Supplementary Information for:

Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity

Authors: Mathias Sablé-Meyer[1,2]*, Joël Fagot[3,4], Serge Caparos[5,6], Timo van Kerkoerle[1], Marie Amalric[7], and Stanislas Dehaene[1,2]*.

Affiliations:

[1] Cognitive Neuroimaging Unit, Commissariat à l'énergie atomique et aux énergies alternatives, INSERM, Université Paris-Saclay, NeuroSpin, 91191 Gif-Sur-Yvette, France;

[2] Chair of Experimental Cognitive Psychology, Collège de France, Université Paris Sciences Lettres (PSL), 75005 Paris, France;

[3] Cognitive Psychology Laboratory, CNRS, Aix-Marseille Université, 13331 Marseille, France;

[4] Station de Primatologie-Celphedia, CNRS UAR846, 13790 Rousset, France

[5] Department of Psychology, Fonctionnement et Dysfonctionnement Cognitifs : les âges de la vie, Université Paris 8, 92000 Nanterre, France;

[6] Human Sciences section, Institut Universitaire de France, 75005 Paris, France;

[7] Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, United States

*Correspondence to: mathias.sable-meyer@ens-cachan.fr or stanislas.dehaene@cea.fr

This PDF file includes:

Supplementary Methods and Results; Figs. S1 to S9; Tables S1 to S4

Additional Methods and Details of each experiment

Adults, experiment 1

Participants. 612 French adults were recruited for an online experiment (395 males, 217 females, age group breakdown: <18 years, 42 subjects; 18-25 years, 127 subjects; 25-60 years, 419 subjects; >60 years, 24 subjects). The experiment was advertised on social media using the lab's social media account. The entire experiment was run on the participant's device and took typically less than 15 minutes. Participants were not compensated for their participation. No personally identifying information was collected in this experiment. This experiment was approved by the ethical committee of Université Paris-Saclay under the reference CER-Paris-Saclay-2019-08.

Procedure & stimuli. This experiment featured only canonical displays (5 reference shapes and 1 deviant shape). It started with two training pairs of geometric shapes, randomly selected from the 3 we used throughout the generalization 2 task for baboons. There were therefore exactly 2 + 11*4 = 46 trials. The experiment was programmed using the jsPsych framework (1), "a JavaScript library for running behavioral experiments in a web browser." Participants first filled a consent form, then a demographic questionnaire, which collected information regarding their sex, age range, and education level. Then they were presented with the task instructions, and finally a sequence of intruder trials. On each trial, they were asked to click on the outlier, either with the mouse or with a touchscreen if their device had one. In this experiment only, the six shapes were organized in a circle as big as the screen permitted. Upon clicking on a shape, participants received visual (highlighting the selected shape in red if incorrect, in green otherwise, and highlighting the correct shape in green) and auditory feedback (rising or falling tone). Shapes were shown in solid black on white background. This experiment can be found at the following address: https://neurospin-data.cea.fr/exp/mathias-sable-meyer/oddball_original/
Statistical analysis: Responses slower than the overall 99th percentile were removed from analysis of this experiment, as well as experiment 2 and the Himbas experiment to match the analyses: during online experiments, some trial took unreasonable durations (e.g. over a minute), strongly suggesting participants taking a break during the experiments. In experiment 1, the 99th percentile thresholding sometimes removed all datapoints from some participants' conditions (e.g. an entire shape); in such case the participant was removed entirely: in total, 7 out of 612 recruited participants were removed.

Adults, experiment 2

Participants. 117 French adults were recruited for an online experiment (45 males and 72 females; age group breakdown: 18-25, 9 subjects; 25-40, 43 subjects; 40-60, 56 subjects; >60, 9 subjects). The recruitment process and ethical approval were identical to that of the first experiment. Because this experiment was longer, participants were incentivized to participate by being offered to participate in a lottery for a 30€ cash prize that three participants would receive. Should they want to participate to the lottery, participants had to disclose an email address, which was collected separately from the experiment's data and could not be linked to it afterwards. 83 out of 117 participants submitted their email for participation.

Procedure and Stimuli. The procedure and stimuli were identical to that of experiment 1 with the following five differences. (1) Participants saw an additional webpage with information about the lottery. (2) Shapes were displayed in white on a black background. (3) Instead of displaying the shapes along a circle they were displayed in two lines of three items, as shown in Fig. 1B. (4) Participants received 10 training trials with images and another 6 with the easy geometric training shapes, in two consecutive blocks. They had to repeat the training blocks if they performed worse than 80% correct. The training stimuli were identical to those used in the baboon experiment (see Fig. 3) and included random rotation[1] and scaling. (5) Half of the displays used a standard presentation (5 reference shapes and 1 deviant), and half used a swapped presentation (5 deviant shapes and 1 reference shape), for a total of 88 experimental trials with geometric shapes.

Compared to experiment 1, the changes listed under points 2-5 were introduced in order to anticipate the changes required to replicate the baboon experiment. The displays in Fig 1B show example stimuli from this version of the experiment. This design was adopted throughout all other experiments. This experiment is available at https://neurospin-data.cea.fr/exp/mathias-sable-meyer/oddball/.

Sequence experiment

Participants. 19 participants were tested in this experiment. It was run at ENS in Paris, in isolated testing booths. The first three participants were pilots whose results were used to tune the difficulty of the experiment. Subjects were recruited through the RISC mailing list, mean age was 23.1 years old (std = 2.55), 9 women and 11 men, with a mean of 3.44 years of post-bachelor education (std = 1.5). All participants signed an informed consent form and received 15€ for their participation. Due a schedule conflict one participant did not complete one condition ("parallelogram") of the experiment: the missing value was replaced in the ANOVA with that participant's overall average error rate, and left missing from all other analysis.

Procedure. The experiment was organized in 9 mini-blocks, each with a fixed geometric shape. In each mini-block, participants were first shown 6 examples of a given sequence (with random scaling and rotation), and were then presented with sequences that could contain a deviant. For each sequence after the sixth example, after the 4th dot was displayed, they had to press a button to indicate whether that sequence followed the reference sequence or not. Following each answer, they received auditory feedback using an ascending pitch if correct and a descending pitch otherwise and were shown the four dots location, as well as a 5th dot at the correct location

[1] Due to a bug in the code of the experiment, the training images were not scaled, though they were properly rotated. This bug only affected the training images. All geometric shapes, in training or testing, were properly scaled. This minor problem affected the training stimuli in experiments with French adults (exp. 2), Himbas adults, French kindergartners, and 1st graders – but was corrected for the baboon experiments.

for deviant trials. After 150 trials, there was a short pause, and then a new mini-block started, with 6 new examples to start with.

Stimuli. The sequences of dots traced the geometric shapes in a top-left, top-right, bottom-left, bottom-right order. Shapes were presented with a random orientation (with angles now ranging from 0 to 359°) and random scaling so that they spanned 150 to 225 pixels on the screen, and they were positioned so that the last position would be at one of 9 possible locations on the screen. In this sequential format, we considered it essential that the last two positions were identical for all shapes. We therefore excluded the two shapes for which we could not match the bottom edge, namely the square and the rhombus. Given the greater difficulty of the task in the sequential presentation mode, we had to adjust the distance of the deviant to the correct location. Pilot participants were run in order to estimate the distance required to obtain a success rate of ~75% overall, and the deviant value used for the remaining N=16 participants was 0.55 times the matched average distance of any two points. The presentation order of the blocks was random with a single block for each shape and 150 trials within each block[2], with half of the trials being outliers. The timing of the sequence was as follows: points appeared for 400 ms followed by a 200 ms empty scree. After the participants' response, the screen stayed black for a random duration ranging from 750 ms to 1250 ms.

Subjective Rating

Participants. 48 French adults were recruited for an online experiment (21 Males and 27 females; age group breakdown: 1-18, 1 subject; 18-25, 3 subjects; 25-60, 41 subjects; >60, 3 subjects). The recruitment process and ethical approval were identical to that of the first experiment.

Stimuli. We presented the participants with our 11 quadrilaterals, in the reference orientation and presented as static images with a white shape on a black background.

Procedure. After the consent and the questionnaire, participants were instructed to give a rating for each shape one the page using a scale from 1 to 100, while trying to be as consistent in the rating as possible. Participants were randomly assigned to one of two conditions: either they were asked to give a rating of "complexity" (27 participants) or to give a rating of "regularity" (21 participants). Participants saw a page with shapes from another study not analyzed here, and then a page with our 11 reference shapes and a slider from 1 to 100 for each shape. They were asked to not transfer the scale from the previous shapes from to the 11 quadrilaterals, but instead to try and use the entire scale again and to be as consistent as possible between the shapes. We merged the data from the two conditions by reversing the scale of the "regularity" condition so that a score of 100 on "regularity" would map on to a score of 1 on "complexity" and conversely.

[2] This was adjusted depending on participants time constraints: min 100, max 170, median 155. Two participants had to stop before the end because of time constraints, one missing one shape and the other two.

Visual Search Paradigm

Participants. 11 French adults were recruited (5 Females, 5 males, age range 21 - 35, mean 27.3 years, one did not complete the demographic form). Participants were not compensated for their participation. This experiment was covered by the ethical committee of Université Paris-Saclay under the reference CER-Paris-Saclay-2019-063.

Stimuli. For each trial, repetitions of a given shape and possibly its deviant were presented in black on light gray (Fig. 2A). Their rotation and scaling were uniformly sampled, similarly to previous experiments, and they were randomly placed inside a gray circle that spanned almost the entire computer screen. The experiment comprised 11 blocks, one per reference shape, each with 24 trials randomly shuffled, using a factorial design with three factors, namely, deviant type (4 possible deviants), numbers of shapes on screen (3 possibilities: 6, 12 or 24) and presence or absence of a deviant shape, for a total of 264 trials. The experiment was programmed using the jsPsych framework and was run online.

