An empirical assay of view-invariant object learning in humans and comparison with baseline image-computable models

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

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

Correspondence: ***dicarlo@mit.edu***

Abstract

How humans learn new visual objects is a longstanding scientific problem. Previous work has led to a diverse collection of models for how it is accomplished, but a current limitation in the field is a lack of empirical benchmarks which can be used to evaluate and compare specific models against each other. Here, we use online psychophysics to measure human behavioral learning trajectories over a set of tasks involving novel 3D objects. Consistent with intuition, these results show that humans generally require very few images (≈ 6) 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 after even one view of each object. We then use those data to develop benchmarks that may be used to evaluate a learning model's similarity to humans. We make these data and benchmarks publicly available [GitHub], and, to our knowledge, they are currently the largest publicly-available collection of learning-related psychophysics data in humans. Additionally, to serve as baselines for those benchmarks, we implement and test a large number of baseline models (n=1,932), 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 find some of these baseline models make surprisingly accurate predictions. However, we also find 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 examples (images) of some new object, their ability to categorize new, unseen views of that object increases, possibly very rapidly (in the limit, from one example). What are the mechanisms that allow an adult human to do so?

Efforts from cognitive science, neuroscience, and machine learning have lead to a diverse array of ideas to understand and replicate this specific 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 (1–6), models which depend on unspecified intermediate computations (i.e. non-image-computable models) (7–10), to end-to-end learning models which take raw pixels as input (11–17).

An important step in determining which (if any) of these ideas might lead to accurate descriptions of human object learning is to implement them in forms that allow for quantitative predictions in novel situations (i.e. predictions for novel images and image sequences from image-computable learning models), then to compare those predictions to behavioral measurements of humans as they perform the same learning tasks.

However, while empirical behavioral benchmarks exist for visual tasks involving known object categories (18–20), the field currently lacks a publicly-available set of benchmarks for comparing models to humans as they learn novel object categories, making it difficult to gauge progress in the field.

To address this gap, we aimed to create and release benchmarks that allow for the standardized assessment of any image-computable object learning model of human learning. We wished to make minimal assumptions about the learning models that might be compared today or in the future, and we required only that they are computable in the following sense: on each test trial, they can take a pixel image as input, choose an action, and receive scalar-valued feedback from the environment (e.g. to drive any changes to their internal state).

Our first and primary experimental goal in this study was to measure human learning trajectories across a variety of elementary binary object discrimination learning tasks involving novel 3D objects rendered at highly varied views (n=64 subtasks). We could then use the resultant dataset to build a benchmark which compares a core signature of any learning system to that of humans: its rate and pattern of learning across a variety of novel

situations.

Based on prior suggestions of where humans may be particularly powerful (13, 21), our secondary goal was to measure human behavior during the special case of "one-shot" learning, where the learner receives a single example from each category before being asked to generalize to unseen examples. We therefore aimed to create a second benchmark which compares a model's and humans' ability to be generalize from a single example of novel objects to a range of image transformations (e.g. translation, scaling, and 3D rotation).

To serve as baselines for these benchmarks, we implemented then tested models from a standard cognitive theory of learning, which posits that adult humans re-represent incoming visual stimuli in a stable, preexisting multi-dimensional Euclidean space, build categorization boundaries in that space by applying some learning rule to exemplars, and apply this learned boundary to categorize new examples (1, 5, 6, 8, 17, 22). To actually build those baseline models, we drew from ongoing efforts in computational cognitive science (23–28) and visual neuroscience (29–31) in building and validating image-computable models of human visual representations, primarily based on intermediate layers of deep convolutional neural networks (DCNNs). We then combined those representational models with trainable decision models that have been considered neurally plausible, namely linear decoders which are adjusted by scalar reward-based update rules (32–34).

In summary, our goal in this work was to take the following scientific steps: measure human behavior across a range of reinforced object choice learning tasks, implement baseline learning models that are capable of tuning their behavior during those same tasks for any novel image sequence, and compare and contrast the behavior of humans to the behavior produced by (i.e. predicted by) each of the models.

We reasoned that, if any such models were found to be statistically indistinguishable from humans, they could then serve as leading scientific hypotheses 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 facilitate a standard evaluation of current and future visual object learning models.

Materials and Methods

1 Overview of experiments

We studied human learning of novel object discrimination tasks. We aimed to assess this ability in two experiments (Experiment 1 and Experiment 2), and we provide a brief overview of these experiments below.

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. Conceptually, we consider such measurements to be an assessment of the human ability to learn noun labels for new objects. Because we are here primarily focused on the *typical* human ability, our primary measures were computed by averaging over many subjects.

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.

2 Behavioral task paradigm

For both experiments, subjects were recruited from Mechanical Turk (35), and ran tasks on their personal computers. Each experiment 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.[1] In this phase, 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).[2]

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).

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

[2]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.

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.[3].

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.

3 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.

3.1 Stimulus image generation

We designed 3D object models (n=128) using the "Mutator" generative design process (36). We generated a collection of images for each of those 3D objects using the POV-Ray rendering program (37). 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 (38). 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.

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

3.2 Human behavioral measurements

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 presented to the subject as described in Methods 2 above. 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 via the Mechanical Turk platform (35) 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 Methods 2 above. If the subject failed to complete this task with an average overall accuracy of at least 85%, we 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 2, 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 (39). 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.

3.3 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 performed 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 $\hat{M}$.

3.4 Comparing model learning with human learning

The learning behavior generated by an image-computable model of human learning ($\hat{M}$) should minimally replicate each of the entries in $\hat{H}$ 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 the matrix of human learning measurements $\hat{H}$ (accuracy estimates for 64 subtasks over 100 trials) and corresponding model measurements ($\hat{M}$), we computed a standard goodness-of-fit metric, the mean-squared error (MSE, where a lower value indicates a better model). The formula for the MSE is given by:

$$\text{MSE}(\hat{M}, \hat{H}) = \frac{1}{64 \cdot 100} \sum_{s=1}^{64} \sum_{t=1}^{100} (\hat{M}_{st} - \hat{H}_{st})^2$$

Because each $\hat{H}_{st}$ and $\hat{M}_{st}$ is a random variable (i.e. sample means), the MSE is a random variable, and has an expected value. The expected value of $\text{MSE}(\hat{M}, \hat{H})$ consists of two conceptual components: the expected difference between the model and humans, and intractable *noise components*:

$$\begin{aligned}\mathbf{E}[\text{MSE}(\hat{M}, \hat{H})] = \frac{1}{64 \cdot 100} \sum_{s=1}^{64} \sum_{t=1}^{100} (\mathbf{E}[\hat{M}_{st}] - \mathbf{E}[\hat{H}_{st}])^2 \\ + \sigma^2\left(\hat{H}_{st}\right) + \sigma^2\left(\hat{M}_{st}\right)\end{aligned}$$

Where $\mathbf{E}[\cdot]$ denotes the expected value, and $\sigma^2(\cdot)$ denotes the variance due to finite sampling (a.k.a. "noise"). It can be seen that the variance terms, which are always positive, create a lower bound on the expected MSE. That is, even if a model is expected to perfectly match the subject-averaged behavior of humans (i.e. $\mathbf{E}[M_{st}] = \mathbf{E}[H_{st}]$, for all subtasks s and trials t), it cannot be expected to achieve an error below this lower bound.

The expression above also shows how the expected MSE for a model depends not only on its expected predictions $\mathbf{E}[\hat{M}_{st}]$, but also its sampling variance $\sigma^2(\hat{M}_{st})$, which depends on the predictions of the model itself.[4]

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.[5]

We corrected for this inferential bias by estimating, then subtracting, these variance terms from the "raw" MSE for each model we tested.[6]. We refer to this bias-corrected error as MSE_n (aka the bias-corrected MSE):

$$\text{MSE}_n(\hat{M}, \hat{H}) = \frac{1}{64 \cdot 100} \sum_{s=1}^{64} \sum_{t=1}^{100} (\hat{M}_{st} - \hat{H}_{st})^2 - \hat{s}^2\left(\hat{M}_{st}\right)$$

Where $s^2\left(\hat{M}_{st}\right)$ is the unbiased estimator for the variance of $\hat{M}_{st}$:

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

Intuitively, MSE_n is an estimate of the mean-squared error that would be achieved by a model, if an infinite number of simulations of that model were to be performed.

We note that it is possible to perform an additional bias-correction step by estimating then removing $\sigma^2\left(\hat{H}_{st}\right)$ terms, which is the error attributable to the experimental variance in our estimate of human performance on subtask s and trial t. We chose not do so here, as it would not affect any inferences on models.

Finally, to aid in the human interpretation of these error scores, one may take the square root of MSE_n to get a rough estimate of the average deviation which would be expected between a model and humans,[7] in the units of the measurements (i.e. accuracies). We refer to this square root statistic using the notation $\sqrt{\text{MSE}_n}$.

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 approximated the distribution for $\text{MSE}_n(\hat{H}, \hat{M})$ that would be expected under the null hypothesis, using bootstrapping (where bootstrap samples of $\hat{H}$ and a null model $\hat{M}$ were created by resampling over individual human sessions).

If a model's actual $\text{MSE}_n(\hat{M}, \hat{H})$ 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). We approximated the null distribution using B=1,000 bootstrap samples.

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 .

[4] 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.

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

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

[7] The root mean-squared error is in general a biased estimator.

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

4 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.

4.1 One-shot behavioral subtasks

We used the same task paradigm described in Methods 2 (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 Subsection 3).

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.

4.2 Stimulus 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 3.1)). 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.

4.3 Human behavioral measurements

Subject recruitment. We used the same two-step subject recruitment procedure described above (3.2), 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.

