in most sessions after the first sessions and therefore both **H1.4** (*Users' perceptions of the usability of the robot decrease over multiple sessions, regardless of the CL model*) and **H2.8** (*Users perceive a robot that retrains on all previous objects to be more useful and easier to use than a robot that does not retrain on all previous objects*) are supported partially. While usability scores are similar among the three approaches in the first session and partially in the second and third sessions, they drop significantly in sessions 3, 4, and 5 only

when the robot uses the FT approach and is thus forgetting more frequently. The robot, however, is perceived to be equally usable over repeated sessions with the CBCL and the JT approaches. Therefore, **H2.7** (*Users perceive a robot that remembers most previous objects to be more useful and easier to use than a robot that forgets*) can only be supported partially. This result is particularly interesting since *usability* is not directly linked to the classification accuracy of the model. This result also shows that users find continual learning robots to be useful even when the underlying model might forget previously learned objects. This could be because some users might only care about the robot's performance on the new objects, as observed for some users in our study that did not test the FT model on many old objects in the later sessions. This might also explain the high variance in accuracy seen for the two models for different participants. Therefore, CL researchers might need to not only focus on developing optimal models but also focus on the needs and tendencies of the participants when designing continual learning robots.

Finally, we observed that JT's classification accuracy dropped by $\sim 15\%$ after the first two sessions, whereas CBCL's accuracy remained consistent over all sessions. This was surprising as JT is a theoretical upper bound and trains on all the data from previous sessions, whereas CBCL only uses the data from the current session. This could be because JT was originally designed to learn from a large number of training images per object class, whereas CBCL can learn from a few images per object class. Note that in the fifth session, the time required for JT to learn new objects was $\sim 20 - 30$ seconds because it had to retrain on the data on all the objects from the previous four sessions, whereas CBCL required < 1 second to learn new objects even in the fifth session. This was a promising result indicating that SOTA CL models that require much less time to learn new objects can perform similarly to the theoretical upper bound for continual learning when applied to real robots interacting with real users. However, we observed that the classification accuracy of both CBCL and JT was much lower ($\sim 45 - 60\%$) than when tested on static datasets or with the experimenters ($> 90\%$) [9, 15]. These results indicate that the performance of the continual learning robots is quite different in the real world and it is drastically affected by the teaching style of their users.

7 Conclusions

In this work, we designed a long-term user study to understand human perceptions of a continual learning robot while teaching and testing the robot over five sessions. We conducted a between-participant study with three CL models and asked participants about their perceptions of the robot in terms of trust, social attributes, task load, and usability of the system, after directly teaching and testing the robot over five sessions. Our results indicate that users' perceptions of trust, competence, and usability of the robot are negatively affected by forgetting of the CL models. Our results also indicate that the performance of even the SOTA CL models is unreliable (only $\sim$50% accuracy) when learning

from the users instead of learning on static datasets. Therefore, with the current SOTA CL models, continual learning robots are not perceived to be very trustworthy or competent by their users. However, an encouraging result was that the performance of the SOTA CL models is comparable to the theoretical upper bound for continual learning which takes a much longer time to learn new objects. Furthermore, the task load for teaching and testing the continual learning robot, and perceptions of warmth and discomfort stayed low and were not affected by the choice of the CL model. These results are encouraging as they indicate the potential feasibility of personalized continual learning robots that might learn from their users over a long period of time. Our results also indicate that future continual learning research should also focus on the task load and the needs and tendencies of the users when designing CL models that learn through human interactions.

Our user study is the first step toward testing machine learning-based CL models in the realm of HRI. We hope that these results can help machine learning and HRI researchers design CL models while considering the perceptions of human users who might interact with these systems over a long period of time. Particularly, researchers need to focus on improving the performance of the CL models when learning from human users, which might also improve users' perceptions of trust, competence, and usability of the continual learning robots. One potential direction could be to integrate CL models with human-centered AI methods, such as interpretability, and fairness, to reduce labeling ambiguities and errors caused by robots' users, which could improve the robots' performance and users' perceptions of the robots. For example, adding simple feedback, such as the number of times an object has been taught and a baseline number of images required to teach an ML model for acceptable performance, could help improve users' perception of their teaching and expectations about the robot.

