Feature blindness: a challenge for understanding and modelling visual object recognition

Gaurav Malhotra*, Marin Dujmović, Jeffrey S Bowers

School of Psychological Sciences, University of Bristol,
12a Priory Rd, Bristol, BS8 1TU, UK

* gaurav.malhotra@bristol.ac.uk

Abstract

A central problem in vision sciences is to understand how humans recognise objects under novel viewing conditions. Recently, statistical inference models such as Convolutional Neural Networks (CNNs) seem to have reproduced this ability by incorporating some architectural constraints of biological vision systems into machine learning models. This has led to the proposal that, like CNNs, humans solve the problem of object recognition by performing a statistical inference over their observations. This hypothesis remains difficult to test as models and humans learn in vastly different environments. Accordingly, any differences in performance could be attributed to the training environment rather than reflect any fundamental difference between statistical inference models and human vision. To overcome these limitations, we conducted a series of experiments and simulations where humans and models had no prior experience with the stimuli. The stimuli contained multiple features that varied in the extent to which they predicted category membership. We observed that human participants frequently ignored features that were highly predictive and clearly visible. Instead, they learned to rely on global features such as colour or shape, even when these features were not the most predictive. When these features were absent they failed to learn the task entirely. By contrast, ideal inference models as well as CNNs always learned to categorise objects based on the most predictive feature. This was the case even when the CNN was pre-trained to have a shape-bias and the convolutional backbone was frozen. These results highlight a fundamental difference between statistical inference models and humans: while statistical inference models such as CNNs learn most diagnostic features with little regard for the computational cost of learning these features, humans are highly constrained by their limited cognitive capacities which results in a qualitatively different approach to object recognition.

Author summary

Any object consists of hundreds of visual features that can be used to recognise it. How do humans select which feature to use? Do we always choose features that are best at predicting the object? In a series of experiments using carefully designed stimuli, we find that humans frequently ignore many features that are clearly visible and highly predictive. This behaviour is statistically inefficient and we show that it contrasts with statistical inference models such as state-of-the-art neural networks. Unlike humans, these models learn to rely on the most predictive feature when trained on the same data. We argue that the reason underlying human behaviour may be a bias to look for

features that are less hungry for cognitive resources and generalise better to novel instances. This may be why human vision overly relies on global features, such as shape, and glosses over many other features that are perfectly diagnostic. Models that incorporate cognitive constraints may not only allow us to better understand human vision but also help us develop machine learning models that are more robust to changes in incidental features of objects.

Introduction

Sometimes we fail to see what's right in front of our eyes.

The seemingly simple task of recognising an object requires contending with a multitude of problems. Humans can recognise something as a "chair" for a vast range of lighting conditions, distances to the retina, viewing angles and contexts. We can recognise chairs made out of wood, metal, plastic and glass. Thus, to classify something as a chair, the brain must take the image of the object projected onto the retina and convert it into an internal representation that remains invariant under all these conditions [8]. A lot of effort in psychology, computational neuroscience and computer vision has gone into understanding how the brain constructs these invariant representations [7, 57].

One hypothesis is that the brain learns these invariant representations from the statistics of natural images [22, 7]. But till recently, it has proved challenging to construct scalable statistical inference models that learn directly from natural images and match human performance. A breakthrough has come in recent years from the field of artificial intelligence. Deep Convolutional Neural Networks (CNNs) are statistical inference models that are able to match, and in some cases exceed, human performance on some image categorisation tasks [34]. Like humans, these models show impressive generalisation to new images and to different translations, scales and viewpoints [23]. And like humans, this capacity to generalise seems to stem from the ability of Deep Networks to learn invariant internal representations [15]. It is also claimed that the learned representations in humans and networks are similar [30, 23, 47]. These results raise the exciting possibility that Deep Networks may finally provide a good model of human object recognition [28, 60, 6, 45] and perhaps even provide important insights into visual information processing in the primate brain [5, 63, 24, 46, 40].

Many reasons could be, and are, given for why CNNs have succeeded where previous models have failed [34, 2]. For example, it is often argued that CNNs excel in image classification because they incorporate a number of key insights from biological vision, including the hierarchical organization of the convolutional and pooling layers [33]. In addition, both systems are thought to implement optimisation frameworks, generating predictions by performing statistical inferences [64, 46]. Indeed, evidence suggests that humans perform some form of statistical optimisation for many cognitive tasks including language learning [49], spatial cognition [11], motor learning [27] and object perception [22]. Due to this architectural and computational overlap between the two systems it might seem reasonable to hypothesise that humans and CNNs end up with similar internal representations.

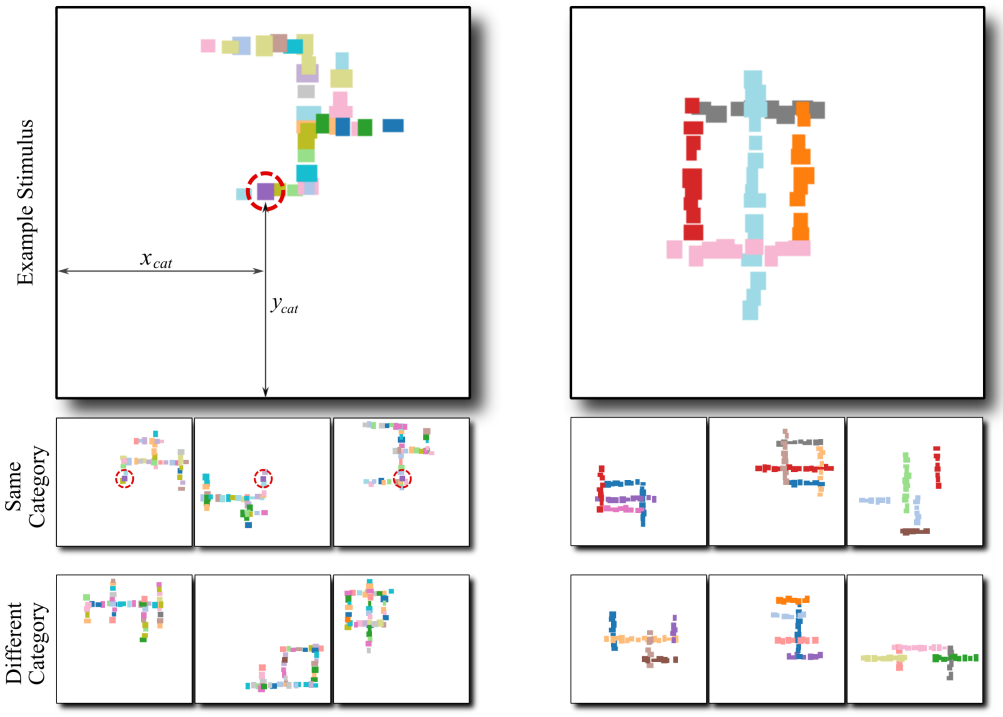
However, the parsimony and promise of this hypothesis is somewhat dampened by recent studies that have shown striking differences between CNNs and human vision. For example, CNNs are susceptible to small perturbations of images that are nearly invisible to the human eye [16, 43, 9]. They often classify images of objects based on statistical regularities in the background [58], or based on texture of objects [36, 12] and even based on single diagnostic pixels present within images [38]. That is, CNNs are prone to overfitting, often relying on predictive features that are idiosyncratic to the training set [13].

To what extent do these findings reflect fundamental differences between statistical inference models and human vision? On the one hand, such differences could be simply down to differences in training data [10, 21, 12]. While human beings have a lifetime of experience in recognising and interacting with 3D objects, CNNs are trained to classify images from fairly homogeneous training sets such as `ImageNet`. On the other hand, these differences may arise out of fundamental differences in resource constraints posed on the two systems. While both systems have the same learning objective, humans have to learn to classify objects under a limited set of cognitive and physiological resources, such as limited working memory [1], attention [62] and metabolic costs [52, 26]. So unlike standard CNNs, the computational time and cost is a key consideration for humans when choosing the stimulus feature for carrying out a task.

In this study, we explore the extent to which humans, like CNNs, are driven by performing statistical inferences when learning to categorise novel objects. To ensure that any difference in performance was not driven by familiarity with objects due to past experience, we developed a novel set of images, where each image contained multiple diagnostic features. These features predicted category membership of an image with a fixed probability. We conducted a series of experiments where we presented these novel images and their category labels to (i) human participants, (ii) an ideal inference model, and (iii) a CNN. After participants (humans or models) had observed a sequence of images and their category memberships, they were asked to predict the category of unseen images containing a subset of the diagnostic features. We used the ideal inference model to predict the best feature for making these decisions based on observations in an experiment. We then compared the features predicted by this model to the ones used by the CNN and humans. Our objective was to see whether human inferences match the statistical inferences predicted by the ideal inference model and the CNN.