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 $\hat{H}^{os}$.

4.4 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, $\hat{M}^{os}$.

Noise-corrected mean-squared error and null hypothesis testing. We followed the same approach as in our primary benchmark (introduced in Methods 3.4) 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).

5 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 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 a action preferences (in this study, $a = 2$). The action with the highest preference score is selected, and ties are broken randomly.

After the model takes an action, 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; 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.

5.1 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 (40), each of which had pretrained parameters for solving the Imagenet object classification task (41).

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 276 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) (42) 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. $\max_i \|x_i\| = 1$).

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 Methods 3.2). 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.

5.2 Tunable decision stage

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

Generating a decision. To output an action, a set of action preferences are calculated using the matrix multiplication Wx, where $W \in \mathbb{R}^{a,d}$ (i.e., action preferences are linear "readouts" of the representation computed by the encoding stage).

Then, the decision stage simply selects the action with the highest preference, breaking ties randomly. In total, the equation for generating an action is:

$$\text{action} = \mathbf{argmax}_i \left(Wx\right)_i$$

Learning from feedback. Once an action is taken, the environment may convey some scalar-valued feedback (e.g. reward or punish signals). The model may use this feedback to change its behavior (i.e., to learn). In this case, behavior is determined by the value of the weights $W \in \mathbb{R}^{a,d}$, so learning consists of changing those weights by some $\delta \in \mathbb{R}^{a,d}$:

$$W_{t+1} = W_t + \delta_t$$

There are many possible choices on how this δ_t 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,[8] the goal of the learner can be understood as *predicting the reward* following the choice of the action i.

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.

Each of these update rules has a single free parameter – the 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 (43). 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.

6 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.

6.1 Effect of model choices on human behavioral similarity

As described above in Section 5, 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:

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

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.

6.2 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.

6.3 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

[8] 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.

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).

Results

We measured human behavior over two variants of an object learning task (Experiments 1 & 2). Consistent with intuition, our main empirical findings show that humans 1) require few images to learn, 2) find some object discriminations easier than others, and 3) generalize well over a range of image transformations after seeing even one view of each object. We then compared those empirical measurements to the predictions made by a suite of learning models.

1 Humans are rapid, but imperfect novel object learners

In our primary experiment (Experiment 1), we measured a population of anonymous human subjects (n=70 subjects) performing 64 learning subtasks, each requiring that the subject learn to discriminate two new novel objects, rendered under high view variation (see Figure 1A).

On average over all 64 subtasks we tested, we found that human discrimination accuracy improved immediately, 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 at the last trial (trial 100) was 0.87 (mean; $[0.85, 0.88]$ 95% CI), and 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) resulted in widely different learning curves. This is illustrated in Figure 1D which shows the estimated average human learning curve for each of the 64 subtasks (i.e. $\hat{H}_s$ for subtask $s = 1, 2, ...64$, see Methods). That is, we observed that some tasks were "easy" for humans to learn, and some were hard. 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 Methods 6.2).

Overall, these results show 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), 2) learning new objects is highly dependent on the 3D shapes of those objects, and 3) many object pairs are far from perfectly learned within 100 trials (e.g. mean accuracy of ≈ 0.65 for the most difficult 10% of subtasks), and the trend lines suggest that they might never be perfectly learned. 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.

2 Computing the high-variation benchmark on a suite of baseline object learning models

Scoring a family of baseline models. We implemented a family of baseline models (Methods 5) to assess each as a possible explanation of human object learning. As shown in Figure 2A, each model was comprised of two components: an encoding stage (a specific intermediate layer of some Imagenet-pretrained(41) DCNN), and a tunable decision stage (a set of linear action preferences, trained by one of 7 update rules).

We implemented a large family of such models (n=1,932 models), each based on a different combination of encoding stage and learning rule, and tested each of these on the same set of subtasks ($n = 64$ subtasks) as humans, simulating 500 behavioral sessions per subtask, per model. Then, to compare the learning behavior of these learning models to humans, we computed a mean-squared-error statistic (MSE_n, see 3.4). The square root of that value ($\sqrt{\text{MSE}_n}$) gives a rough estimate of the average deviation that would be expected between a model's prediction and a human behavioral measurement.

This model comparison metric is conceptually simple: it is an estimate of the average squared error between the predictions of the model and the measurements of humans (Methods 3.4). A lower value of MSE_n indicates a better alignment of the model's behavior to human behavior (i.e. lower error); higher values are worse.

In principle, the value of MSE_n can be no lower than the experimental "noise floor" σ_h^2, which is equivalent to the sampling variance associated with our measurements of human behavior. We made an unbiased estimate of this noise floor ($\sigma_h^2 \approx 0.003$, see Methods 3.4). The square root of this value gives a rough estimate of the average deviation that would be expected upon an experimental repeat ($\approx \pm 0.05$).

We summarize scores for each of the models we tested in Figure 4. Many of the models were far from the noise floor (median $\sqrt{\text{MSE}_n} = 0.30$), but to our surprise, we found that a small subset of models achieved relatively low error: the best 1% of the models (which we term "strong baseline models") had predictions which were on average within $\approx \pm 0.08$ of human measurements, coming relatively close to the limits of experimental noise ($\approx \pm 0.05$).

Model components affecting the score of a model. We wished to analyze how the two components making up each learning model – the encoding stage and tunable decision stage – affected its alignment with humans.

One general trend we observed was that models built with encoding stages from deeper layers of DCNNs tended to produce more human-like patterns of learning (see Figure 2B). On the other hand, the different update rules appeared to have little effect in the model's ability to support human-like learning ((see Figure 5A for example).

Specifically, a two-way ANOVA over all model scores (Methods 6.1) revealed that the choice of update rules explained only 0.1% of the variation in model scores. By contrast, 99.4% 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.

Still, though some models we tested made relatively accurate predictions of humans, we found that *all* models were behaviorally distinguishable from humans; all models were rejected with a significance level of at least $\alpha < 0.001$ (see Methods 3.4). Our next step in this study was to ask in what aspects of learning behavior differences lay between models and humans.

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

To gain insight into where the behavior of these models diverged from humans, we compared models to humans along two summary statistics of learning behavior: 1) the *overall accuracy* over all subtasks and trials tested, and 2) Spearman's rank correlation coefficient (ρ) across the trial-averaged accuracy values for all of the 64 subtasks between humans and models, which we refer to as *consistency* (see Figure 6A).

Overall accuracy over a single session is a gross measure of a learning model's rate of learning, ranging from 0.5 (chance, no gain in performance) to 0.995 (perfect learning after just one trial). Consistency quantifies the extent to which a model finds the same subtasks easy and hard as humans. It ranges from -1 (perfectly anti-correlated pattern of performance) to 1 (perfectly correlated pattern of performance). A value of $\rho = 0$ indicates no correlation between the pattern of performance between two learning systems.

These metrics are theoretically independent of each other; e.g. a model may have high overall accuracy, but have low consistency with humans.[9]

Nevertheless, we found a positive relationship between these two metrics: models with high overall accuracy tended to also have high consistency (Figure 6B).

Though many of the strong baseline models matched or exceed human-level overall accuracy (over 100 trials), none of the models had full consistency with humans, indicating that part of their failure to fully explain human behavior is due to the fact they differ in their patterns of performance across different learning situations.

[9]Except for models which are either at chance or perform perfect one-shot learning in all situations; then the correlation coefficient is not defined.

Moreover, even though many strong baseline models matched (or exceeded) humans in terms of their overall accuracy over this experiment, it was possible they had systematic differences from humans at specific trials that was masked by trial-averaging. We therefore next examined models' learning curves against humans.

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

We noticed that the strong baseline models' accuracy early on in learning appeared to be slightly below that of humans (see Figure 7A). We tested for this by comparing the average accuracy (over subtasks) early on in learning (which we defined as the average accuracy over trials 1-5).

We found that *all* of the baseline models were significantly worse than humans early on (see Figure 7B).

We wondered whether this gap persisted across difficulty levels (e.g., that models tended to perform particularly poorly on "hard" subtasks relative to humans, but were human-level for other subtasks). We therefore performed this analysis again across four different difficulty levels of subtasks (where each level consisted of 16 out of the 64 total subtasks we tested, grouped by human difficulty levels), and found models were 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 7B).

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).

5 Experiment 2: Characterizing one-shot object learning in humans

Our results above suggest 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 in which we measured the extent and pattern of human generalization to nine kinds of image variations following experience with a single image of each object category (i.e. one-shot generalization).

Our motivation here was to test whether the strong baseline models differed from humans in ability to generalize to any of the five kinds of image variation present in our original experiment (i.e. in-plane translation, scale, random backgrounds, in-plane object rotation, and out-of-plane object rotation). We also tested four additional kinds of image variation (contrast shifts, pixel deletion, blur, and shot noise) that were not present in our original experiment, but could nonetheless serve as informative comparisons for identifying functional deficiencies in the strong baseline models relative to humans.

Following 10 trials of training in which subjects repeatedly performed trials with respect to a single image from each object category, we then presented subjects with random "test" trials consisting of transformed versions of the support set (see Figure 8A). For each kind of transformation, we tested four possible "levels" of variation. For example, to test generalization to scale, we tested images where the object was resized to 12.5%, 25%, 50%, and 150% of the original support image (see Methods 4.2 for details).

We found humans showed varied patterns of generalization (see Figure 8B). For some kinds of image variation, humans had showed nearly perfect generalization (e.g. translation, backgrounds, contrast, in-plane-rotation) over the ranges we tested. In others, we observed varied patterns of generalization. Overall, these patterns of generalization were measured with a high degree of experimental precision, as quantified by our estimates of the human noise floor (standard error of $\approx \pm 0.02$).

