An empirical assay of visual object learning in humans and baseline image-computable models

Michael J. Lee[1,2] **and James J. DiCarlo**[1,2,3]

***For correspondence:** dicarlo@mit.edu (JJD)

Present address: Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

[1]Department of Brain and Cognitive Sciences, MIT; [2]MIT Quest for Intelligence and Center for Brains, Minds and Machines; [3]McGovern Institute for Brain Research, MIT

Abstract How humans learn new visual objects is a longstanding scientific problem. Previous work has led to a diverse collection of models for how object learning may be accomplished, but a current limitation in the field is a lack of empirical benchmarks that evaluate the predictive validity of specific, image-computable models and facilitate fair comparisons between competing models. Here, we used online psychophysics to measure human learning trajectories over a set of tasks involving novel 3D objects, then used those data to develop such benchmarks. We make all data and benchmarks publicly available, and, to our knowledge, they are currently the largest publicly-available collection of visual object learning psychophysical data in humans. Consistent with intuition, we found that humans generally require very few images (<10) to approach their asymptotic accuracy, find some object discriminations more easy to learn than others, and generalize quite well over a range of image transformations, even after just one view of each object. To serve as baseline reference values for those benchmarks, we implemented and tested a large number of baseline models (n=2,408), each based on a standard cognitive theory of learning: that humans re-represent images in a fixed, Euclidean space, then learn linear decision boundaries in that space to identify objects in future images. We found some of these baseline models make surprisingly accurate predictions, but also identified reliable prediction gaps between all baseline models and humans, particularly in the few-shot learning setting.

Introduction

People readily learn to recognize new visual objects. As an individual receives views of a new object – new spatial patterns of photons striking their eyes – their ability to correctly categorize new views of that object increases, possibly very rapidly. What are the mechanisms that allow an adult human to do so?

Efforts from cognitive science, neuroscience, and machine learning have led to a diverse array of ideas to understand and replicate this human ability, and human example-based learning in general. These works range in levels of specification, from conceptual frameworks that do not directly offer quantitative predictions (***Shepard, 1987***; ***Biederman, 1987***; ***Nosofsky, 1992***; ***Tenenbaum and Griffiths, 2001***; ***Ashby and Maddox, 2005***), models which depend on unspecified intermediate computations (i.e. non-image-computable models; ***Reed*** (***1972***); ***Kruschke*** (***1992***); ***Bülthoff and Edelman*** (***1992***); ***Mckinley and Nosofsky*** (***1996***); ***Maddox and Ashby*** (***1993***)), to end-to-end learning models which take images as input (***Poggio and Edelman, 1990***; ***Duvdevani-Bar and Edelman, 1999***; ***Li et al., 2006***; ***Salakhutdinov et al., 2013***; ***Lake et al., 2015***; ***Erdogan and Jacobs, 2017***; ***Sorscher et al., 2022***).

An important step in determining which (if any) of these ideas might lead to accurate descrip-

tions of human object learning is to instantiate them in forms that can generate actual behavioral outputs over arbitrary object learning tasks (i.e. computable models), then to quantitatively compare each model's behavior to human behavior over a common set of such tasks. Broadly speaking, there are two components needed to enable such a comparison: 1) a dataset of human behavioral measurements over a set of experimental conditions (e.g. specific learning tasks) and 2) a procedure to compare those measurements to the behavioral predictions made by the model in the same experimental conditions. Taken together, we refer to those two components as a "behavioral benchmark" (***Schrimpf et al., 2020***).

While behavioral benchmarks exist for visual tasks involving known object categories (e.g. ***Rajalingham et al.*** (***2018***); ***Geirhos et al.*** (***2021***); ***Hebart et al.*** (***2022***)), the field currently lacks a publicly-available set of benchmarks for comparing models to humans for learning tasks involving novel objects, making it difficult to gauge progress in the field.

To address this gap, we aimed to create and release benchmarks for human object learning. We wished to make minimal assumptions about the learning models that might be compared today or in the future, and we only required that a model is computable in the following sense: on each trial, it can take a pixel image as input, make a multiple-choice selection, and receive scalar-valued feedback from the environment (e.g. to drive any changes to its internal state).

Experimentally, our primary goal was to measure human learning trajectories across a variety of binary object discrimination learning tasks involving highly varied views of novel 3D objects. We could then use the resultant dataset to make our primary benchmark, which compares a model to humans on the basis of core quantitative behavioral signatures associated with any learning system: its learning curves across a variety of novel tasks.

Based on prior suggestions of where humans may be particularly powerful (***Li et al., 2006***; ***Lake et al., 2011***), our secondary experimental goal was to measure human behavior during the special case of "one-shot" learning, where the subject receives a single view of each object before being asked to generalize to unseen views. Here, we sought to measure human generalization over a variety of tests involving identity-preserving transformations (e.g. translation, scaling, and 3D rotation) in the one-shot setting, then use the resultant measurements to create our secondary benchmark, which evaluates the extent to which a model can replicate human-like patterns of generalization across the same tests.

To serve as references for these two benchmarks, we implemented models from a standard cognitive theory of learning, which posits that adult humans re-represent incoming visual stimuli in a stable, multi-dimensional Euclidean space, build categorization boundaries in that space by applying some learning rule to exemplars, and use this learned boundary to categorize new examples (***Shepard, 1987***; ***Kruschke, 1992***; ***Tenenbaum and Griffiths, 2001***; ***McClelland et al., 2010***; ***Ashby and Maddox, 2005***; ***Sorscher et al., 2021***). To actually build those baseline models, we drew from ongoing efforts in computational cognitive science (***Morgenstern et al., 2021***, ***2019***; ***Nosofsky et al., 2018***; ***Singh et al., 2020***; ***Battleday et al., 2017***; ***Peterson et al., 2018***) and visual neuroscience (***Yamins et al., 2014***; ***Kubilius et al., 2019***; ***Schrimpf et al., 2020***) in building and validating image-computable models of human visual representations, based on intermediate layers of deep convolutional neural networks (DCNNs). We took a large sample of those networks and converted them into testable learning models by combining them with simple downstream learning modules that have been considered neurally plausible, namely linear decoders which are adjusted by scalar reward-based update rules (***Rosenblatt, 1958***; ***Law and Gold, 2009***; ***Niv, 2009***; ***Frémaux and Gerstner, 2015***).

In summary, our goal in this work was to take the following scientific steps: measure human behavior across a range of object learning tasks, develop procedures for comparing those measurements to the predictions made by any image-computable object learning model (i.e. behavioral benchmarks), and score a standard family of such models on those benchmarks.

At the outset of this study, we reasoned that, if the behavior produced by any such model was found to be statistically indistinguishable from that of humans, it could serve as a leading scientific

hypothesis to drive further experiments. If they were not found, predictive gaps could be used to guide future work in improving models of human object learning. Either way, the benchmarks created in this work could facilitate a standard evaluation of current and future visual object learning models.

Results

We measured human behavior over two variants of an object learning task (Experiments 1 & 2). In Experiment 1, we measured human subjects learning to discriminate pairs of novel objects as subjects were provided with an increasing number of views of those objects, and feedback on their choices. In Experiment 2, we also measured humans learning to discriminate between pairs of objects, but provided only one view per object before assessing subjects' accuracy on a variety of generalization tests. The results of each experiment are presented below, along with quantitative comparisons of those results with a large set of baseline models. Further details of the experiments and the models are provided in the Materials and Methods.

Experiment 1: Humans are rapid, but imperfect novel object learners

In Experiment 1, we measured a population of anonymous human subjects (n=70) performing 64 learning subtasks. Each subtask required that the subject learn to discriminate a different pair of two novel objects, rendered under high view variation (see Figure 1A).

These subtasks proceeded in a trial-by-trial fashion. At the beginning of a trial, a test image containing one of two possible objects was briefly presented (at ≈6°of the visual field for ≈ 200 milliseconds). Then, the subject was asked to report which of the two objects was present "in" that image through a button press, and evaluative feedback (correct or incorrect) was delivered based on their choice (see Figure 1B). The subtask then proceeded to the next trial (for a total of 100 trials).

The core measurement we sought to obtain for each subtask was the discrimination accuracy of a typical subject as a function of the number of previously performed trials (i.e. the *learning curve* for each subtask). To estimate the learning curve for a particular subtask, we recorded the sequence of corrects and incorrects achieved by achieved by multiple subjects (n=50 subjects) performing n=100 randomly sampled trials (see Figure 1C), then averaged across subjects to estimate the learning curve for that subtask. We estimated learning curves in this manner for all n=64 subtasks in this experiment (depicted in Figure 1D).

Upon examination of these learning curves, we found that on average (over subjects and subtasks), human discrimination accuracy improved immediately – i.e. after a single image example and accompanying positive or negative feedback. By construction, accuracy on the first trial is expected to be 50% (random guessing); but on the following trial, humans had above-chance accuracy (mean 0.65; [0.63, 0.67] 95% bootstrapped CI), indicating behavioral adaptation occurred immediately and rapidly. Average discrimination accuracy continued to rise across learning: the subject-averaged, subtask-averaged accuracy on the last trial (trial 100) was 0.87 (mean; [0.85, 0.88] 95% CI). The subject-averaged, subtask-averaged accuracy over all 100 trials was 0.82 (mean; [0.81, 0.84] 95% CI).

As anticipated, we found that *different* subtasks (i.e. different pairs of objects) could have widely different learning curves. This is illustrated in Figure 1D, which shows the estimated average human learning curve for each subtask. That is, we observed that some tasks were "easy" for humans to learn, and some were harder (e.g. mean accuracy of ≈0.65 for the most difficult 10% of subtasks). These variations were not artifacts of experimental variability, which we established by estimating the value of Spearman's rank correlation coefficient between average subtask performances that would be expected upon repetitions of the experiment ($\rho = 0.97$; see Additional Analyses).

Overall, these observations indicate that 1) humans can acquire a significant amount of learning with respect to novel visual object concepts with a small number of examples (e.g. ~4 training examples to reach 75% correct, ~6 to reach 90% of their final performance), and 2) learning new

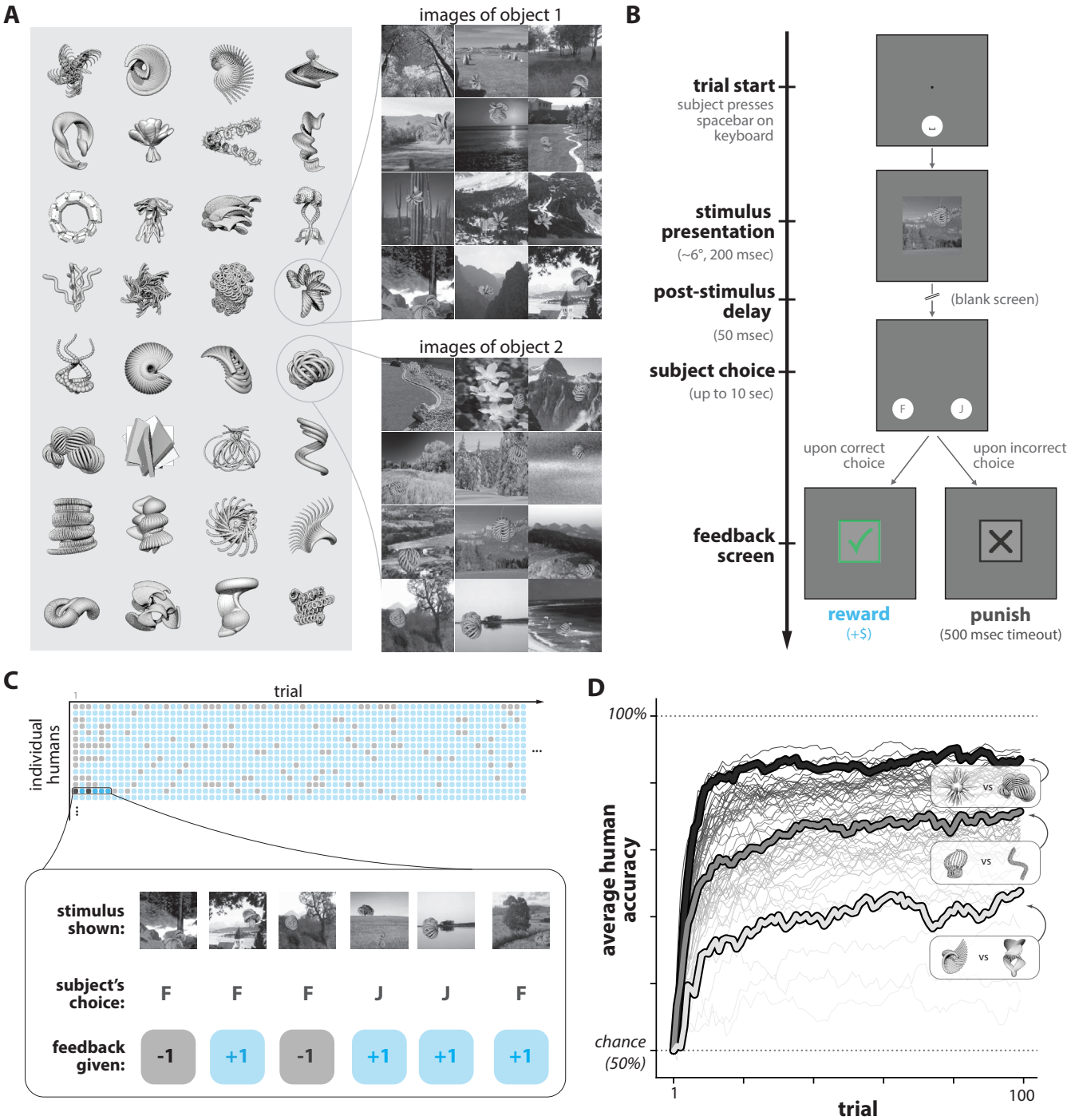


Figure 1. Humans learning novel objects. **A. Images of novel objects.** Images of synthetic 3D object models were created using random viewing parameters (background, location, scale, and rotational pose). **B. Task paradigm.** On each trial, a randomly selected image (of one of two possible objects) was briefly shown to the subject. The subject had to report the identity of the object by making one of two possible choices ("F" or "J"). Positive reinforcement was delivered if the subject choice was "correct", per an object-choice contingency that the subject learned through trial-and-error (e.g., object 1 corresponds to "F", and object 2 corresponds to "J"). **C. Example subject-level learning data.** Each subject performing a subtask performed a randomly sampled sequence of 100 trials (i.e. images and their choice-reward contingencies), and we measured their sequence of correct (blue) and incorrect (gray) choices. Image stimuli were never repeated, ensuring each trial tests the subject's ability to generalize to unseen views of the objects. **D. Human learning curves**. We averaged across human subjects to estimate accuracy as a function of trials for n=64 subtasks (each consisting of a distinct pair of objects). We found that some subtasks were found to be reliably harder for humans than others; three example subtasks across the range of difficulty are highlighted. Learning curves shown are smoothed with a moving window filter (for visualization only).

objects is highly dependent on the 3D shapes of those objects, with many object pairs being far from perfectly learned within 100 trials. We next asked how well a family of baseline models based on a standard cognitive theory of learning are – or are not – able to explain these behavioral measurements.