Procedure. When connecting to the shared online URL, participants clicked to start and were prompted with instructions. For each display, they had to press the left arrow key if they thought that one of the shapes differed from the others, and the right arrow key if they thought that all shapes were identical. After pressing one of the arrow keys, the experiment started: the screen displayed a light-gray circle spanning the maximum available area with 15px padding at the top and the bottom, inside which items were placed randomly. After each response, subjects received both auditory and visual feedback, which explicitly indicated the location of the deviant shape if one was present (the deviant was colored green if answered correctly, red otherwise). The experiment was structured in blocks of similar shapes and lasted about 20 minutes in total.

Analysis and results. For each shape, each number of displayed item, and each target presence condition, we removed responses whose response time exceeded the mean response time plus three standard deviation. Detailed analyses of the visual search available in the supplementary materials.

Himbas

Participants. 44 native Himba adults were recruited for an experiment taking place on a tablet computer (mean age 24.5 years, minimum 14 years old and maximum 62 years old, 13 Male and 31 Females). The Himba of Northern Namibia (Southern Africa) are a population living a traditional lifestyle in rural settlements, with little exposure to Western society. All the participants were native speakers of (and monolingual in) Otjihimba, a dialect of the Otjiherero language, which does not have vocabulary for most geometric shapes (though they refer to "squares", for example, with a very direct metaphor akin to "a shape with four angles"). Out of the 44 participants, we analyzed data of 22 participants who did not attend a single year of schooling (15 Females, 7 Males, age range 14-62, mean 26); additional analyses of the effect of schooling below. Ethical approval was obtained from the ethics committee of Goldsmiths University of London (REISC_1390, 4 june 2018).

Procedure & Stimuli. The experiment was rigorously identical to experiment 2, but the instructions were given verbally by a translator. Participants were compensated in kind (1Kg of sugar, 1Kg of flour, and 500mg of soap).

A typical testing day with the Himba unfolds as follows. On arrival at a village, we park outside the village boundary. The interpreter speaks to the village chief or his representative if he is absent for more than a day. The chief is informed of the general purpose of our visit and asked if he can inform the village that they may participate in our tasks in return for a small gift of flour, sugar and soap (value ~USD 3). We do not offer money for which, in any case, the remote villagers would have little use. If the chief agrees (there has never been a case when he has not) we set up our equipment. We never approach any individual Himba, but our translator welcomes them if they ask to take part. Occasionally, people are too busy or reluctant to take part, but normally the only reason for obtaining small samples of participants is the absence of a large part of the population away from the village with their herds. In general, the word gets round and people volunteer, sometimes coming from other nearby villages.

In all cases, participants are told that they can refuse to take part in the study or withdraw at any point. We do not collect the names of the participants. We collect information of gender, estimated age, and reported level of education. We explain the purpose of the study in words that can be understood by the participant. Explanations are translated from English to Otjihimba by the local guide. We obtain oral consent, and inform the participants that they will receive the gift in any case, even if they decide to terminate the task. Although we decided to always terminate a testing session if a participant shows signs of distress, this never happened given the trivial nature of the tasks. Beyond acquiring approval that conforms to our professional Code of Practice, we always bear in mind codes of conduct appropriate for the Himba.

The translator explains the following to each participant:

"You are here to participate in a vision task which is a bit like a game. You do not have to participate if you do not want to and can stop at any time if you feel uncomfortable. The task is not difficult and will last for about 30 to 45 minutes. You will be given instructions and do a short practice first. The task is harmless and does not cause any pain. You can ask us not to use your results after you have participated. At the end of the task, you will receive three presents (flour, sugar, and soap). Before we start, you must confirm that you agree with these things. You can now ask any question if something is unclear. If you do not like the task, you can stop at any time and leave. You will receive the presents anyway."

All these elements (plus some simple explanations about the aim of the test, that is, to study “how we see the world”) are also given to the chief when we arrive in a traditional village. We hope and expect that the Himba will be direct and indirect beneficiaries, and that the project will contribute to the national and international database on endangered languages and cultures, and to the preservation of the Himba language and culture. We take seriously the responsibilities and the mutualities of benefit that accrue from cross-cultural research with remote peoples, and we believe that we can demonstrate that we have actively furthered remote peoples’ interests in our previous research. Issues of identity, belonging and exclusion are currently highly prominent and our project contributes to inter-cultural understanding in a non-trivial way. The intellectual property rights of the Himba in their language and culture is explicitly respected.

Kindergartners

Participants. 28 French kindergartners (mean age 64 months; range 59-70 months; 15 boys, 13 girls) from two classrooms were tested individually in their school, by groups of two, in a quiet room. Each participant was accompanied by one experimenter. They were not compensated for their participation. This experiment was approved by the ethical committee of Université Paris-Saclay under the reference CER-Paris-Saclay-2019-08 after a specific amendment was submitted. Parents were contacted and had to give their consent beforehand. The participants gave oral consent on the day of the experiment.

Procedure & Stimuli. The experiment was identical to experiment 2 except for the fact that we removed the swapped trials to make the experiment shorter.

First graders

Participants. 156 French first participated in this study. Parents were sent letters beforehand, and could request that children not participate in the project. Participants were tested individually on tables in a quiet room in their school. The data collection was part of the Bien Joué project, approved by the ethical committee of Université Paris-Saclay under the reference CER-Paris-Saclay-2019-042-A1.

Procedure & Stimuli. The experiment was completely identical to the kindergartners' experiment.

Baboons

General set-up.

Participants were 26 Guinea baboons (Papio papio, 18 females, age range 1.5-23 years, mean age 11 years) from the CNRS primate facility (Rousset-sur-Arc, France). Baboons lived in a 700 m2 outdoor enclosure with access to indoor housing and had a permanent access to ten Automated Learning Devices for Monkeys equipped with a 19-inch touch screen and a food dispenser. Note that the baboons' environment contains a mixture of natural features (e.g. trees, congeners) and artificial tools and buildings with rectangular shapes (e.g. prefabricated rooms, testing booths, computer screens, etc).

A key feature of ALDM is a radio-frequency identification (RFID) reader that can identify individual baboons through microchips implanted in their arm (2). The baboons therefore participate in the research at will, without having to be captured, as the test programs can recognize them automatically. The experiment was controlled using EPrime software (Version 2.0, Psychology Software Tools, Pittsburgh). Ethical Standards: the baboon experiment received ethical approval from the French Ministry of Education (approval APAFIS 2717-2015111708173794 v3).

Training scheme. The baboon experiment required several steps of training to ensure that, stimuli set aside, the primates understood the intruder task and could generalize rapidly to new stimuli from different domains. Because we were not sure about the outcome of each of the steps, the entire experiment presented in Fig. 3A was run over three different batches of about one week: a pilot mid-October 2018, a first test of generalization late November 2018, and the test with the quadrilaterals in May 2019.

In the first pilot batch, we tested only 6 primates (Cauet, Dora, Dream, Flute, Hermine and Articho, although the latter animal was not interested in the task and stopped early on). We attempted to start training with displays containing 6 shapes with one intruder. While all baboons except Articho succeeded after 2000 to 3200 trials, the low reinforcement level (chance at one in six) made the early exploration of the task unrewarding and we feared baboons might become disinterested before starting to grasp the task. Therefore, for the two other batches of training, we introduced progressive learning steps with only 3, then 4, 5 and ultimately 6 shapes on display for each trial (see Fig. 3).

In the second batch, we tested all available primates (22 animals) following the structure of Fig. 3A up to and including generalization 1, i.e., the first generalization task. Each primate automatically moved to the next step whenever the error rate fell under 20%. Out of 22 baboons, 18 learned the task to the criterion up to stage 5 and progressed to the generalization task. Out of these 18, all generalized successfully: the percentage of errors was significantly better than chance on the first block with 10 novel images in both presentation modes (binomial test against chance, separately for each baboon: all *p*'s ≤ .001). With further training, all animals again reached the 20% error threshold. Out of the remaining four, three did not reach the end of the first training task at all, and one reached the second training task and stopped. On Fig. 3B, the data reported in the "initial training" and "generalization 1" plots are taken from this batch of data.

The third and final batch tested all available primates (25 animals), following the structure of Fig. 3A. All animals were restarted from the first training task and followed the entire training scheme, only skipping generalization 1 and going straight to generalization 2, then on to the main test. Out of 25 baboons, 20 baboons reached generalization 2. Testing for significant generalization on only 6 different trials could not be done for each animal individually, but we verified that performance was better than chance when grouping the 20 animals together (binomial test, 42 errors in 120 trials, chance at 83.3%, $p < .0001$). After further training on those stimuli, all of them successfully reached 20% error threshold on generalization 2 and moved on to the test task where they stayed either until they reached 100 blocks of 88 trials (11 primates) or until they stopped performing the task.

Among the 5 baboons who did not participate to the final test, 4 never reached the 20% error threshold on the first training task (three of them stopped being interested in the task early on, one stayed at chance for more than 7500 trials but kept trying). Finally, one primate progressed very slowly over 8800 trials in the first training task, reached the 20% error threshold on block 88 (after having performed 5700 trials in session one and reaching 54% errors), and stopped performing the task. The data reported in Fig. 3B ("generalization 2") and Fig. 3C are taken from this batch of data, i.e. from the 20 primates that reached generalization 2. For reference, Fig. S3 shows the evolution of performance over successive training stages for each of those 20 animals (first 20 rows), and the performance for the remaining 6 animals who could not be successfully trained (last 6 rows).

Method. The stimuli were identical to those used with French adults in the second version of the intruder task, except (i) the experiment itself was reprogrammed using custom software specific to the baboon lab, and (ii) baboons received a drop of dry wheat for every correct response. Incorrect responses were followed by a 3-sec time-out indicated by a green screen.