8 Limitations and Future Work

Although we used realistic household objects and allowed participants to bring their own objects, the study was performed in a robotics lab and not in a household environment. In the future, we plan to conduct a study in a smart home with the same robot and the same learning models to determine if the household environment affects user perceptions of the continual learning robot. Further, we did not add any social cues to the robot, such as gaze or affective expressions, which might affect users' perceptions of the robot and promote more human-robot engagement. This might even improve the performance of the model through better teaching by the users. Furthermore, the available mobile manipulator robot, Fetch, has a very 'functional' appearance, compared to other highly expressive social robots. In the future, we hope to expand on this study and add social capabilities to the continual learning robot. Although we conducted the first user study with a mix of experts and non-experts , they were all university students between the ages of 18 and 37 years. In the future,

we plan to conduct this study with older adults, who might be less familiar with robots and technology in general, to understand the effectiveness of continual learning robots for assistive applications. Additionally, the study was conducted with one particular robot and with two CL models. Expanding this work in comparative studies involving more interactive and social robots with other CL models can help us understand the larger design space of continual learning robots and users' perceptions of these robots.

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

Declarations

- Consent for publication: All authors whose names appear on the submission approved the version to be published.
- Consent to participate: Informed consent was obtained from all individual participants included in the study.
- Code availability: Our code is available at https://github.com/aliayub7/cl_hri.
- Availability of data and materials: The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.
- Authors' contributions: Conceptualization: Ali Ayub, Patrick, Chrystopher L. Nehaniv, Kerstin Dautenhahn; Methodology: Ali Ayub, Patrick, Chrystopher L. Nehaniv, Kerstin Dautenhahn; Formal analysis and investigation: Ali Ayub, Zachary De Francesco; Writing - original draft preparation: Ali Ayub; Writing - review and editing: Ali Ayub, Patrick, Chrystopher L. Nehaniv, Kerstin Dautenhahn; Funding acquisition: Kerstin Dautenhahn; Resources: Kerstin Dautenhahn; Supervision: Chrystopher L. Nehaniv, Kerstin Dautenhahn.

Compliance with Ethical Standards

- Funding: This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.
- Conflict of Interest: The authors have associations or collaborations with the following domains: uwaterloo.ca, psu.edu, herts.ac.uk. The authors declare that they have no other conflicts of interest.

References

[1] Matarić, M.J.: Socially assistive robotics: Human augmentation versus automation. Science Robotics **2**(4) (2017)

[2] Petrecca, L.: How Robot Caregivers Will Help an Aging U.S. Population (2018). https://www.aarp.org/caregiving/home-care/info-2018/new-wave-of-caregiving-technology.html Accessed 2020-10-05

[3] Saunders, J., Syrdal, D.S., Koay, K.L., Burke, N., Dautenhahn, K.: "Teach Me–Show Me"—End-user personalization of a smart home and companion robot. IEEE Transactions on Human-Machine Systems **46**(1), 27–40 (2016)

[4] Koay, K.L., Webster, M., Dixon, C., Gainer, P., Syrdal, D., Fisher, M., Dautenhahn, K.: Use and usability of software verification methods to detect behaviour interference when teaching an assistive home companion robot: A proof-of-concept study. Paladyn, Journal of Behavioral Robotics **12**(1), 402–422 (2021)

[5] Reiser, U., Jacobs, T., Arbeiter, G., Parlitz, C., Dautenhahn, K.: Care-o-bot® 3 – vision of a robot butler. Your Virtual Butler, 97–116 (2013)

[6] Shah, J., Ayub, A., Nehaniv, C.L., Dautenhahn, K.: Where is my phone? towards developing an episodic memory model for companion robots to track users' salient objects. In: Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23, pp. 621–624. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3568294.3580160

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

[8] Valipour, S., Quintero, C.P., Jägersand, M.: Incremental learning for robot perception through hri. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2772–2777 (2017)

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

[10] Thomaz, A.L., Cakmak, M.: Learning about objects with human teachers. In: 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 15–22 (2009)

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[12] Simonyan, K., Zisserman, A.: Two-stream convolutional networks for

action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1. NIPS'14, pp. 568–576. MIT Press, Cambridge, MA, USA (2014)

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

[14] Mcclelland, J.L., Mcnaughton, B.L., Oreilly, R.C.: Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. Psychological Review **102**(3), 419–457 (1995). https://doi.org/10.1037/0033-295x.102.3.419