Comparing baseline object learning models to humans

As described above, the core set of behavioral measurements we obtained in Experiment 1 consisted of subject-averaged learning curves for each subtask (accuracy values for 64 subtasks across 100 trials). Our next step was to to specify a procedure that assesses how well any given image-computable model of object learning can quantitatively reproduce those curves.

Given a model, this procedure consisted of 1) simulating the same set of subtasks in the model and estimating its learning curves for those subtasks (exactly analogous to how they were estimated in humans), and 2) computing a scalar-valued error score (MSE_n) which summarizes the extent to which the learning curves of the model quantitatively matches (or not) the corresponding learning curves in humans.

To create baselines for this benchmark, we sought to score models drawn from a standard model family (see Figure 2A). Each model in this family consists of two conceptual stages: an *encoding stage* which re-represents incoming images as points in a representational space, and a tunable *decision stage*, which generates a choice by computing (linear) choice preferences using that representation. The learning of new object-choice associations is guided by an *update rule*, which processes environmental feedback to adjust the linear weights of the tunable decision stage. Learning takes place only in the decision stage; the encoding stage is held completely fixed.

Here, we implemented a large set of such learning models, each based on a different combination of encoding stage and update rule. We considered n=344 encoding stages based on specific intermediate layers of Imagenet-pretrained DCNNs (***Deng et al., 2009***), and n=7 update rules drawn from statistical learning theory and reinforcement learning (see Baseline model family for details). In total, we implemented models based on all possible combinations of these encoding stages and update rules (n=2,408 learning models).

We then tested each model on the same set of subtasks as humans, simulating n=32,000 behavioral sessions per model (n=500 simulations per subtask). We estimated each model's learning curves for each subtask (i.e. accuracy values over 64 subtasks and 100 trials), and scored its similarity to humans using a mean-squared-error statistic (MSE_n; details in Bias-corrected mean squared error). We show an illustration of this procedure for an example model shown in Box 1, and a histogram of MSE_n scores for all tested models is shown in Figure 3.

A higher value of MSE_n means the model is a worse predictor of human learning behavior; lower values are better. In principle, no model can be expected to have an MSE_n lower than a "noise floor" σ_h^2, which comes from the uncertainty in our experimental estimates of each human learning curve (i.e. from the finite amount of human data collected). Intuitively, the value of σ_h^2 can be understood as the MSE_n score that can be expected from the "perfect" model of human learning behavior. We made an unbiased estimate of this noise floor ($\hat{\sigma}_h^2 \approx 0.003$, see Noise floor estimation), then compared this to the MSE_n scores achieved by the models we tested. We note that the *root* value of the noise floor (and MSE_n) is a rough estimate of the expected error in the native units of the measurements (accuracy values).

Many of the models were far from the noise floor, but we found that a subset of models achieved relatively low error. For example, we found the best 1% of the models (which we refer to as "strong baseline models") had root-mean squared errors of $\sqrt{\mathrm{MSE}_n} \approx 0.08$, coming relatively close to the noise floor ($\hat{\sigma}_h \approx 0.05$). Still, all models, including these ones, were statistically distinguishable from humans; all models were rejected as having expected behavior identical to humans with significance level of at least $p < 0.001$ (see Null hypothesis testing).

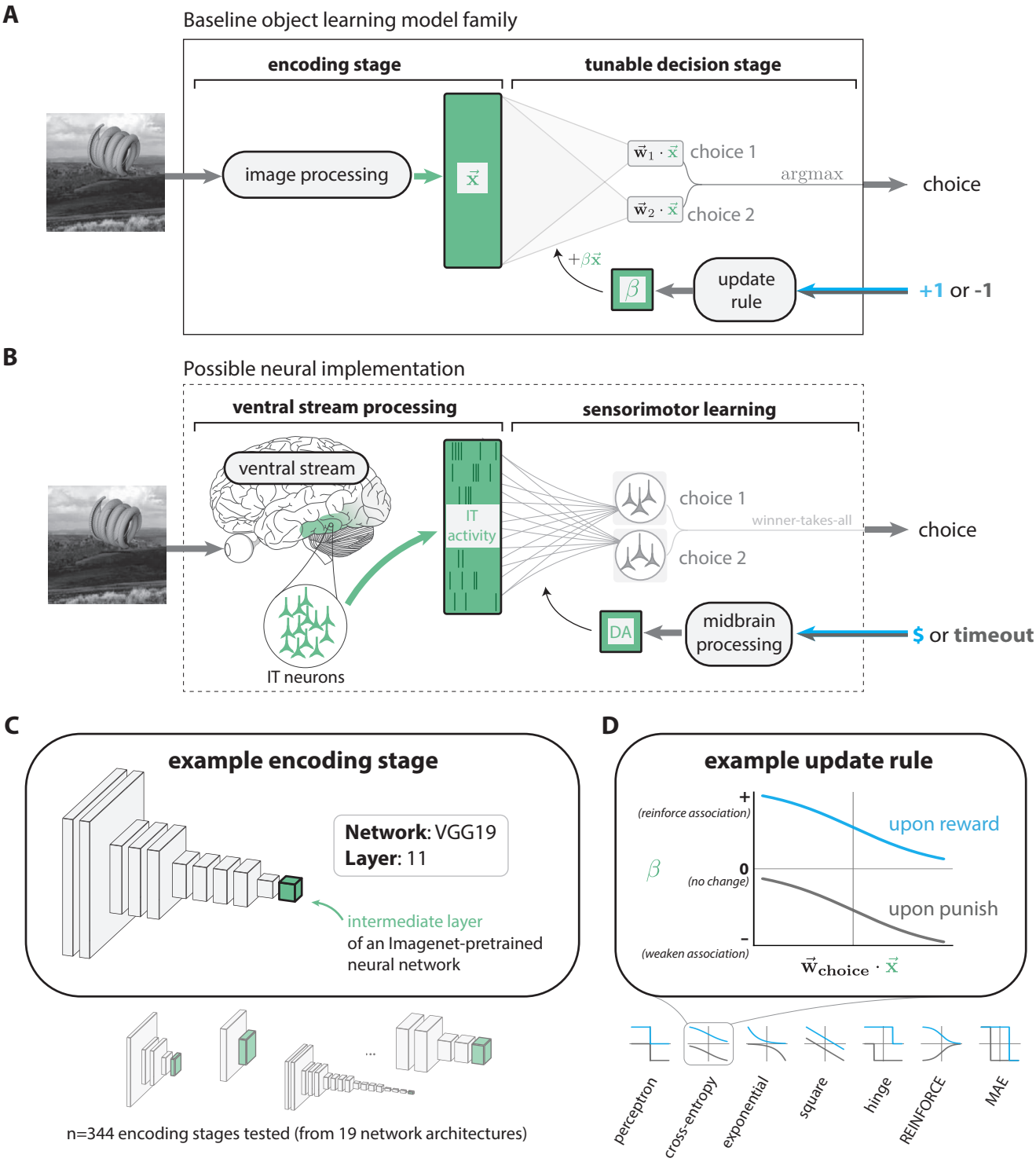
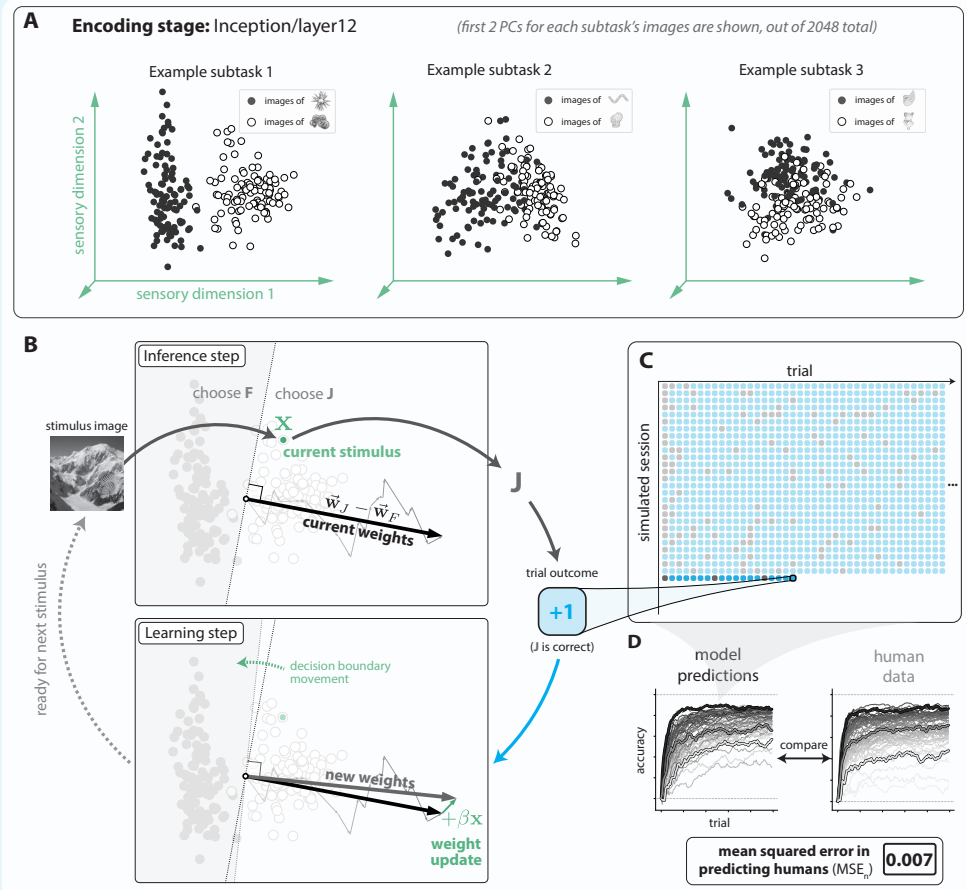


Figure 2. Baseline model family of object learning in humans. **A. Baseline model family.** Each model in this family had two stages: an *encoding stage* which re-represents an incoming pixel image as a vector ($\vec{x}$), and a tunable *decision stage*, which uses $\vec{x}$ to generate a choice by computing linear choice preferences ($\vec{w}_c \cdot \vec{x}$), then selecting the most preferred choice. Subsequent environmental feedback is processed to update the parameters of the decision stage. Specific models in this family correspond to specific choices of the encoding stage and update rule. **B. Possible neural implementation.** The functionality of the encoding stage in **A** could be implemented by the ventral stream, which re-represents incoming retinal images into a pattern of distributed population activity in high level visual areas, such as area IT. Sensorimotor learning might be mediated by plasticity in a downstream association region. Midbrain processing of environmental feedback signals could guide plasticity via dopaminergic (DA) projections. **C. Encoding stages.** We built learning models based on several encoding stages, each based on a specific intermediate layer of an Imagenet-pretrained deep convolutional neural network. **D. Update rules**. We drew update rules from statistical learning theory and reinforcement learning (Table 1). Each rule aims to achieve a slightly different optimization objective (e.g., the "square" rule attempts to minimize the squared error between the choice preference and the subsequent magnitude of reward).

Box 1. Model simulations of human learning.



A. Example encoding stage representations of novel object images. Each subtask consists of images of two novel objects (indicated in black and white dots). The first two principal components of a 2048-dimensional encoding stage (Inception/layer12) are shown here (computed separately on the images for each subtask, for clarity). Linear separability can be observed, to varying degrees. **B. Simulating a single trial.** Clockwise, starting from top left: the incoming stimulus image is re-represented by the encoding stage into a location in a representational space, $\vec{\mathbf{x}}$. Preferences for each choice are computed ($\vec{\mathbf{x}} \cdot \vec{\mathbf{w}}_J$ and $\vec{\mathbf{x}} \cdot \vec{\mathbf{w}}_F$), and the most preferred choice is selected, which amounts to making the choice based on where $\vec{\mathbf{x}}$ falls with respect to a linear decision boundary ($\vec{\mathbf{w}}_J - \vec{\mathbf{w}}_F$). Here, the choice "J" is selected, which happens to be the correct choice for this image. The subsequent reward causes the decision boundary to change based on the update rule. **C. Simulated model behavioral data.** For each learning model, we simulated a total of n=32,000 behavioral sessions (64 subtasks, 500 simulations each), and recorded its behavior (correct or incorrect) on each trial. **D. Comparing model and human behavior.** We averaged across simulations to obtain the model's learning curves for each subtask, then compared them to subject-averaged human learning curves, using a bias-corrected mean-squared error metric (MSE_n; see Bias-corrected mean squared error) to quantify the (dis)similarity of the model to humans.

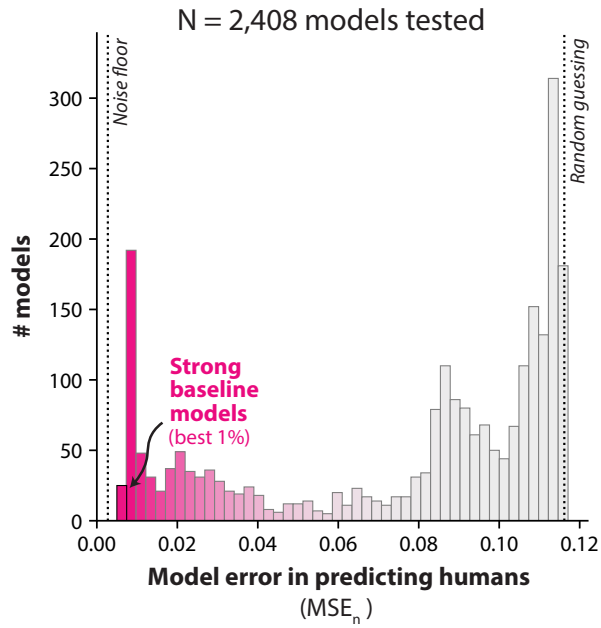


Figure 3. Bias-corrected mean-squared errors (MSE_n) vs. humans for all models tested. The n=2,408 models we tested varied widely in the extent of their alignment with human learning (i.e. their average squared prediction error). We denote the best 1% of such models as *"strong baseline models"*. The noise floor corresponds to an estimate of the lowest possible error achievable ($\sigma_h^2 = 0.003$), given the experimental power in this study. The vertical line labeled "random guessing" marks the error incurred by a model which produces a random behavioral output on each trial.

Model components affecting the score of a model

Given the range of MSE_n scores we observed, we next wished to perform a secondary analysis on how each of the two components defining a baseline model (its encoding stage and update rule) affected its error score on the benchmark above. One general trend we observed was that models built with encoding stages from deeper layers of DCNNs tended to produce more human-like learning behavior (see Figure 2B). On the other hand, the choice of update rule (which defined by the tunable decision stage) appeared to have little effect in a model's ability to generate human-like learning behavior (see Figure 4A).

We quantified these observations by performing a two-way ANOVA over all model scores (see Additional Analyses), treating the update rule and encoding stage as the two factors. This analysis showed that the choice of update rules explained less than 0.1% of the variation in model scores; by contrast, 99.8% of the variation was driven by the encoding stage, showing that the predominant factor defining the behavior of the learning model was the encoding stage.

Strong baseline models are largely, but not perfectly, correlated with human performance patterns

The set of MSE_n scores indicated that *all* baseline models tested had significant differences from humans, based on their respective learning curves. There are several ways in which these differences could originate – for example, a model might have "fast" learning curves for tasks that humans learn slowly, and "slow" learning curves for tasks that humans learn rapidly. Alternatively, a model might have the same pattern of difficulty across subtasks as humans, but simply be slower at learning, overall.