Additional analyses. To evaluate the heterogeneity across primates, Fig. S4 presents the cross-correlation matrix of the error rates of the 20 baboons that reached the testing task, separately for early (first 33 blocks), middle (blocs 34 to 66) and late (blocs 67 to 99) parts of the experiment. Of note, baboons were free to take different numbers of blocks – this explains why there are fewer primates in the "late" category. Within a category, all primates are comparable in that they performed the same number of blocks. We can see that as baboons progressed in their training (and fewer remain), their behavior became increasingly consistent across animals.

Definition of the symbolic model

The symbolic model assumes that participants extract the discrete geometric properties of shapes while abstracting away from superficial changes in size, location, orientation and display type (static or sequential). As a result, the model predicts that outlier detection difficulty should depend only on the symbolic distance between the lists of features of the standard and outlier shapes. The more geometric properties a shape has, the more properties a deviant might break, therefore the easier it should be to detect. Because the distance is computed pairwise, this model does currently not account for any difference between canonical and swapped conditions, although a penalty could easily be added.

This model has a single free parameter: a perceptual threshold θ below which the model fails to discriminate lengths or angles and therefore considers them equal. The model considers that two lengths are equal by looking at their ratio: two lengths l_1 and l_2, with $l_1 > l_2$, are considered equal whenever $\frac{l_1}{l_2} - 1 < \theta$. For simplicity, the same threshold is used for angles: two angles are considered equal whenever they differ by less than $\theta \times \frac{\pi}{2}$

For any given quadrilateral, and for a given threshold, the model computes a vector of bits of length 22, representing the following properties: (i) 6 bits, one per pair of edges, coding whether their lengths are equal or different, (ii) 6 bits, one per pair of edges, coding whether their directions are parallel or not, (iii) 6 bits, one per pair of angles, coding whether their angles are equal or not, and (iv) 4 bits, one per angle, coding whether the angles are right angles or not.

For all reference shapes and all deviants, the model computes the distance between the shapes by counting the number of symbolic properties on which the two shapes differ, and returns a list of 11x4 distances.

The threshold θ was fitted by maximizing the r^2 fit between the symbolic model and the behavioral data of French adults, Exp. 2. For the figures and the analyses, we used the value of 12.5%, but a good fit ($r^2 = .37$) was already obtained with θ=0, and any value between 3% and 20% yielded similar r^2 values (Fig. S5), indicating that our results do not hinge on a particular choice of behavioral tolerance threshold but rather on any reasonable ability to detect similarity lengths and angles.

Definition of the neural network model and its variants

We used the CORnet neural network, variant S, whose architecture is schematically depicted in figure 4B. We used the weights made available by the authors of (3) after training on the ImageNet-1000 dataset, where the task of the network was to assign each image of the dataset a label among 1000 possible categories . We did not modify the network or weights, but simply retrieved the activity of units in the internal layers (roughly matching brain areas V1, V2, V4 and IT). To simulate a behavioral trial, we fed the six shapes separately to the network, and retrieved the six vector outputs of the penultimate layer, corresponding to inferotemporal cortex (IT) and which yielded the best performance (Fig. S8 shows the predictions when other layers are used). We considered the vector most distant from the average of the others to be the outlying shape, and repeated this process 10000 times to approximate the error rate of the network. We also report the performance obtained from layers V1, V2 and V4, as well as that obtained by simply picking the outlier on dimensions such as the perimeter or area. The same procedure was repeated using two other top-scoring networks of brain-score.org: DenseNet and ResNet (see Fig. S7).

Variational auto-encoder (VAE) model

For the VAE, we used PyTorch (4)'s off-the-shelf implementation of the canonical model (5) (ReLUs and the adam optimizer replaced of sigmoids and adagrad, as recommended by PyTorch's implementation to make the network converge faster.) For each of the 11 reference shape, we generated 6 rotated times 6 scaled images of size 24x24. These 36 images were randomly split in a training set and a testing set, both of size 18. The VAE was then trained over the course of 150 epochs to minimize the loss on the training set, with an evaluation on the testing set at each epoch (Fig. S9A shows the loss on the testing set across epochs for each shape). This gave us access to the VAE's performance across the course of learning for each shape (details in Fig. S9B) and we correlated the performance for each shape with the behavior of both humans (exp.2) and baboons in Fig. S9C. To make the comparison with CNNs more straightforward, for each of our shapes (references and deviants), we extracted the output of the innermost layers of the fully trained VAE, the latent mean and the latent standard deviation layer, from which we replicated the methodology using with the CNNs in order to simulating behavioral outlier detection. The results are summarized in Fig. S9D: overall, the output of the innermost layers varied very little across shapes, and those variations did not capture the variance of either any of the human population, or any of the baboons.

Additional analyses, results and discussions

Detailed analysis of the visual search experiment

The error rates and mean response times of the visual search experiment were entered into an ANOVA with shapes (11-level factor), number of items (as a numerical factor in 6, 12 or 24), target presence (present or absent), and their interaction, and participant as the random factor. For error rates, there was a significant effect of shape (F(10,100) = 16.15, $p < .0001$), of number of items (F(1,10) = 20.33, p = .0011), of target presence (F(1,10) = 31.45, p = .0002), of shape and target presence (F(10, 100) = 2.03, p = .0375), but little interaction between number of items and target presence (F(1,10) = 4.91, p = .0509), no significant interaction between shape and number of items (F(10,100) = 0.87, p = .564), nor a three-way interaction between shape, number of items and target presence (F(10,100) = .42, p = .93). For response times, there was a significant effect of shape (F(10,100) = 9.89, $p < .0001$), of number of items (F(1,10) = 26.70, p = .0004), of target presence (F(1,10) = 29.58, p = .0003), of the interaction between shape and number of items (F(10,100) = 3.71, p = .0003), but no significant interaction between number of items and target presence (F(1,10) = 3.78, p = .080), no significant interaction between shape and target presence (F(10, 100) = 0.90, p = .54) nor a three-way interaction between shape, number of items and target presence (F(10,100) = .52, p = .88).

The error rates closely followed the classical geometric regularity effect observed in the intruder task, as there was a significant correlation between the mean error rates in visual search and the French adults error rates in the intruder task (experiment 2), both overall (R^2=.98, $p < 0.0001$, Fig. 2B) and regardless of the number of items on the screen (6 items, R^2 = .86, $p < .0001$, 12 items R^2 = .93, $p < .0001$, 24 shapes R^2 = .96, $p < .0001$). The mean RTs also followed a geometric regularity effect overall (R^2 = 0.88, $p < 0.0001$) and for each number of items (6 items, R^2 = .90, $p < .0001$, 12 items R^2 = .89, $p < .0001$, 24 shapes R^2 = .85, $p < .0001$).

To test for the seriality of visual search, the mean response time within each subject was entered in separate ANOVAs for each shape, with number of items (a numerical factor equal to 6, 12 or 24), target presence (present or absent), and their interaction as factors, and participants as a random factor. All shapes elicited a serial visual search (all $p < 0.05$ for the effect of the number of items; Fig. 2C).

For each shape and participants, we computed the slope of the visual search for both present and absent condition by fitting a linear model on the median of the response times per item number. We then tested whether the slope of the visual search in the “absent” condition was twice the slope of the “present” condition, as expected from serial search (6). For each shape, we used a paired Student’s test to compare, across subjects, the distribution of slopes in the absent condition and the distribution of twice the slope in the present condition. None of those differences except for one shape were significant at the .05 level (right-kite: p = .044; all other shapes $p > 0.05$). Additionally, the best fit of a linear model across subjects that predicts the slope, as computed above, when the item is absent from the slope when it is present had a significant (p = .0003) coefficient of 1.66, SE = .30, not significantly different from 2 (p = .29).

Finally, the slope of the visual search exhibited a geometric regularity effect: it correlated with the error rates observed in experiment 2, both overall ($R^2 = .70$, $p = .0013$), and when the target was present ($R^2 = .60$, $p = .0047$; Fig. 2C) and absent ($R^2 = .68$, $p = .0019$).

Possible role of feedback in human and non-human primates

It could be argued that, in the intruder task, human subjects were treated differently from baboons because on error trials, the visual feedback the correct responses was highlighted in green (surrounded with a green square for training images, filled in green for geometric shapes), thus giving an additional indication about the task. To examine whether this made any difference in humans, we analyzed the data from each participant's very first trial with a given shape, before they received any feedback. In both experiments 1 and 2, such analysis produced results that were indistinguishable from the results of the full dataset analysis. The error rates were strongly correlated with those of the full dataset (exp. 1: $r^2 = .99$, $p < .0001$; exp. 2: $r^2 = .94$, $p < .0001$); the best fit of a linear regression "full data ~ $\beta_0 + \beta_1$ * first_trial" had an intercept β_0 not significantly different from 0 and a slope β_1 not significantly different from 1 (all p's > .1), suggesting that little or no learning took place in human participants over the course of the 88 trials.

Retraining of the neural networks with geometric shapes

A possible reason for the failure of neural networks to mimic human data could be that the geometric shapes differed from the network's training data (colored photographs). Perhaps our stimuli ended up on the extremities of the feature hyperspace, thus leading to inconsistent or chaotic behavior of the network. Here we present several arguments that mitigate this possibility. First, the labels that were attributed to the shapes were highly consistent and suggested that the network did recognize them. Table S2 provides details of the labels given by CorNet without retraining. The network overwhelmingly categorized the shapes as "envelopes," and its next choices were mostly "Band-Aids" or "binders", with a few interesting deviations (e.g. trapezoids were classified as "lampshades"). This result was replicated almost perfectly with DenseNet, while ResNet primarily categorized the shapes as envelopes, followed by noisier categories. Second, the three convolutional neural networks we tested were highly consistent in the error rates that they predicted; and, as showing in figures 4C, 4D and S7, these predictions were not random, but tightly correlated with baboon behavior.