We observed that (i) human participants frequently ignored features that were clearly visible and highly predictive and instead relied on reasonably diagnostic global features – such as overall shape or colour – when these features were present, (ii) when global feature were absent, participants struggled to learn some tasks entirely, even though they contained other highly predictive features, (iii) when multiple global features were concurrently present (e.g. overall shape as well as colour), participants frequently selected only one of the predictive features even though the optimal policy was to learn multiple features simultaneously, (iv) even when the relevant features were pointed out at the beginning of the experiment, participants still struggled to classify objects based on these features, highlighting that limitations in cognitive resources play a fundamental role in how humans learn the task, (v) in contrast, both statistical inference models placed a large emphasis on the most predictive feature, regardless of the computational resources required to learn the feature, and (vi) even when CNNs were trained to have a shape-bias [12, 20], this bias is lost as soon as they were trained on a new dataset with a different bias and CNNs learned whatever feature was most diagnostic.

These results highlight important differences in how human participants and statistical inference models learn to extract features from novel objects. Instead of an optimisation approach that underlies many machine learning models, we argue that human behaviour is much more in line with a *satisficing account* [53], where features are selected because they allow participants to perform reasonably on the task while taking into account their limited cognitive resources. While performing statistical inferences is certainly important, models of vision must also consider the cognitive costs and biases in order to be realistic theories of human object recognition.



(a) Patch location (b) Segment colour

Fig 1. Examples of training images from Experiments 1 and 2. Each stimulus contains one or two predictive features and was constructed by varying the colour, size and position of patches and segments. (a) Two features predict stimulus category: global shape and location (x_{cat}, y_{cat}) of one of the patches. For illustration, the predictive patch is circled. Stimuli in the same category (middle row, reduced size) have a patch with the same colour at the same location, while none of the stimuli in any other category (bottom row) have a patch at this location. (b) Global shape and colour of one of the segments predict stimulus category. Only stimuli in the same category (middle row) but not in any other category (bottom row) have a segment of this colour (red). The right-most stimulus in the middle row shows an example of a training image containing a non-shape feature (red segment) but no shape feature. For further illustration of stimuli used in these and other experiments, see Appendix E and Movies S1 and S2 in Supplementary Information.

Results

Behavioural tasks and Simulations

The behavioural tasks mimicked the process of learning object categorisation through supervised learning. In each experiment, participants were trained to categorise artificially generated images into one of five categories. Each image consisted of coloured patches that were organised into segments. These segments were, in turn, organised so that they appeared to form a solid structure. The exact location, size and colour of patches as well as segments varied from images to image, making each stimulus unique and avoiding any unintentional diagnostic features. See Figure 1 for some example images.

For each experiment, we constructed a dataset of images where one or more generative factors – *features* – predicted the category labels. In Experiments 1 to 4, images were drawn from datasets with two predictive features. One of these features

Table 1. Feature combinations examined in different experiments

Experiment	Features					% Shape
	Global Shape	Patch Location	Segment Colour	Average Size	Global Colour	
Exp 1a	■	■				100%
Exp 1b	■	■				80%
Exp 2a	■		■			100%
Exp 2b	■		■			80%
Exp 3a	■			■		100%
Exp 3b	■			■		80%
Exp 4a	■				■	100%
Exp 4b	■				■	80%
Exp 5		■	■	■	■	0%
Exp 6		■	■	■		0%

Rows correspond to experiments and columns correspond to features. A shaded cell indicates that the feature in that column was used in the experiment in that row. The last column shows the proportion of training trials that contain a diagnostic shape. In Experiments 1–4 each participant saw stimuli that consisted of the combination of features shown in that row. Experiments 5 and 6 were between-subject designs so that participants were allocated to four (Experiment 5) or three (Experiment 6) groups and each participant saw stimuli with only one non-shape diagnostic feature.

was shape (the global configuration of segments) while the other feature was different in each experiment. In Experiment 1, the second feature was the location of a single patch in the image – that is, all images of a category contained a patch of a category-specific colour at a particular location (and none of the images from other categories contained a patch at this location). In Experiment 2, this feature was the colour of one of the segments – that is, all images assigned to a category contained a segment of a particular colour (and none of the images from other categories contained a segment of this colour). In Experiment 3, the second feature was the average size of patches – all patches in an image had similar sizes and the average size was diagnostic of the category. In Experiment 4, this feature was the colour of patches – all patches in an image had the same colour and images of different categories had different colours. In Experiment 5 and 6, all images had only one predictive feature. This was either patch location, segment colour, patch size or overall colour; but none of the categories had a predictive shape.

Table 1 summarises the different combinations of features that were examined in the behavioural tasks in this study. Examples for all Experiments are shown in Appendix E and two movies illustrating the predictive features in Experiments 1 and 2 can be seen in Movies S1 and S2 in Supporting Information.

In Experiments 1 to 4, training blocks were interleaved with test blocks which presented novel images that had not been seen during training. Each test block contained four types of test trials – `Both`, `Conflict`, `Shape` and `Non-shape` – that were designed to reveal the feature(s) used by the participant to categorise images. Trials in the `Both` condition contained the same combination of features that predicted an image's category during training. `Conflict` trials contained images with shape feature from one category and the second feature was swapped from another category.

`Shape` trials contained images with only the shape feature and a non-predictive value of the second feature. Finally, the `Non-Shape` trials contained images where the five segments were placed at random locations on the canvas, giving the stimulus no coherent shape, but each image contained the second predictive feature. Examples of these test trials are shown in Appendix E.

We can infer the features that a participant uses by looking at the pattern of performance across the test conditions. There are four possible patterns. If a participant relies on shape, they should perform well in trials where shape predicts the image category. Thus their pattern of performance should be high, high, high, and low in the `Both`, `Conflict`, `Shape`, and `Non-shape` conditions, respectively. In contrast, if the participant relies on the non-shape feature, this pattern should be high, low, low, high. If a participant uses both (shape and non-shape) features, the pattern should be high, medium, high, high, where a "medium" performance in the `Conflict` condition is indicative of the fact that the two cues (features) learnt by the participant will compete with each other in these trials. Finally, if a participant does not learn either feature, their performance should be low in all four conditions. For a similar methodological approach for determining features used to categorise novel stimuli see [14].

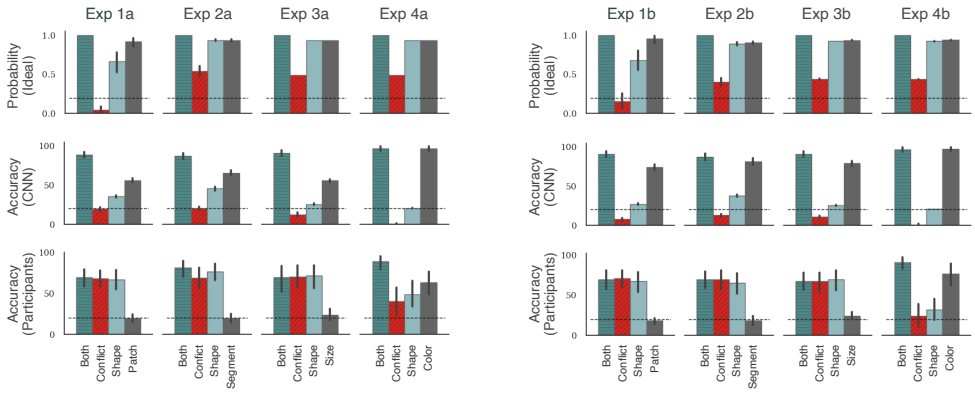
In each experiment, we compared the behaviour of participants with two statistical inference models: an ideal inference model and a CNN. The ideal inference model computes what should a participant do if they wanted to be statistically as efficient as possible and use all the information available during training trials. It uses a sequential Bayesian updating procedure to compute the probability distribution over category labels given the training data and a test image. Similarly, the CNN computes the most-likely category-label for an image by learning a mapping between images in the training set and their category labels. Thus, it makes an approximate statistical inference by approximating a regressive model [18, p85–89]. Both models are described in Materials and Methods below.

Both features equally predictive

Figure 2a shows the pattern of performance in the final test block in Experiments 1a, 2a, 3a, and 4a. In these tasks, both shape and non-shape features perfectly and independently predict the category label during training. Thus the learner could use either (or both) features to learn an image's category. The top row shows the pattern of performance for the ideal inference model. In all four tasks, this model predicts that the probability of choosing the correct category is high in the `Both`, `Shape` and `Non-shape` conditions. This indicates that there is enough information in the training trials for all four experiments to predict the category label based on either the shape or the non-shape feature.

The middle row shows the pattern of performance for the CNN model. In all four tasks, the network showed high accuracy in the `Both` condition – showing an ability to generalise to novel (test) stimuli, as long as both shape and non-shape features were preserved in the stimuli. It showed a low accuracy in the `Conflict` condition, but high accuracy in the `Non-shape` condition. Its performance in the `Shape` condition was above chance in Experiments 1a, 2a and 3a and at chance in Experiment 4a. The above-chance performance in the `Shape` condition implies that this network is able to pick up on shape cues. However, its performance is significantly lower in the `Shape` condition compared to the `Non-shape` condition. When these two cues competed with each other, in the `Conflict` condition, the network favoured the non-shape cue and the accuracy was at or below chance. These results indicate that the CNN learns to categorise using a combination of shape and non-shape features.