To gain insight into these possibilities, we performed an additional analysis in which we compared each model to humans along two more granular statistics: 1) its *overall accuracy* over all subtasks and trials tested, and 2) its *consistency* with the patterns of difficulty exhibited by humans

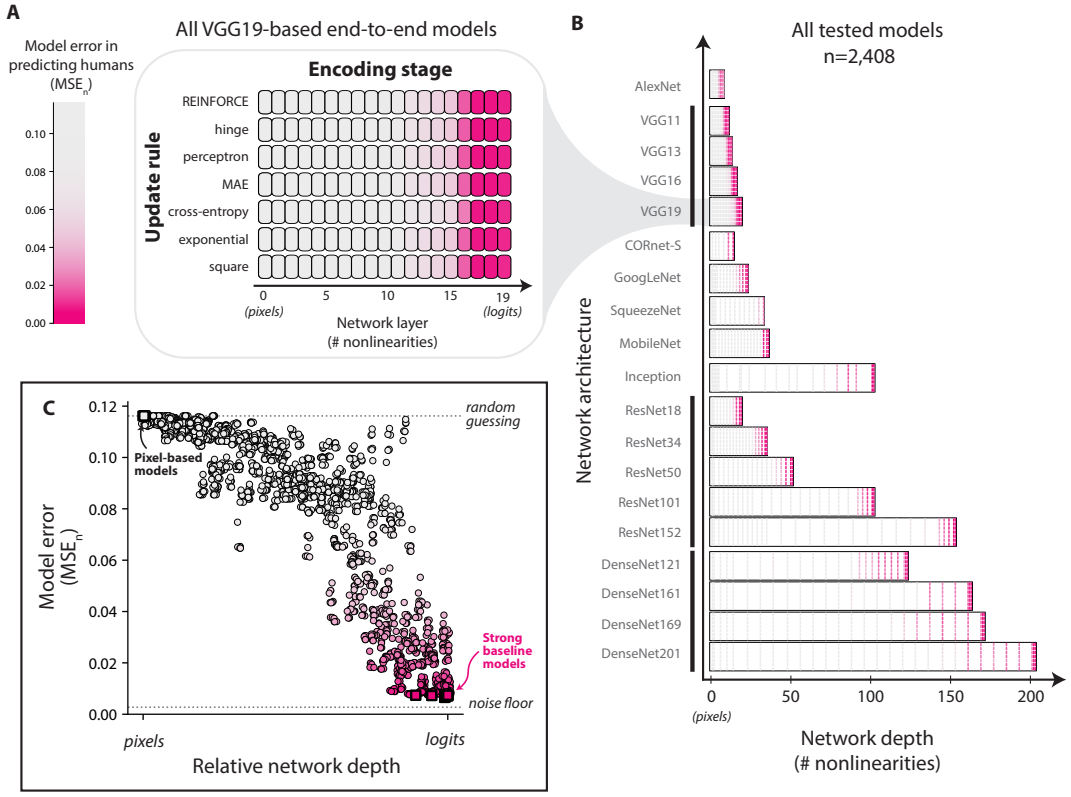


Figure 4. Evaluating the effect of model design choices on predictive accuracy of human learning. **A. Example model scores across encoding stages and update rules.** A typical example of the relative contributions of update rule (y-axis) and encoding stage (x-axis) on model scores (MSE_n, encoded by color). **B. Overview of all models tested.** In total, we tested encoding stages drawn from a total of n=19 DCNN architectures, varying widely in depth. A model's similarity to humans was highly affected by the choice of encoding stage; those based on deeper layers of DCNNs showed the most human-like learning behavior. On the other hand, the choice of update rules had a minuscule effect. **C. Predictive accuracy increases as a function of relative network depth.** Learning models with encoding stages based on DCNN layers closer to the final layer of the architecture tended to be better.

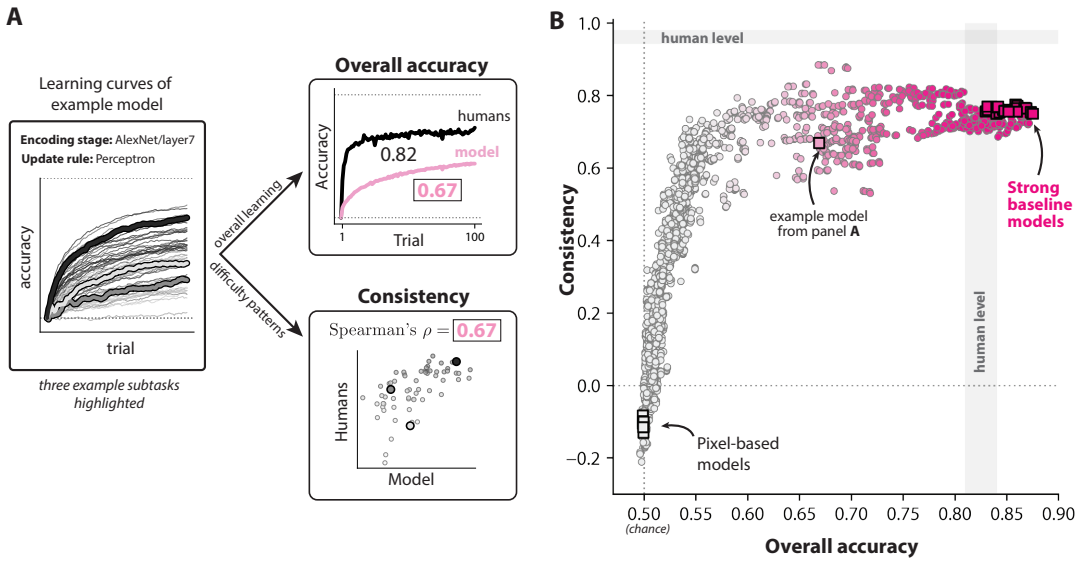


Figure 5. Comparing models to humans along more granular behavioral signatures of learning. **A. Decomposing model behavior into two metrics.** We examined model behavior along two specific aspects of learning behavior: **overall accuracy** (top right), which is the average accuracy of the model over the entire experiment (i.e. averaging across all 100 trials and all 64 subtasks), and **consistency** (bottom right), which conveys how well a model's pattern of trial-averaged performance over different subtasks rank-correlates with that of humans (Spearman's rank correlation coefficient). **B. Consistency and overall accuracy for all models.** Strong baseline models (top-right) matched (or exceeded) humans in terms of overall accuracy, and had similar (but not identical) patterns of performance with humans (consistency). The gray regions are the bootstrap-estimated 95% range of each statistic between independent repetitions of the behavioral experiment. The color map encodes the overall score (MSE_n) of each model (colorbar in Figure 4A).

across subtasks (see Figure 5A).

Intuitively, overall accuracy is a gross measure of a learning system's overall ability to learn, ranging from 0.5 (chance, no learning occurs) to 0.995 (learning completed after just one trial). Consistency (ρ) quantifies the extent to which a model finds the same subtasks easy and hard as humans, and ranges from $\rho = -1$ (perfectly anticorrelated pattern of performance) to $\rho = 1$ (perfectly correlated pattern of performance). A value of $\rho = 0$ indicates no correlation between the patterns of difficulty across subtasks in a model and humans.

These metrics are theoretically unrelated to each other; given any overall accuracy,[1] a model may have a high or low consistency with humans, and vice versa. Nevertheless, we observed that these two metrics strongly covaried for these models; models with high overall accuracy also tended to have high consistency (see Figure 5B).

Humans learn new objects faster than all tested baseline models in low-sample regimes

Though many baseline models matched or exceeded human-level overall accuracy (i.e. accuracy averaged over trials 1-100 for all subtasks), we noticed that all models' accuracy *early on* in learning consistently appeared to be below that of humans (see Figure 6A). We tested for this by comparing the accuracy of models and humans in an initial phase of learning (trials 1-5 for all subtasks), and indeed found that *all* of the baseline models were significantly worse than humans in the early phase of learning (all p<0.05, bootstrap hypothesis test, see Figure 6B).

We wondered whether this gap was present across levels of difficulty (e.g., that models tended to perform particularly poorly on "hard" subtasks relative to humans, but were human-level for other subtasks), and repeated this analysis across four different difficulty levels of subtasks (where

[1] Except for models which are either at chance or perform perfect one-shot learning in all situations; then the correlation coefficient must be undefined.

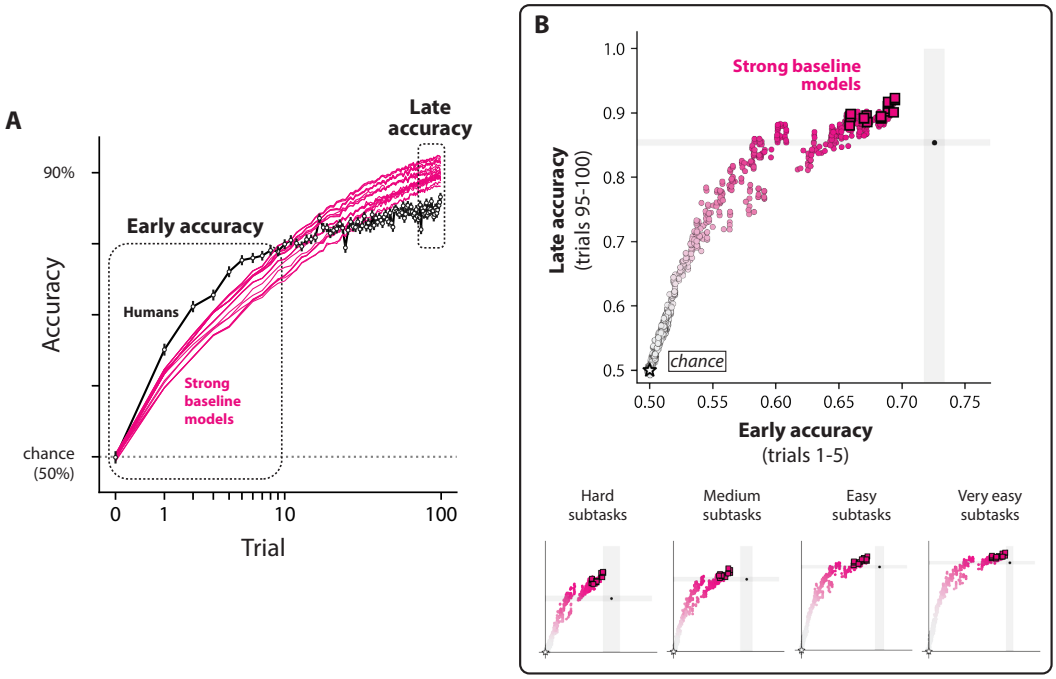


Figure 6. Humans outperform all strong baseline models in low-sample regimes. **A. Subtask-averaged learning curves for humans and strong baseline models.** The y-axis is the percent chance that the subject made the correct object report (chance is 50%). The x-axis is the total number of image examples shown prior to the test trial (log scale). The average learning curves for humans (black) and models (magenta) show that humans outperform all models in low-sample learning regimes. Errorbars on the learning curves are the bootstrapped SEM; model errorbars are not visible. **B. No model achieves human-level early accuracy.** We tested all models for whether they could match humans early on in learning. Several models (including all strong baseline models) were capable of matching or exceeding late accuracy in humans (average accuracy over trials 95-100), but no model reached human-level accuracy in the early regime (average over trials 1-5). This trend was present in subtasks across different levels of difficulty (bottom row). The gray region shows the 95% bootstrapped CI for each statistic; 95% CIs for models are too small to be shown.

each level consisted of 16 out of the 64 total subtasks we tested, grouped by human difficulty levels). We found models were consistently slower than humans across the difficulty range, though we could not reject a subset (11/20) of the strong baseline models at the easiest and hardest levels (see Figure 6B).

Lastly, though all models failed to match humans in the early regime, many models readily matched or exceeded human performance late in learning (i.e. the average accuracy on trials 95-100 of the experiment).

Experiment 2: Characterizing one-shot object learning in humans

Our observation above suggested baseline models learn more slowly than humans in few-shot learning regimes involving random views of novel objects. To further characterize possible differences between models and humans in this “early learning" regime, we performed an additional behavioral experiment (Experiment 2) in which we measured the ability of humans to generalize following experience with only a single image of each object category.

Experiment 2 followed the same task paradigm as Experiment 1 (binary discrimination learning with evaluative feedback). Each behavioral session was based on one of 32 possible subtasks (i.e. 32 possible pairs of novel objects), and began with a "training phase" of 10 trials in which the subject acquired an object-response contingency using a single canonical image for each of the two objects. After the training phase, we then asked subjects to perform trials with "test images" consisting of transformed versions of the two training images (see One-shot behavioral testing for further details).

These test images were generated by applying the five kinds of image variations present in our original experiment (translation, scale, random backgrounds, in-plane object rotation, and out-of-plane object rotation) to the test images. We also generated test images using four additional kinds of image variation that were not present in the original experiment (contrast shifts, pixel deletion, blur, and shot noise), but might nonetheless serve as informative comparisons for identifying functional deficiencies in a model relative to humans. For each kind of transformation, we tested four "levels" of variation. For example, in measuring humans’ one-shot test accuracy to scale, we showed subjects images where the object was resized to 12.5%, 25%, 50%, and 150% of the size of the object in the original training image (see One-shot stimulus image generation for details).

Under this experimental setup, the core measurements we sought to obtain was the subject-averaged discrimination accuracy for n=36 generalization tests (9 transformation types, 4 levels of variation each). Intuitively, each of the n=36 measurements is an estimate of a typical human’s ability to successfully generalize to a specific kind and magnitude of view transformation (e.g. down-scaling the size of an object by 50%), after exposure to a single positive and negative example of a new object. Unlike Experiment 1, here we combined observations across the 32 subtasks used in this experiment, ignoring the fact that there may be variation in these measurements based on the specific objects involved. We also attempted to correct for any memory or attentional lapses in these estimates (see One-shot behavioral statistics in humans for details).

Across these 36 test conditions, we found that humans had varied patterns of generalization (Figure 7B). For example, we observed that accuracy varied systematically based on the level of variation applied with respect to scale, out-of-plane rotation, and blur. On the other hand, human subjects had nearly perfect generalization across all tested levels of variation for translations, backgrounds, contrast shifts, and in-plane rotations. Overall, these diverse patterns of generalization were estimated with a relatively high degree of experimental precision, as quantified by our estimates of the human noise floor for this experiment (root noise floor of $\hat{\sigma}_h \approx 0.02$).

We next used these measurements to create a benchmark that could be used to compare any image-computable object learning model – including the baselines considered in this study – against human object learning in this one-shot setting.