Third, we examined how CorNet would perform if it received additional training with geometric shapes (similar perhaps to a young child being exposed to geometric shapes and toys). Our results are summarized in Fig. S8. We trained different versions of CorNet to categorize either all of our 11 shapes ("All shapes"), or the subset of 5 shapes that have a common name in English ("Nameable shapes", i.e. square, rectangle, rhombus, parallelogram and trapezoid).

Our goal was to keep the properties that made the original network successful in image recognition, but also familiarize it with our shape space. We proceeded as follows: (i) we added either 5 or 11 output unit to the output (decoder) units; those were fully connected to the previous layer and randomly initialized, while keeping the rest of the network intact; (ii) we trained the network to categorize solely our shapes (solid white on black images, one shape per image, same rotation and scaling factors as for behavioral experiments), and allowed the backpropagation to modify either the entire network ("All layers"), or only the last main group of layers ("IT only"), with training on 80% of the images per shape and validation on the remaining 20% (plotted on Fig. S8A); the learning optimizer was Adam with a learning rate of 1.0E-6; (iii) we checked, for each training step, the performance of the updated network on the original dataset, ImageNet. After sufficient training, all conditions lead to perfect categorization of all geometric shapes, including on the validation set of shapes. Meanwhile, performance on ImageNet remained high, with a higher loss when the entire network was allowed to change in order to accommodate the new geometric shapes (Fig. S8A); (iv) finally, using our multiple-regression methodology, we compared the predictive power of each of the four types of retrained network with that of our symbolic model.

The results appear in Fig. S8B. None of the four training schemes significantly improved the predictive power of the neural network model on human participants. As for baboons, the various training conditions either did not change anything or worsened the predictive power.

Possible effect of non-matched visual properties

We matched our 11 shapes on several important size variables (see the section on "Stimuli" above). However, those constraints imposed that we could not match them for other visual properties. In particular, the shapes were not strictly equalized in area and perimeter (see table S1). Given the random scaling we added to each of the six shapes, choosing the outlier based on area or perimeter could not give rise to the high level of performance observed in humans. Furthermore, although the error rate predicted by such strategies varied across shapes, regressions indicated it could not explain the geometric regularity effect observed in humans (all p's > .05). In Fig. S6 we show the predictions and correlation with baboons: the area-based strategy significantly correlates with the observed behavior in baboons ($p < .0001$) while the perimeter-based strategy does not ($p = .5$). Both strategies elicit more errors than the baboons, indicating that these strategies do not suffice to explain the baboons' behavior.

Possible effect of education in Himba participants

The Himba population we sampled was heterogeneous in its formal education background. Out of the 44 participants we tested, 22 never attended school (those subjects are reported in the main text), and the 22 others ranged from 1 to 8 years of school, with otherwise comparable general demographic information.

This variability provided an opportunity to test for the effect of the number of years of schooling on the geometric regularity effect. The error rates were entered into an ANOVA with geometric regularity (a numerical factor determined by the error rate in French subjects in experiment 2), years of schooling (a numerical factor ranging from 0 to 8), their interaction and participants as random factors. There was a significant effect of shapes ($F(1, 42) = 229.21$, p <

.0001) but no significant effect of the years of schooling (F(1, 42) = 0.05, p = .82) and no significant interaction (F(1,42) = .65, p = 0.43). This negative finding does not exclude that, with more participants, an effect of education would be observed. However, this additional analysis confirms the universality of the geometric regularity effect.

Possible impact of a "carpentered world"

The Western environment has been called a "carpentered world" (7), where vision is bombarded with many rectilinear objects (e.g. buildings, tables, books, etc.). Could such a difference in the statistics of the environment explain the geometric regularity effect? We believe that this is unlikely for several reasons explained in the discussion part of the main text. The main reason is that we replicated the effect in the Himba, but failed to observed it in baboons. The rural settlements of the Himba are quite unlike industrialized societies and their environment is relatively free of rectilinear objects (for photographs, see e.g. https://en.wikipedia.org/wiki/Himba_people). Conversely, the baboons were not wild animals, but grew up and lived in an environment comprised of both natural objects (trees, rocks) and man-made, rectilinear objects (buildings, doors, testing booths, computer screens… see inset picture). Arguably, the baboon's environment is equally or even more "carpentered" than the Himbas (see photo).

Convention for plots

When error bars are presented on a graphics, they show the standard error computed over all relevant data points without intermediary averaging at the participant level. It follows that they represent the confidence over the accuracy of given value rather than the variance across subjects.

Supplementary Figures and Legends

Fig. S1. Correlation between error rates and response times in adult experiments 1 and 2.

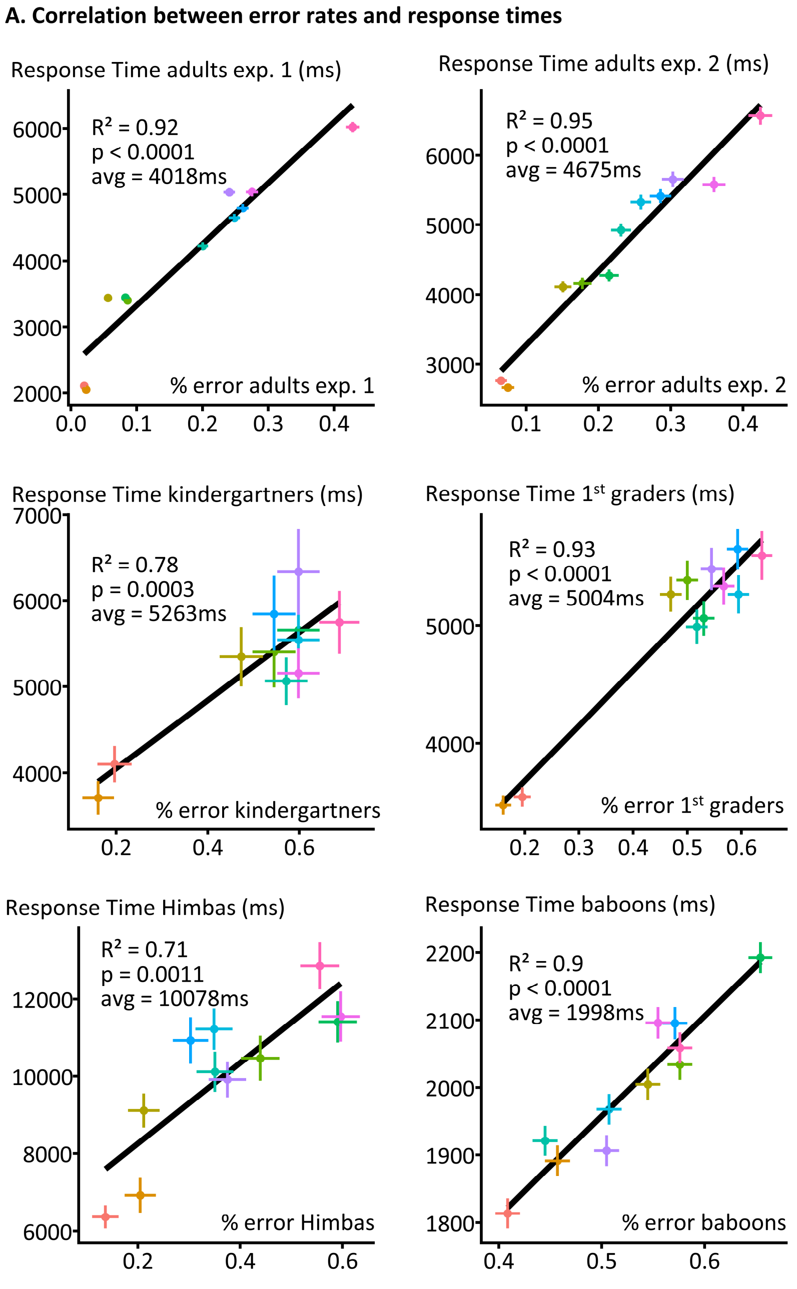


Figure S1

Correlation between averaged participants’ error rate (x axis) and response times in milliseconds (y axis) across all 11 shapes for each test group. From left to right, from top to bottom: French Adults exp. 1, then exp. 2, then kindergartners and 1st graders, then Himbas, and finally baboons.

Fig. S2. Detailed results of the geometric intruder tests in children.