We next used these data to create a benchmark that could be used to compare any candidate object learning model – including the ones we considered here – against human object learning.

6 Baseline models show weaker one-shot generalization compared to humans

We simulated the same experiment (i.e. 10 trials of training followed by 10 trials of randomly selected test images) in all of our models, and compared each of their behavioral predictions to humans under our error metric (MSE_n, see Methods 4).

Similar to our results from our previous experiment, we found that models varied widely in their alignment with human learning behavior. Specific models achieved relatively low error – within ± 0.06 experimental error, where the noise floor is approximately ± 0.02), but *all* models had statistically significant differences in their behavior relative to humans.

We found that part of these differences lay in systematic generalization failures in models, relative to humans. For example, we observed that strong baseline models had lower one-shot accuracy than humans with respect to four kinds of image variation: pixel deletion, blur, shot noise, and scale (see Figure 8C).

We also observed a positive relationship between the scores of the two benchmarks: models that were most human-like in the high-variation object learning setting (Experiment 1) also tended to be the most human-like in one-shot generalization (Experiment 2) (Figure 8D) – though we emphasize that *no* model explained human behavior to the limits of statistical noise.

7 Specific individual humans outperform all baseline models

Both of the benchmarks we developed in this study tested the ability of a model to predict human object learning at the "subject-averaged" level; any individual differences in learning behavior are ignored (by design) and not considered in those benchmarks.

To gauge the extent to which those individual differences are present (if at all) over the subtasks we tested, we performed a *post-hoc* analysis on our behavioral data from Experiment 1. We first identified subjects who performed all 64 subtasks in that experiment (22 out of 70 total) . We then attempted to reject the null hypothesis that there was no significant variation in their overall learning ability (see Methods 6.3). If rejected, this would indicate that individuals indeed systematically vary, at least in terms of their overall performance on these tasks. We indeed found that some subjects were reliably better object learners than others ($p <$ 1e-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 (*ResNet152/avgpool*, with the square loss update rule). To do so, we performed Welch's t-test on overall learning performance (Methods 6.3) between each individual human's overall performance and this model's overall performance.

Using this analysis, we identified $n = 5$ individuals whose overall accuracy significantly exceeded that of this model (all p<1e-5, 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 behavior. In this study, we collected a set of such behavioral measurements across a variety of object learning settings (n=371k trials), which allowed us to quantify the speed of human object learning (~6 trials to achieve close-to-asymptotic accuracy), the distinct pattern of learning difficulty they have for different objects, and the extent of generalization to specific image transformations after a single image example.

We then developed procedures to compare those measurements with the predictions made by a model in those same settings (a.k.a., behavioral benchmarks). We implemented and tested a set of baseline object learning models (n=1,932 models) on those benchmarks. Each of these models consist of two stages: 1) a fixed *encoding stage* that re-represents an incoming pixel image as a point in a representational space, followed by 2) a tunable *decision stage* that generates an action choice by computing choice preferences which are weighted sums of that representation. Plasticity only occurs in the decision stage, and is done through an *update rule* that guides changes to the weights using the scalar reinforcement signal provided on each trial – the exact same type of signal provided to human subjects. For each model, the encoding stage was based on an intermediate representation of a contemporary deep convolutional neural network model (DCNN), and the update rule in each tunable decision stage was taken from a set of standard alternatives in the field.

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 focus our discussion on the observed predictive accuracy of these baseline models, but we highlight that our 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's behavior to human behavior under high view-variation learning conditions, a subset of baseline models produced relatively accurate predictions of human learning behavior. Importantly, these models are not accurate simply because they learn new objects as rapidly as humans – they also strongly (though not perfectly) predicted the patterns of difficulty observed in humans (Figure 6B). That is, they successfully predicted object discriminations that humans will learn rapidly and those that human will fail to learn rapidly.

We were surprised by how similar the baseline models were to human, because many authors have suggested that current DCNNs are likely to be inadequate models of human learning (15, 44–47) (see below). Contrary to this belief, the results reported here suggest that some DCNN models, though imperfect, may be 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 in this study 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 (32). This makes them not only plausible descriptions for the computations executed by the brain over object learning, but, with some additional assumptions (Fig 2B), makes predictions of neural phenomena.

For example, if this interpretation is taken at face

value, the strong baseline models make a few qualitative predictions: first, if we assume 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 several prior studies showing adult ventral stream changes are typically moderate and much slower that that timescale, at least in the conditions used here (see (48) for review).

The complementary prediction of the strong baseline models mapped in this way is that the neural changes that underlie learning are not distributed over the entire visual processing stream, but are focused 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 motor initiation, such as the caudate nucleus, as one such possible candidate (49, 50).

Baseline models at predicting human few-shot learning. Despite the overall strength of the baseline models we tested, we emphasize that none of the models were able to fully explain human behavior on either benchmark. Here, we point out one specific source of these deficiencies, which is the consistent accuracy gap between models and humans in low-sample regimes.

In Experiment 1, we found humans acquired performance rapidly (in terms of the number of examples). By comparison, none of the models we tested could rival humans when given an equivalent, small number of exemplars (< 10, see Figure 7).

In Experiment 2, we found similar kinds of deficiencies in the few-shot regime (see Figure 8. For example, here we replicated previous reports (51) of deep neural networks failing to generalize as well to scale as humans, and found other accuracy gaps.

Taken together, these results point to a central inference of this work: all 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

Conceptually, there are several ways to improve the predictive accuracy of the models we tested in this study. For example, it is possible that this model family (fixed representations combined with linear decoders trained by reward signals) is in fact sufficient, and we simply did not test specific models in this family which accurately predicted the human learning benchmarks. If that is the case, there are two conceptual components that one could change: the encoding stage, and the update rule which defines the tunable decision stage. Here, we found the choice of update rule had little effect on the predictive power of these models (see Figure 5A), and did not interact significantly with the choice of representational model. Still, we note there are some notable update rules which we did not consider here – namely exemplar (52) and prototype-based rules (7, 17, 53).

On the other hand, we found the image-computable representation that was used for each model's encoding stage had a dramatic effect on its overall predictive power as an object learning model (on both benchmarks), and it is therefore likely that alternate encoding stages could lead to more accurate models of human learning. Here, we only considered encoding stages based on Imagenet-pretrained DCNN representations; these have both similarities to, and also well-characterized divergences from, internal primate visual representations as measured by electrophysiological studes (31) as well as similarities and differences as measured by behavioral studies (18, 54). If image-computable representations that more closely adhere to human visual representations are built and/or identified, they might lead to image-computable models of learning which close the prediction gap on the benchmarks we developed here.

Stepping back, it is also possible that *no* model of a fixed representations followed by linear learning could lead to fully accurate predictions on these benchmarks (or future benchmarks), but perhaps 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 (e.g. object identity) (4, 11, 13, 15, 16, 55). It is possible these alternate approaches would lead to more accurate accounts of human behavior in the learning tasks we tested here, and are possibly the critical steps needed to close the predictive gaps on the human object learning benchmarks collected here and beyond. Implementing and testing these alternative models on a common set of benchmarks (such as the ones developed here) is therefore an important direction for future work.

Future extensions of benchmarks to evaluate object learning models

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 learn about multiple objects simulta-

neously. Scaling to tasks involving multiple objects is one straightforward extension of the task paradigm and analyses utilized here, and we note that simple variants of the baseline models we tested here naturally scale to task paradigms involving multiple objects (i.e. by the addition of additional linear action preferences to the decision stage).

As well, an important future direction is to design task paradigms which measure how humans might encode, organize, and later recall knowledge of previously learned novel objects, and to test the ability of image-computable models for doing the same, which is a nontrivial problem for DCNN-based models models of the kind we have tested here (56, 57).

Extending stimulus presentation time. For presenting stimuli, we followed conventions used in previous visual neuroscience studies (18, 38) 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 (58). 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 work (which was designed to prevent such saccades from our subjects) would be *underestimating* the number of images needed by humans to achieve learning on new objects, relative to conditions in which subjects had unlimited viewing time. However, 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 exemplars.

Still, measuring human behavior in even more naturalistic learning contexts (e.g. longer viewing times, watching objects in motion, physical interaction with novel objects), will be an important extension of this work.

Differences in individual subjects. In this study, 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 1). 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 9B). We did not attempt to model this individual variability in this study. Whether these differences can be explained in terms of individual differences in underlying sensory representations, learning rules (e.g. the learning rate), random weight initialization, or are unexplainable by any model in our baseline family remains an area for future study.

Furthermore, performing subject-averaging leads to the masking of learning dynamics that may be identifiable only at the level of single subjects, such as delayed learning or "jumps" in accuracy at potentially unpredictable points in the session (59). Performing analyses to compare any such learning dynamics between individual humans and learning models is an 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 (60), 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.

ACKNOWLEDGEMENTS

This work was funded in part by the Semiconductor Research Corporation (SRC).