It is also worth noting that, unlike the ideal inference model, the CNN showed a bias towards relying on non-shape features in all experiments, even though it would be ideal



(a) Both features equally predictive **(b)** Non-shape feature more predictive

Fig 2. Results in Experiments 1–4. Each column corresponds to an experiment and each row corresponds to the type of learner (ideal inference model, CNN or human participants). The top row shows the posterior probability of choosing the labelled class for a test trial given the training data. The bottom two rows show categorisation accuracy for this labelled class. Each panel shows four bars that correspond to the four types of test trials. `Patch`, `Segment`, `Size` and `Colour` refer to the `Non-shape` test trials in Experiments 1, 2, 3 and 4, respectively. Error bars show 95% confidence and dashed black line shows chance performance. In any panel, a large difference between the `Both` and `Conflict` conditions shows that participants rely on the non-shape cue to classify stimuli. Both models show this pattern while humans show no significant difference.

(from an information-theoretic perspective) to learn both features in parallel. A similar result was observed by Hermann and Lampinen [20], who found that when multiple features predict the category, CNNs preferentially represent one of them and suppress the other.

The bottom row shows the average accuracy in the four experiments for human participants (N=25 in each task). Like the ideal inference model and the neural network model, participants showed high accuracy in the `Both` condition (mean accuracy was between 70% (in Experiment 1a) and 89% (in Experiment 4a). This indicates an ability to generalise to novel (test) stimuli as long as shape and non-shape features were preserved. However, their pattern of performance across the other three conditions were in sharp contrast to the two models. In Experiments 1a, 2a, and 3a, participants showed a high-high-high-low pattern in the `Both-Conflict-Shape-Non-shape` conditions, indicating that they strongly preferred the shape cue over the non-shape cue. In fact, performance in the `Non-shape` trials was at chance in all three tasks with mean accuracy ranging from 20% to 24%. Single sample t-tests confirmed that performance was statistically at chance in all three tasks (largest $t(24) = 0.99, p > .05$). Thus, unlike the ideal inference model, which learnt both predictive cues, participants chose one of these cues. And unlike the neural network model, which favoured the non-shape cue, participants preferred to rely on shape.

The behaviour of participants was different in Experiment 4a, where the non-shape cue was the colour of the entire figure. Performance was again high in the `Both` condition, but significantly lower in the `Conflict`, `Shape` and `Non-shape` conditions ($F(3, 72) = 8.18, p < .01, \eta_p^2 = .25$). So, on average, participants seemed to be using both shape and non-shape (colour) cues to make their decisions, but neither feature was strongly preferred over the other. This behaviour seemed to be qualitatively similar to the ideal inference model, which learnt to use both predictive cues simultaneously.

However, examining each participant separately, we found that participants could be grouped into two types, those that primarily relied on shape (N=12) and those that relied on colour (N=13). Participants were categorised as relying on colour if performance in the `Non-shape` condition was above performance in the `Shape` condition. Figure S4 in Appendix B shows the average pattern of performance for each of these groups. The first group shows a high-low-low-high pattern, indicating that they were predominantly using the colour cue to classify test images. The second group shows a high-high-high-low pattern, indicating that they were predominantly using the shape cue. Mixing these two groups of participants results in the high-medium-medium-medium pattern shown in Figure 2a.

One feature more predictive than the other

Our next step was to check what happens when one of the features predicts the category *better* than the other. If CNNs and humans are driven by performing statistical inferences, we expect both systems to start relying on the feature that is better at predicting the category label. In Experiments 1b, 2b, 3b, and 4b the shape feature predicted the category label in only 80% of the training trials. The remaining 20% images contained horizontal and vertical segments placed at random locations on the canvas so that these images contained no coherent shape. The second feature (patch location, segment colour, patch size or overall colour) predicted the category label in 100% of training trials. See Figures 1 and S12 for some examples of training images that do not contain a shape feature but contain a non-shape feature. Figure 2b shows the performance for the two models as well as human participants (N=25 in each task). The ideal inference model (top row) showed a very similar performance, again predicting that a participant should learn both features simultaneously. Its accuracy on non-shape feature was slightly better. This is a consequence of larger number of samples containing non-shape cues. In contrast, the performance for the CNN model was significantly different. In all experiments, the model now showed a high–low–low–high pattern, with performance in the `Shape` condition close to chance in most experiments. Thus, the CNN model started relying almost exclusively on the (more predictive) non-shape feature.

In contrast to both models, participants continued showing a high–high–high–low pattern in Experiments 1b, 2b, and 3b, indicating a clear preference for relying on shape. It should be noted that this happens even though shape is *not* the most predictive feature. In fact, performance in the `Non-shape` condition was at chance (mean accuracy ranged from 18% to 24%, largest $t(24) = 1.74, p > .05$ when compared to chance level), showing that participants completely ignored the most predictive feature.

The behaviour of participants was again different in the experiment using colour of entire figure as the non-shape cue (Experiment 4b). Average accuracy across participants was high in the `Both` condition, but significantly lower in the `Conflict`, `Shape`, and `Non-shape` conditions ($F(3, 72) = 22.68, p < .01, \eta_p^2 = .49$). Like Experiment 4a, examining each participant separately in Experiment 4b showed that participants could be divided into two groups – those that learnt to rely on shape and those that learnt to rely on colour. However, the ratio of participants in these groups changed. While 12 participants (out of 25) relied on shape in Experiment 4a, 7 participants (out of 25) relied on it in Experiment 4b (see Figure S5 in Appendix B).

Effect of previous training on CNN behaviour

In the above experiments, we observed that the participants systematically deviated from the two statistical inference models. This contrast was particularly noteworthy in Experiments 1b-4b. Here, the non-shape feature was more predictive than shape but

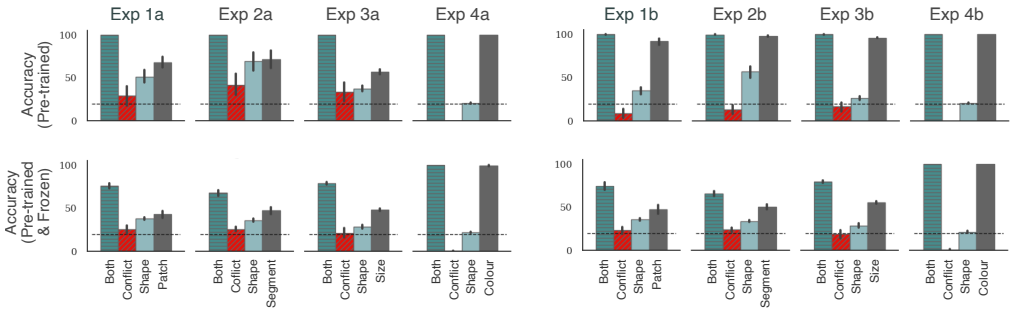
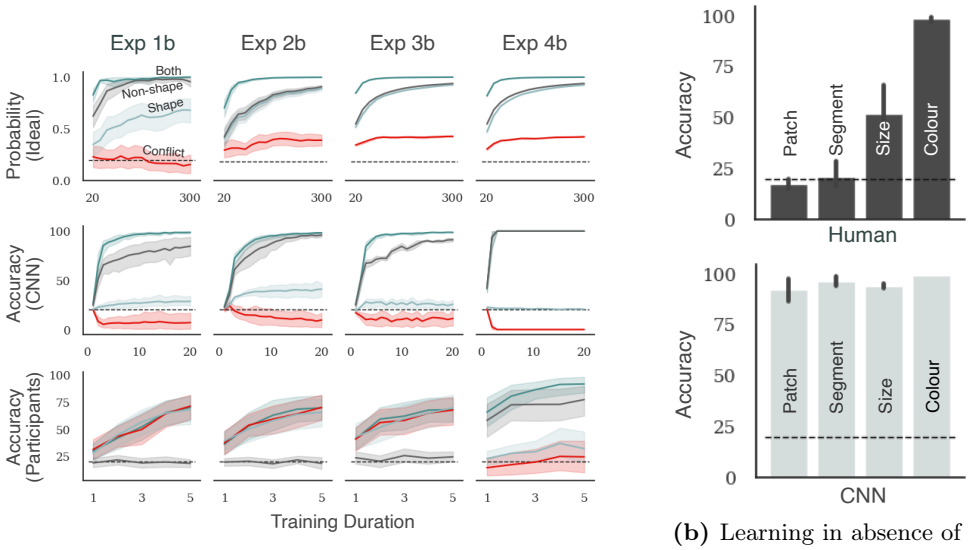


Fig 3. Results for pre-training on a dataset with a shape-bias. The first row shows results when the CNN was pre-trained on the Style-transfer ImageNet [12] and allowed to learn throughout the network. The second row shows results of the same network when weights for all convolution layers are frozen. First column shows results when both features are equally likely (Experiments 1a, 2a, 3a and 4a) while the second column shows results when the non-shape cue is more predictive (Experiments 1b, 2b, 3b and 4b). In all panels, we again observed a large difference between the `Both` and `Conflict` conditions, indicating that despite pre-training, models relied heavily on the non-shape cue to classify stimuli.

participants still focused on global features like shape. In contrast, the CNN preferred to rely on the more predictive (non-shape) feature. So we wanted to explore whether CNNs can be made to behave like humans through training. A recent set of studies have suggested that CNNs indeed start showing a shape-bias if they are pre-trained on a dataset that contains such a bias [12, 21]. However, after the network had been pre-trained on the first set with a shape-bias, these studies did not systematically manipulate how well each feature predicted category membership in the new set of images. This is a crucial manipulation in the above studies that allowed us to more directly assess the feature biases of CNNs, and our results suggest that the CNN learns to rely on the most diagnostic feature in this new set.