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

[16] Costanzi, M., Cianfanelli, B., Santirocchi, A., Lasaponara, S., Spataro, P., Rossi-Arnaud, C., Cestari, V.: Forgetting unwanted memories: Active forgetting and implications for the development of psychological disorders. Journal of Personalized Medicine **11**(4), 241 (2021). https://doi.org/10.3390/jpm11040241

[17] Mack, M.L., Love, B.C., Preston, A.R.: Building concepts one episode at a time: The hippocampus and concept formation. Neuroscience Letters **680**, 31–38 (2018)

[18] Kirkpatrick, J., Pascanu, R., Rabinowitz, N.C., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences of the United States of America **114**(13), 3521–3526 (2017)

[19] Kemker, R., Kanan, C.: Fearnet: Brain-inspired model for incremental learning. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=SJ1Xmf-Rb

[20] Mundt, M., Hong, Y.W., Pliushch, I., Ramesh, V.: A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. CoRR **abs/2009.01797** (2020)

[21] Mundt, M., Lang, S., Delfosse, Q., Kersting, K.: CLEVA-compass: A continual learning evaluation assessment compass to promote research transparency and comparability. In: International Conference on Learning Representations (2022). https://openreview.net/forum?id=rHMaBYbkkRJ

[22] Smith, J., Hsu, Y.-C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9374–9384 (2021)

[23] Lomonaco, V., Maltoni, D.: Core50: a new dataset and benchmark for continuous object recognition. In: Proceedings of the 1st Annual Conference on Robot Learning, vol. 78, pp. 17–26 (2017)

[24] Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.S.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: The European Conference on Computer Vision (ECCV) (2018)

[25] Hayes, T.L., Kafle, K., Shrestha, R., Acharya, M., Kanan, C.: Remind your neural network to prevent catastrophic forgetting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 466–483. Springer, Cham (2020)

[26] Smith, J.S., Seymour, Z., Chiu, H.-P.: Incremental learning with differentiable architecture and forgetting search. In: 2022 International Joint Conference on Neural Networks (IJCNN), pp. 01–08 (2022). IEEE

[27] Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)

[28] Ayub, A., Wagner, A.: Eec: Learning to encode and regenerate images for continual learning. In: International Conference on Learning Representations (2021). https://openreview.net/forum?id=lWaz5a9lcFU

[29] Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., Nabi, M.: Learning to remember: A synaptic plasticity driven framework for continual learning. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11321–11329 (2019)

[30] Ayub, A., Wagner, A.R.: Cognitively-inspired model for incremental learning using a few examples. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)

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

[32] Lesort, T., Lomonaco, V., Stoian, A., Maltoni, D., Filliat, D., Díaz-Rodríguez, N.: Continual learning for robotics: Definition, framework, learning strategies, opportunities and challenges. Information fusion **58**, 52–68 (2020)

[33] Tao, X., Chang, X., Hong, X., Wei, X., Gong, Y.: Topology-preserving class-incremental learning. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) Computer Vision – ECCV 2020, pp. 254–270. Springer, Cham (2020)

[34] Zhang, C., Song, N., Lin, G., Zheng, Y., Pan, P., Xu, Y.: Few-shot incremental learning with continually evolved classifiers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12455–12464 (2021)

[35] Paetzel, M., Perugia, G., Castellano, G.: The persistence of first impressions: The effect of repeated interactions on the perception of a social robot. In: Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. HRI '20, pp. 73–82. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3319502.3374786. https://doi.org/10.1145/3319502.3374786

[36] Lyons, J.B., aldin Hamdan, I., Vo, T.Q.: Explanations and trust: What happens to trust when a robot partner does something unexpected? Computers in Human Behavior **138**, 107473 (2023)

[37] De Visser, E.J., Peeters, M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., Neerincx, M.A.: Towards a theory of longitudinal trust calibration in human–robot teams. International journal of social robotics **12**(2), 459–478 (2020)

[38] Rossi, A., Dautenhahn, K., Koay, K.L., Saunders, J.: Investigating human perceptions of trust in robots for safe hri in home environments. In: Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. HRI '17, pp. 375–376. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3029798.3034822. https://doi.org/10.1145/3029798.3034822