Bibliography

1. R N Shepard. Toward a universal law of generalization for psychological science. *Science (New York, N.Y.)*, 237(4820):1317–1323, 1987. ISSN 0036-8075. doi: 10.1126/science.3629243.
2. Irving Biederman. Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147, 1987. ISSN 0033295X. doi: 10.1037/0033-295X.94.2.115.
3. Robert M Nosofsky. Similarity scaling and cognitive process models. *Annu. Rev. Psychol*, 43:25–53, 1992.
4. Heinrich H. Bülthoff and Shimon Edelman. 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*, 89(1):60–64, 1992. ISSN 00278424. doi: 10.1073/PNAS.89.1.60.
5. Joshua B. Tenenbaum and Thomas L. Griffiths. Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4):629–640, 8 2001. ISSN 0140-525X. doi: 10.1017/S0140525X01000061.
6. F Gregory Ashby and W Todd Maddox. Human category learning. *Annu. Rev. Psychol*, 56:149–78, 2005. doi: 10.1146/annurev.psych.56.091103.070217.
7. Stephen K. Reed. Pattern recognition and categorization. *Cognitive Psychology*, 3(3):382–407, 7 1972. ISSN 0010-0285. doi: 10.1016/0010-0285(72)90014-X.
8. John K Kruschke. ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 1992.
9. Stephen C Mckinley and Robert M Nosofsky. Selective Attention and the Formation of Linear Decision Boundaries. *Journal of Experimental Psychology*, 22(2):294–317, 1996.
10. W. Todd Maddox and F. Gregory Ashby. Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics 1993 53:1*, 53(1):49–70, 1 1993. ISSN 1532-5962. doi: 10.3758/BF03211715.
11. T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature 1990 343:6255*, 343(6255):263–266, 1990. ISSN 1476-4687. doi: 10.1038/343263a0.
12. Sharon Duvdevani-Bar and Shimon Edelman. Visual Recognition and Categorization on the Basis of Similarities to Multiple Class Prototypes. *International Journal of Computer Vision 1999 33:3*, 33(3):201–228, 1999. ISSN 1573-1405. doi: 10.1023/A:1008102413960.
13. Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 4 2006. ISSN 01628828. doi: 10.1109/TPAMI.2006.79.
14. Ruslan Salakhutdinov, Joshua B. Tenenbaum, and Antonio Torralba. Learning with

hierarchical-deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1958–1971, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2012.269.

15. B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 12 2015. ISSN 0036-8075. doi: 10.1126/science.aab3050.
16. Goker Erdogan and Robert A Jacobs. Visual Shape Perception as Bayesian Inference of 3D Object-Centered Shape Representations. *Psychological Review*, 124(6):740–761, 2017. doi: 10.1037/rev0000086.
17. Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. The Geometry of Concept Learning. *bioRxiv*, page 2021.03.21.436284, 1 2021. doi: 10.1101/2021.03.21.436284.
18. Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. 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*, 38(33):7255–7269, 8 2018. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0388-18.2018.
19. Robert Geirhos, Kantharaju Narayanappa, Benjamin Mitzkus, Tizian Thieringer, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Partial success in closing the gap between human and machine vision. *arXiv*, 2021.
20. M.N. Hebart, O. Contier, L. Teichmann, A.H. Rockter, C.Y. Zheng, A. Kidder, A. Corriveau, M. Vaziri-Pashkam, and C.I. Baker. THINGS-data: A multimodal collection of large-scale datasets for investigating object representations in brain and behavior. *bioRxiv*, page 2022.07.22.501123, 7 2022. doi: 10.1101/2022.07.22.501123.
21. Brenden M. Lake, Ruslan R. Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society (CogSci 2011)*, 172:2568–2573, 2011.
22. James L. McClelland, Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith. Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8):348–356, 2010. ISSN 13646613. doi: 10.1016/j.tics.2010.06.002.
23. Yaniv Morgenstern, Frieder Hartmann, Filipp Schmidt, Henning Tiedemann, Eugen Prokott, Guido Maiello, and Roland W. Fleming. An image-computable model of human visual shape similarity. *PLOS Computational Biology*, 17(6):e1008981, 6 2021. ISSN 1553-7358. doi: 10.1371/JOURNAL.PCBI.1008981.
24. Yaniv Morgenstern, Filipp Schmidt, and Roland W. Fleming. One-shot categorization of novel object classes in humans. *Vision Research*, 165:98–108, 12 2019. ISSN 0042-6989. doi: 10.1016/J.VISRES.2019.09.005.
25. Robert M. Nosofsky, Craig A. Sanders, Brian J. Meagher, and Bruce J. Douglas. Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*, 50(2):530–556, 4 2018. ISSN 15543528. doi: 10.3758/S13428-017-0884-8/TABLES/8.
26. Pulkit Singh, Joshua C Peterson, Ruairidh M Battleday, and Thomas L Griffiths. End-to-end Deep Prototype and Exemplar Models for Predicting Human Behavior. *arXiv*, 2020.
27. Ruairidh M Battleday, Joshua C Peterson, and Thomas L Griffiths. Modeling Human Categorization of Natural Images Using Deep Feature Representations. *arXiv*, 2017.
28. Joshua C. Peterson, Joshua T. Abbott, and Thomas L. Griffiths. Evaluating (and Improving) the Correspondence Between Deep Neural Networks and Human Representations. *Cognitive Science*, 42(8):2648–2669, 11 2018. ISSN 15516709. doi: 10.1111/COGS.12670.
29. Daniel L K Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–24, 2014. ISSN 1091-6490. doi: 10.1073/pnas.1403112111.
30. Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J Majaj, Elias B Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L K Yamins, and James J Dicarlo. Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. In *33rd Conference on Neural Information Processing Systems*, 2019.
31. Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence. *Neuron*, 108(3):413–423, 11 2020. ISSN 0896-6273. doi: 10.1016/J.NEURON.2020.07.040.
32. Chi-Tat Law and Joshua I Gold. Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature neuroscience*, 12(5):655–63, 2009. ISSN 1546-1726. doi: 10.1038/nn.2304.
33. Yael Niv. Reinforcement learning in the brain. *Journal of Mathematical Psychology*, 53(3):139–154, 6 2009. ISSN 0022-2496. doi: 10.1016/J.JMP.2008.12.005.
34. Nicolas Frémaux and Wulfram Gerstner. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in Neural Circuits*, 9(JAN2016):85, 1 2015. ISSN 16625110. doi: 10.3389/FNCIR.2015.00085/BIBTEX.
35. Gabriele Paolacci, Jesse Chandler, Panagiotis G Ipeirotis, and Leonard N Stern. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.
36. Stephen Todd and William Latham. *Evolutionary art and computers*. Academic Press Inc., 1992. ISBN 978-0124371859.
37. Persistence of Vision Pty. Ltd. Persistence of Vision Raytracer, 2004.
38. R. Rajalingham, K. Schmidt, and J. J. DiCarlo. Comparison of Object Recognition Behavior in Human and Monkey. *Journal of Neuroscience*, 35(35):12127–12136, 2015. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.0573-15.2015.
39. Harold Stanislaw and Natasha Todorov. Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, 31(1), 1999. ISSN 07433808. doi: 10.3758/BF03207704.
40. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury Google, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf Xamla, Edward Yang, Zach Devito, Martin Raison Nabla, Alykhan Tejani, Sasank Chilamkurthy, Qure Ai, Benoit Steiner, Lu Fang Facebook, Junjie Bai Facebook, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *33rd Conference on Neural Information Processing Systems*, 2019.
41. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Institute of Electrical and Electronics Engineers (IEEE), 3 2009. doi: 10.1109/CVPR.2009.5206848.
42. William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.
43. Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: from theory to algorithms*. Cambridge University Press, 2014. ISBN 978-1-107-05713-5.
44. Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 2017. doi: 10.1017/S0140525X16001837.
45. Gary Marcus. Deep Learning: A Critical Appraisal. *arXiv*, 2018.
46. Shimon Ullman. Using neuroscience to develop artificial intelligence. *Science*, 363 (6428):692–693, 2 2019. ISSN 10959203. doi: 10.1126/SCIENCE.AAU6595/ASSET/52C673F4-40F5-452C-B626-A0E2769AC25F/ASSETS/GRAPHIC/363{_}692{_}F1.JPEG.
47. Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience 2020 22:1*, 22(1):55–67, 11 2020. ISSN 1471-0048. doi: 10.1038/s41583-020-00395-8.
48. Hans P. Op de Beeck and Chris I. Baker. The neural basis of visual object learning. *Trends in Cognitive Sciences*, 14(1):22–30, 2010. ISSN 13646613. doi: 10.1016/j.tics.2009.11.002.
49. C. A. Seger. The Roles of the Caudate Nucleus in Human Classification Learning. *Journal of Neuroscience*, 25(11):2941–2951, 2005. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.3401-04.2005.
50. Hyoung F. Kim, Ali Ghazizadeh, and Okihide Hikosaka. Separate groups of dopamine neurons innervate caudate head and tail encoding flexible and stable value memories. *Frontiers in Neuroanatomy*, 8(October):120, 2014. ISSN 1662-5129. doi: 10.3389/fnana.2014.00120.
51. Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific Reports 2020 10:1*, 10(1):1–13, 1 2020. ISSN 2045-2322. doi: 10.1038/s41598-019-57261-6.
52. Robert M Nosofsky. The generalized context model: An exemplar model of classification. In Emmanuel M Pothos, editor, *Formal approaches in categorization*, chapter 2, pages 18–39. Cambridge University Press, 2011.
53. Douglas L. Medin and Edward E. Smith. Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, 7(4):241–253, 7 1981. ISSN 00961515. doi: 10.1037/0278-7393.7.4.241.
54. Jeffrey S Bowers, Gaurav Malhotra, Marin Dujmović, Milton Llera Montero, Christian Tsvetkov, Valerio Biscione, Guillermo Puebla, Federico G Adolfi, John Hummel, Rachel Flood Heaton, Benjamin Evans, Jeff Mitchell, and Ryan Blything. Deep Problems with Neural Network Models of Human Vision. *PsyArXiv*, 2022. doi: 10.31234/OSF.IO/5ZF4S.
55. Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, 2010. ISSN 13646613. doi: 10.1016/j.tics.2010.05.004.
56. Robert Ajemian, Alessandro D'Ausilio, Helene Moorman, and Emilio Bizzi. A theory for how sensorimotor skills are learned and retained in noisy and nonstationary neural circuits. *Proceedings of the National Academy of Sciences of the United States of America*, 2013. doi: 10.1073/pnas.1320116110/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1320116110.
57. German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 5 2019. ISSN 0893-6080. doi: 10.1016/J.NEUNET.2019.01.012.
58. Dale Purves, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel Lamantia, Jomes O. McNamara, and S. Mark Williams. Types of Eye Movements and Their Functions. In Dale Purves, George J. Augustine, David Fitzpatrick, Lawrence C. Katz, Anthony-Samuel LaMantia, James O. McNamara, and S. Mark Williams, editors, *Neuroscience*. Sinauer Associates, Sunderland, MA, 2nd edition edition, 2001. ISBN 0-87893-742-0.
59. Charles R. Gallistel, Stephen Fairhurst, and Peter Balsam. The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 101(36):13124–13131, 9 2004. ISSN 00278424. doi: 10.1073/PNAS.0404965101.
60. Harry F. Harlow. The formation of learning sets. *Psychological Review*, 56(1):51–65, 1 1949. ISSN 0033295X. doi: 10.1037/H0062474.
61. Anne C. Smith and Emery N. Brown. Estimating a State-Space Model from Point Process Observations. *Neural Computation*, 15(5):965–991, 5 2003. ISSN 0899-7667. doi: 10.1162/089976603765202622.