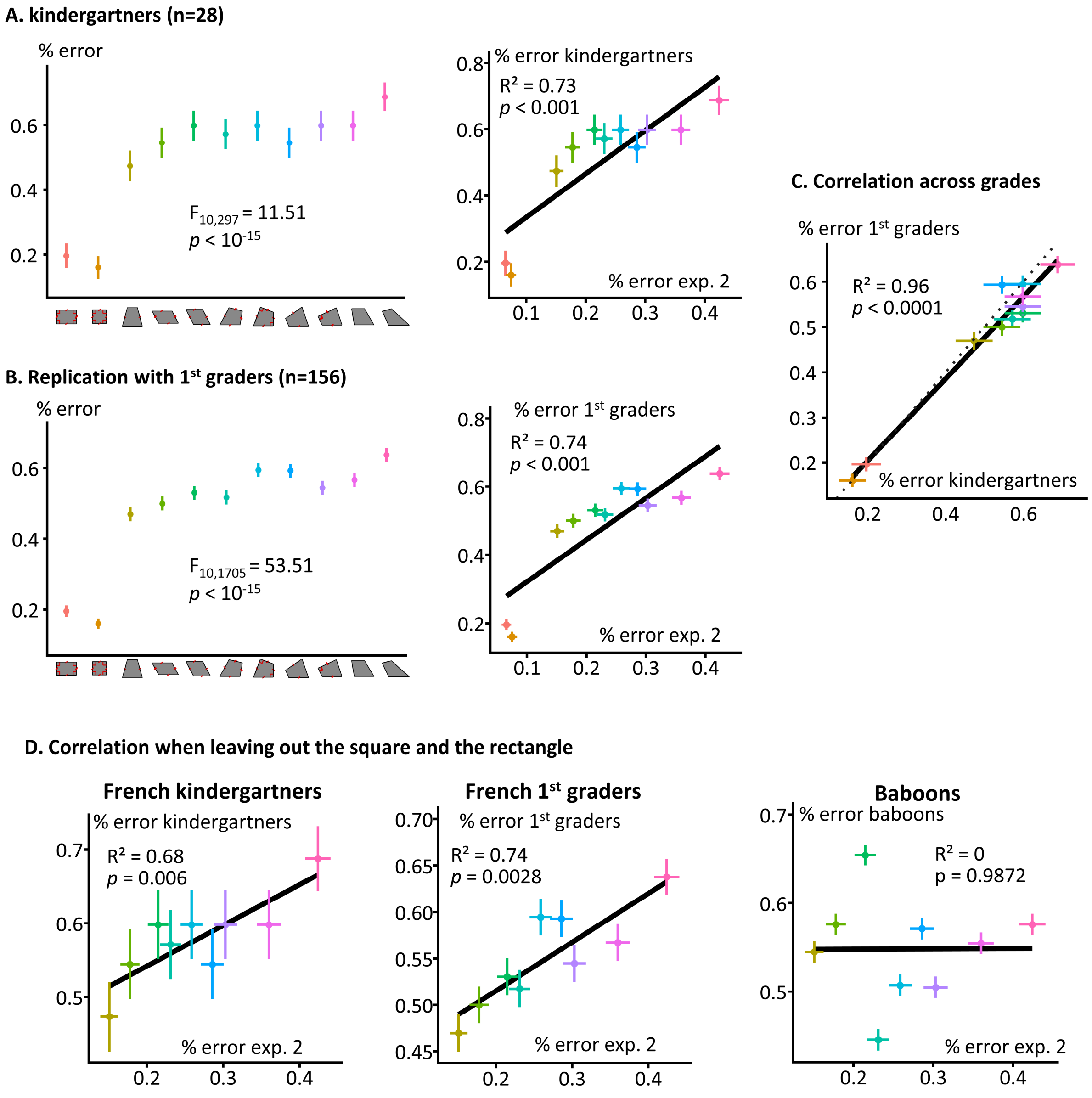


Figure S2

A, Kindergartner study and comparison with adults. Left: Main effect of quadrilaterals on performance in the intruder task. **Right:** Correlation between French kindergartners and French adults. Colors match the left plot and indicate the shape. **B, 1st graders study and comparison with adults**. **C, Comparison between kindergartners and 1st graders.** The dotted line indicates a slope of 1 while the solid line indicates the best fit (slope = .91, SE = .06). **D, Geometric regularity effects after exclusion of square and rectangle.** Although the data from kindergartners and 1st graders suggested that the square and rectangle shapes were outliers, their performance continued to exhibit a geometric regularity effect and remained correlated with that of French adults even when square and rectangle shapes were excluded from the analysis. In baboons, by contrast, the correlation remained nonsignificant.

Fig. S3. Details of the training and testing in each baboon

Each graph shows the average error rate as a function of the number of trials that the animal took, split according to the different phases of the training and testing (as defined in Fig. 3). Each line corresponds to a baboon: the first 20 lines show all animals that produced data in the final test of geometric figures, and the last 6 rows show all animals that dropped at various stages of training. The x-axis is a logarithmic axis (Log10), so that generalization blocks (which typically contain far fewer trials) can be seen. When a plot is missing, it means that the baboon did not take that particular block. Baboons with names in bold pursued the task until after block 81 and were therefore included in the main analyses.

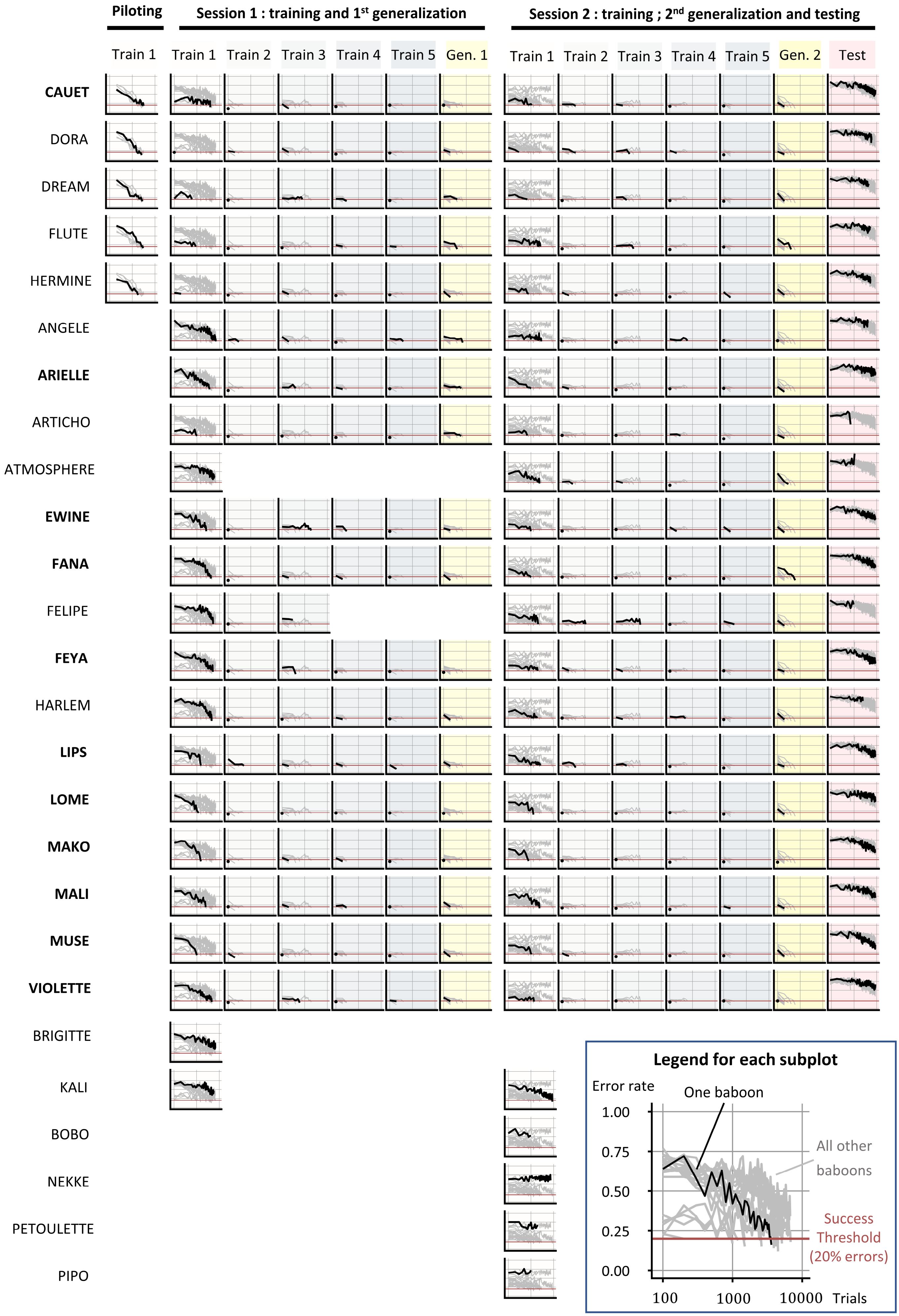

Piloting
Session 1 : training and 1st generalization
Session 2 : training ; 2nd generalization and testing
Train 1
Train 1
Train 2
Train 3
Train 4
Train 5
Gen. 1
Train 1
Train 2
Train 3
Train 4
Train 5
Gen. 2
Test
CAUET
DORA
DREAM
FLUTE
HERMINE
ANGELE
ARIELLE
ARTICHO
ATMOSPHERE
EWINE
FANA
FELIPE
FEYA
HARLEM
LIPS
LOME
MAKO
MALI
MUSE
VIOLETTE
BRIGITTE
KALI
BOBO
NEKKE
PETOULETTE
PIPO
Legend for each subplot
Error rate
One baboon
1.00
0.75
0.50
0.25
0.00
All other baboons
Success Threshold (20% errors)
100
1000
10000
Trials

Figure S3

Fig. S4. Consistency across baboons and across different training periods.

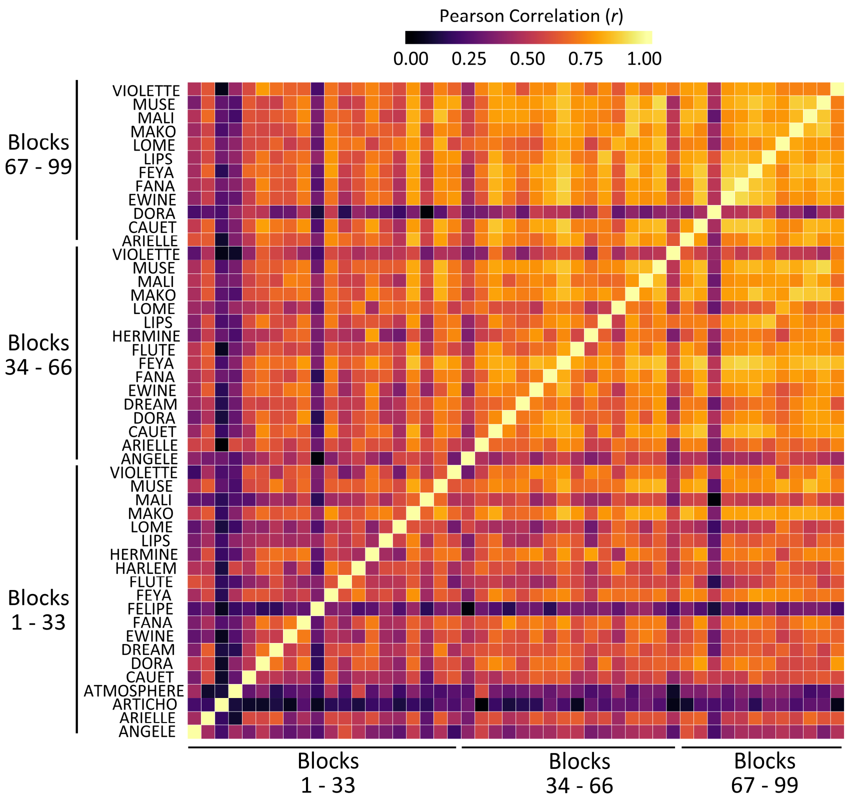


Figure S4

Cross-correlation matrix of the performance of each individual baboon over the course of testing, across 44 data points (11 shapes X 4 deviant types).