[39] Nayyar, M., Wagner, A.R.: When should a robot apologize? understanding how timing affects human-robot trust repair. In: Ge, S.S., Cabibihan, J.-J., Salichs, M.A., Broadbent, E., He, H., Wagner, A.R., Castro-González, Á. (eds.) Social Robotics, pp. 265–274. Springer, Cham (2018)

[40] Andras, P., Esterle, L., Guckert, M., Han, T.A., Lewis, P.R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S.T., *et al.*: Trusting intelligent machines: Deepening trust within socio-technical systems. IEEE

Technology and Society Magazine **37**(4), 76–83 (2018)

[41] Esterwood, C., Robert, L.P.: Do you still trust me? human-robot trust repair strategies. In: 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), pp. 183–188 (2021). IEEE

[42] Chi, V.B., Malle, B.F.: People dynamically update trust when interactively teaching robots. In: Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction. HRI '23, pp. 554–564. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3568162.3576962

[43] Scheunemann, M.M., Salge, C., Polani, D., Dautenhahn, K.: Human perception of intrinsically motivated autonomy in human-robot interaction. Adaptive Behavior **30**(5), 451–472 (2022)

[44] Scassellati, B., Boccanfuso, L., Huang, C.-M., Mademtzi, M., Qin, M., Salomons, N., Ventola, P., Shic, F.: Improving social skills in children with asd using a long-term, in-home social robot. Science Robotics **3**(21), 7544 (2018)

[45] De Graaf, M.M., Ben Allouch, S., van Dijk, J.A.: Long-term evaluation of a social robot in real homes. Interaction studies **17**(3), 462–491 (2016)

[46] De Graaf, M., Ben Allouch, S., Van Dijk, J.: Why do they refuse to use my robot? reasons for non-use derived from a long-term home study. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 224–233 (2017)

[47] Kosch, T., Karolus, J., Zagermann, J., Reiterer, H., Schmidt, A., Woźniak, P.W.: A survey on measuring cognitive workload in human-computer interaction. ACM Comput. Surv. (2023). https://doi.org/10.1145/3582272

[48] Zhang, J., Yang, K., Constantinescu, A., Peng, K., Müller, K., Stiefelhagen, R.: Trans4trans: Efficient transformer for transparent object and semantic scene segmentation in real-world navigation assistance. IEEE Transactions on Intelligent Transportation Systems **23**(10), 19173–19186 (2022)

[49] Wise, M., Ferguson, M., King, D., Diehr, E., Dymesich, D.: Fetch and freight: Standard platforms for service robot applications. In: IJCAI, Workshop on Autonomous Mobile Service Robots (2016)

[50] Wu, Y., Chen, Y., Wang, L., Ye, Y., Liu, Z., Guo, Y., Fu, Y.: Large scale incremental learning. In: The IEEE Conference on Computer Vision and

Pattern Recognition (CVPR) (2019)

[51] Castro, F.M., Marin-Jimenez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: The European Conference on Computer Vision (ECCV) (2018)

[52] Kang, M., Park, J., Han, B.: Class-incremental learning by knowledge distillation with adaptive feature consolidation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16071–16080 (2022)

[53] Hou, S., Pan, X., Loy, C.C., Wang, Z., Lin, D.: Learning a unified classifier incrementally via rebalancing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[54] Hayes, T.L., Krishnan, G.P., Bazhenov, M., Siegelmann, H.T., Sejnowski, T.J., Kanan, C.: Replay in deep learning: Current approaches and missing biological elements. Neural computation **33**(11), 2908–2950 (2021)

[55] Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Advances in Neural Information Processing Systems 30, pp. 2990–2999 (2017)

[56] Wu, C., Herranz, L., Liu, X., wang, y., van de Weijer, J., Raducanu, B.: Memory replay gans: Learning to generate new categories without forgetting. In: Advances in Neural Information Processing Systems 31, pp. 5962–5972 (2018)

[57] Bhunia, A.K., Gajjala, V.R., Koley, S., Kundu, R., Sain, A., Xiang, T., Song, Y.-Z.: Doodle it yourself: Class incremental learning by drawing a few sketches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2293–2302 (2022)

[58] Hersche, M., Karunaratne, G., Cherubini, G., Benini, L., Sebastian, A., Rahimi, A.: Constrained few-shot class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9057–9067 (2022)