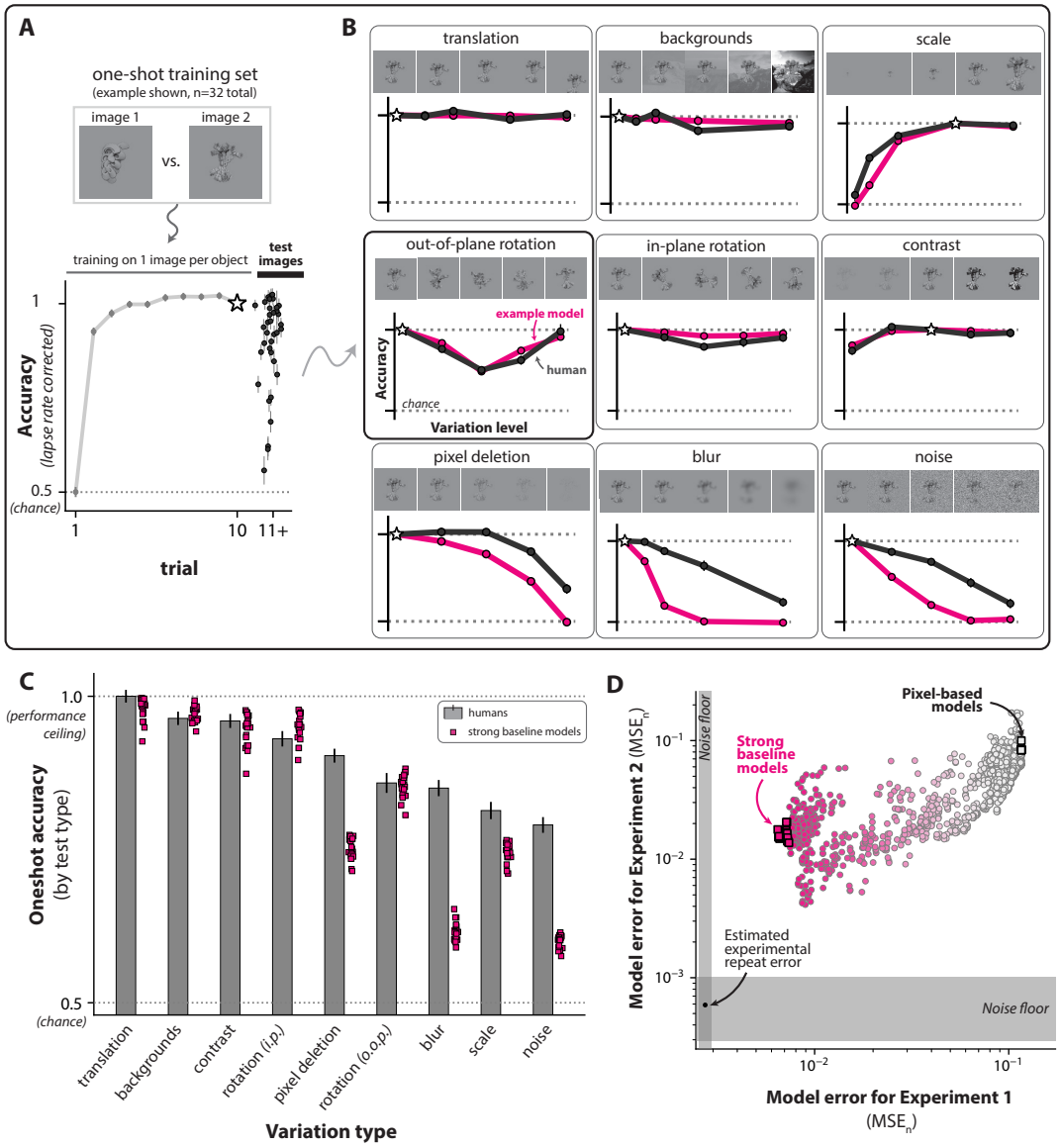


Figure 7. One-shot learning in humans. **A. One-shot learning task paradigm.** We performed an additional study (Experiment 2) to characterize human one-shot learning abilities (using the same task paradigm in Figure 1). The first 10 trials were based on two images (n=1 image per object) that were resampled in a random order. On trials 11-20, humans were tested on transformed versions of those two images (nine types of variation, four variation levels, n=36 total generalization tests) **B. Human and example model one-shot accuracy for all generalization tests.** An example strong baseline model's pattern of generalization (magenta) is shown overlaid against that of humans. **C. Humans outperform strong baseline models on some kinds of image variations.** We averaged human one-shot accuracy (gray) on each type of image variation, and overlaid all strong baseline models (magenta). The errorbars are the the 95% CI (basic bootstrap). **D. Comparison of MSE_n scores for Experiment 1 and 2.** No strong baseline model could fully explain the pattern of one-shot generalization observed in humans (Experiment 2), nor their behavior on the first benchmark (Experiment 1). The error scores are shown on the log scale.

Baseline models show weaker one-shot generalization compared to humans

As in Experiment 1, the benchmarking procedure for Experiment 2 consisted of 1) generating predictions of behavior from a model by having it perform the same experiment conducted in humans, then 2) scoring the similarity of that behavior to humans using an error statistic (MSE_n). Thus, for each baseline model, we replicated the one-shot behavioral experiment (n=16,000 simulated sessions per model), measured their accuracy on each of the 36 generalization tests described above, then compared those behavioral predictions to humans using MSE_n.

Similar to our results from Experiment 1, here we found that models varied widely in their alignment with human learning behavior, and again found the top 1% subset of models achieved relatively low error ($\sqrt{\mathrm{MSE}_n} \approx 0.06$, where the root noise floor is approximately $\hat{\sigma}_h \approx 0.02$). And as in Experiment 1, we found that *all* models had statistically significant differences in their behavior relative to humans, for this experiment. We also observed a positive relationship between the scores of the two benchmarks: models that were most human-like as evaluated by the benchmark based on Experiment 1 also tended to be the most human-like here, in the one-shot setting (see Figure 7D) – though *no* model explained human behavior in either experiment to the limits of statistical noise.

Part of the prediction failures we observed here lay in a failure to generalize as well as humans to several kinds of image variation. For example, we observed that all strong baseline models (identified from the benchmark from Experiment 1) had lower one-shot accuracy than humans in the presence of object pixel deletions, blur, shot noise, and scale shifts (see Figure 7C).

Specific individual humans outperform all baseline models

Both benchmarks we developed in this study tested the ability of a model to predict human object learning at the "subject-averaged" level, where behavioral measurements drawn from several subjects are averaged together. This approach, by design, ignores any individual differences in learning behavior that may exist.

We wished to gauge the extent to which any such individual differences were present, and we performed an analysis on our behavioral data from Experiment 1. We identified subjects who performed all 64 subtasks in that experiment (22 out of 70 subjects total) . We then attempted to reject the null hypothesis that there was no significant variation in their overall learning ability (see Additional Analyses). If this hypothesis were to be rejected, it would indicate that individuals must systematically vary in their learning behavior, at least in terms of their overall performance on these tasks. We indeed found that some subjects were reliably better object learners than others ($p < 1\text{e-}4$, permutation test).

Given this was the case, we next asked whether any of these individuals had an overall performance level higher than that of the highest performing model we identified in Experiment 1 (an encoding stage based on *ResNet152/avgpool*, and a tunable decision stage using the square update rule). We identified n=5 individuals whose overall accuracy significantly exceeded that of this model (all $p<1\text{e}{-5}$, Welch's t-test, Bonferroni corrected). On average, this subset of humans had an overall accuracy of 0.92 ± 0.01 (SEM over subjects); this was around ~4% higher than this model's average of 0.88.

Discussion

An understanding of how humans accomplish visual object learning remains an open scientific problem. A necessary step to solve this problem is evaluating the predictive validity of alternative models with respect to measurements of human object learning behavior. In this study, we collected a set of such measurements across a variety of object learning settings (n=371k trials), which allowed us to quantify the speed of human object learning (<10 trials to achieve close-to-asymptotic accuracy), the distinct pattern of learning difficulty they have for different objects, and their extent of generalization to specific image transformations after a single image example.

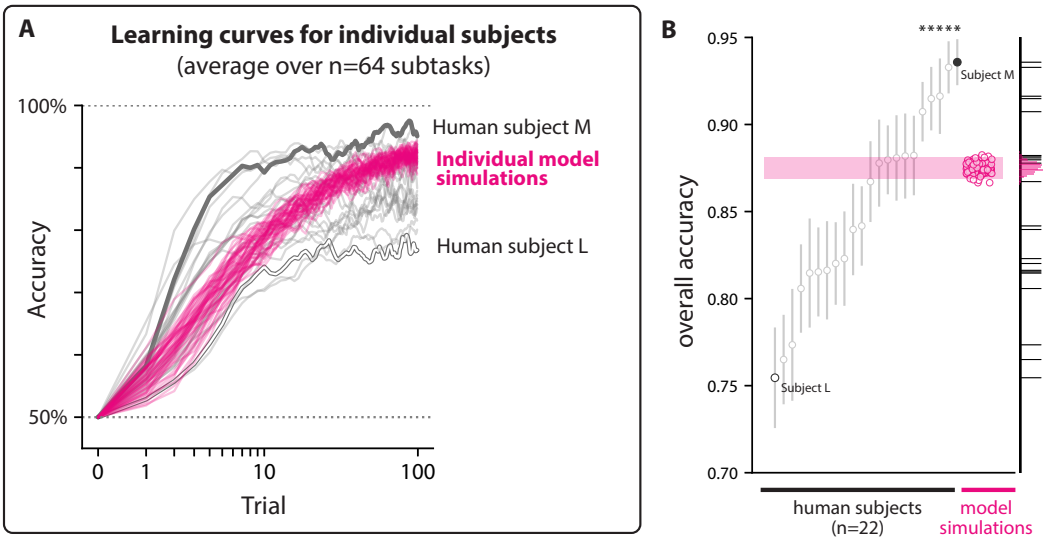


Figure 8. Individual differences in learning ability. **A. Individual-level learning curves.** We identified 22 subjects who performed all 64 subtasks in Experiment 1, and computed their subtask-averaged learning curves. Each gray curve corresponds to the learning curve for a different individual subject (smoothed using state-space estimation from ***Smith and Brown*** (***2003***)). In humans, a range of overall learning performance is seen: some subjects consistently outperformed others (e.g. Subject M, highest accuracy over all trials and subtasks), while others consistently underperformed (e.g. Subject L, lowest average accuracy). In magenta are subtask-averaged learning curves corresponding to individual model simulations from the highest-performing model e tested in this study (encoding stage = ResNet152/avgpool, update rule = square). **B. Some individual humans outperform all baseline models.** Five out of 22 subjects had significantly higher overall performance than the highest performing model we tested (one-tailed Welch's t-test, Bonferroni corrected, $p<0.05$).

We then developed procedures to evaluate any given image-computable object learning model over those same learning settings (i.e., behavioral benchmarks for learning), and tested a set of baseline object learning models (n=2,408 models) on those benchmarks. Each of these models consisted of two stages: 1) a fixed *encoding stage* based on an intermediate representation contained in a deep convolutional neural network model (DCNN), followed by 2) a tunable *decision stage* which learns new object-choice contingencies by adjusting weighted sums of that representation.

Prior to this study, we did not know if some or any of these baseline models might be capable of explaining human object learning as assessed here. As such, we center our discussion on these baseline models, but highlight that our raw behavioral data (and associated benchmarks) are now a publicly available resource for testing image-computable object learning models beyond those evaluated here [GitHub].

Strengths and weaknesses of current baseline object learning models

Linear learning on deep representations as strong baseline models of human object learning

On our first benchmark, which compares a learning model to humans under high view-variation learning conditions, we found a subset of baseline models produced relatively accurate predictions of human learning behavior. We point out the observed accuracy of these models does not simply originate from the fact they can successfully learn new objects – they also *fail* to rapidly learn certain objects that humans also find difficult to learn (Figure 5B), suggesting they have nontrivial similarities with humans, at least behaviorally.

We were surprised by the extent of similarity we observed between these baseline models and humans, because some have suggested that DCNNs are unlikely to provide adequate descriptions of human learning (e.g. ***Lake et al.*** (***2015***, 2017); ***Marcus*** (***2018***); ***Ullman*** (***2019***); ***Saxe et al.*** (***2020***)). Contrary to this belief, the results reported here suggest that specific models based on current DCNNs, though imperfect, are a reasonable starting point to quantitatively account for the ability (and inability) of humans to learn specific, new objects.

It is worth noting the models we considered are composed only of operations that closely hew to those executed by first-order models of neurons – namely, linear summation of upstream population activity, ramping nonlinearities, and adjustment of associational strengths at a single visuomotor interface. This makes them not only plausible descriptions for the computations executed by the brain over object learning, but, with some additional assumptions (Figure 2B), they make predictions of neural phenomena.

For example, if the interpretation suggested in Figure 2B is taken at face value, the baseline models make a couple of qualitative predictions. First, given the assumption that the encoding stage corresponds to the output of the ventral visual stream, these models predict that ventral stream representations used by humans over object learning need not undergo plastic changes to mediate behavioral improvements over the duration of the experiments we conducted (seconds-to-minutes timescale). This prediction is in line with prior studies showing adult ventral stream changes are typically moderate and take place on longer timescales (see ***Op de Beeck and Baker*** (***2010***) for review).

Even if the neural changes that underlie learning are not distributed over the entire visual processing stream, it remains possible that any changes driving learning are distributed across a series of downstream associational and premotor areas. However, we note that the learning mechanisms utilized by the baseline models here (linear reweighting of activity in an upstream representation) could plausibly be conducted at a single visuomotor synaptic interface, where reward-based signals are available. Several regions downstream of the ventral visual stream are possible candidates for this locus of plasticity during invariant object learning; we point to striatal regions receiving both high-level visual inputs and midbrain dopaminergic signals and involved in premotor processing, such as the caudate nucleus, as one set of candidates (***Seger, 2005***; ***Kim et al., 2014***).

Baseline models at predicting human few-shot learning

Despite the relative strength of some baseline models we tested, all models we tested were unable to fully explain human behavior on either benchmark. One consistent prediction failure we observed in all models was a failure to learn new objects as rapidly as humans in low-sample regimes. We found this to be the case in both Experiment 1 (see Figure 6) and in Experiment 2, where we found that all tested models had lower accuracy than humans after one-shot across a variety of generalization tests (see Figure 7C). For example, we found that these models cannot one-shot generalize as well to scale shifts as humans, replicating previous work (***Han et al., 2020***).

Taken together, these observations support a central inference of this work: all baseline learning models we tested are currently unable to account for the human ability to learn in the few-shot regime. We next discuss possible next steps to close these (and other) predictive gaps.

Future object learning models to be tested

There are several ways to improve the predictive accuracy of the models we tested in this study (i.e. to find more human-like models). For example, it is possible that another model from this model family could fully predict human learning on the benchmarks in this study, and we simply failed to implement and test it here. If that is the case, such a model could differ from the ones we tested here along two main components: its encoding stage, and/or its update rule. Here, we found the choice of update rule had little effect on the predictive power of these models (see Figure 4A), and did not interact significantly with the choice of encoding stage. Still, we note there are some update rules which we did not consider here – namely exemplar (***Nosofsky, 2011***) and prototype-based rules (***Reed, 1972***; ***Medin and Smith, 1981***; ***Sorscher et al., 2022***).

On the other hand, we found the image-computable representation that was used for each model's encoding stage had a large effect on its overall predictive power as an object learning model (on both benchmarks), which suggests alternate encoding stages could lead to more accurate models of human learning. Here, we only considered encoding stages based on current Imagenet-pretrained DCNN representations; these are known to only partially match primate visual representations as measured by electrophysiological studies (***Schrimpf et al., 2020***) and behavioral studies (***Rajalingham et al., 2018***; ***Jacob et al., 2021***; ***Bowers et al., 2022***). If image-computable representations that more closely adhere to human visual representations are built and/or identified, they might lead to better object learning models that close the prediction gap on the benchmarks we developed here.