Fig. S5. Influence of the tolerance parameter on the symbolic model.

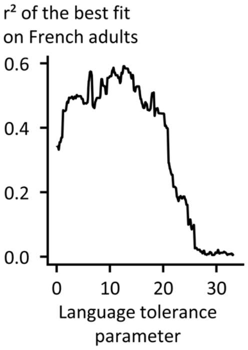


Figure S5

Correlation (R^2) between the behavioral data of Experiment 2 (with French adults) and the predictions of the symbolic model, as a function of the tolerance threshold for accepting two sides or two angles as approximately equal. Any tolerance threshold between ~3% and ~20% yielded roughly similar fit, indicating that the model is robust to the exact choice of its only free parameter.

Fig. S6. Detailed predictions arising from various models of visual perception

Each row displays the prediction from a given model of visual perception, with the predicted error rates across shapes (left; displayed over the data from baboons in dark gray and humans in light gray) and the correlation with the aggregate of baboons' data after the 80th trial (right). The first four rows show the prediction of each major layer of CorNet in order (V1, V2, V4, and IT used throughout this document), followed by a model that picks the shape with area most distant from the average of the other shape's area, and an equivalent model with the perimeter. All reach significant levels at the $p < .05$ levels except the perimeter, and the R^2 increase with the layers in CorNet.

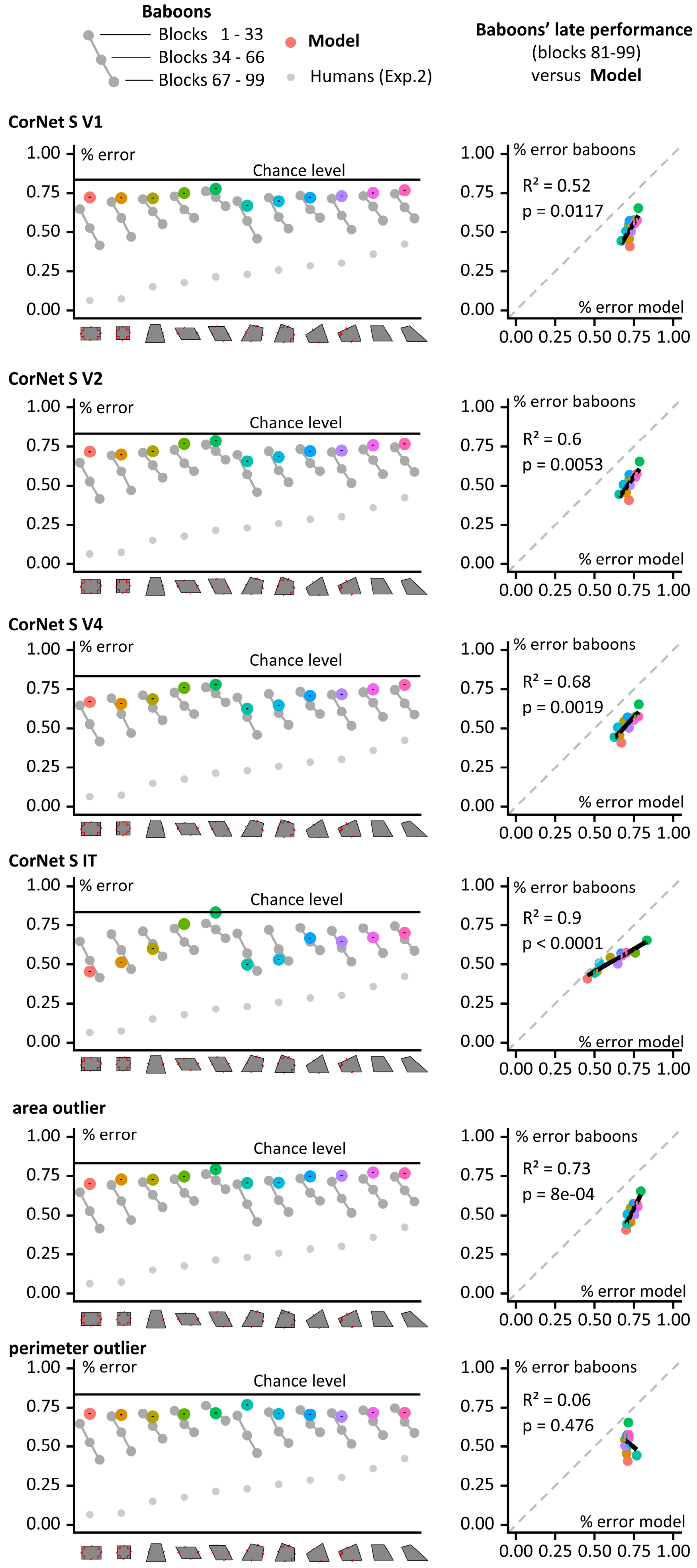

Baboons
Blocks 1 - 33
Blocks 34 - 66
Blocks 67 - 99
Model
Humans (Exp.2)
Baboons' late performance
(blocks 81-99)
versus Model
CorNet S V1
% error
Chance level
1.00
0.75
0.50
0.25
0.00
% error baboons
R² = 0.52
p = 0.0117
% error model
0.00 0.25 0.50 0.75 1.00
CorNet S V2
R² = 0.6
p = 0.0053
CorNet S V4
R² = 0.68
p = 0.0019
CorNet S IT
R² = 0.9
p < 0.0001
area outlier
R² = 0.73
p = 8e-04
perimeter outlier
R² = 0.06
p = 0.476

Figure S6

Fig. S7. Replication of analyses with two other classical neural-network models of image recognition

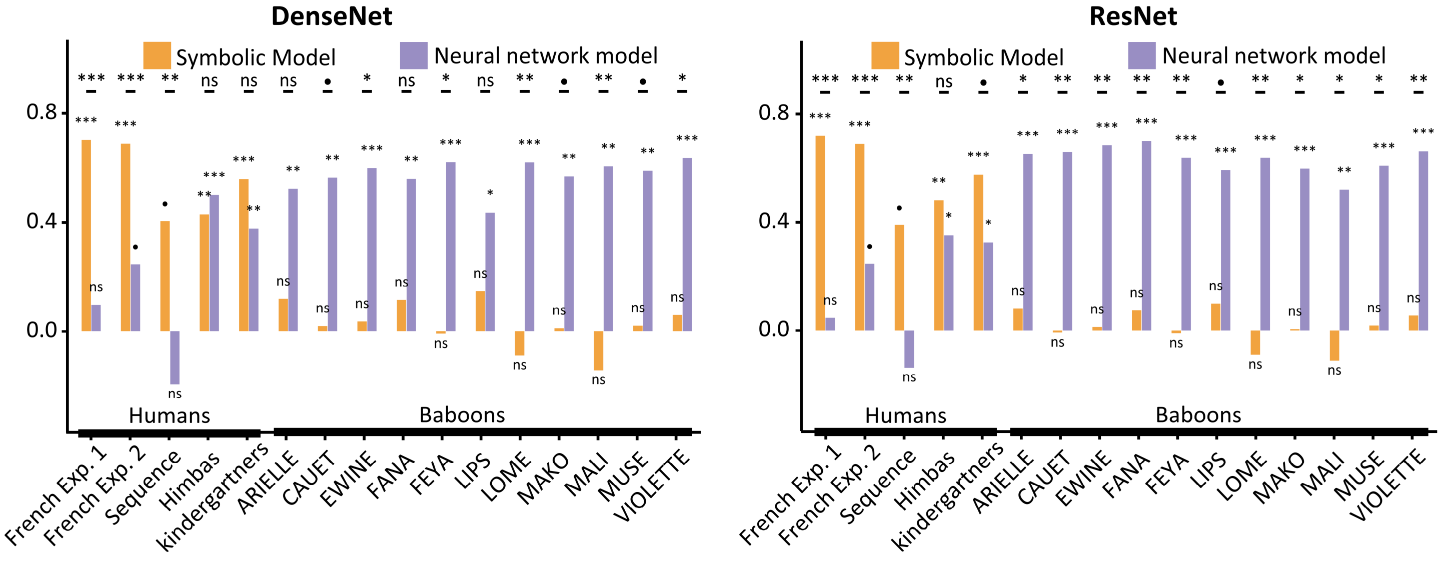
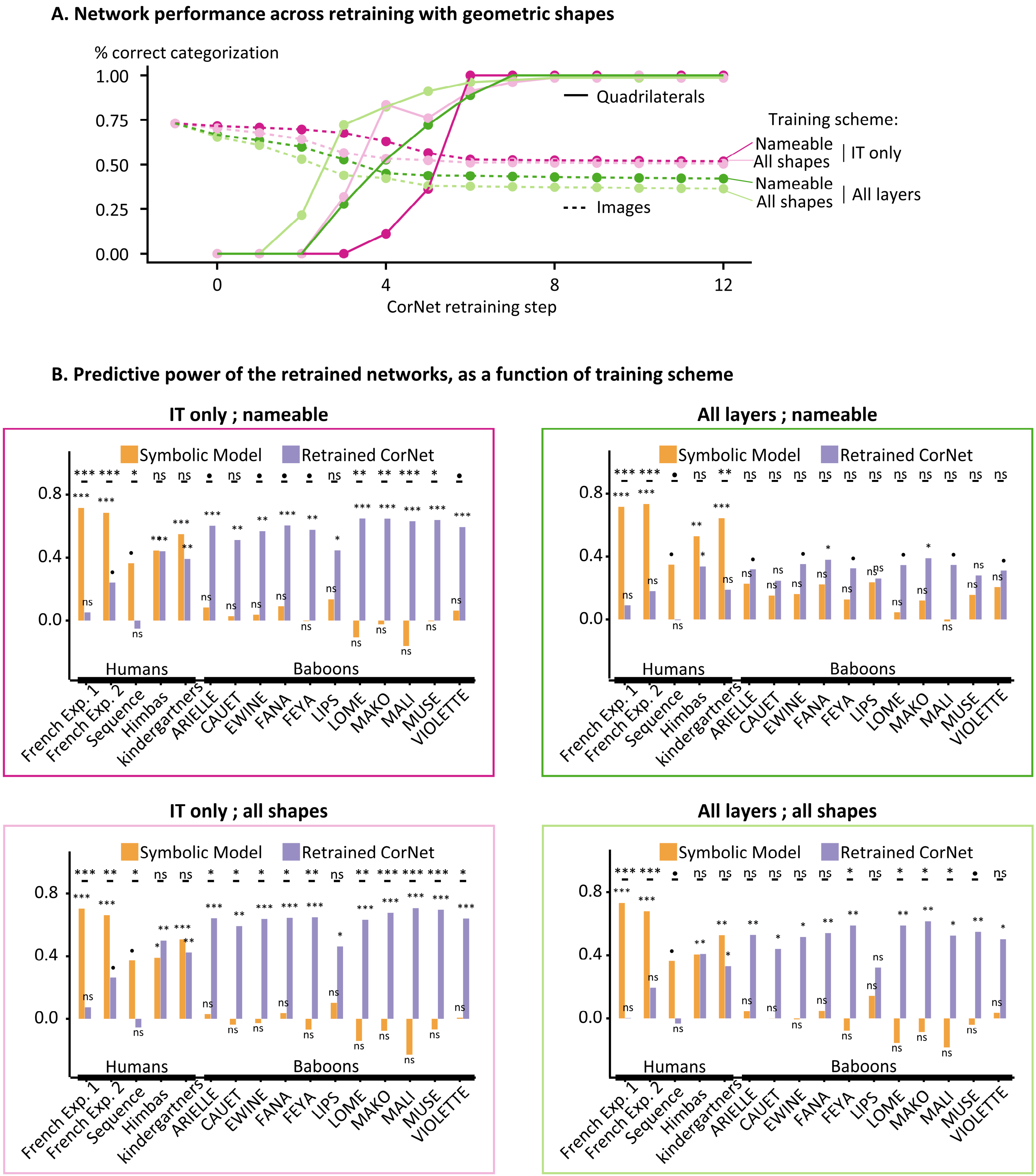