To test the effect of pre-training, we used the same CNN as above – `ResNet50` – but this time pre-trained on the Style-transfer ImageNet database created by [12] to encourage a shape-bias. We then trained this network on our task under two settings: (i) the same setting as above, where we retuned the weights of the network at a reduced learning rate, and (ii) an extreme condition where we froze the weights of all convolution layers (that is 49 out of 50 layers) limiting learning to just the top (linear) layer.

The results under these two settings are summarised in Figure 3. In line with previous results [12, 21], we observed that this network had a larger shape-bias – for example, it predicts the target category better in the `Shape` condition than the network pre-trained on ImageNet (compare with the middle row in Figure 2). In some cases, this makes the network behave more like the ideal inference model, where it is able to predict the category based on either shape or non-shape features. But this pattern is still in contrast with participants who were at chance when predicting based on non-shape features in Experiments 1–3. Crucially, when the non-shape feature is made more predictive, the network shows a bias towards this feature, showing the same high-low-low-high pattern observed above (Figure 3, top right). Even under the extreme condition, where we froze the weights of all except the final layer, the network preferred the non-shape feature as long as this feature was more predictive (Figure 3, bottom right). That is, CNNs do not learn to preferentially rely on shape when learning new categories even when pre-trained to have a shape bias on other categories.



(a) Dynamics of learning

(b) Learning in absence of shape

Fig 4. Each panel in (a) shows how accuracy on the four types of test trials changes with experience. The top, middle and bottom row correspond to ideal inference model, CNN and human participants respectively. Columns correspond to different experiments. The scale on the x-axis represents the number of training trials in the top row, the number of training epochs in the middle row and the index of the test block in the bottom row. The two panels in (b) show accuracy in test blocks for humans and CNN, respectively, when they are trained on images that lack any coherent shape. Each bar corresponds to the type of non-shape feature used in training.

Dynamics of learning

We probed the learning strategy used by models and participants by examining performance at regular intervals during training. If a participant (or model) learns multiple features in parallel, they should show an above-chance performance on both the `Shape` and `Non-shape` test trials at the probed interval. If they focus on a single feature, their performance on that feature should be above-chance and match the performance on the `Both` trials. If they switch between different features over time, their relative performance on `Shape` and `Non-shape` trials should also switch over time.

Figure 4a shows the performance under the four test conditions over time for Experiments 1b, 2b, 3b and 4b (results for Experiments 1a, 2a, 3a and 4a show a similar pattern and are shown in Appendix C). The ideal inference model shows an above-chance performance on the `Shape` as well as `Non-shape` trials throughout learning. This confirms the expectation that the ideal inference model should keep track of both features in parallel. However, this is neither what the CNN nor what human participants do. The CNN shows a bias towards learning the most predictive (non-shape) feature from the outset, with performance on the `Non-shape` trials closely following performance on the `Both` trials. Human participants showed the opposite bias, with performance on the `Shape` trials closely following performance on the `Both` trials. We did not observe any case where the relative performance on the `Shape` and `Non-shape` trials switched over time. This suggests that participants did not systematically explore different features and choose one – rather they continued learning a feature as long as it yielded enough reward. Even in Experiment 4b, where some participants used the colour cue while others used the shape cue, no participant in either group showed any evidence for switching form one feature to the other.

Learning in the absence of shape

The above experiments always pit a highly predictive feature against shape. We wanted to know whether participants struggle to learn the predictive local feature even when a diagnostic shape was absent. If participants only fail to learn this feature when a diagnostic shape is present, it indicates a difference in the bias between participants and CNNs (humans prefer global shape, while CNNs prefer more local features). On the other hand, if participants struggle to learn this feature even when it is clearly visible and a diagnostic shape is absent, it indicates a more fundamental limitation in human (but not CNN) capacity to extract these features. To test this, we designed a behavioural task (Experiment 5) where a shape feature was absent from the training set. Like the above experiments, each training stimulus still contained a set of patches and segments, but the segments were not consistently organised in a spatial structure (see Figure S12 in Appendix E for examples of this stimuli). Instead, every training trial contained a non-shape predictive feature. We used the same features as above – patch location, segment colour, patch size or overall colour. Participants were divided into four groups based on the type of predictive feature they were shown in the training trials. The test block consisted of novel images (that were not seen in training) but had the same diagnostic feature as training (equivalent to the `Non-shape` condition in the above experiments).

The average accuracy in test trials for each type of diagnostic feature is shown in Figure 4b. There was a large difference in performance depending on the type of diagnostic feature. When the colour of the entire figure predicted the category, accuracy on test trials was high ($M = 98.67\%$). The responses collected for training trials indicated that participants learned this feature quickly (performance reached 94.40% after 100 training trials). Accuracy in the test block was lower (though still significantly above chance) when the size of patches predicted the category ($M = 52.40\%$) and participants learned this feature at a slower rate. In contrast to these two conditions, participants were unable to learn the other two diagnostic local features. Performance was at chance in test trials both when the colour of a segment predicted the category ($M = 21.47\%$) and when the location and colour of a single patch predicted the category ($M = 17.47\%$). Thus participants seemed sensitive to the computational complexity of the diagnostic feature. They extracted simple features like the colour of the entire figure or the size of patches, but did not extract more complex features like colour of single segment or patch. Figure 4b also shows the performance of the CNN on this task. In contrast to human participants, the network learnt all four types of non-shape stimuli and showed high accuracy on test trials in all four conditions.

Identifying versus Learning features

In order to discover the correct diagnostic features in the experiments above, a participant must perform two distinct operations: they must identify a diagnostic feature (from a list of all possible features) and match the correct value of this feature to each category. For example, in Experiment 2, the participant must first realise that the diagnostic feature is the colour of each segment. That is, they must find this feature in the space of all possible features (shape, number of patches, location, size, etc.). Secondly, they must map the stimulus on a given trial to the correct category, extracting the colour of all five segments, working out which segment is diagnostic and what the mapping is between the diagnostic colour and category. The second operation – mapping a diagnostic value to a category – is a computationally demanding task as it requires the participant to remember several pieces of information, comparing the features observed in a given stimulus with the features and outcomes of past stimuli. One reason why participants might fail when the CNN succeeds is that humans and

CNNs have very different computational resources available to them. For example, while humans are limited by the capacity of their working memories (the number of features they can process at the same time), CNNs have no such limitations. If this was the case – i.e., if participants were failing because of their limited cognitive resources and not because they were unable to identify the correct feature – we hypothesised that helping the participants identify the diagnostic feature will not improve their performance on these tasks.

We checked this hypothesis in Experiment 6 that repeated the design of Experiment 5, where participants saw stimuli that had only the non-shape diagnostic feature and no coherent shape. Instead of letting participants figure out which feature was diagnostic, we informed them of the diaganostic feature in each task and showed them two examples of stimuli with the diagnostic feature (see Materials and Methods for details). Additionally, we increased the duration of each stimulus from 1s to 3s to ensure that participants do not underperform because of the time constraint. Finally, we gave participants an added incentive to learn the task, increasing the possible bonus reward based on their performance in the test block. Participants then completed 6 training blocks (50 trials each) where they saw random samples of stimuli from each category. We already know that participants can solve the task when the diagnostic feature was the colour of the entire figure (see Figure 4b above). Therefore, we tested three groups of participants, where each group was trained on stimuli with one of the other three non-shape features – patch location, segment colour or average size – being diagnostic of the category.

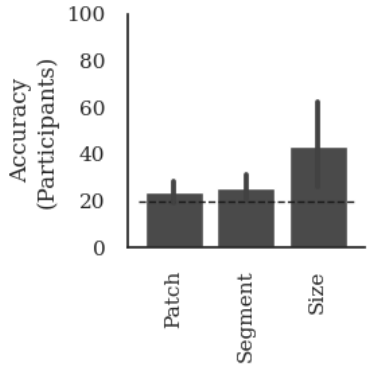


Fig 5. Results of telling participants the diagnostic cue. Each bar shows mean accuracy across 10 participants in the test block. Participants were divided into three groups based on the diagnostic cue – patch location, segment colour, or average size – used to train the participants.