[59] LeChun, Y.: The mnist database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/

[60] Bobu, A., Wiggert, M., Tomlin, C., Dragan, A.D.: Feature expansive reward learning: Rethinking human input. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21, pp. 216–224. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3434073.3444667. https://doi.org/10.1145/3434073.3444667

[61] Chao, C., Cakmak, M., Thomaz, A.L.: Transparent active learning for robots. In: 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 317–324 (2010)

[62] Leite, I., Martinho, C., Pereira, A., Paiva, A.: As time goes by: Long-term evaluation of social presence in robotic companions. In: RO-MAN 2009-the 18th IEEE International Symposium on Robot and Human Interactive Communication, pp. 669–674 (2009). IEEE

[63] Babel, F., Hock, P., Kraus, J., Baumann, M.: It will not take long! longitudinal effects of robot conflict resolution strategies on compliance, acceptance and trust. In: 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 225–235 (2022). IEEE

[64] Sung, J., Christensen, H.I., Grinter, R.E.: Robots in the wild: understanding long-term use. In: Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction, pp. 45–52 (2009)

[65] Hoffman, G., Ju, W.: Designing robots with movement in mind. J. Hum.-Robot Interact. **3**(1), 91–122 (2014). https://doi.org/10.5898/JHRI.3.1.Hoffman

[66] Anzalone, S.M., Boucenna, S., Ivaldi, S., Chetouani, M.: Evaluating the engagement with social robots. International Journal of Social Robotics **7**(4), 465–478 (2015). https://doi.org/10.1007/s12369-015-0298-7

[67] Takayama, L., Dooley, D., Ju, W.: Expressing thought: Improving robot readability with animation principles. In: 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 69–76 (2011)

[68] Aliasghari, P., Ghafurian, M., Nehaniv, C.L., Dautenhahn, K.: Effects of gaze and arm motion kinesics on a humanoid's perceived confidence, eagerness to learn, and attention to the task in a teaching scenario. In: Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction. HRI '21, pp. 197–206. Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3434073.3444651. https://doi.org/10.1145/3434073.3444651

[69] Robinette, P., Howard, A.M., Wagner, A.R.: Effect of robot performance on human–robot trust in time-critical situations. IEEE Transactions on Human-Machine Systems **47**(4), 425–436 (2017). https://doi.org/10.1109/THMS.2017.2648849

[70] Rossi, A., Dautenhahn, K., Koay, K.L., Walters, M.L., Holthaus, P.: Evaluating people's perceptions of trust in a robot in a repeated interactions study. In: Wagner, A.R., Feil-Seifer, D., Haring, K.S., Rossi, S., Williams,

T., He, H., Sam Ge, S. (eds.) Social Robotics, pp. 453–465. Springer, Cham (2020)

[71] Gadre, S.Y., Rosen, E., Chien, G., Phillips, E., Tellex, S., Konidaris, G.: End-user robot programming using mixed reality. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 2707–2713 (2019). https://doi.org/10.1109/ICRA.2019.8793988

[72] Solanes, J.E., Muñoz, A., Gracia, L., Martí, A., Girbés-Juan, V., Tornero, J.: Teleoperation of industrial robot manipulators based on augmented reality. The International Journal of Advanced Manufacturing Technology **111**(3-4), 1077–1097 (2020). https://doi.org/10.1007/s00170-020-05997-1

[73] Louie, W.-Y.G., Nejat, G.: A social robot learning to facilitate an assistive group-based activity from non-expert caregivers. International Journal of Social Robotics **12**(5), 1159–1176 (2020). https://doi.org/10.1007/s12369-020-00621-4

[74] Schrum, M.L., Hedlund-Botti, E., Gombolay, M.C.: Towards Improving Life-Long Learning Via Personalized, Reciprocal Teaching. Workshop on Lifelong Learning and Personalization in Long-Term Human-Robot Interaction (LEAP-HRI), 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (2022)

[75] Liu, B.: Learning on the job: Online lifelong and continual learning. Proceedings of the AAAI Conference on Artificial Intelligence **34**(09), 13544–13549 (2020). https://doi.org/10.1609/aaai.v34i09.7079

[76] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)

[77] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

[78] Hosang, J., Benenson, R., Schiele, B.: Learning non-maximum suppression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