Stepping back, it is also possible that *no* model from this model family could lead to fully accurate predictions on these benchmarks (or future benchmarks), but other types of models might do so. In that regard, we highlight that we did not test the full array of available cognitive theories of object learning in this study, and there may be other promising approaches that score well on the benchmarks collected here. For example, one influential class of cognitive theories posits that the brain learns new objects by building structured, internal models of those objects from image exemplars, then uses those internal models to infer the latent content of each new image (***Bülthoff and Edelman, 1992***; ***Erdogan and Jacobs, 2017***; ***Griffiths et al., 2010***; ***Li et al., 2006***; ***Lake et al., 2015***; ***Poggio and Edelman, 1990***). It is possible that models based on these alternate approaches could generate more human-like learning over the tasks we tested here, and could be the key to achieving a full computational description of human object learning. In any case, implementing and testing these models on the benchmarks here is an important direction for future work.

Future extensions of object learning benchmarks

Extensions of task paradigm

The two benchmarks we developed here certainly do not encompass all aspects of object learning. For example, each benchmark focused on discrimination learning between two novel objects, but humans can potentially learn and report on many more objects simultaneously. Moreover, humans can readily learn object *categories* at different levels of abstraction, each of which may

encompass multiple specific objects (***Rosch et al., 1976***). Scaling up the number of objects being learned simultaneously, and/or the complexity of the categories being learned, is a natural extension of the work done here. The baseline models tested here scale naturally to task paradigms involving additional objects (via the incorporation of new linear choice preferences to the decision stage), and are potentially capable of learning multi-object categories; comparing them to humans in those richer learning settings (and identifying any of their limits in those settings) would strongly motivate the consideration of more complex models.

Extending stimulus presentation time

For presenting stimuli, we followed conventions used in previous visual neuroscience studies (***Rajalingham et al., 2015***, ***2018***) of object perception: achromatic images containing single objects rendered with high view uncertainty on random backgrounds, presented at <10 degrees of visual field and for <200 milliseconds.

The chosen stimulus presentation time of 200 milliseconds is too short for a subject to initiate a saccadic eye movement based on the content of the image (***Purves et al., 2001***). Such a choice simplifies the input of any model (i.e., to a single image, rather than the series of images induced by saccades); on the other hand, active viewing of an image via target-directed saccades might be a central mechanism deployed by humans to mediate learning of new objects.

We note that if this is the case, our task paradigm (which would disrupt any such saccade-based mechanisms from being used) would be *underestimating* the number of images needed by humans to achieve learning on new objects, compared to a scenario in which subjects had unlimited viewing time on each trial. Thus, removing such a bias in our experimental design would only strengthen our central inference relating to accuracy, which is that none of the models we tested learn as rapidly as humans given a small number of image examples.

Beyond extending viewing time, designing tasks which more closely hew to typical object learning contexts for humans (e.g. involving colored images and/or movies of potentially multiple objects physically embedded in natural scenes) will be an important direction for future work.

Differences between individual subjects

We primarily focused on studying human learning at the subject-averaged level, where behavioral measurements are averaged across several individuals (i.e. subject-averaged learning curves; see Figure 1D). However, individual humans may have systematic differences in their learning behavior that are (by design) ignored with this approach.

For example, we found that individual subjects may differ in their overall learning abilities: we identified a subpopulation of humans who were significantly more proficient at learning compared to other humans (see Figure 8B). We did not attempt to model this individual variability in this study; whether these differences can be explained by alterations to this model family, if at all, (e.g. through the introduction of random effects to the parameters of the encoding stage and/or update rules) remains an area for future study.

Performing subject-averaging also leads to the masking of learning dynamics that may only be identifiable at the level of single subjects, such as "delayed rises" or "jumps" in accuracy at potentially unpredictable points in a learning session (***Gallistel et al., 2004***). Performing analyses to compare any such learning dynamics between individual humans and learning models is another important extension of our work.

Lastly, we did not attempt to model any systematic increases in a subject's learning performance as they performed more and more subtasks available to them (in either Experiment 1 or 2). This phenomenon (learning-to-learn, learning sets, or meta-learning) is well-known in psychology (***Harlow, 1949***), but to our knowledge has not been systematically measured or modeled in the domain of human object learning. Expanding these benchmarks (and models) to measure and account for such effects is an important future step of building models of the work done here.

Materials and Methods

Overview of experiments

For both experiments, the core measurement we sought to obtain was the discrimination performance of a typical subject as they received increasing numbers of exposures to images of the to-be-learned (i.e. new) objects.

We assumed that different pairs of objects result in potentially different rates of learning, and we wanted to capture those differences. Thus, in Experiment 1, we aimed to survey the empirical landscape of this human ability by acquiring this learning curve measurement for many different pairs of objects (n=64 pairs). Specifically, for each pair of to-be-learned objects (referred to as a "subtask"), we aimed to measure (subject-averaged) human learning performance across 100 learning trials, where each trial presented a test image generated by one of the objects under high viewpoint uncertainty (e.g. random backgrounds, object location, and scale). We refer to this 100-dimensional set of measurements as the *learning curve* for each subtask.

In Experiment 2, we aimed to measure the pattern of human learning that results from their experience with just a *single* canonical example of each of the to-be-learned objects (a.k.a. "one-shot learning"). Specifically, we wished to measure the pattern of human discrimination ability over various kinds of identity-preserving image transformations (e.g, object scaling, transformation, and rotation). In total, we tested nine kinds of transformations. We anticipated that humans would show distinct patterns of generalization across these transformations, and we aimed to measure the human commonalities in those patterns (i.e. averages across subjects).

Experiments 1 and 2 both utilized a two-way object learning task paradigm that is conceptually outlined in Figure 1B. The two experiments differed only in the manner in which test images were generated and sampled for presentation, and we describe those differences in detail in their respective sections. Before that, we provide more detail on the specific procedures and parameters we used to implement the common two-way object learning task paradigm.

Task paradigm

For both experiments, human subjects were recruited from Mechanical Turk (***Paolacci et al., 2010***), and ran tasks on their personal computers. Demographic information (age, sex, gender, or ethnicity) was not collected; all subjects were anonymous. We designed these experiments in accordance to a protocol approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (Protocol # 0812003043A017).

Each experiment (Experiments 1 & 2) consisted of a set of subtasks. For each subtask, we asked a population of human subjects to learn that subtask, and we refer to the collection of trials corresponding to a specific subject in a subtask as a "session".

At the beginning of each session, the subject was instructed that there would be two possible objects – one belonging to the "F" category and the other belonging to the "J" category. The subject's goal was to correctly indicate the category assignment for each test image. The specific instructions were: *"On each trial, you'll view a rapidly flashed image of an object. Your task is to figure out which button to press (either "F" or "J" on your keyboard) after viewing a particular image. Each button corresponds to an object (for example, a car might correspond to F, while a dog might correspond to J)."*

Subjects were also informed that they would receive a monetary bonus (in addition to a base payment) for each correctly indicated test image, incentivizing them to learn. We next describe the structure of a single trial in detail below.

Test image presentation

Each trial began with a display start screen that was uniformly gray except for a small black dot at the center of the screen, which reliably indicated the future center of each test image.[2] In this phase,

[2]The center of the test image is not necessarily the same as the center of the object in the test image.

the subject could initiate the trial by pressing the space bar on their keyboard. Once pressed, a test image (occupying ~6° of the visual field) belonging to one of the two possible object categories immediately appeared. That test image remained on the screen for ~200 milliseconds before disappearing (and returning the screen to uniform gray).[3]

For each subject and each trial, the test image was selected by first randomly picking (with equal probability) one of the two objects as the generator of the test image. Then, given that selected object, an image of that object was randomly selected from a pool of pre-rendered possible images. Test images were always selected without replacement (i.e. once selected, that test image was removed from the pool of possible future test images for that behavioral session).

Subject choice reporting

Fifty milliseconds after the disappearance of the test image, the display cued the subject to report the object that was "in" the image. The display showed two identical white circles – one on the lower left side of the fixation point and the other on the lower right side of the fixation point. The subject was previously instructed to select either the "F" or "J" keys on their keyboard. We randomly selected one of the two possible object-to-key mappings prior to the start of each session, and held it fixed throughout the entire session. This mapping was not told to the subject; thus, on the first trial, subjects were (by design) at chance accuracy.

To achieve perfect performance, a subject would need to associate each test image of an object to its corresponding action choice, and not to the other choice (i.e., achieving a true positive rate of 1 and a false positive rate of 0).

Subjects had up to 10 seconds to make their choice. If they failed to make a selection within that time, the task returned to the trial initiation phase (above) and the outcome of the trial was regarded as being equivalent to the selection of the incorrect choice.[4].

Trial feedback

As subjects received feedback which informed them whether their choice was correct or incorrect (i.e. corresponding to the object that was present in the preceding image or not), they could in principle learn object-to-action associations that enabled them to make correct choices on future trials.

Trial feedback was provided immediately after the subject's choice was made. If they made the correct choice, the display changed to a feedback screen that displayed a reward cue (a green checkmark). If they made an error, a black "x" was displayed instead. Reward cues remained on the screen for 50 milliseconds, and were accompanied by an increment to their monetary reward (see above). Error cues remained on the screen for 500 milliseconds. Following either feedback screen, a 50 millisecond delay occurred, consisting of a uniform gray background. Finally, the display returned to the start screen, and the subject was free to initiate the next trial.

Experiment 1: Learning objects under high view variation

Our primary human learning benchmark (Experiment 1) was based on measurements of human learning curves over subtasks involving images of novel objects rendered under high view-variation. We describe our procedure for generating those images, collecting human behavioral measurements, and benchmarking models against those measurements below.

High-variation stimulus image generation

We designed 3D object models (n=128) using the "Mutator" generative design process (***Todd and Latham, 1992***). We generated a collection of images for each of those 3D objects using the POV-Ray

[3]We assumed our subjects used computer monitors with a 16:9 aspect ratio, and naturally positioned themselves so the horizontal extent of the monitor occupied between 40°-70° degrees of their visual field. Under that assumption, we estimate the visual angle of the stimulus would vary between a minimum and maximum of $\approx 4^\circ - 8^\circ$. Given a monitor has a 60 Hz refresh rate, we expect the actual test image duration to vary between $\approx 183 - 217$ milliseconds.

[4]In practice, this was quite rare and corresponded to ~0.04% of all trials that are included in the results in this work.

rendering program (***Persistence of Vision Pty. Ltd., 2004***). To generate each image, we randomly selected the viewing parameters of the object, including its projected size on the image plane (25%-50% of total image size, uniformly sampled), its location ($\pm 40\%$ translation from image center for both x and y planes, uniformly sampled), and its pose relative to the camera (uniformly sampled random 3D rotations). We then superimposed this view on top of a random, naturalistic background drawn from a database used in a previously reported study (***Rajalingham et al., 2015***). All images used in this experiment were grayscale, and generated at a resolution of 256x256 pixels. We show an example of 32 objects (out of 128 total) in Figure 1A, along with example stimulus images for two of those objects on the right.

Design of subtasks

We randomly paired the 128 novel objects described above into pairs (without replacement) to create n=64 subtasks for Experiment 1, each consisting of a distinct pair of novel objects. Each behavioral session for a subtask consisted of 100 trials, regardless of the subject's performance. On each trial of a session, one of the two objects was randomly selected, and then a test image of that object was drawn randomly without replacement from a pre-rendered set of 100 images of that object (generated using the process above). That test image was then presented to the subject (as described in Test image presentation). We collected 50 sessions per subtask and all sessions for each subtask were obtained from separate human subjects, each of whom we believe had not seen images of either of the subtask's objects before participation.

Subject recruitment and data collection

Human subjects were recruited on the Mechanical Turk platform (***Paolacci et al., 2010***) through a two-step screening process. The goal of the first step was to verify that our task software successfully ran on their personal computer, and to ensure our subject population understood the instructions. To do this, subjects were asked to perform a prescreening subtask with two common objects (elephant vs. bear) using 100 trials of the behavioral task paradigm (described in Task paradigm above). If the subject failed to complete this task with an average overall accuracy of at least 85%, we intentionally excluded them from all subsequent experiments in this study.

The goal of the second step was to allow subjects to further familiarize themselves with the task paradigm. To do this, we asked subjects to complete a series of four "warmup" subtasks, each involving two novel objects (generated using the same "Mutator" software, but distinct from the 128 described above). Subjects who completed all four of these warmup subtasks, regardless of accuracy, were enrolled in Experiment 1. Data for these warmup subtasks were not included in any analysis presented in this study. In total, we recruited n=70 individual Mechanical Turk workers for Experiment 1.

Once a subject was recruited (above), they were allowed to perform as many of the 64 subtasks as they wanted, though they were not allowed to perform the same subtask more than once (median n=61 total subtasks completed, min=1, max=64). We aimed to measure 50 sessions per subtask (i.e. 50 unique subjects), where each subject's session consisted of an independently sampled, random sequence of trials. Each of these subtasks followed the same task paradigm (described in Methods Task paradigm), and each session lasted 100 trials. Thus, the total amount of data we aimed to collect was 64 subtasks $\times$ 100 trials $\times$ 50 subjects $= 320k$ measurements.

Behavioral statistics in humans

We aimed to estimate a typical subject's accuracy at each trial, conditioned on a specific subtask. We therefore computed 64×100 accuracy estimates (*subtask* $\times$ *trial*) by taking the sample mean across subjects. We refer to this $[64, 100]$ matrix of point statistics as $\hat{H}$. Each row vector $\hat{H}_s$ has 100 entries, and corresponds to the mean human "learning curve" for subtask $s = \{1, 2, ...64\}$.

Because each object was equally likely to be shown on any given test trial, each of these 100 values of $\hat{H}_s$ may be interpreted as an estimate of the average of the true positive and true negative

rates (i.e. the balanced accuracy). The balanced accuracy is related to the concept of *sensitivity* from signal detection theory – the ability for a subject to discriminate two categories of signals (***Stanislaw and Todorov, 1999***). We note that an independent feature of signal detection behavior is the *bias* – the prior probability with which the subject would report a category. We did not attempt to quantify or compare the bias in models and humans in this study.

Simulating behavioral sessions in computational models

To obtain the learning curve predictions of each baseline computational model, we required that each model perform the same set of subtasks that the humans performed, as described above. We imposed the same requirements on the model as we did on the human subjects: that it begins each session without knowledge of the correct object-action contingency, that it should generate a action choice based solely on a pixel image input, and that it can update its future choices based on the history of scalar-valued feedback ("correct" or "incorrect"). If the choices later in the session are more accurate than those earlier in the session, then we colloquially say that the model has "learned", and comparing and contrasting the learning curves of baseline models with those of humans was a key goal of Experiment 1.

We ran n=32,000 simulated behavioral sessions for each model (500 simulated sessions for each of the 64 subtasks), where on each simulation a random sequence of trials was sampled in an identical fashion as in humans (see above). During each simulation, we recorded the same raw "behavioral" data as in humans (i.e. sequences of correct and incorrect choices), then applied the same procedure we used to compute $\hat{H}$ (see above) to compute an analogous collection of point statistics on the model's raw behavior, which we refer to as $\hat{M}$.