The results of Experiment 6 are shown in Figure 5. Like Experiment 5, mean performance across participants was above chance in the Size condition but at chance in the Patch and Segment conditions. The overall pattern of results for the three conditions was statistically indistinguishable from the results of Experiment 5. In other words, even when participants were told the diagnostic features and given additional time and incentive to learn the task, they struggled to classify stimuli based on patch location or segment colour. These results confirm the hypothesis that the difficulty of these tasks for human participants is not limited to identifying the diagnostic features. Instead, the cognitive resources required to extract the diagnostic feature value and mapping it to the correct category may play a critical role in how humans select features for object classification.

Discussion

In a series of experiments we repeatedly observed that participants failed to pick up highly predictive features of the training set in a visual recognition task. This behaviour contrasted with the ideal inference model as well as the CNN, both of which always learnt to categorise based on the predictive non-shape features. These results pose a challenge for the hypothesis that humans and CNNs have similar internal representations of visual objects. CNNs are designed to learn the statistical dependencies between features and categories via backpropagation. So it is not surprising that they learned to classify objects based on highly predictive features. In contrast, participants were blind to many perfectly predictive statistical regularities – which goes to show that humans do not select features purely through this mechanism of internalising statistical dependencies. This is in keeping with many psychological studies which show that humans prefer to categorise visual objects based on global features, such as overall shape [42, 3, 32, 35].

We also observed that participants diverged from ideal inference model in interesting ways. Not only did they ignore many diagnostic features they could clearly see, they frequently selected only one of several possible features available to learn an input-output mapping (e.g. in Experiment 4 participants chose to classify either based on colour or shape but almost never both, even though this was the optimal policy). Furthermore, in Experiment 6, we observed that even when the relevant features are pointed out, participants still did not learn to classify objects relying on these features. These results highlight that limitations in cognitive resources play a fundamental role in how humans learn the task and suggest that participants are looking for a satisfactory rather than optimal solution to the categorisation problem. In other words, participant behaviour is better explained by a "satisficing" account [53] than an optimising account of object recognition.

The contrast in the behaviour of humans and CNNs speaks to an ongoing debate about the role of inductive biases in humans and Deep Neural Networks. Many researchers have suggested that innate inductive biases, such as a shape-bias, are needed to generalize under challenging conditions [31, 50, 13, 66, 17]. But building-in inductive biases is a controversial approach. Instead of imposing constraints on the architectures or learning algorithms of CNNs, some researchers have argued that current models may not display similar biases to humans simply because of limitations in standard training sets (such as `ImageNet`) where shape is not the most diagnostic feature of object classes. And indeed, studies have shown that standard networks can develop a shape-bias when the training set is designed to make shape the most diagnostic feature for object classification [10, 12, 21, 20]. On this view, the lack of a shape-bias and a corresponding limitation of generalisation in many CNNs reflects a limitation of the training sets, rather than a more fundamental limitation of architecture or learning algorithm.

Our findings pose a challenge to this claim. We designed a behavioural task with novel objects that participants had never classified before. We did this to preclude the possibility that participants can use previously learnt diagnostic features of these objects to classify them. When learning to classify these objects, participants systematically relied on shape, even when shape was *not* the most diagnostic feature (Figure 2b). In other words, unlike CNNs, participants did not need to be trained on a dataset with a shape-bias in order to display a shape-bias. These results support the hypothesis that humans do indeed have an inductive bias to classify objects by shape even when this is not the ideal statistical feature for classification in a given task. Our findings are consistent with recent study that also found DNNs learn to classify images on the basis of simple diagnostic features and ignore more complex features [51]. Interestingly, the focus of this study was not on comparing DNNs to humans, but rather, showing how a simplicity bias limits generalization.

Of course, the participants in our studies had a life-time of exposure to a natural world where shape may be the most diagnostic feature. Accordingly, it is possible that they had acquired a shape-bias early on in life [55] that constrained how the new objects in our experiments were learned. But we observed that CNNs did not adopt a shape-bias in classifying our objects even when we induced a shape-bias in pre-training and when we froze the weights in an attempt to preserve the shape-bias when classifying our new objects. Instead they simply learned whatever features of new object categories were most diagnostic. In other words, even when a network has learned a shape-bias, this bias is wiped out when a new training set contains a non-shape features that are diagnostic of object category.

It should also be noted that the behaviour of participants observed here highlights a more extreme form of shape-bias than has been reported before. In a typical shape-bias experiment, the term shape-bias indicates the inductive-bias to rely on shape in the presence of alternative features that are equally good at predicting the target category [32, 55]. In our experiments, we observed that participants relied on shape even in the presence of features that were *better* at predicting the target category. Furthermore, in two of our experiments (Experiments 5 and 6) there was no consistent shape at all that could be used to predict category membership. In these experiments, participants failed to pick some perfectly predictive statistical features (like location of patch or colour of segment) even in the absence of a diagnostic shape. This blindness towards certain features cannot be explained as a shape-bias as there is no competing shape feature to learn.

An important outstanding question is *why* participants in our study relied on global features such as shape or overall colour whereas CNNs and the ideal inference model learned on the basis of whatever features in the dataset were most predictive. One relevant difference between the models and participants is that the models do not suffer from the same resource limitations as humans. A striking example of this is that CNNs not only succeed in learning to classify millions of images in `ImageNet` into 1000 categories, they can also learn to classify the same number of random patterns of TV static-like noise into 1000 different categories [67], something far beyond the capacity of humans [59]. This capacity was no doubt exploited by the CNNs in the current learning context. By contrast, our participants had to learn the object categories in the face of many well documented cognitive limitations of humans, such as limited capacity of visual short-term memory [1], visual crowding [61, 39] and selective attention [37, 68].

Another difference between humans and CNNs is how they perceive sensory data. While CNNs work with raw pixels, we know that the human visual system has a limited acuity for colour, contrast, location, etc. [48, 56]. However, we do not believe that this limited acuity is the key factor underlying the results above. First of all, we tested acuity to location and colour in a pilot study (see Material and Methods) by asking participants to remember the location and colour of a given patch embedded in a typical stimulus. We found that participants were very good at performing this task, indicating that they were capable of perceiving and remembering the locations of single patches. Indeed, as can be verified by looking at Movies S1 and S2 of the stimuli in Supplementary Information, once the diagnostic patch in a category is pointed out, it is difficult to miss it. Secondly, Experiments 5 and 6 show that participant performance was at chance not only in the Patch condition, but also in the Segment condition, where the diagnostic feature was the colour of a whole segment. There can be no doubt that participants can perceive the colour of the five segments, especially as they were able to use the same colours to classify stimuli in the global Colour condition. Based on the results of Experiment 6, the most plausible explanation of participant behaviour is that limited cognitive resources play a critical role in which features they select for classification. Indeed, limited cognitive resources may also explain why humans have a

shape-bias in the first place. Shape may be a compact code – a low-dimensional representation that ignores many details of the object – that is not only highly diagnostic of object categories, but a well suited feature for a brain with limited resources.

Whatever the origin of the shape-bias, the results here should give pause for thought to researchers interested in computational models of visual object recognition. These results show that humans are blind to a wide range of non-shape predictive features when classifying objects, and if models are going to be used as theories of human vision, they should be blind to these features as well. This may result not only in models that are more psychologically relevant, but also capture the robustness and generalisability of the human visual system that is lacking in current models [13, 50, 9].

Materials and Methods

Experimental Details

Materials We constructed nine datasets of training and test images. There were 2000 training images and 500 test images in each dataset. Each image consisted of 30–55 coloured patches on a white background. The colours of patches were sampled from a palette of 20 distinct colours so that they were clearly discernible. These patches were organised into five segments. There were four short segments (consisting of 5–10 patches) and one long segment (consisting of 10–15 patches). Each segment was oriented either vertically or horizontally. Images were grouped into five target categories and each category was paired with a unique spatial configuration of segments. It is this spatial configuration of segments that we refer to as *shape*. All images in a category also contained a second diagnostic feature, which was the location and colour of a patch in Experiment 1, the colour of a segment in Experiment 2, the average size of patches in Experiment 3 and the colour of all the segments in Experiment 4.

Within each category, images were randomly generated and varied in the number, colour, location and size of patches. This variability ensured that (i) participants (human and CNN) had to generalise over images to learn the category mappings, and (ii) there were no incidental local features that could be used to predict the category. The exact number of patches in each segment was sampled from a uniform distribution; the size and location of each patch was jittered (around 30%); and the colour of each patch (Experiments 1 and 3) or each segment (Experiment 2) was randomly sampled from the set of (non-diagnostic) colours. In addition, each figure was translated to a random location on the canvas and could be presented in one of four different orientations (0, $\pi/2$, π and $3\pi/4$ radians).