[79] Chai, J.Y., Gao, Q., She, L., Yang, S., Saba-Sadiya, S., Xu, G.: Language to action: Towards interactive task learning with physical agents. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, pp. 2–9. International Joint Conferences on Artificial Intelligence Organization, Stockholm (2018). https:

//doi.org/10.24963/ijcai.2018/1

[80] Qualtrics. https://www.qualtrics.com (2005)

[81] Madsen, M., Gregor, S.: Measuring human-computer trust. In: Proceedings of the 11 Th Australasian Conference on Information Systems, pp. 6–8 (2000)

[82] Carpinella, C.M., Wyman, A.B., Perez, M.A., Stroessner, S.J.: The Robotic Social Attributes Scale (RoSAS): Development and Validation. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 254–262. ACM, Vienna Austria (2017). https://doi.org/10.1145/2909824.3020208. https://dl.acm.org/doi/10.1145/2909824.3020208 Accessed 2022-09-29

[83] Hart, S.G.: Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting **50**(9), 904–908 (2006). https://doi.org/10.1177/154193120605000909. Accessed 2022-09-29

[84] Brooke, J.: Sus: A quick and dirty usability scale. Usability Eval. Ind. **189** (1995)

[85] Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52**(3/4), 591–611 (1965)

[86] Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945)

[87] Armstrong, R.A.: When to use the Bonferroni correction. Ophthalmic and Physiological Optics **34**(5), 502–508 (2014)

Appendix A FT as a Baseline Approach for the Study

It can be argued that FT is a sub-optimal model that users would perceive negatively, and thus it should not be tested. However, although it is expected that users might perceive FT negatively, it is unknown if there is a statistically significant improvement in human perceptions when the robot uses a SOTA CL model (CBCL). Therefore, to quantify human perceptions of CL models, it is necessary to compare them against a baseline (FT) in the context of CL through HRI. Further, a unique aspect of FT is that, although it forgets objects from previous sessions, it can accurately classify new objects. In the context of our study, it is unknown if users would even care about the performance of the model on old objects and perhaps be more forgiving of the model as it is able to learn new objects. Note that in SGCL participants are not told to test the

Fig. A1: The twenty-five objects used in our study. Note that the visual similarity of some objects and their size variation make this a challenging task.

robot on old objects. Instead, they have flexibility regarding which objects to test in each session. Results for FT (Section 5) show the effect of this flexibility.

Finally, if we chose a baseline that simply does not learn anything, that would be a trivial comparison, since users would be expected to perceive this model negatively. Another choice of a baseline model is to pre-train a model on a set of objects. However, our study is not only about CL but also about human-robot teaching in an unconstrained manner. In our study, users are given high flexibility in teaching the objects. Particularly, users can name the objects however they like, teach an object as many times as they like in a session, and bring in a few of their own objects. In such an unconstrained environment, it would be impossible to choose a baseline that is pre-trained on some objects. For these reasons, we believe it is important to test the FT model in our study as a baseline.

Appendix B Objects used in the Study

Figure A1 shows the twenty-five objects used in our study. We used realistic objects from daily life with significant variations in size, color, and shape, which made the object-learning task quite challenging. Participants taught these objects in five sessions with five objects per session. Participants were also allowed to bring objects of their own choice after the first two sessions. If the participants brought their own objects, we used them to replace some of the twenty-five objects. Overall, each participant taught twenty-five objects.

Appendix C Overall Results of the Study

Table C1 shows the overall statistics of the study across all 60 participants. Tables C2, C5, C4, C3 show the overall results for the four questionnaires for the three conditions across all 60 participants.

Table C1: Overall results of the study irrespective of the CL model chosen.

Session	1		2		3		4		5	
Value	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
Trust	2.81	0.86	2.68	0.88	2.31	0.89	2.29	0.88	2.29	0.92
Warm.	2.93	1.34	×	×	×	×	×	×	2.71	1.36
Comp.	4.91	1.30	×	×	×	×	×	×	4.14	1.39
Disc.	1.88	0.89	×	×	×	×	×	×	1.98	0.91
TLX	26.5	9.63	25.7	11.4	26.8	12.2	27.6	13.4	27.5	14.3
SUS	74.4	16.8	72.9	17.1	68.9	15.5	67.9	18.8	67.5	16.3

Table C2: Detailed results for trust in the three conditions. NS stands for not significant.