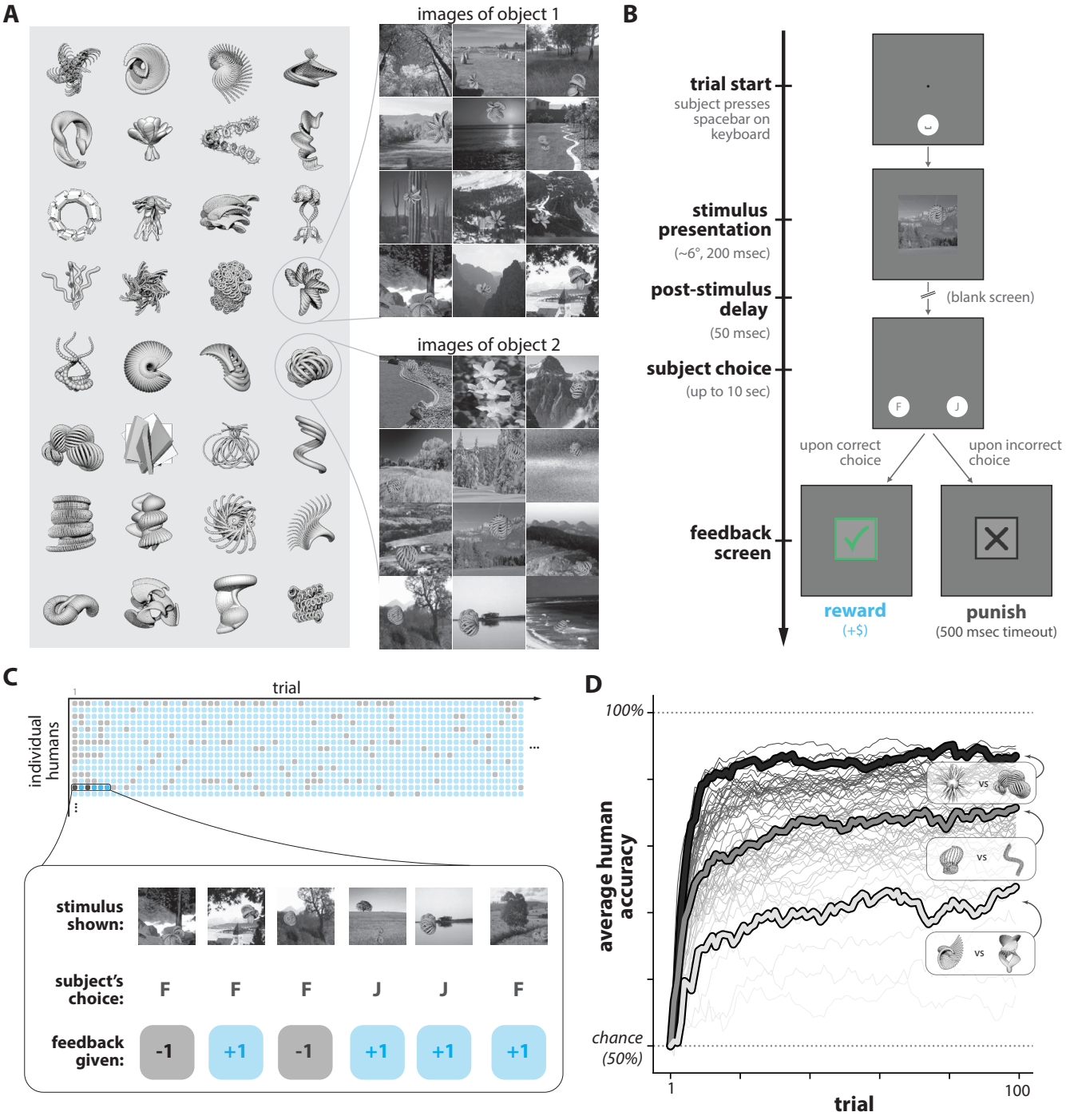

Fig. 1. Humans learning novel objects. **A. Images of novel objects.** Views of synthetically-generated 3D object models were created with random viewing parameters (i.e. background, location, scale, and rotational pose). **B. Task paradigm.** On each trial, a randomly selected image (of one of two possible objects) was briefly shown to the subject. The subject then had to report the identity of the object using a left/right choice. Positive reinforcement was delivered if the subject choice was "correct", based on an object-choice contingency that the subject learned through trial-and-error (e.g., object 1 corresponds to "right", and object 2 corresponds to "left"). **C. Example subject-level learning data.** Each behavioral session consisted of a randomly sampled sequence of 100 trials (i.e. images and their choice-reward contingencies). Image stimuli were never repeated, ensuring each trial gauges the ability of the subject to generalize to unseen views of the new objects. Each behavioral session resulted in a sequence of corrects / incorrects, for each subject. **D. Human learning curves**. We averaged across human subjects to obtain an estimate of average accuracy as a function of trials on $n = 64$ subtasks (each consisting of a distinct pair of objects). We found that subjects were reliably less accurate on some subtasks than others; three example subtasks across that range of accuracy are highlighted.