The original size of images was 600x600 pixels. This was reduced to 224x224 pixels for the simulations with CNNs. For the behavioural experiments, the stimuli size was scaled to 90% of the screen height (e.g. if the screen resolution was 1920x1080 the image size would have been 972x972). This ensured that participants could clearly discern the smallest feature in an image (a single patch) which we confirmed in a pilot study (see Procedure below).

Participants Participants were recruited and reimbursed through Prolific. In Experiments 1-4 there were N = 25 participants per experiment (total N = 200), and in Experiments 5 and 6 there were 10 participants per experimental condition (N = 40 in Experiment 5, and N = 30 in Experiment 6). In the first 5 experiments participants received 4 GBP for participating in the experiment and could earn an additional 2 GBP depending on average accuracy in the test blocks. In Experiment 6 the incentive was increased to 5.30 GBP and a possible bonus of 3 GBP based on performance in the test

block. Calculated as payment per hour, the average payout per participant in our experiments was 7.53 GBP per hour.

Procedure All experiments consisted of blocks of training trials, where participants learned the categorisation task, followed by test trials, where their performance was observed. During training trials participants saw an image for 1000ms and were asked to predict its category label. After each training trial, participants were told whether their choice was correct and received feedback on the correct label if their choice was incorrect. In Experiments 1 to 5, participants had to discover the predictive features themselves, while in Experiment 6, they were explicitly told what the predictive feature was at the beginning of the experiment. In this experiment, they were given textual instructions describing the target feature and shown exemplars where the target feature was highlighted. Participants saw 5 blocks of 60 training trials in Experiments 1–4 and 10 blocks of 50 trials in Experiments 5 and 6. The number of training trials was chosen based on a pilot experiment and ensured that participants learnt the behavioural task. In Experiments 1 to 4, each training block was followed by a test block containing 40 trials (10 per condition). In Experiments 5 and 6, one test block was presented at the end of training consisting of 75 trials. Test trials followed the same procedure as training, except participants were not given any feedback. As we were interested in object recognition rather than visual problem solving, all trials (training as well as test) used a short presentation time of 1000ms. In a follow-up experiment (as well as Experiment 6), we also tried a longer presentation time of 3000ms and observed a similar pattern of results (see Appendix D in Supplementary Information).

All experiments were designed in PsychoPy and carried out online on the Pavlovia platform. We ensured that participants could clearly see the location of each patch by conducting a pilot study. In this study, participants were shown an image from one of our datasets and asked to attend to a highlighted patch. After a blank screen they were shown a second image from the same dataset and asked to click on the patch which was in the same position as the highlighted patch in the first image. We found that the median location indicated by participants deviated from the center of the target patch by only a quarter of the width of a patch - meaning that participants were able to attend, keep in working memory and point out a specific patch location. This indicates that even the smallest of the local features used in this study was perceivable for human participants.

Data Analysis In all experiments chance performance was 20% since there is a 1 in 5 chance of randomly picking the correct category. Single sample t-tests were conducted in order to determine whether participants were above chance level performance. Repeated measures analyses of variance (ANOVA) were conducted when determining whether there was an effect of condition (Both, Conflict, Shape, Non-shape) on performance in an experiment. Follow-up comparisons were conducted with the Tukey HSD correction for multiple comparisons.

Ethics All studies adhered to the University of Bristol ethics guidelines and obtained an ethics approval from the University of Bristol ethics approval board.

Simulation Details

Neural Network model During a supervised learning task (like the task outlined in this study), a neural network performs an approximate statistical inference by constructing an input-output mapping between a random vector $\mathbf{X}$ and a dependent variable Y. The training set consists of N realisations of this random vector,

$\{\mathbf{x_1} \ldots, \mathbf{x_N}\}$ and N category labels $\{c_1 \ldots, c_n\}$. For a CNN, the vectors $\mathbf{x_i}$ can simply be an image (i.e. a vector of pixel values). That is, $\mathbf{X}$ lies in a high-dimensional image space. The neural network learns a non-linear parametric function $\hat{c_i} = F(\mathbf{x_i}, \mathbf{w})$ by finding the connection weights $\mathbf{w}$ which minimise the difference between the outputs produced by the network $\hat{c_i}$ and the given category labels, c_i. During a test trial, the network performs an approximate statistical inference by deducing the class of a test vector $\mathbf{x_{test}}$ by applying the learnt parametric function to this vector: $c = F(\mathbf{x_{test}}, \mathbf{w})$.

Since our task involved image classification, we evaluated three state-of-the-art deep convolutional neural networks, ResNet50 [19], VGG-16 [54] and AlexNet [29] which performs image classification on some image datasets to a near-human standard. We obtained qualitatively similar results with all three architectures. Therefore, we focus on the results of ResNet50 in the main text and describe the results of the other two architectures in SI Appendix A. Since participants had a lifetime experience of classifying naturalistic objects prior to the experiment, we pre-trained our networks on a set of naturalistic images (ImageNet). In each experiment, this pre-trained network was fine-tuned to classify the 2000 images sampled from the corresponding dataset into 5 categories. This fine-tuning was performed in the standard manner [65] by replacing the final (fully-connected) layer of the network to reflect the number of target classes in each dataset. The models learnt to minimise the cross-entropy error by using the Adam optimiser [25] with a mini-batch size of 32 and learning rate of 10^{-5}, which was reduced by a factor of 10 on plateau using the Pytorch scheduler function `ReduceLROnPlateau`. In one simulation study (Figure 3), we used a network that was pre-trained on a variation of ImageNet that induces a shape bias [12] and then froze the weights in all but the final classification layer to ensure that the learned bias was present during the training on the new images. In all simulations, learning continued till the loss function had converged. Generally this meant that accuracy in the training set was $> 99\%$, except in the case where we froze all convolution weights where accuracy converged to a value $> 70\%$. Each model was tested on 500 images drawn from `Both`, `Conflict`, `Shape` and `Non-Shape` conditions outlined above. The results presented here are averaged over 10 random seed initialisations for each model. All simulations were perfomed using the `Pytorch` framework [44] and we used `torchvision` implementation of all models.

Ideal inference model The goal of this model is to answer the following question: what class, $Y \in \{1, \ldots, C\}$, should a decision-maker assign to a test image, given a set of mappings from images to class labels (training trials). For the purpose of statistical inference, each image can be treated as a vector of features and each training trial assigns a feature vector, $\mathbf{x_i} = (x_i^1, \ldots, x_i^F)$, to a class label, $Y = c$. In our behavioural task, each feature (colour / location / size) can take a discrete set of values, so we treat each feature as a categorical random variable, $X^f \in \{1, \ldots, K\}$. The decision-maker infers the class label for a test image, $\mathbf{x_{test}}$, in two steps. Like the neural network, it first learns a set of parameters $\boldsymbol{\theta}$ that encode the dependencies between class labels and feature values in the training data. It then uses these parameters to predict class label for a given test image, $\mathbf{x_{test}}$.

We start at the end. Our goal is to compute $p(Y = c|\mathbf{X} = \mathbf{x_{test}}, \mathcal{D})$, the probability distribution over class labels given the training data, $\mathcal{D}$, and a test image, $\mathbf{x_{test}}$. Using Bayes' law, we have:

$$p(Y = c|\mathbf{X} = \mathbf{x_{test}}, \mathcal{D}) \propto p(\mathbf{X} = \mathbf{x_{test}}|Y = c, \mathcal{D}) \;\; p(Y = c) \tag{1}$$

where $p(Y = c)$ is the class prior and $p(\mathbf{X} = \mathbf{x_{test}}|Y = c, \mathcal{D})$ is a joint class-conditional density – the probability of observing the set of features, $\mathbf{x_{test}}$, for a given class, c. In our behavioural tasks, each feature is independently sampled. This means that the joint

distribution factorises as a product of class-conditional densities for each feature:

$$p(\mathbf{X} = \mathbf{x_{test}}|Y = c, \mathcal{D}) = \prod_{f=1}^{F} p(X^f = x^f_{test}|Y = c, \mathcal{D})$$

Our approach is to estimate these class-conditional densities by constructing a generative model $p(X^f = x^f_{test}|Y = c, \boldsymbol{\theta})$. Here $\boldsymbol{\theta}$ are the parameters of the model that need to be estimated based on training data. Since X^f is a categorical variable, a suitable form for this parametric distribution is the multinomial distribution, $Mult(x^f_{test}|1, \boldsymbol{\theta})$. The Bayesian method of estimating these parameters is to start with the prior distribution $p(\boldsymbol{\theta})$ and update it based on training data, $\mathcal{D}$, to obtain the posterior $p(\boldsymbol{\theta}|\mathcal{D})$. An appropriate prior for the multinomial is the Dirichlet distribution, $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ are the hyper-parameters of the Dirichlet distribution. For this Dirichlet-multinomial model, the update step involves counting the number of times each feature value occurs in the training data and adding these counts to the hyper-parameters [4].