Figure S7

The figure shows the standardized regression weights (beta) of a multiple regression of the average performance from various human and non-human primate groups across 44 data points (11 shapes X 4 deviant types), using the symbolic and neural-network models as predictors. Stars indicate significance level (●, p<.05; *, p<.01; **, p<.001; ***, p<.0001). **Left,** using the output of the penultimate layer of densenet196 pretrained on ImageNet. **Right,** using the output of the penultimate layer of resnet101 pretrained on ImageNet.

Fig. S8. Effect of training the CorNet neural network model on geometric shapes.

A, Evolution of network performance across different retraining schemes. We started with CorNet-S trained on ImageNet and retrained it by adding new output (decoder) units for geometric shapes and presenting it with only quadrilaterals for 13 epochs. For each epoch, we tested the network on new unseen views of the quadrilaterals (solid lines) and on images from ImageNet (dashed lines). We studied the effects of 4 different training schemes, defined by (1) retraining either on all 11 shapes (darker colors), or only on a subset of 5 nameable shapes (rectangle, square, rhombus, parallelogram, trapezoid; lighter colors), and (2) either freezing all layers but the penultimate one, corresponding to inferotemporal cortex IT (green), or backpropagating the error through the entire network (pink). **B, Correlation with experimentally observed performance.** Same format as Fig. 4 in the main text. The figure shows the standardized regression weights (beta) of a multiple regression of the average performance from various human and non-human primate groups across 44 data points (11 shapes X 4 deviant types), using the symbolic and neural-network models as predictors. Stars indicate significance level (●, $p<.05$; *, $p<.01$; **, $p<.001$; ***, $p<.0001$). Each subplot corresponds to a specific training scheme, with color-matching panel A.



A. Network performance across retraining with geometric shapes

B. Predictive power of the retrained networks, as a function of training scheme

IT only ; nameable

All layers ; nameable

IT only ; all shapes

All layers ; all shapes

Figure S8

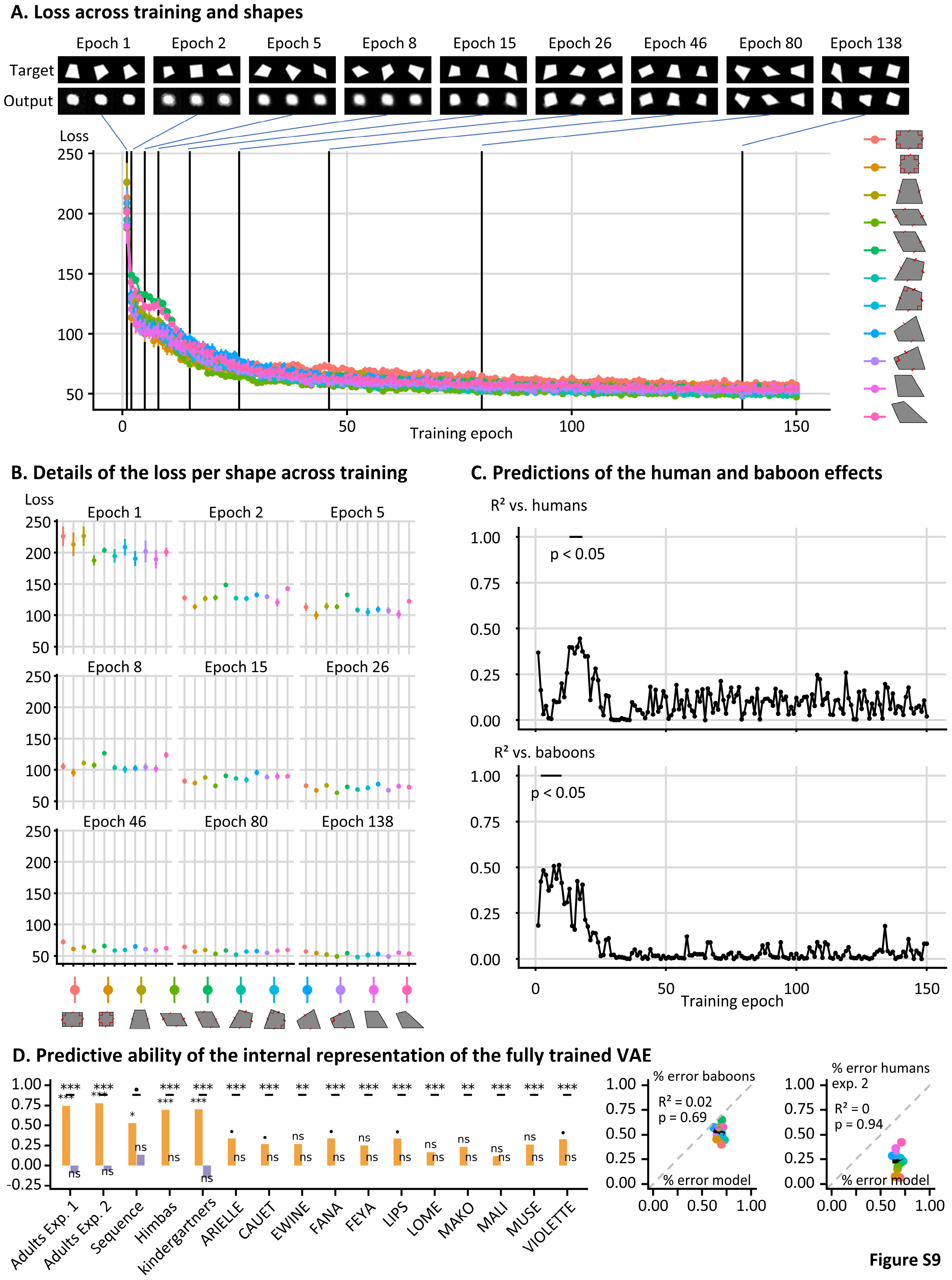
Fig. S9. Variational Auto-Encoder performances.

A, Evolution of network performance across shapes and training epochs. We trained PyTorch's off the shelf VAE to produce all of our reference shapes in six possible orientations and scaling. This plot shows the network's loss on the testing dataset across training epoch for each shape – 50% of each shape was set aside for testing and the network was never trained onthese shapes. At the top of the graph, exemplars of the target shape, and the network's output, are produced, to show that the network does reproduces some fine-grained elements of the shapes, and does not just approximate a single shape that would minimize distances for all of our target shapes.

B, Details of the loss per shape across training. At exponentially spaced epochs, detail of the loss (y axis) for each refence shape (x axis).

C, Prediction of the human and baboon effect. 11 points pearson's R^2 of the correlation between the loss across shapes and the average error rates for humans (top) and baboons (bottom).