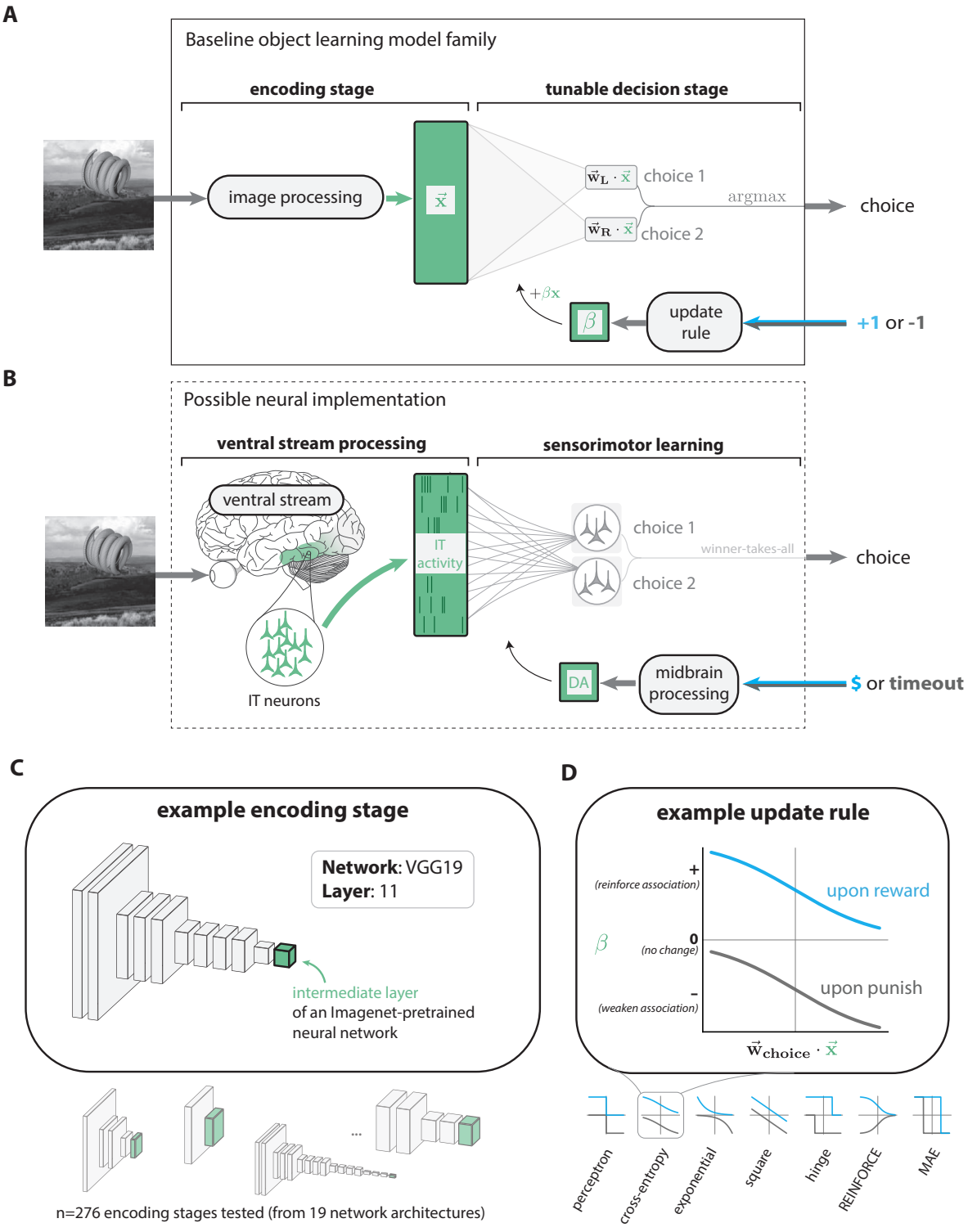

Fig. 2. Baseline model family of new object learning in humans. **A. Baseline model family.** Each model in this family consisted of two stages: an image-computable *encoding stage* which re-represents an incoming pixel image as a vector ($\vec{x}$), and a tunable *decision stage*, which uses $\vec{x}$ to generate a choice and learn from feedback. To generate a choice, choice preferences ($\vec{w}_c \cdot \vec{x}$) are computed, and the most preferred choice is selected. Subsequent scalar-valued environmental feedback is processed to generate an update to (only) the parameters of the decision stage. Specific models in this family correspond to specific choices of the encoding stage (n=276 encoding stages) and the update rule (n=7 update rules). **B. Possible neural implementation of each baseline model in the brain.** We suggest the functionality of the encoding stage in **A** is performed by the visual ventral stream, where distributed population activity high level visual areas (approximately located in green regions), such as area IT, re-represents incoming retinal images. Sensorimotor learning could be mediated by synaptic plasticity in a visuomotor association region, downstream of those areas. Midbrain processing of environmental reinforcement signals could guide plasticity at sensorimotor synapses via dopaminergic (DA) projections. **C. Encoding stages.** We tested 276 sensory models, based on leading image-computable models of ventral stream processing (intermediate layers over 19 Imagenet-pretrained deep convolutional neural networks). **D. Update rules**. We tested seven update rules, drawn from statistical learning theory and reinforcement learning. Each corresponds to a different optimization objective (e.g., the square rule attempts to minimize the squared error between the choice preference and the subsequent magnitude of reward).

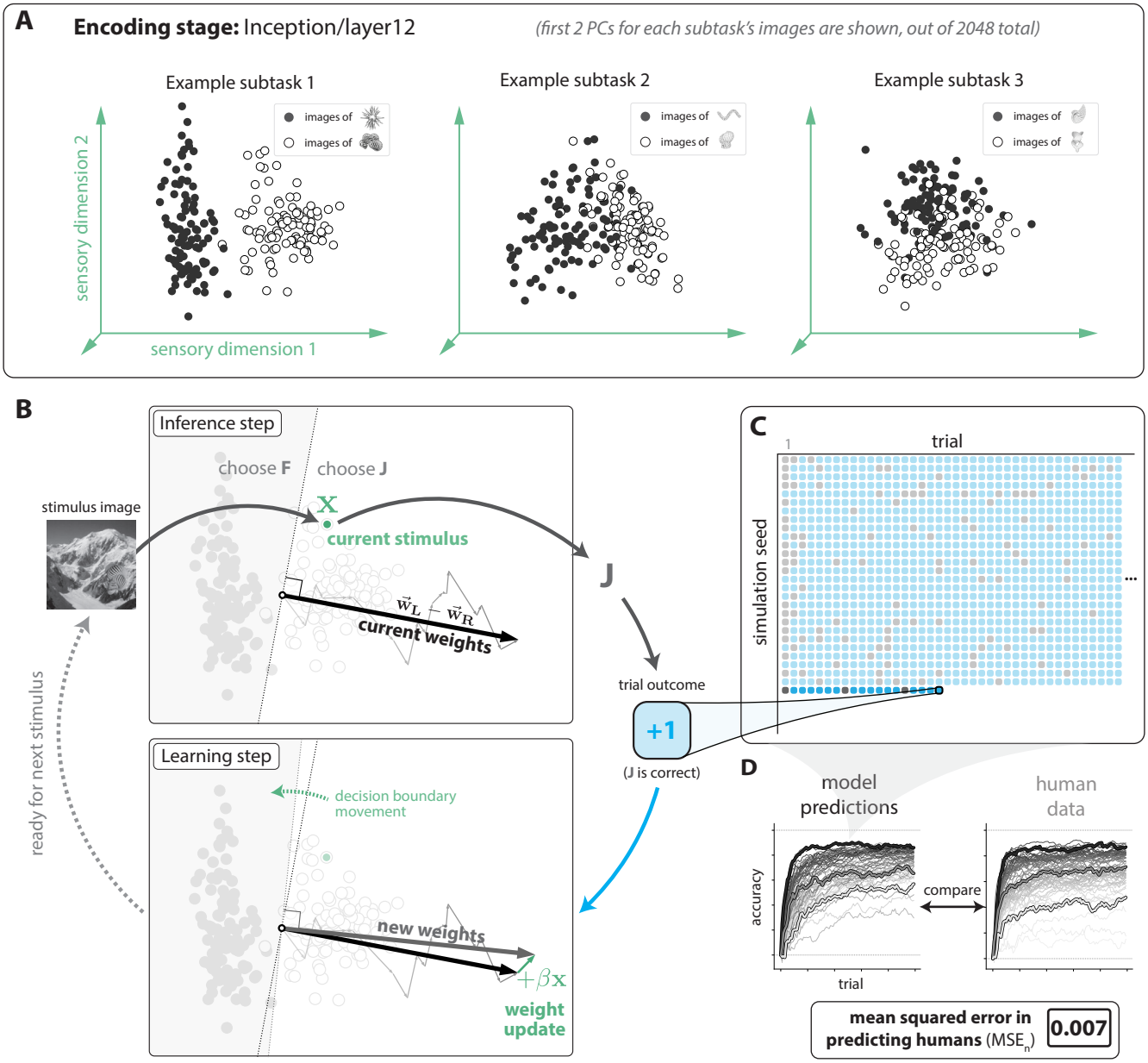

Fig. 3. Model simulations of human learning. **A. Example encoding stage representations of novel object images.** Each subtask consists of images drawn from two object categories (indicated in black and white dots). The first two principal components of an 2048-dimensional encoding stage (Inception/layer12) are shown here. Note that, for clarity, these are not the same two encoding dimensions for each subtask. 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, $\mathbf{x}$. This location falls on the "choose J" side of the current decision boundary ($\vec{\mathbf{w}}_\mathbf{L} - \vec{\mathbf{w}}_\mathbf{R}$), leading to the choice "J", 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 (500 simulated sessions for each of the 64 subtasks), and recorded the responses (correct or incorrect) of the model. **D. Comparing model learning to human learning.** We compared the simulated model predictions to the subject-averaged human learning data, by asking how well a model could predict the performance of a typical human, given the same subtask and the same number of trials. We used a mean-squared error metric (MSE_n; see Methods 3.4) to quantify the goodness-of-fit of these predictions.

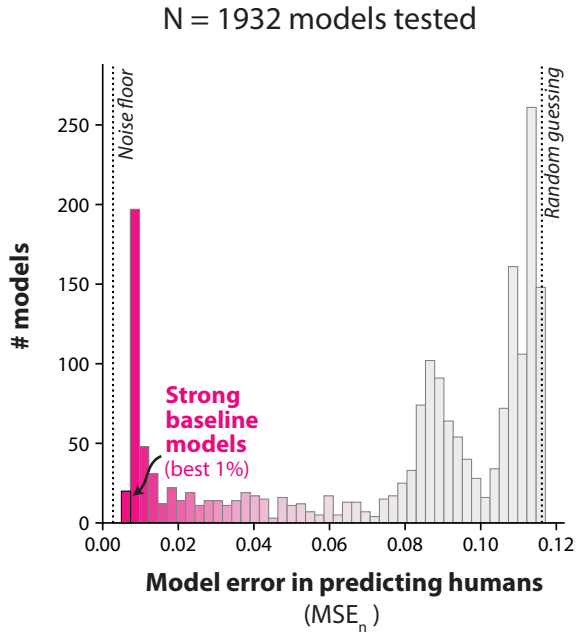

Fig. 4. Bias-corrected mean-squared errors (MSE_n) vs. humans for all models tested. The n=1,932 models we tested varied widely in the strength of their alignment with human learning (i.e. their average squared predictive error). We denote the best 1% of such models as *"strong baseline models"*. The 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.