Comparing model learning with human learning

The learning behavior generated by an image-computable model of human learning ($\hat{M}$) should minimally replicate the measured learning behavior of humans (i.e. $\hat{H}$), to the limits of statistical noise. To identify any such models, we developed a scoring procedure to compare the similarity of the learning behavior in humans with any candidate learning model. We describe this procedure below.

Bias-corrected mean squared error

Given a collection of human measurements $\hat{H}$ (here, a matrix of accuracy estimates for $S = 64$ subtasks over $T = 100$ trials) and corresponding model measurements ($\hat{M}$), we computed a standard goodness-of-fit metric, the mean-squared error (MSE; lower is better). The formula for the MSE is given by:

$$\mathrm{MSE}(\hat{M}, \hat{H}) = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} (\hat{M}_{st} - \hat{H}_{st})^2 \tag{1}$$

Because $\hat{H}_{st}$ and $\hat{M}_{st}$ are random variables (i.e. sample means), the MSE itself is a random variable. It can be seen that the expected value of $\mathrm{MSE}(\hat{M}, \hat{H})$ consists of two conceptual components: the expected difference between the model and humans, and *noise components*:

$$\mathbf{E}[\mathrm{MSE}(\hat{M}, \hat{H})] = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} \left(\mathbf{E}[\hat{M}_{st}] - \mathbf{E}[\hat{H}_{st}]\right)^2 + \sigma^2\left(\hat{H}_{st}\right) + \sigma^2\left(\hat{M}_{st}\right) \tag{2}$$

Where $\mathbf{E}[\cdot]$ denotes the expected value, and $\sigma^2(\cdot)$ denotes the variance due to finite sampling (a.k.a. "noise").

Equation (2) shows that the expected MSE for a model depends not only on its expected predictions $\mathbf{E}[\hat{M}_{st}]$, but also its sampling variance $\sigma^2(\hat{M}_{st})$. In the present case where $\hat{M}_{st}$ is the mean over

independent (but not necessarily identically distributed) Bernoulli variables, the value of $\sigma^2(\hat{M}_{st})$ happens to depend on the expected prediction of the model itself, $\mathbf{E}[\hat{M}_{st}]$. [5]

Because the sampling variance of the model depends on its predictions, it is therefore conceptually possible that a model with worse (expected) predictions could achieve a lower expected MSE, simply because its associated sampling variance is lower.[6]

We corrected for this inferential bias by estimating, then subtracting, these variance terms from the "raw" MSE for each model we tested.[7]. We refer to this bias-corrected error as MSE_n.

$$\mathrm{MSE}_n(\hat{M}, \hat{H}) = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} (\hat{M}_{st} - \hat{H}_{st})^2 - \hat{\sigma}^2 \left(\hat{M}_{st}\right) \tag{3}$$

Where $\hat{\sigma}^2 \left(\hat{M}_{st}\right)$ is an unbiased estimator of the variance of $\hat{M}_{st}$. We write the equation for $\hat{\sigma}^2 \left(\hat{M}_{st}\right)$ below, where k_{st} is the number of observed correct choices over the n_{st} model simulations conducted for subtask s and trial t:

$$\hat{\sigma}^2 \left(\hat{M}_{st}\right) = \frac{\frac{k_{st}}{n_{st}} \left(1 - \frac{k_{st}}{n_{st}}\right)}{n_{st} - 1} \tag{4}$$

Because $\mathbf{E}\left[\hat{\sigma}^2 \left(\hat{M}_{st}\right)\right] = \sigma^2 \left(\hat{M}_{st}\right)$, the expected value of MSE_n can be shown to be:

$$\mathbf{E}[\mathrm{MSE}_n(\hat{M}, \hat{H})] = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} \left(\mathbf{E}[\hat{M}_{st}] - \mathbf{E}[\hat{H}_{st}]\right)^2 + \sigma^2 \left(\hat{H}_{st}\right) \tag{5}$$

Intuitively, MSE_n is an estimate of the mean-squared error that would be achieved by a model if we had a noiseless estimate of its predictions (i.e. had an infinite number of simulations of that model been performed). We note that the value of its square root, $\sqrt{\mathrm{MSE}_n}$, gives a rough[8] estimate of the average deviation between a model's prediction and human measurements, in units of the measurements (in this study, units of accuracy).

Noise floor estimation

It can be in Equation (5) that there are terms $\sigma^2 \left(\hat{H}_{st}\right)$, originating from the uncertainty in our experimental estimates of human behavior. These terms are always positive, and create a lower bound on the expected MSE_n for all models. That is, even if a model is expected to perfectly match the subject-averaged behavior of humans (where $\mathbf{E}[M_{st}] = \mathbf{E}[H_{st}]$, for all subtasks s and trials t), it cannot be expected to achieve an error below this lower bound. We call this lower bound the "noise floor", and use the symbol σ_h^2 to refer to it:

$$\sigma_h^2 = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} \sigma^2(\hat{H}_{st}) \tag{6}$$

It is possible to make an unbiased estimate of the noise floor (σ_h^2) if one can make unbiased estimates of each $\sigma^2(\hat{H}_{st})$ term. We did so by using the unbiased estimator from Equation (4). We write the full expression for our estimate of the noise floor, $\hat{\sigma}_h^2$, below:

$$\sigma_h^2 = \frac{1}{S \cdot T} \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{\frac{k_{st}}{n_{st}} \left(1 - \frac{k_{st}}{n_{st}}\right)}{n_{st} - 1} \tag{7}$$

[5] This can be seen by the expression for the variance of $\hat{M}_{st}$, which is a mean over independent (but not necessarily identically distributed) Bernoulli variables: $\sigma^2 \left(\hat{M}_{st}\right) = \frac{p_{st}(1-p_{st})}{n_{st}}$. The value of $\mathbf{E}[\hat{M}_{st}]$ is the expected behavior of the model on trial t of subtask s, and n_{st} is the number of model simulations.

[6] And/or because more model simulations were performed – though in this study, all tested models performed the same number of simulations, n=500.

[7] In practice, this correction was relatively small, because of the high number of simulations that were conducted.

[8] This estimator is biased.

Where k_{st} is the number of human subjects (out of n_{st} total subjects) that made a correct choice on subtask s and trial t. The square root of this value, σ_h, gives a rough estimate of the average deviation one would expect in our subject-averaged measurements of behavior ($\hat{H}_{st}$) over repetitions of the experiment (i.e. upon another resampling of human subjects and behavioral sessions).

Null hypothesis testing

For each model we tested, we attempted to reject the null hypothesis that $\mathbf{E}[\hat{H}_{st}] = \mathbf{E}[\hat{M}_{st}]$, for all subtasks s and trials t. To do so, we first approximated the distribution for $\mathrm{MSE}_n(\hat{H}, \hat{M})$ that would be expected under this null hypothesis, using bootstrapping.

To do so, we first computed bootstrap replicates of $\hat{H}$ and approximated samples of the null model $\hat{M}^\star$ (where $E[\hat{H}_{st}] = \mathbf{E}[\hat{M}^\star_{st}]$). A bootstrap replicate of $\hat{H}$ was constructed by first resampling individual human sessions without replacement, taking the same number of resamples per subtask as in the original experiment. We then computed the replicate $\tilde{H}$ using the same procedure described in Behavioral statistics in humans. Behavior from the null model cannot be sampled directly (i.e. we do not have the "true model" of human learning), but by definition shares the same expected behavior as a randomly sampled, individual human. We therefore created a bootstrap sample of the null model $\tilde{M}^\star$ by (also) taking resamples of individual human sessions, setting the number of resamples per subtask to the number of model simulations conducted per subtask (here, n=500 simulations per subtask). We then computed and saved $\mathrm{MSE}_n(\tilde{H}, \tilde{M}^\star)$ for that iteration, and repeated this process for B=1,000 iterations to obtain an approximate null distribution for MSE_n.

If a model's actual $\mathrm{MSE}_n(\hat{M}, \hat{H})$ score fell above the α-quantile of the estimated null distribution, we rejected it on the basis of having significantly more error than what would be expected from a "true" model of humans (with estimated significance level α).

Lapse rate correction

Lastly, we corrected for any lapse rates present in the human data. We defined the lapse rate as the probability with which a subject would randomly guess on a trial, and we assumed this rate was constant across all trials and subtasks. To correct for any such lapse rate in the human data, we fit a simulated lapse rate γ parameter to each model, prior to computing its MSE_n. Given a lapse rate parameter of γ (ranging between 0 and 1), a model would, on each trial, guess randomly with probability γ. For each model, we identified the value of γ that minimized its empirical MSE_n .

We note that fitting γ can only drive the behavior of a model toward randomness; it cannot artificially introduce improvements in its learning performance.

Experiment 2: One-shot human object learning benchmark

For the second benchmark in this study, we compared one-shot generalization in humans and models. Our basic approach was to allow humans to learn to distinguish between two novel objects using a single image per object, then test them on new, transformed views of the support set.

One-shot behavioral testing

We used the same task paradigm described in Task paradigm (i.e. two-way object discrimination with evaluative feedback). We created 64 object models for this experiment (randomly paired without replacement to give a total of 32 subtasks). These objects were different from the ones used in the previous benchmark (described in Experiment 1: Learning objects under high view variation).

At the beginning of each session, we randomly assigned the subject to perform one of 32 subtasks. Identical to Experiment 1, each trial required that the subject view an image of an object, make a choice ("F" or "J"), and receive feedback based on their choice. Each session consisted of 20 trials total, which was split into a "training phase" and "testing phase", which we describe below.

Training phase

The first ten trials (the "training phase") of the session were based on a single image for each object object (i.e. n=2 distinct images were shown over the first 10 trials). We ensured the subject performed trial with each training image five times total in the training phase; randomly permuting the order in which these trials were shown.

Testing phase

On trials 11-20 of the session (the "testing phase"), we presented trials containing new, transformed views of the two images used in the training phase. For each trial in the test phase, we randomly sampled an unseen test image, each of which was a transformed version of one of the training images. There were 36 possible transformations (9 transformation types, with 4 possible levels of strength). We describe how we generated each set of test images in the next section (see Figure 1B for examples). On the 15th and 20th trial, we presented "catch trials" consisting of the original training images. Throughout the test phase, we continued to deliver evaluative feedback on each trial.

One-shot stimulus image generation

Here, we describe how we generated all of the images used in Experiment 2. First, we generated each 3D object model using the Mutator process (see High-variation stimulus image generation). Then, for each object (n=64 objects), we generated a single canonical training image – a 256x256 grayscale image of the object occupying $\approx 50\%$ of the image plane, centered on a gray background. We randomly sampled its three axes of pose from the uniform rotational distribution.

For each training image, we generated a corresponding set of test images by applying different kinds of image transformations we wished to measure human generalization on. In total, we generated test images based on 9 transformation types, and we applied each transformation type at 4 levels of "strength". We describe those 9 types with respect to a single training image, below.

Translation

We translated the object in the image plane of the training image. To do so, we randomly sampled a translation vector in the image plane (uniformly sampling an angle from $\theta \in [0^\circ, 360^\circ]$), and translated it r pixels in that direction. We repeated this process (independently sampling θ each time) for $r = 16, 32, 64$, and 96 pixels (where the total image size 256×256 pixels), for two iterations (for a total of eight translated images).

Backgrounds

We gradually replaced the original, uniform gray background with a randomly selected, naturalistic background. Each original background pixel b_{ij} in the training image was gradually replaced with a naturalistic image c using the formula $b'_{ij} = (1-\alpha)b_{ij} + \alpha c_{ij}$. We varied α at four logarithmically spaced intervals, $\alpha = 0.1, 0.21, 0.46, 1$. Note that at $\alpha = 1$, the original gray background is completely replaced by the new, naturalistic background. We generated two test images per α level, independently sampling the background on each iteration (for a total of eight images per object).

Scale

We rescaled the object's size on the image to 12.5%, 25%, 50%, and 150% of the original size (four images of the object at different scales).

Out-of-plane rotations

We rotated the object along equally spaced 45° increments, rendering a test image at each increment. We did so along two separate rotational axes (horizontal and vertical), leading to n=13 test images total based on out-of-plane rotations.

In-plane rotation

We rotated the object inside of the image-plane, along 45° increments. This resulted in n=7 test images based on in-plane rotations.

Contrast

We varied the contrast of the image. For each pixel p_{ij} (where pixels range in value of 0 and 1), we adjusted the contrast using the equation $p'_{ij} = 10^c(p_{ij}) + 0.5(1 - 10^c)$, varying c from $-0.8, -0.4, 0.4$ and 0.8.

Pixel deletion

We removed pixels corresponding to the object in the training image, replacing them with the background color (gray). We removed 25%, 50%, 75%, and 95% of the pixels, selecting the pixels randomly for each training image.

Blur

We blurred the training image using a Gaussian kernel. We applied blurring with kernel radii of 2, 4, 8, and 16 pixels (with an original image resolution of 256×256 pixels) to create a total of 4 blurred images.

Gaussian noise

We applied Gaussian noise to the pixels of the training image. For each pixel p_{ij}, we added $i.i.d.$ Gaussian noise:

$$p'_{ij} = p_{ij} + \mathcal{N}(0, \sigma)$$

We applied noise with $\sigma = 0.125,\ 0.25,\ 0.375$ and 0.5 (where pixels range in luminance value between 0 and 1). We then clipped the resultant pixels to lie between 0 and 1.

Human behavioral measurements for Experiment 2

Subject recruitment

We used the same two-step subject recruitment procedure described above (see Subject recruitment and data collection), and recruited n=170 human subjects. Some of these subjects overlapped with those in Experiment 1 (n=9 subjects participated in both experiments).

All recruited subjects were invited to participate in up to 32 behavioral sessions. We disallowed them from repeating subtasks they had performed previously. Subjects were required to perform a minimum of four such behavioral sessions. In total, we collected n=2,547 sessions ($\approx 51k$ trials) for Experiment 2.

One-shot behavioral statistics in humans

We aimed to estimate the expected accuracy of a subject on each of the 36 possible transformations, correcting for attentional and memory lapses.

To do so, we combined observations across the eight test trials in the testing phase to compute accuracy estimate for each of the 36 transformations; that is, we did not attempt to quantify how accuracy varied across the testing phase (unlike the previous benchmark). We also combined observations across the 32 subtasks in this experiment. In doing so, we were attempting to measure the *average* generalization ability for each type of transformation (at a specific magnitude of transformation change from the training image), ignoring the fact that generalization performance likely depends on both the objects to be discriminated (i.e. the appearance of the objects in each subtask), the specific training images that were used, and the testing views of each object (e.g. the specific way in which an object was rotated likely affects generalization – not just the absolute magnitude of rotation). In total, we computed 36 point statistics (one per transformation).

Estimating performance relative to catch performance

Here we assumed that each human test performance measurement was based on a combination of the subject's ability to successfully generalize, a uniform guessing rate (i.e. the probability with which a subject executes a 50-50 random choice), and the extent to which the subject successfully acquired and recalled the training image-response contingency (i.e. from the first 10 trials). We attempted to estimate the test performance of a human subject that could 1) fully recall the association between each training image and its correct choice during the training phase, and 2) had a guess rate of zero on the test trials.