D, Predictive ability of the internal representation of the fully trained VAE. Left: Standardized regression weights (beta) in a multiple regression of the data from various human and non-human primate groups across 44 data points (11 shapes X 4 outlier types) using the symbolic and VAE models as predictors. Stars indicate significance level (●, p<.05; *, p<.01; **, p<.001; ***, p<.0001). Right: detail of the correlation with the behavior from baboons and humans (exp. 2)



A. Loss across training and shapes

B. Details of the loss per shape across training

C. Predictions of the human and baboon effects

D. Predictive ability of the internal representation of the fully trained VAE

Figure S9

Table S1. Precise definition of the 11 shapes

Shape	topLeft x	topLeft y	topRight x	topRight y	botRight x	botRight y	Avg pairs	Perimeter	Area	Number of properties
rectangle	0	1	1.5	1	1.5	0	1.434	1	1	15
square	0	1.26	1.26	1.26	1.26	0	1.434	1.008	1.059	19
iso-trapezoid	0.365	1.362	1.109	1.362	1.5	0	1.433	1.014	1.019	5
parallelogram	-0.517	0.896	0.983	0.896	1.5	0	1.434	1.014	0.896	7
rhombus	-0.908	0.931	0.392	0.931	1.3	0	1.434	1.04	0.807	9
kite	0.766	1.29	1.77	1.007	1.5	0	1.434	1.017	1.007	5
right-kite	0.529	1.404	1.5	1.038	1.5	0	1.434	1.015	1.038	7
hinge	-0.248	0.533	0.98	1.393	1.5	0	1.434	1.015	0.986	1
right-hinge	-0.296	0.634	1.064	1.268	1.5	0	1.434	1.008	0.984	2
trapezoid	-0.227	1.2	0.724	1.2	1.5	0	1.434	1.02	0.98	1
Irregular	-0.45	1.058	0.227	1.24	1.5	0	1.434	1.025	0.885	0

For reproducibility we provide here the precise coordinates of the vectors defining the three corners of each shape. With the bottom left vertex at coordinates (0,0), the first six columns define the three vectors required to locate the top-left, top-right and bottom-right vertices. When presented, the reference orientation (0°) of each shape was the one where the top edge was horizontal, around which the random orientations (-25°, -15°, -5°, 5°, 15°, 25°) took place. All bottom-right vectors, except for the square and the rhombus, are matched for length. The "Avg pairs" column gives the average distance between all pairs of points, another metric matched across shapes. The Perimeter and Area columns give respectively the perimeter and area relative to that of the rectangle: for lack of enough degrees of freedom, these properties are not matched across shapes. Neither explain the human behavior (area: p = .32, perimeter: p = .13) or the symbolic model (area: p = .28, perimeter p = .14). See additional discussion of this in the Additional Analysis section.

Table S2. Significance and effect size of various predictors for each population

	Shape			Outlier Pos			Outlier Type			Outlier Scale			Outlier Rotation			Symbolic Model	
	F , p , η^2G	p	η^2G	F	p	η^2G	F	p	η^2G	F	p	η^2G	F	p	η^2G	p	r^2
French Adults 1	F(10, 6040) = 292.88	**<0.01**	0.3	F(5, 3020) = 4.96	**<0.01**	<0.01	F(3, 1812) = 114.09	**<0.01**	0.1	F(5, 3020) = 4.46	**<0.01**	<0.01	F(5, 3020) = 21.19	**<0.01**	0	**<0.01**	0.537
French Adults 2	F(10, 1160) = 70.96	**<0.01**	0.3	F(5, 580) = 2.26	0.05	<0.01	F(3, 348) = 53.60	**<0.01**	0.1	F(5, 580) = 2.16	0.06	<0.01	F(5, 580) = 9.66	**<0.01**	0	**<0.01**	0.591
Himbas	F(10, 210) = 19.61	**<0.01**	0.4	F(5, 105) = 0.32	0.9	<0.01	F(3, 63) = 10.99	**<0.01**	0.1	F(5, 105) = 2.07	0.07	0.04	F(5, 105) = 1.77	0.13	0	**<0.01**	0.351
Preschoolers	F(10, 270) = 14.90	**<0.01**	0.3	F(5, 135) = 1.92	0.1	0.04	F(3, 81) = 12.03	**<0.01**	0.2	F(5, 130) = 2.47	0.04	0.05	F(5, 135) = 0.75	0.59	0	**<0.01**	0.463
1st graders	F(10, 1550) = 76.93	**<0.01**	0.2	F(5, 775) = 3.51	**<0.01**	0.01	F(3, 465) = 53.38	**<0.01**	0.1	F(5, 775) = 9.60	**<0.01**	0.03	F(5, 775) = 8.38	**<0.01**	0	**<0.01**	0.514
baboons	F(10, 100) = 24.68	**<0.01**	0.4	F(5, 50) = 3.50	**<0.01**	0.08	F(3, 30) = 102.97	**<0.01**	0.6	F(5, 50) = 2.98	0.02	0.05	F(5, 50) = 44.82	**<0.01**	0.4	0.12	0.0568

For each tested population, we ran five separate ANOVAs to measure the significance and effect size on performance of five different aspects of the stimuli: geometric shape (11 shapes), position of the outlier (6 positions), type of outlier (4 types of deviants, as defined in figure 1), scale of the outlier (6 scale changes), and rotation of the outlier (6 angles). The table reports, for each ANOVA, the p-value and generalized eta-squared value (proportion of variance accounted for). On all human populations, there was a main effect of the shape (i.e. the geometric regularity effect), and a significant but smaller effect of outlier type. Other predictors were either not significant or had extremely small effect size. By contrast, three variables impact baboons' behavior: the shape, the type of outlier, and the rotation of the outlier. The shape effect (different from the human geometric regularity effect) is described in the main text. As for the outlier type and rotation effects, baboons fared better on trials where the deviants were smaller due to an inward displacement of the bottom right vertex, and fared better when the outlier was maximally rotated in one direction or the other.

Table S3. Effects of geometric properties on participant's errors

term	estimate	std.error	statistic	df	p.value
Intercept	0.44	0.01	29.54	332.51	<10e-8
right-angle	-0.03	0.01	-3.33	1166	0.00091
parallels	-0.1	0.01	-11.09	1166	<10e-8
symmetry	-0.07	0.01	-5.94	1166	<10e-8
equal-sides	-0.13	0.02	-8.7	1166	<10e-8

To quantify the contribution of each geometric property to our symbolic model, we ran a mixed-effect linear regression on the data from our French adult experiment 2. The model predicted the error rate of participants on 11 shapes, given the presence or absence of exact property in each shape, with participants as a random effect. The intercept corresponds to the predicted error rate for a shape without any regularity (44%), and each additional property significantly improves the prediction of the performance of participants. Equal sides had the greatest impact (13% gain overall), followed by parallelism (10%), symmetry (7%), and finally right angles (3%).

Table S4. Labels attributed by a convolutional neural network (CorNet) to the 11 geometric shapes used in our experiments.

Shape	label1	label2	label3	label4	label5
Rectangle	envelope, 71.68%	band aid, 7.22%	band aid, 2.02%	spatula, 2.28%	letter opener, 1.61%
Square	envelope, 74.62%	envelope, 33.07%	switch, 2.11%	book jacket, 1.66%	face powder, 1.57%
iso-trapezoid	envelope, 62.05%	envelope, 37.8%	lampshade, 6.98%	lampshade, 4.51%	binder, 2.04%
Parallelogram	envelope, 64.91%	band aid, 7.64%	cleaver, 3.75%	binder, 2.7%	table lamp, 2.57%
Rhombus	envelope, 58.4%	band aid, 9%	letter opener, 4.61%	book jacket, 2.8%	wing, 3.96%
Kite	envelope, 68.63%	band aid, 10.18%	binder, 2.04%	carton, 1.64%	switch, 1.68%
right-kite	envelope, 74.28%	band aid, 7.51%	face powder, 2.25%	binder, 1.8%	binder, 1.89%
Hinge	envelope, 77.49%	band aid, 3.91%	table lamp, 2.08%	band aid, 1.58%	cleaver, 1.95%
right-hinge	envelope, 77.09%	band aid, 4.69%	letter opener, 2.38%	binder, 1.71%	table lamp, 1.55%
Trapezoid	envelope, 72.57%	band aid, 8.41%	binder, 2.62%	face powder, 2.46%	face powder, 1.63%
Irregular	envelope, 59.55%	envelope, 28.07%	binder, 3.3%	carton, 2.6%	table lamp, 2.02%

For each shape, columns show the first five top predictions and the associated average confidence level, for CorNet trained on ImageNet. Each shape was presented in 36 slightly different variants (6 rotations X 6 scaling factors). We averaged these predictions for each shape, and put in each column the prediction whose average associated confidence level was the highest, and the corresponding average confidence level.

References cited in Appendix SI

1. J. R. de Leeuw, jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behav Res* **47**, 1–12 (2015).
2. J. Fagot, E. Bonté, Automated testing of cognitive performance in monkeys: Use of a battery of computerized test systems by a troop of semi-free-ranging baboons (Papio papio). *Behavior Research Methods* **42**, 507–516 (2010).
3. J. Kubilius, *et al.*, "CORnet: Modeling the Neural Mechanisms of Core Object Recognition" (Neuroscience, 2018) https:/doi.org/10.1101/408385 (February 8, 2020).
4. A. Paszke, *et al.*, "PyTorch: An Imperative Style, High-Performance Deep Learning Library" in *Advances in Neural Information Processing Systems 32*, H. Wallach, *et al.*, Eds. (Curran Associates, Inc., 2019), pp. 8024–8035.
5. D. P. Kingma, M. Welling, Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]* (2014) (October 26, 2020).
6. J. M. Wolfe, What Can 1 Million Trials Tell Us About Visual Search? *Psychol Sci* **9**, 33–39 (1998).
7. M. H. Segall, D. T. Campbell, M. J. Herskovits, Cultural Differences in the Perception of Geometric Illusions. *Science* **139**, 769–771 (1963).