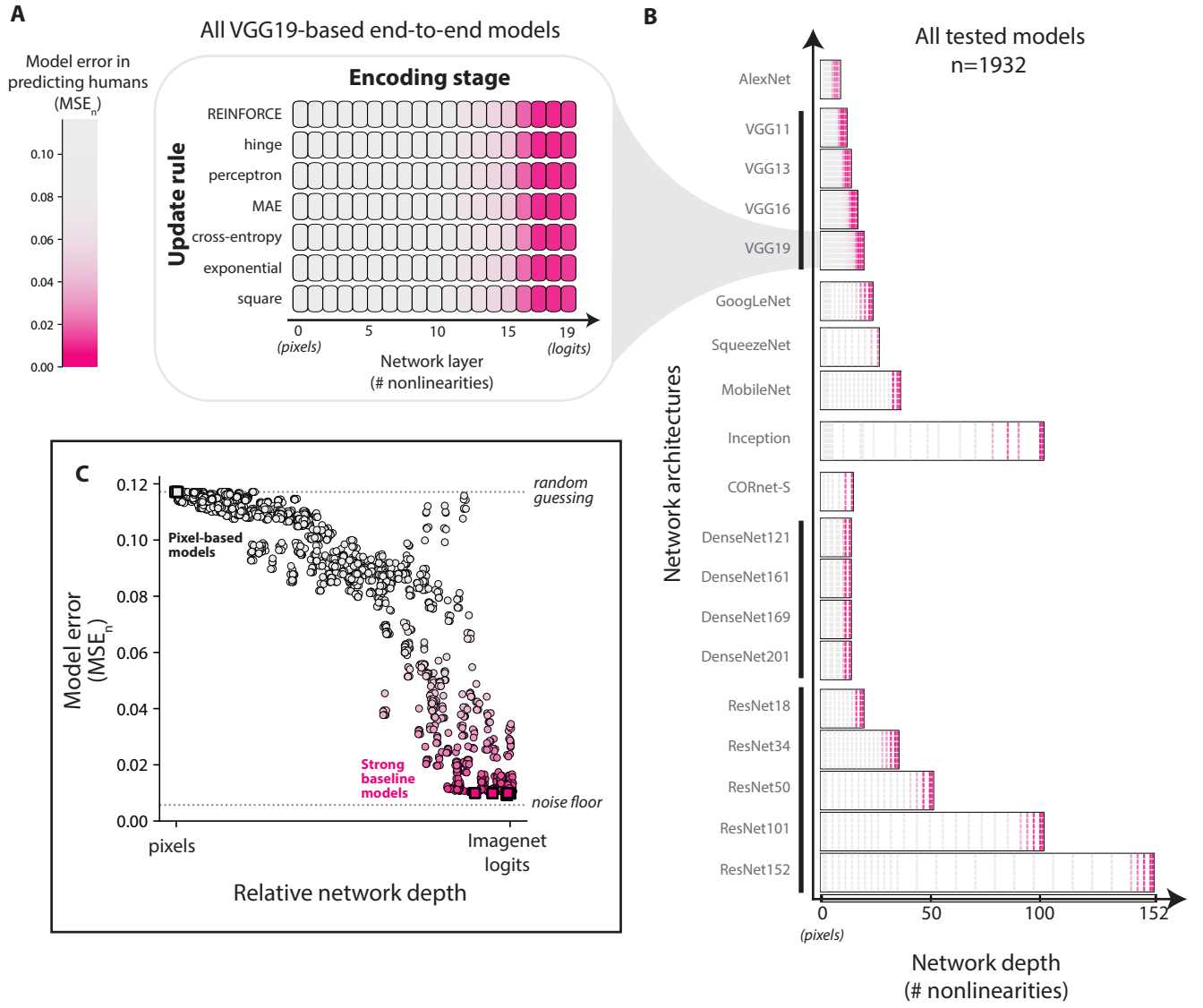

Fig. 5. Evaluating the effect of model design choices on predictive accuracy of human learning. **A. Example model scores across encoding stages and update rules.** An example of the effect of update rule (y-axis) and encoding stage (x-axis) on model scores (color). Model predictive accuracy was highly affected by the choice of encoding stage; those based on deeper layers of DCNNs had the most human-like learning behavior. On the other hand, the choice of update rules had a minuscule effect. **B. Overview of all models tested.** In total, we tested encoding stages drawn from a total of n=19 DCNN architectures. **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.

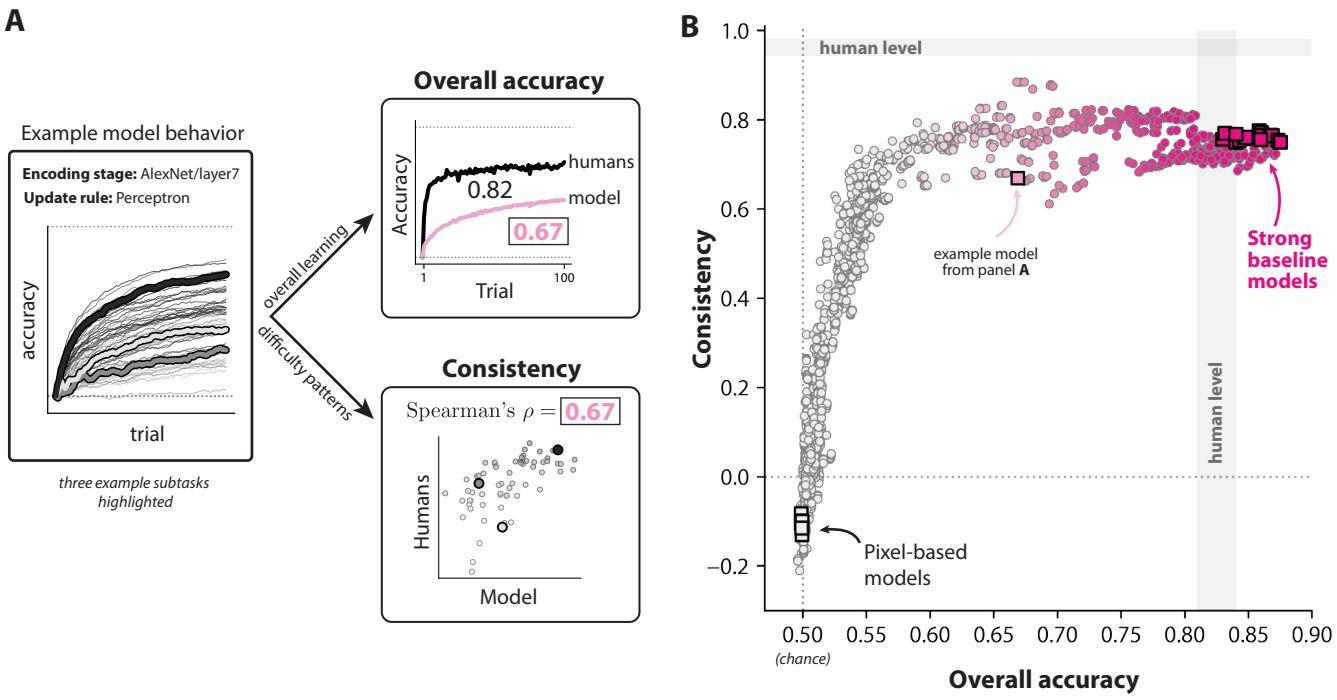

Fig. 6. Comparing more granular behavioral signatures of human learning. **A. Decomposing model behavior into two metrics.** We examined model behavior along two specific aspects of learning behavior: **overall accuracy** (top right), which was the average accuracy of the model over the entire experiment (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 matches that of humans. We quantified consistency using Spearman's rank correlation coefficient between the vector of subtask-wise accuracies in humans and in models. **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 noise ceilings are the estimated range of metrics expected upon repetitions of the behavioral experiment. The color map encodes the overall score (MSE_n) of the model (colorbar in Figure 5A).

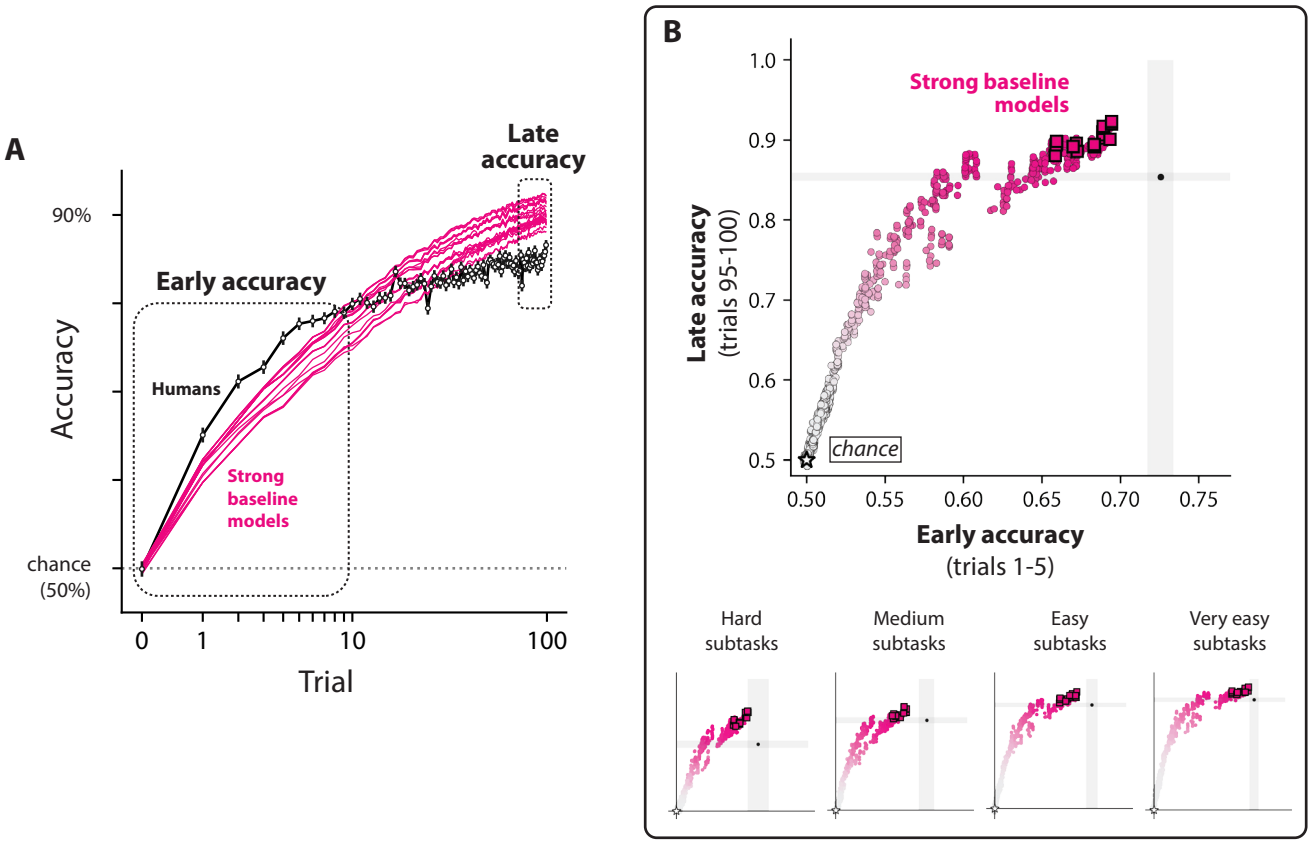

Fig. 7. Humans outperform all strong baseline models in low-sample regimes. **A. Subtask-averaged learning curves for humans and strong baseline models.** The y-axis is the percent chance that the subject made the correct object report (chance is 50%). The x-axis is the total number of image examples shown prior to the test trial (log scale). Overlaying the average learning curves for humans (black) and models (magenta) reveals that humans have an advantage in performance in low-sample regimes, compared to all strong baseline models. 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 late accuracy in humans (which we defined as accuracy at the end of the experiment, over trials 95-100), but no model reached human-level accuracy in the early regime (which we defined as the average accuracy over trials 1-5). This trend was not due to specific subtasks, as it was present in subtasks of different levels of difficulty (bottom row).