To do so, we used trials 15 and 20 of each session, where one of the two training images was presented to the subject ("catch trials"). Our main assumption here was that performance on these trials would be 100% assuming the subject had perfect recall, and had a guess rate of zero. Under that assumption, the actual, empirically observed accuracy p_{catch} would be related to any overall guess and/or recall failure rate γ by the equation $\gamma = 2 - 2p_{\text{catch}}$. We then adjusted each of the point statistics (i.e. test performances) to estimate their values had γ been equal to zero, by applying the following formula:

$$p' = \frac{p}{1-\gamma} - \frac{\gamma}{2-2\gamma}$$

We refer to the collection of 36 point statistics (following lapse rate correction) as $\hat{H}^{os}$.

Comparing model one-shot learning with human one-shot learning

Model simulation of Experiment 2

For this benchmark, we required that a model perform a total of 16,000 simulated behavioral sessions (500 simulated sessions for each of the 32 possible subtasks). Each simulated session proceeded using the same task paradigm as in humans (i.e. 10 training trials, followed by a test phase containing 8 test trials and 2 catch trials). Based on the model's behavior over those simulations, we computed the same set of point statistics described above, though we did not correct for any attentional lapses or recall lapses in the model, which we assumed was absent in models. In this manner, for each model, we obtained a collection of point statistics reflecting their behavior on this experiment, $\hat{M}^{os}$.

Bias-corrected mean-squared error and null hypothesis testing

We used the same statistical approach for our primary benchmark (introduced in Comparing model learning with human learning) to summarize the alignment of a model with humans. That is, we used the bias-corrected error metric MSE_n as our metric of comparison:

$$\text{MSE}_n(\hat{M}^{os}, \hat{H}^{os}) = \frac{1}{36}\sum_{i=1}^{36}\left(\hat{M}_i^{os} - \hat{H}_i^{os}\right)^2 - \hat{s}^2\left(\hat{M}_i^{os}\right)$$

We estimated the null distribution for MSE_n using bootstrap resampling, following the same procedure outlined in the first benchmark (bootstrap-resampling individual sessions).

Baseline model family

For a model to be scored on the benchmarks we described above, it must fulfill only the following three requirements: 1) it takes in any pixel image as its only sensory input (i.e. it is image computable), 2) it can produce an action in response to that image, and 3) it can receive scalar-valued feedback (rewards). Here, we implemented several baseline models which fulfill those requirements.

All models we implemented consist of two components. First, there is an *encoding stage* which re-represents the raw pixel input as a vector $\vec{x}$ in a multidimensional Euclidean space.The parameters of this part of the model are held fixed (i.e., no learning takes place in the encoding stage).

The second part is a *tunable decision stage*, which takes that representational vector and produces a set of C choice preferences (in this study, $C = 2$). The preference for each choice is computed through a dot product $\vec{w}_c \cdot \vec{x}$, where $\vec{w}_c$ is a vector of weights for choice c. The choice with the highest preference score is selected, and ties are broken randomly.

After the model makes its choice, the environment may respond with some feedback (e.g. positive or negative reward). At that point, the decision stage can process that feedback and use it to change its parameters (i.e. to learn). All learning in the models tested here takes place only in the parameters of the decision stage (all weight vectors $\vec{w}_1 ... \vec{w}_C$); the encoding stage has completely fixed parameters.

In total, any given model in this study is defined by these two components – the encoding stage and the decision stage. We provide further details for those two components below.

Encoding stages

The encoding stages were intermediate layers of deep convolutional neural network architectures (DCNNs). We drew a selection of such layers from a pool of 19 network architectures available through the PyTorch library (***Paszke et al., 2019***), each of which had pretrained parameters for solving the Imagenet object classification task (***Deng et al., 2009***).

For each architecture, we selected a subset of these intermediate layers to test in this study, spanning the range from early on in the architecture to the final output layer (originally designed for Imagenet). We resized pixel images to a standard size of 224x224 pixels using bilinear interpolation. In total, we tested n=344 intermediate layers as encoding stages.

Dimensionality reduction

Once an input image is fed into a DCNN architecture, each of its layers produces a representational vector of a dimensionality specified by the architecture of the model. Depending on the layer, this dimensionality may be relatively large ($>10^5$), making it hard to efficiently perform numerical calculations on contemporary hardware. We therefore performed dimensionality reduction as a preprocessing step. We performed dimensionality reduction using random Gaussian projections to a standard size of 2048, if the original dimensionality of the layer was greater than this number. This procedure approximately preserves the original representational structure of the layer (i.e., pairwise distances between points in that space) (***Johnson and Lindenstrauss, 1984***) and is similar to to computing and retaining the first 2048 principal components of the representation.

Feature normalization

Once dimensionality reduction was performed, we performed another standardization step. We computed centering and scaling parameters for each layer, so that its activations fit inside a sphere of radius 1 centered about the origin (i.e. $\|\vec{x}\| \leq 1$, for all $\vec{x}$).

To do so, we computed the activations of the layer over using the images from the "warmup" tasks human subjects were exposed to prior to performing any task in this study (i.e. 50 randomly selected images of 8 objects, see Subject recruitment and data collection). We computed the sample mean of those activations, and set this as the new origin of the encoding stage (i.e. the centering parameter). Then, we took the 99th quantile of the activation norms (over those same images) to calculate the approximate radius of the representation, and set this as our scaling parameter (i.e. dividing all activations by this number). Any activations with a norm greater than this radius were scaled to have a norm of 1.

Other kinds of feature standardization schemes are possible: for instance, one could center and scale the sensory representations for each subtask separately. However, such a procedure would expose models to the statistics of subtasks that are meant to be independent tests of their ability to learn new objects – statistics which we considered to be predictions of the encoding stage.

Tunable decision stage

Once the encoding stage re-represents an incoming pixel image as a multidimensional vector $\vec{x} \in \mathbb{R}^d$, a *tunable decision stage* takes that vector as an input, and produces a choice as an output.

Generating a decision

To select a choice, the tunable decision stage first generates choice preferences for each of the C possible actions, using the dot products $\vec{w}_i \cdot \vec{x}$ for $i = 1...C$. Then, the choice with the highest preference is selected ($c = \mathbf{argmax}_i \left(\vec{w}_i \cdot \vec{x}\right)$). If all choices have the same preference value, a choice is randomly selected.

Learning from feedback

Once a choice is selected, the environment may convey some scalar-valued feedback (e.g. reward or punish signals). The model may use this feedback to change its future behavior (i.e., to learn). For all models considered here, this may be accomplished (only) by changing its weight vectors $\vec{w}_1, ...\vec{w}_C$. Thus, learning on each trial can be summarized by the changes to each weight vector by some $\vec{\delta}_i \in \mathbb{R}^d$:

$$\vec{w}_i \leftarrow \vec{w}_i + \vec{\delta}_i$$

There are many possible choices on how each $\vec{\delta}_i$ may be computed from feedback; here, we focused on a set of seven rules based on the stochastic gradient descent algorithm for training a binary classifier or regression function. In all cases except one,[9] the underlying strategy of each decision stage can be understood as *predicting the reward* following the selection of each possible choice, and using these predictions to select the choice it believes will lead to the highest reward.

Specifically, we tested the update rules induced by the gradient descent update on the perceptron, cross-entropy, exponential, square, hinge, and mean absolute error loss functions (shown in Figure 2D), as well as the REINFORCE update rule. They are summarized in Table 1.

Each of these update rules has a single free parameter – a learning rate. For each update rule, there is a predefined range of learning rates that guarantees the non-divergence of the decision stage, based on the smoothness or Lipschitz constant of each of the update rule's associated loss function (***Shalev-Shwartz and Ben-David, 2014***). We did not investigate different learning rates in this study; instead, we simply selected the highest learning rate possible (such that divergence would not occur) for each update rule.

Data Availability Statement

Behavioral data for both experiments, and computer code for generating the main findings of the study, can be accessed under the MIT License at github.com/himjl/hobj.

Acknowledgments

This work was supported in part by the Semiconductor Research Corporation, DARPA, Simons Foundation grant SCGB-542965 (MJL, JJD), and the Friends of McGovern Fellowship from the McGovern Institute for Brain Research at MIT (MJL).

References

Ashby FG, Maddox WT. Human category learning. Annu Rev Psychol. 2005; 56:149–78. www.annualreviews.org, doi: 10.1146/annurev.psych.56.091103.070217.

Battleday RM, Peterson JC, Griffiths TL. Modeling Human Categorization of Natural Images Using Deep Feature Representations. arXiv. 2017; .

[9]The REINFORCE update rule is a "policy gradient" rule that optimizes parameters directly against the rate of reward; it does not aim to predict reward.

Table 1. Summary of update rules. Each update rule can be understood by the update $\vec{\delta}_i$ it generates for each weight vector $\vec{w}_i \leftarrow \vec{w}_i + \vec{\delta}_i$, based on the current input $\vec{x} \in \mathbb{R}^d$, the selected choice $c \in \{0, 1, ...C\}$, and the subsequent environmental reward $r \in [-1, 1]$. Each update rule is parameterized by a learning rate α, between 0 and 1. These equations assume that the input $\vec{x}$ has a bounded norm of $\|\vec{x}\| \leq 1$.

Update rule	**Weight update for selected choice** ($\vec{\delta}_c$)
Perceptron	$\vec{\delta}_c = \begin{cases} \alpha r \vec{x} & \text{if } r\left(\vec{w}_c \cdot \vec{x}\right) \leq 0 \\ 0 & \text{else} \end{cases}$
Hinge	$\vec{\delta}_c = \begin{cases} \alpha r \vec{x} & \text{if } r\left(\vec{w}_c \cdot \vec{x}\right) < 1 \\ 0 & \text{else} \end{cases}$
MAE	$\vec{\delta}_c = \begin{cases} \alpha \vec{x} & \text{if } (\vec{w}_c \cdot \vec{x}) < r \\ -\alpha \vec{x} & \text{else} \end{cases}$
Square	$\vec{\delta}_c = \alpha\left(r - \vec{w}_c \cdot \vec{x}\right)\vec{x}$
Exponential	$\vec{\delta}_c = \alpha r \exp\left(-r\vec{w}_c \cdot \vec{x}\right)\vec{x}$
Cross-entropy	$\vec{\delta}_c = \alpha r \left(\frac{4}{1+\exp\left(r\vec{w}_c \cdot \vec{x}\right)}\right)\vec{x}$
REINFORCE	$\vec{\delta}_c = \alpha r \left(1 - \exp\left(\vec{w}_c \cdot \vec{x} - \log\left(\sum_{j=1}^{C} \exp\left(\vec{w}_j \cdot \vec{x}\right)\right)\right)\right)\vec{x}$ $\vec{\delta}_{i \neq c} = -\alpha r \exp\left(\vec{w}_i \cdot \vec{x} - \log\left(\sum_{j=1}^{C} \exp\left(\vec{w}_j \cdot \vec{x}\right)\right)\right)\vec{x}$

Except for REINFORCE, each rule above only changes the weight vector for the selected choice c; the other weight vector(s) are left unchanged ($\vec{\delta}_{i \neq c} = 0$). Additionally, the Exponential update rule performs a weight normalization step (not shown): if the norm of the weights exceed a certain threshold ($\|w_i\| > B$) , the weights are projected to the closest weight vector $w_i' = \mathbf{argmin}_{u;\|u\|=B}\left(\|u - w_i\|\right)$ with $\|w_i'\| = B$. In this study, we chose $B = 10$.

Op de Beeck HP, Baker CI. The neural basis of visual object learning. Trends in Cognitive Sciences. 2010; 14(1):22–30. doi: 10.1016/j.tics.2009.11.002.

Biederman I. Recognition-by-Components: A Theory of Human Image Understanding. Psychological Review. 1987; 94(2):115–147. doi: 10.1037/0033-295X.94.2.115.

Bowers JS, Malhotra G, Dujmović M, Montero ML, Tsvetkov C, Biscione V, Puebla G, Adolfi FG, Hummel J, Heaton RF, Evans B, Mitchell J, Blything R. Deep Problems with Neural Network Models of Human Vision. PsyArXiv. 2022; https://psyarxiv.com/5zf4s/, doi: 10.31234/OSF.IO/5ZF4S.

Bülthoff HH, Edelman S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. Proceedings of the National Academy of Sciences of the United States of America. 1992; 89(1):60–64. https://www.pnas.org, doi: 10.1073/PNAS.89.1.60.

Deng J, Dong W, Socher R, Li LJ, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: *IEEE Conference on Computer Vision and Pattern Recognition* Institute of Electrical and Electronics Engineers (IEEE); 2009. p. 248–255. doi: 10.1109/CVPR.2009.5206848.

Duvdevani-Bar S, Edelman S. Visual Recognition and Categorization on the Basis of Similarities to Multiple Class Prototypes. International Journal of Computer Vision 1999 33:3. 1999; 33(3):201–228. https://link.springer.com/article/10.1023/A:1008102413960, doi: 10.1023/A:1008102413960.

Erdogan G, Jacobs RA. Visual Shape Perception as Bayesian Inference of 3D Object-Centered Shape Representations. Psychological Review. 2017; 124(6):740–761. http://dx.doi.org/10.1037/rev0000086740, doi: 10.1037/rev0000086.

Frémaux N, Gerstner W. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. Frontiers in Neural Circuits. 2015 1; 9(JAN2016):85. doi: 10.3389/FNCIR.2015.00085/BIBTEX.

Gallistel CR, Fairhurst S, Balsam P. The learning curve: Implications of a quantitative analysis. Proceedings of the National Academy of Sciences of the United States of America. 2004 9; 101(36):13124–13131. www.pnas.orgcgidoi10.1073pnas.0404965101, doi: 10.1073/PNAS.0404965101.

Geirhos R, Narayanappa K, Mitzkus B, Thieringer T, Bethge M, Wichmann FA, Brendel W. Partial success in closing the gap between human and machine vision. arXiv. 2021; https://github.com/bethgelab/model-vs-human/.

Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB. Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences. 2010; 14(8):357–364. http://dx.doi.org/10.1016/j.tics.2010.05.004, doi: 10.1016/j.tics.2010.05.004.

Han Y, Roig G, Geiger G, Poggio T. Scale and translation-invariance for novel objects in human vision. Scientific Reports 2020 10:1. 2020 1; 10(1):1–13. https://www.nature.com/articles/s41598-019-57261-6, doi: 10.1038/s41598-019-57261-6.

Harlow HF. The formation of learning sets. Psychological Review. 1949 1; 56(1):51–65. /record/1949-03097-001, doi: 10.1037/H0062474.

Hebart MN, Contier O, Teichmann L, Rockter AH, Zheng CY, Kidder A, Corriveau A, Vaziri-Pashkam M, Baker CI. THINGS-data: A multimodal collection of large-scale datasets for investigating object representations in brain and behavior. bioRxiv. 2022 7; p. 2022.07.22.501123. https://www.biorxiv.org/content/10.1101/2022.07.22.501123v1https://www.biorxiv.org/content/10.1101/2022.07.22.501123v1.abstract, doi: 10.1101/2022.07.22.501123.

Jacob G, Pramod RT, Katti H, Arun SP. Qualitative similarities and differences in visual object representations between brains and deep networks. Nature Communications 2021 12:1. 2021 3; 12(1):1–14. https://www.nature.com/articles/s41467-021-22078-3, doi: 10.1038/s41467-021-22078-3.

Johnson WB, Lindenstrauss J. Extensions of Lipschitz mappings into a Hilbert space. Contemp Math. 1984; 26:189–206. https://cir.nii.ac.jp/crid/1571135650630973824.bib?lang=en.