Once we have a posterior distribution on the model parameters, $p(\boldsymbol{\theta}|\mathcal{D})$, we can obtain the required class-conditional densities, $p(X^f = x^f_{test}|Y = c, \mathcal{D})$ by integrating over these parameters. This leads to the following expression (see [41]):

$$p(X^f = x^f_{test}|Y = c, \mathcal{D}) = \frac{N_k + \alpha_k}{\sum_v N_v + \alpha_v}$$

Here N_k is the number of times X^f takes the value k in the training data and the sum in the denominator is carried out over all possible values $\{1, \ldots, K\}$ of X^f. Thus this model predicts that the class-conditional density of observing a feature value during a test trial depends on the relative frequency with which the given feature value occurs during the training data. These class-conditional densities can be plugged back into Equation 1 to give the probability distribution over all classes given the test image, $\mathbf{x_{test}}$. In our Results, we report this probability for the labelled class averaged over all the test images in a test condition.

References

[1] Alan Baddeley. Working memory. *Current biology*, 20(4):R136–R140, 2010.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988.

[4] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[5] Charles F Cadieu, Ha Hong, Daniel LK Yamins, Nicolas Pinto, Diego Ardila, Ethan A Solomon, Najib J Majaj, and James J DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, 10(12):e1003963, 2014.

[6] Radoslaw M Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317, 2019.

[7] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[8] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[9] Marin Dujmović, Gaurav Malhotra, and Jeffrey S Bowers. What do adversarial images tell us about human vision? *Elife*, 9:e55978, 2020.

[10] Reuben Feinman and Brenden M Lake. Learning inductive biases with simple neural networks. *arXiv preprint arXiv:1802.02745*, 2018.

[11] József Fiser and Richard N Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, 12(6):499–504, 2001.

[12] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[13] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[14] Micah B Goldwater, Hilary J Don, Moritz JF Krusche, and Evan J Livesey. Relational discovery in category learning. *Journal of Experimental Psychology: General*, 147(1):1, 2018.

[15] Ian Goodfellow, Honglak Lee, Quoc Le, Andrew Saxe, and Andrew Ng. Measuring invariances in deep networks. *Advances in neural information processing systems*, 22:646–654, 2009.

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[17] Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2020.

[18] Simon Haykin and N Network. *Neural Networks: A comprehensive foundation*, volume 2. 1999.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33, 2020.

[21] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

[22] Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annu. Rev. Psychol.*, 55:271–304, 2004.

[23] Saeed Reza Kheradpisheh, Masoud Ghodrati, Mohammad Ganjtabesh, and Timothée Masquelier. Deep networks can resemble human feed-forward vision in invariant object recognition. *Scientific reports*, 6(1):1–24, 2016.

[24] Tim C Kietzmann, Patrick McClure, and Nikolaus Kriegeskorte. Deep neural networks in computational neuroscience. *BioRxiv*, page 133504, 2018.

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[26] Wouter Kool and Matthew Botvinick. Mental labour. *Nature human behaviour*, 2 (12):899–908, 2018.

[27] Konrad P Körding and Daniel M Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(6971):244–247, 2004.

[28] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1:417–446, 2015.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[30] Jonas Kubilius, Stefania Bracci, and Hans P Op de Beeck. Deep neural networks as a computational model for human shape sensitivity. *PLoS computational biology*, 12(4):e1004896, 2016.

[31] 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*, 40, 2017.

[32] Barbara Landau, Linda B Smith, and Susan S Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

[33] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[34] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[35] E Charles Leek, Mark Roberts, Zoe J Oliver, Filipe Cristino, and Alan J Pegna. Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. *Neuropsychologia*, 89:495–509, 2016.

[36] Bria Long and Talia Konkle. The role of textural statistics vs. outer contours in deep cnn and neural responses to objects. In *Conference on Computational Cognitive Neuroscience*, page 4, 2018.

[37] Arien Mack. Inattentional blindness: Looking without seeing. *Current Directions in Psychological Science*, 12(5):180–184, 2003.

[38] Gaurav Malhotra, Benjamin D Evans, and Jeffrey S Bowers. Hiding a plane with a pixel: examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, 174:57–68, 2020.

[39] Mauro Manassi, Bilge Sayim, and Michael H Herzog. Grouping, pooling, and when bigger is better in visual crowding. *Journal of Vision*, 12(10):13–13, 2012.

[40] Johannes Mehrer, Courtney J Spoerer, Emer C Jones, Nikolaus Kriegeskorte, and Tim C Kietzmann. An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, 118(8), 2021.

[41] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[42] David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3):353–383, 1977.

[43] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[44] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[45] Pietro Perconti and Alessio Plebe. Deep learning and cognitive science. *Cognition*, 203:104365, 2020.

[46] Blake A Richards, Timothy P Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, et al. A deep learning framework for neuroscience. *Nature neuroscience*, 22(11):1761–1770, 2019.

[47] Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pages 2940–2949. PMLR, 2017.

[48] John G Robson. Spatial and temporal contrast-sensitivity functions of the visual system. *Josa*, 56(8):1141–1142, 1966.

[49] Jenny R Saffran, Richard N Aslin, and Elissa L Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.

[50] Thomas Serre. Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, 5:399–426, 2019.

[51] Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *arXiv preprint arXiv:2006.07710*, 2020.

[52] Amitai Shenhav, Sebastian Musslick, Falk Lieder, Wouter Kool, Thomas L Griffiths, Jonathan D Cohen, and Matthew M Botvinick. Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, 40:99–124, 2017.

[53] Herbert A Simon. Rational choice and the structure of the environment. *Psychological review*, 63(2):129, 1956.

[54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[55] Linda B Smith, Susan S Jones, Barbara Landau, Lisa Gershkoff-Stowe, and Larissa Samuelson. Object name learning provides on-the-job training for attention. *Psychological science*, 13(1):13–19, 2002.

[56] John A Swets, Wilson P Tanner Jr, and Theodore G Birdsall. Decision processes in perception. *Psychological review*, 68(5):301, 1961.

[57] Andrea Tacchetti, Leyla Isik, and Tomaso A Poggio. Invariant recognition shapes neural representations of visual input. *Annual review of vision science*, 4:403–422, 2018.

[58] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[59] Christian Tsvetkov, Gaurav Malhotra, Benjamin D Evans, and Jeffrey S Bowers. Adding biological constraints to deep neural networks reduces their capacity to learn unstructured data. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*, 2020.

[60] Rufin VanRullen. Perception science in the age of deep neural networks. *Frontiers in psychology*, 8:142, 2017.

[61] David Whitney and Dennis M Levi. Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in cognitive sciences*, 15(4): 160–168, 2011.

[62] Christopher D Wickens. Processing resources and attention. *Multiple-task performance*, 1991:3–34, 1991.

[63] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356–365, 2016.

[64] Daniel LK 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*, 111(23):8619–8624, 2014.

[65] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *arXiv preprint arXiv:1411.1792*, 2014.

[66] Alan L Yuille and Chenxi Liu. Deep nets: What have they ever done for vision? *International Journal of Computer Vision*, pages 1–22, 2020.

[67] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[68] Li Zhaoping. A new framework for understanding vision from the perspective of the primary visual cortex. *Current opinion in neurobiology*, 58:1–10, 2019.

Supporting Information

A Robustness across CNN architectures

The results in Figures 2 and 4 in the main text show the behaviour of one CNN architecture – `ResNet50` – in our experiments. Here we evaluated the robustness of these results by training and testing two other CNN architectures – AlexNet [29] and VGG-16 [54] – on our tasks. The training and testing procedure remained the same as for `ResNet50`, and as described in the Section above. As can be seen from Figures S1, S2 and S3 below, both these architectures show qualitatively similar results to `ResNet50`: both models pick up on the `Non-shape` feature in all experiments and clearly favour this feature when it is more predictive that the `Shape` feature in Experiments 1b–4b.

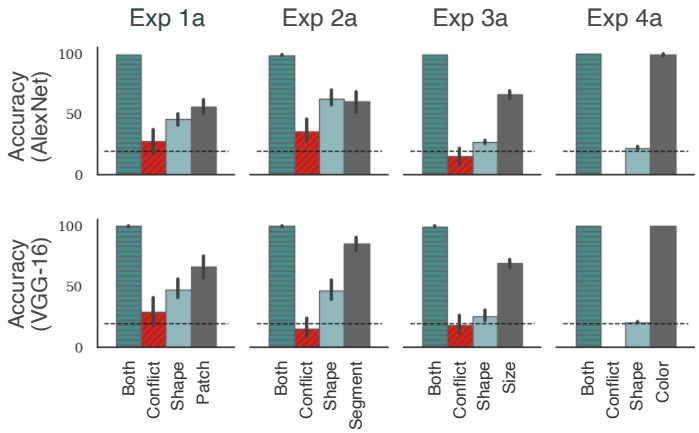


Fig S1. Results when both features are equally predictive. Each panel shows the accuracy under the four test conditions for AlexNet (top row) or VGG-16 (bottom row). Each column corresponds to a different experiment. Both models were pre-trained on `ImageNet` and fine-tuned by reshaping the final layer to reflect the number of target classes in each experiment and trained on 2000 images from the training set (see Section for details). A comparison with Figure 2a shows that both architectures showed the same pattern of results as ResNet50: models were able to learn the task (high accuracy in the Same condition), learned both the Shape and Non-shape features (above chance accuracy in Shape and Non-shape conditions) and preferred to rely on the Non-shape feature (low accuracy in the Swap condition).