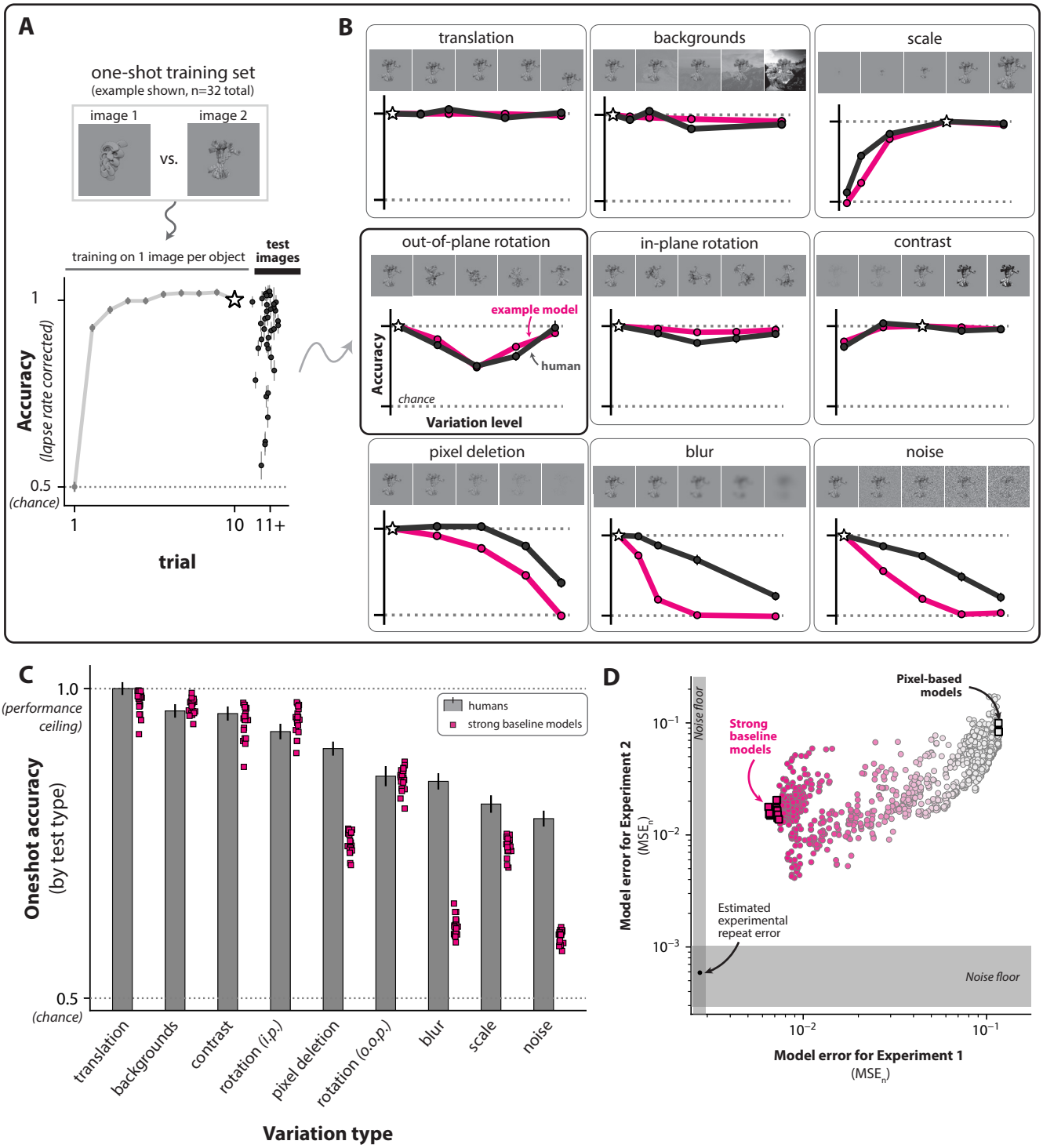

Fig. 8. One-shot learning in humans. **A. One-shot learning task paradigm.** We performed an additional study (Experiment 2) to characterize human one-shot learning abilities (using the same task paradigm in Figure 1). The first 10 trials were based on two images (n=1 image per object) that were resampled in a random order. On trials 11-20, humans were tested on transformed versions of those two images (nine types of variation, with four levels of each, n=36 total generalization tests) **B. Human and model one-shot generalization to nine types of image variation.** 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 one-shot tests.** We averaged human one-shot accuracy (gray) on each type of image variation, and overlaid all strong baseline models (magenta, n=20 models). The errorbars are the the 95% CI (basic bootstrap). **D. Comparison of one-shot and high-variation MSE$_n$ scores.** No strong baseline model could fully explain the pattern of one-shot generalization observed in humans (Experiment 2), nor their behavior on the high-variation benchmark (Experiment 1). The error scores are shown on the log scale.

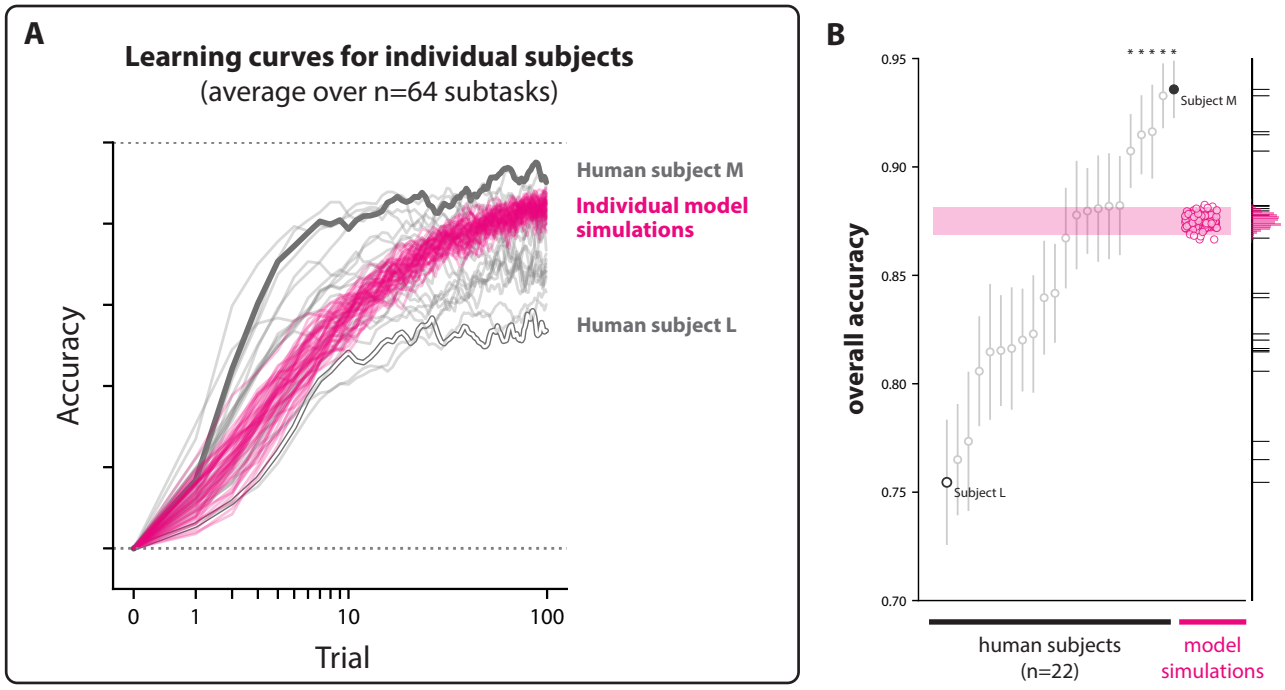

Fig. 9. Individual human subjects vary widely in their performance. We analyzed 22 subjects who performed all 64 subtasks in Experiment 1, and tested for differences in their overall learning abilities. **A. Individual-level learning curves.** Each gray curve corresponds to the subtask-averaged learning curve for particular human subject (using state-space smoothing (**?**)). In humans (top row), some subjects (e.g. **Subject M**, highest average performance over all subtasks) consistently outperformed the subject-averaged human learning curve (in gray), while others consistently underperformed (e.g. **Subject L**, lowest average performance over all subtasks). In magenta are learning curves taken from the top performing model. **B. Some individual humans outperform all baseline models.** Five out of 22 subjects ($\approx 22\%$ of the population) had significantly higher overall performance than the highest performing model we tested (one-tailed Welch's t-test, Bonferroni corrected, $p<0.05$).