Kim HF, Ghazizadeh A, Hikosaka O. Separate groups of dopamine neurons innervate caudate head and tail encoding flexible and stable value memories. Frontiers in Neuroanatomy. 2014; 8(October):120. doi: 10.3389/fnana.2014.00120.

Kruschke JK. ALCOVE: An exemplar-based connectionist model of category learning. Psychological Review. 1992; 99(1).

Kubilius J, Schrimpf M, Kar K, Rajalingham R, Hong H, Majaj NJ, Issa EB, Bashivan P, Prescott-Roy J, Schmidt K, Nayebi A, Bear D, Yamins DLK, Dicarlo JJ. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In: *33rd Conference on Neural Information Processing Systems*; 2019. .

Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015 12; 350(6266):1332–1338. http://science.sciencemag.org/content/350/6266/1332.abstract, doi: 10.1126/science.aab3050.

Lake BM, Salakhutdinov RR, Gross J, Tenenbaum JB. One shot learning of simple visual concepts. Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011). 2011; 172:2568–2573.

Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. Building machines that learn and think like people. Behavioral and Brain Sciences. 2017; http://cims.nyu.edu/~brenden/http://www.mit.edu/~tomeru/http://web.mit.edu/cocosci/josh.htmlhttp://gershmanlab.webfactional.com/index.html, doi: 10.1017/S0140525X16001837.

Law CT, Gold JI. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. Nature neuroscience. 2009; 12(5):655–63. http://dx.doi.org/10.1038/nn.2304, doi: 10.1038/nn.2304.

Li FF, Fergus R, Perona P. One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2006 4; 28(4):594–611. doi: 10.1109/TPAMI.2006.79.

Maddox WT, Ashby FG. Comparing decision bound and exemplar models of categorization. Perception & Psychophysics 1993 53:1. 1993 1; 53(1):49–70. https://link.springer.com/article/10.3758/BF03211715, doi: 10.3758/BF03211715.

Marcus G. Deep Learning: A Critical Appraisal. arXiv. 2018; http://www.nytimes.com/2012/11/24/science/scientists-see-advances-in-deep-learning-a-part-of-artificial-.

McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. Trends in Cognitive Sciences. 2010; 14(8):348–356. doi: 10.1016/j.tics.2010.06.002.

Mckinley SC, Nosofsky RM. Selective Attention and the Formation of Linear Decision Boundaries. Journal of Experimental Psychology. 1996; 22(2):294–317.

Medin DL, Smith EE. Strategies and classification learning. Journal of Experimental Psychology: Human Learning and Memory. 1981 7; 7(4):241–253. /record/1982-02709-001, doi: 10.1037/0278-7393.7.4.241.

Morgenstern Y, Hartmann F, Schmidt F, Tiedemann H, Prokott E, Maiello G, Fleming RW. An image-computable model of human visual shape similarity. PLOS Computational Biology. 2021 6; 17(6):e1008981. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008981, doi: 10.1371/JOURNAL.PCBI.1008981.

Morgenstern Y, Schmidt F, Fleming RW. One-shot categorization of novel object classes in humans. Vision Research. 2019 12; 165:98–108. doi: 10.1016/J.VISRES.2019.09.005.

Niv Y. Reinforcement learning in the brain. Journal of Mathematical Psychology. 2009 6; 53(3):139–154. doi: 10.1016/J.JMP.2008.12.005.

Nosofsky RM. Similarity scaling and cognitive process models. Annu Rev Psychol. 1992; 43:25–53. www.annualreviews.org.

Nosofsky RM. The generalized context model: An exemplar model of classification. In: Pothos EM, editor. *Formal approaches in categorization* Cambridge University Press; 2011.p. 18–39.

Nosofsky RM, Sanders CA, Meagher BJ, Douglas BJ. Toward the development of a feature-space representation for a complex natural category domain. Behavior Research Methods. 2018 4; 50(2):530–556. https://link.springer.com/article/10.3758/s13428-017-0884-8, doi: 10.3758/S13428-017-0884-8/TABLES/8.

Paolacci G, Chandler J, Ipeirotis PG, Stern LN. Running experiments on Amazon Mechanical Turk. Judgment and Decision Making. 2010; 5(5):411–419.

Paszke A, Gross S, Massa F, Lerer A, Bradbury Google J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Xamla AK, Yang E, Devito Z, Raison Nabla M, Tejani A, Chilamkurthy S, Ai Q, Steiner B, Facebook LF, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *33rd Conference on Neural Information Processing Systems*; 2019. .

Persistence of Vision Pty Ltd , Persistence of Vision Raytracer. Williamstown, Victoria, Australia: Persistence of Vision Pty. Ltd.; 2004.

Peterson JC, Abbott JT, Griffiths TL. Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. Cognitive Science. 2018 11; 42(8):2648–2669. doi: 10.1111/COGS.12670.

Poggio T, Edelman S. A network that learns to recognize three-dimensional objects. Nature 1990 343:6255. 1990; 343(6255):263–266. https://www.nature.com/articles/343263a0, doi: 10.1038/343263a0.

Purves D, Augustine GJ, Fitzpatrick D, Katz LC, Lamantia AS, McNamara JO, Williams SM. Types of Eye Movements and Their Functions. In: Purves D, Augustine GJ, Fitzpatrick D, Katz LC, LaMantia AS, McNamara JO, Williams SM, editors. *Neuroscience*, 2nd edition ed. Sunderland, MA: Sinauer Associates; 2001.https://www.ncbi.nlm.nih.gov/books/NBK10991/.

Rajalingham R, Schmidt K, DiCarlo JJ. Comparison of Object Recognition Behavior in Human and Monkey. Journal of Neuroscience. 2015; 35(35):12127–12136. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.0573-15.2015, doi: 10.1523/JNEUROSCI.0573-15.2015.

Rajalingham R, Issa EB, Bashivan P, Kar K, Schmidt K, DiCarlo JJ. Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. Journal of Neuroscience. 2018 8; 38(33):7255–7269. https://www.jneurosci.org/content/38/33/7255https://www.jneurosci.org/content/38/33/7255.abstract, doi: 10.1523/JNEUROSCI.0388-18.2018.

Reed SK. Pattern recognition and categorization. Cognitive Psychology. 1972 7; 3(3):382–407. doi: 10.1016/0010-0285(72)90014-X.

Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P. Basic objects in natural categories. Cognitive Psychology. 1976; 8(3):382–439. doi: 10.1016/0010-0285(76)90013-X.

Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review. 1958 11; 65(6):386–408. /record/1959-09865-001, doi: 10.1037/H0042519.

Salakhutdinov R, Tenenbaum JB, Torralba A. Learning with hierarchical-deep models. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013; 35(8):1958–1971. doi: 10.1109/TPAMI.2012.269.

Saxe A, Nelli S, Summerfield C. If deep learning is the answer, what is the question? Nature Reviews Neuroscience 2020 22:1. 2020 11; 22(1):55–67. https://www.nature.com/articles/s41583-020-00395-8, doi: 10.1038/s41583-020-00395-8.

Schrimpf M, Kubilius J, Lee MJ, Ratan Murty NA, Ajemian R, DiCarlo JJ. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. Neuron. 2020 11; 108(3):413–423. doi: 10.1016/J.NEURON.2020.07.040.

Seger CA. The Roles of the Caudate Nucleus in Human Classification Learning. Journal of Neuroscience. 2005; 25(11):2941–2951. http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3401-04.2005, doi: 10.1523/JNEUROSCI.3401-04.2005.

Shalev-Shwartz S, Ben-David S. Understanding machine learning: from theory to algorithms. Cambridge University Press; 2014.

Shepard RN. Toward a universal law of generalization for psychological science. Science (New York, NY). 1987; 237(4820):1317–1323. doi: 10.1126/science.3629243.

Singh P, Peterson JC, Battleday RM, Griffiths TL. End-to-end Deep Prototype and Exemplar Models for Predicting Human Behavior. arXiv. 2020; .

Smith AC, Brown EN. Estimating a State-Space Model from Point Process Observations. Neural Computation. 2003 5; 15(5):965–991. https://direct.mit.edu/neco/article/15/5/965/6734/Estimating-a-State-Space-Model-from-Point-Process, doi: 10.1162/089976603765202622.

Sorscher B, Ganguli S, Sompolinsky H. The Geometry of Concept Learning. bioRxiv. 2021 1; p. 2021.03.21.436284. http://biorxiv.org/content/early/2021/05/16/2021.03.21.436284.abstract, doi: 10.1101/2021.03.21.436284.

Sorscher B, Ganguli S, Sompolinsky H. Neural representational geometry underlies few-shot concept learning. Proceedings of the National Academy of Sciences of the United States of America. 2022 10; 119(43):e2200800119. https://www.pnas.org/doi/abs/10.1073/pnas.2200800119, doi: 10.1073/PNAS.2200800119/SUPPL{_}FILE/PNAS.2200800119.SAPP.PDF.

Stanislaw H, Todorov N. Calculation of signal detection theory measures. Behavior Research Methods, Instruments, and Computers. 1999; 31(1). doi: 10.3758/BF03207704.

Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences. 2001 8; 24(4):629–640. https://www.cambridge.org/core/product/identifier/S0140525X01000061/type/journal_article, doi: 10.1017/S0140525X01000061.

Todd S, Latham W. Evolutionary art and computers. Academic Press Inc.; 1992.

Ullman S. Using neuroscience to develop artificial intelligence. Science. 2019 2; 363(6428):692–693. https://www.science.org/doi/10.1126/science.aau6595, doi: 10.1126/SCIENCE.AAU6595/ASSET/52C673F4-40F5-452C-B626-A0E2769AC25F/ASSETS/GRAPHIC/363{_}692{_}F1.JPEG.

Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(23):8619–24. http://www.pnas.org/content/111/23/8619.short, doi: 10.1073/pnas.1403112111.

Appendix 1

Additional Analyses

Our core experimental aim in this study was to create benchmarks which produce an error score summarizing a model's (dis)similarity with humans (i.e. MSE_n values for each model). However, we conducted additional analyses to provide further insight into what specific aspects of behavior a model might diverge from humans, and we describe those here.

Effect of model choices on human behavioral similarity

As described in Section , each model in this study was defined by two components (the encoding stage and the update rule). We wished to evaluate the effect of each of these components in driving the similarity of the model to human behavior. For example, it was possible that all models with the same encoding stage had the same learning score, regardless of which update rule they used (or vice versa).

To test for these possibilities, we performed a two-way ANOVA over all observed model scores (in MSE_n) computed in this study, using the encoding stage and update rule as the two factors, and MSE_n as the dependent variable. By doing so, we were able to estimate the amount of variation in model scores that could be explained by each individual component, and thereby gauge their relative importance. We briefly describe the procedure for this analysis below. First, we wrote the MSE_n score of each model as a combination of four variables:

$$\mathrm{MSE}_n(\text{encoding stage } i, \text{rule } j) = \mu + e_i + r_j + \gamma_{ij}$$

Where μ is the average MSE_n score, over all models. The variables e_i and r_j encode the value of the average difference from μ given encoding stage i and rule j, respectively. Any remaining residual is assigned to γ_{ij} (i.e. corresponding to any interaction between rule and encoding stage). The importance of each model component could be assessed by calculating the proportion of variation in model scores that could be explained by the selection of component alone.

Subtask consistency

In our primary benchmark, we measured human learning over 64 distinct subtasks, each consisting of 100 trials. For each subtask, the trial-averaged accuracy is a measure of the overall "difficulty" of learning that subtask, ranging from chance (0.5; no learning occurred over 100 trials) to perfect one-shot learning (0.995, perfect performance after a single example). For each of the 64 subtasks, one may estimate their trial-averaged performances (obtaining a length 64 "difficulty vector"), and use this as the basis of comparison between two learning systems (e.g. humans and a specific model).

To do so, we computed Spearman's rank correlation coefficient (ρ) between a model's difficulty vector and the human's difficulty vector. The value of ρ may range between -1 and 1. If $\rho = 1$, the model has the same ranking of difficulty between the different subtasks (i.e., finds the same subtasks easy and hard). If $\rho = 0$, there is no correlation in the rankings.

In addition to computing ρ between each model and humans, we estimated the ρ that would be expected between two independent repetitions of the experiment we conducted here (i.e., an estimate of experimental reliability in measuring this difficulty vector). To do this, we took two independent bootstrap resamples of the experimental data, calculated their respective difficulty vectors, and computed the ρ between them. We repeated this

process for $B = 1,000$ bootstrap iterations, and thereby obtained the expected distribution of experimental-repeat ρ.

Individual variability in overall learning ability

In this work, we focused primarily on *subject-averaged* measurements of human learning. However, individual subjects may also systematically differ from each other. We aimed to investigate whether any such differences existed in learning behavior for the subtasks we tested in this study.

Here, we attempted to reject the null hypothesis that all subjects had the same learning behavior. To do so, we tested whether there were statistically significant differences in *overall learning performance* between individuals – that is, whether some individuals were "better" or "worse" learners. If this was the case, this implies individuals differ (at least in terms of overall learning performance), and the null hypothesis could be rejected.

Permutation test for individual variability in overall learning ability

To test this null hypothesis, we identified a subset of human subjects who conducted all 64 subtasks in the primary, high-variation benchmark (n=22 subjects). For each subject, we computed their "overall learning performance", which was their empirically observed average performance over all n=64 subtasks. That is, for subject s, we computed:

$$\hat{G}_s = \frac{1}{64} \sum_{i=1}^{64} \hat{g}_{is}$$

Where $\hat{g}_{is}$ is the trial-averaged performance on subtask i, for subject s. The value of $\hat{G}_s$ is a gross measure of the subject's ability to learn the objects in this study, ranging from 0.5 (no learning on all subtasks) to 0.995 (perfect one-shot learning on all subtasks). In total, we computed n=22 estimates of $\hat{G}_s$ (one for each subject in this analysis).

We then computed the sample variance over the various $\hat{G}_s$:

$$\hat{\sigma}^2 = \frac{1}{S-1} \sum_{s=1}^{S} (\hat{G}_s - \bar{G})^2$$

Where $\bar{G}$ is the mean of overall lifetime performances. Intuitively, $\hat{\sigma}^2$ is high if individuals differ in their overall learning performance, and is low if all individuals have the same overall learning performance (as would be the case under the null hypothesis).

We performed a permutation test on $\hat{\sigma}^2$ to test whether it was significantly higher than would be expected under the null hypothesis, permuting the assignments of each $\hat{g}_{is}$ to each subject s. For each permutation, we computed the replication test statistic $\tilde{\sigma}^2$ (using the same formulas above, on the permuted data). We performed $P = 10,000$ permutation replications, then computed the one-sided achieved significance level by counting the number of replication test statistics greater than the actual, experimentally observed value $\hat{\sigma}^2$.

Testing whether specific humans outperform a model

To test whether a specific human has significantly higher overall learning abilities than a specific model (over the subtasks tested in this study), we performed Welch's t-test for unequal variances on the overall learning performance, $\hat{G}$ (defined above). That is, for a specific subject s and model m, we attempted to reject the null hypothesis that $\hat{G}_s \leq \hat{G}_m$.

We adjusted for multiple comparisons using the Bonferroni correction (using the total number of pairwise comparisons we made between a model m and specific subjects s).