Session	CBCL		FT		JT	
Value	μ	σ	μ	σ	μ	σ
1	3.07	0.53	2.52	0.96	2.85	0.95
2	3.09	0.49	1.92	0.79	2.99	0.80
3	2.76	0.67	1.66	0.81	2.49	0.82
4	2.64	0.68	1.75	0.88	2.41	0.87
5	2.74	0.77	1.51	0.66	2.61	0.82
all	2.86	0.65	1.88	0.89	2.67	0.87

Session	FT-CBCL		CBCL-JT		FT-JT	
Value	p	W	p	W	p	W
1	0.0230	130	NS	NS	NS	NS
2	3.3×10^{-5}	42.5	NS	NS	0.0003	316.5
3	0.0001	62.5	NS	NS	0.0046	305
4	0.0030	83.5	NS	NS	0.0369	240
5	6.3×10^{-5}	56.5	NS	NS	0.0001	354
all	2.2×10^{-14}	1930.5	NS	NS	3.5×10^{-9}	7344

Table C3: Detailed results for usability in the three conditions. NS stands for not significant.

Session	CBCL		FT		JT	
Value	μ	σ	μ	σ	μ	σ
1	75.9	20.6	73.3	14.0	73.9	15.6
2	77.4	11.3	64.5	21.2	76.6	14.9
3	71.9	17.5	62.6	15.9	72	10.9
4	73.5	16.5	58.2	22.3	70.9	14.5
5	72.7	16.7	59.0	17.4	70.2	11.7
all	74.3	16.6	63.7	18.8	72.7	13.6

Session	FT-CBCL		CBCL-JT		FT-JT	
Value	p	W	p	W	p	W
1	NS	NS	NS	NS	NS	NS
2	0.0386	130.5	NS	NS	NS	NS
3	NS	NS	NS	NS	0.0321	279.5
4	0.0353	114	NS	NS	0.0442	237.5
5	0.0145	116	NS	NS	0.02624	295.5
all	2.0×10^{-5}	3441	NS	NS	0.0002	6521

Table C4: Detailed results for task load in the three conditions. No statistically significant difference was seen for any condition, therefore p and W values are not reported.

Session	CBCL		FT		JT	
Value	μ	σ	μ	σ	μ	σ
1	24.4	7.94	28.9	9.72	27.2	11.7
2	25.1	13.3	29.5	11.5	22.9	7.98
3	25.8	11.9	27.9	14.6	27.1	10.9
4	28.1	14.3	26.8	14.7	27.5	11.9
5	28.3	16.8	27.9	16.2	26.4	9.38
all	26.4	13.0	28.3	13.3	26.2	10.3

Table C5: Detailed results for robot's social attributes (warmth, competence, discomfort) in the three conditions. NS stands for not significant.

	Session	CBCL		FT		JT	
	Value	μ	σ	μ	σ	p	W
Warmth	1	2.62	1.19	3.22	1.35	2.97	1.46
	5	2.69	1.46	2.55	1.28	2.86	1.37
	all	2.65	1.32	2.89	1.35	2.92	1.40
Competence	1	5.04	1.12	4.72	1.39	4.97	1.41
	5	4.49	1.41	3.27	1.17	4.63	1.19
	all	4.77	1.29	4.01	1.47	4.8	1.30
Discomfort	1	1.91	0.89	1.99	0.83	1.74	0.97
	5	2.02	0.82	2.00	0.92	1.94	1.01
	all	1.96	0.84	1.99	0.86	1.84	0.98

	Session	FT-CBCL		CBCL-JT		FT-JT	
	Value	p	W	p	W	p	W
Warmth	1	NS	NS	NS	NS	NS	NS
	5	NS	NS	NS	NS	NS	NS
	all	NS	NS	NS	NS	NS	NS
Competence	1	NS	NS	NS	NS	NS	NS
	5	0.0065	105.5	NS	NS	0.0013	333.5
	all	0.0132	604.5	NS	NS	0.0155	1127
Discomfort	1	NS	NS	NS	NS	NS	NS
	5	NS	NS	NS	NS	NS	NS
	all	NS	NS	NS	NS	NS	NS