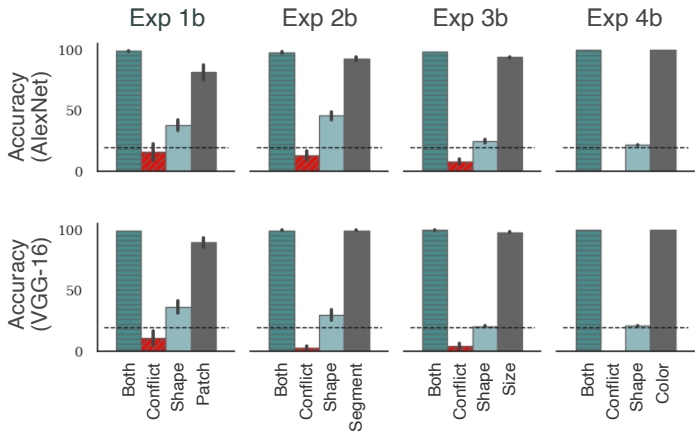


Fig S2. Results when non-shape feature is more predictive. Each panel again shows the accuracy under the four test conditions for AlexNet (top row) or VGG-16 (bottom row). Each column corresponds to a different experiment. A comparison with Figure 2b shows that both architectures showed the same pattern of results as ResNet50: models showed a strong preference to rely on the non-shape feature in this case (a high-low-low-high pattern in the Same-Swap-Shape-Non-shape conditions) and this preference became larger than the experiments where both features were equally predictive (compare with Figure S1 above).

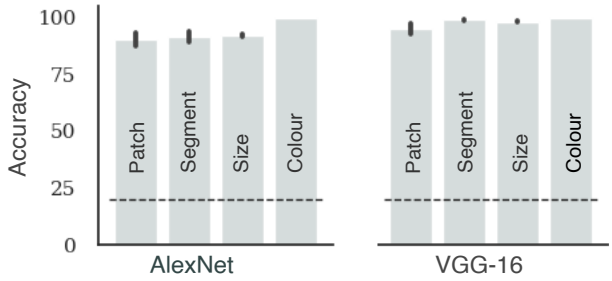


Fig S3. Results for learning without shape feature. The two panels show accuracy in test blocks for AlexNet and VGG-16, respectively, when these models were trained on images that lack any coherent shape (Experiment 5). Each bar corresponds to the type of non-shape feature used in training. Like ResNet50, but unlike human participants (compare with Figure 4b), both models were able to learn all types of non-shape features.

B Two groups in Experiment 4

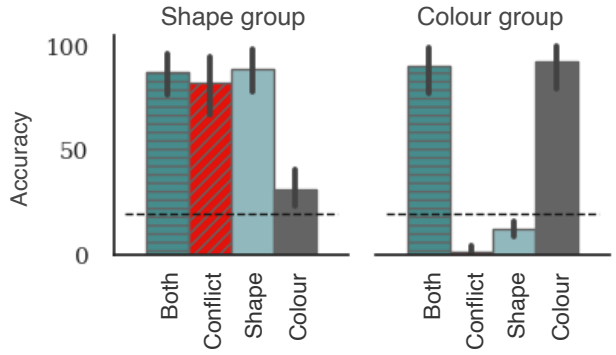


Fig S4. Two groups in Experiment 4a. Each panel shows the accuracy under the four test conditions for a subgroup of participants. Participants were split based on whether they performed better in the shape or colour conditions. The first group contained N=12 participants and the second group contained N=13 participants.

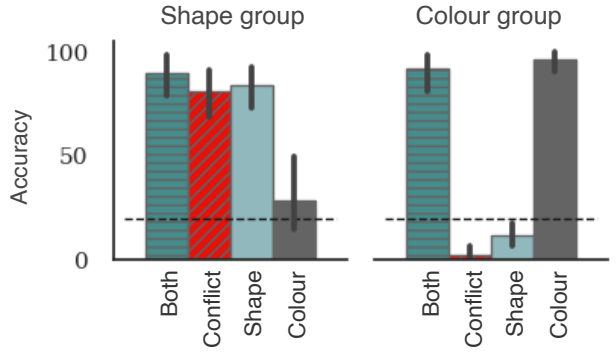


Fig S5. Two groups in Experiment 4b. Each panel again shows accuracy under the four test conditions for the subgroups of participants who prefer to rely on shape and colour, respectively. In this case, the first group consisted of N=7 participants and the second group consisted of N=18 participants.

C Learning dynamics in Experiments 1a, 2a, 3a, and 4a

Figure 4a in the main text shows the change in performance under the four test conditions in Experiment 1b, 2b, 3b and 4b, where the non-shape feature and more predictive than shape features in training. Here we have plotted how performance changes in Experiments 1a, 2a, 3a and 4a, where both features are equally likely. A comparison of Figure S6 and Figure 4a from the main text shows a very similar pattern in all experiments and for humans as well as the two types of models. The two models predict that a difference between `Both` and `Swap` conditions emerges early and grows with learning. In contrast, human participants show no difference in the two conditions throughout the experiment in Experiments 1a, 2a and 3a. Further analysis of individual participants showed that, like Experiments 1b, 2b, 3b and 4b, no participant switched from using one feature to another during the experiment.

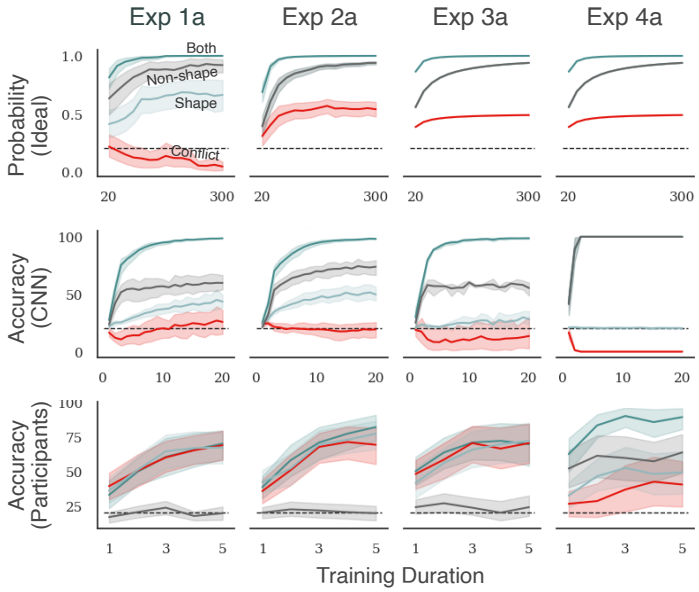


Fig S6. Change in test performance with training in Experiments 1a, 2a, 3a, and 4a. Each panel shows how accuracy on the four types of test trials changes with experience. The top, middle and bottom row correspond to optimal decision model, CNN and human participants respectively. Columns correspond to different experiments. The scale on the x-axis represents the number of training trials in the top row, the number of training epochs in the middle row and the index of the test block in the bottom row.

D Longer presentation time

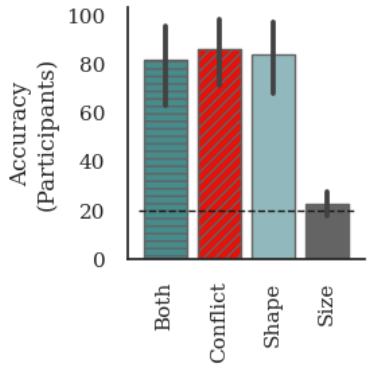


Fig S7. Results of giving participants more time in Experiment 3a. Accuracy in the four conditions when participants are shown the stimuli for 3s instead of 1s. In this experiment, every trial has two diagnostic features – global shape and average size. Despite the increase in the duration of the stimulus, participants performed well in the `Both`, `Conflict` and `Shape` conditions, but performed at chance in the non-shape (`Size`) condition, indicating that they still preferred to learn based on shape. Notice, we used Experiment 3 (non-shape cue = average size) to test this because this is experiment in which the participants were most likely to pick on the non-shape (`Size`) cue based on results in Experiment 5, where mean performance in the Size condition was above chance, while mean performance in Segment or Patch conditions was at chance, even when there was no competing shape feature.

E Example stimuli

In this section, we show an examples of images from all experiments. The reader may also want to look at Movies S1 and S2 in Supplementary Materials, which best illustrate the features used in Experiment 1 and 2.

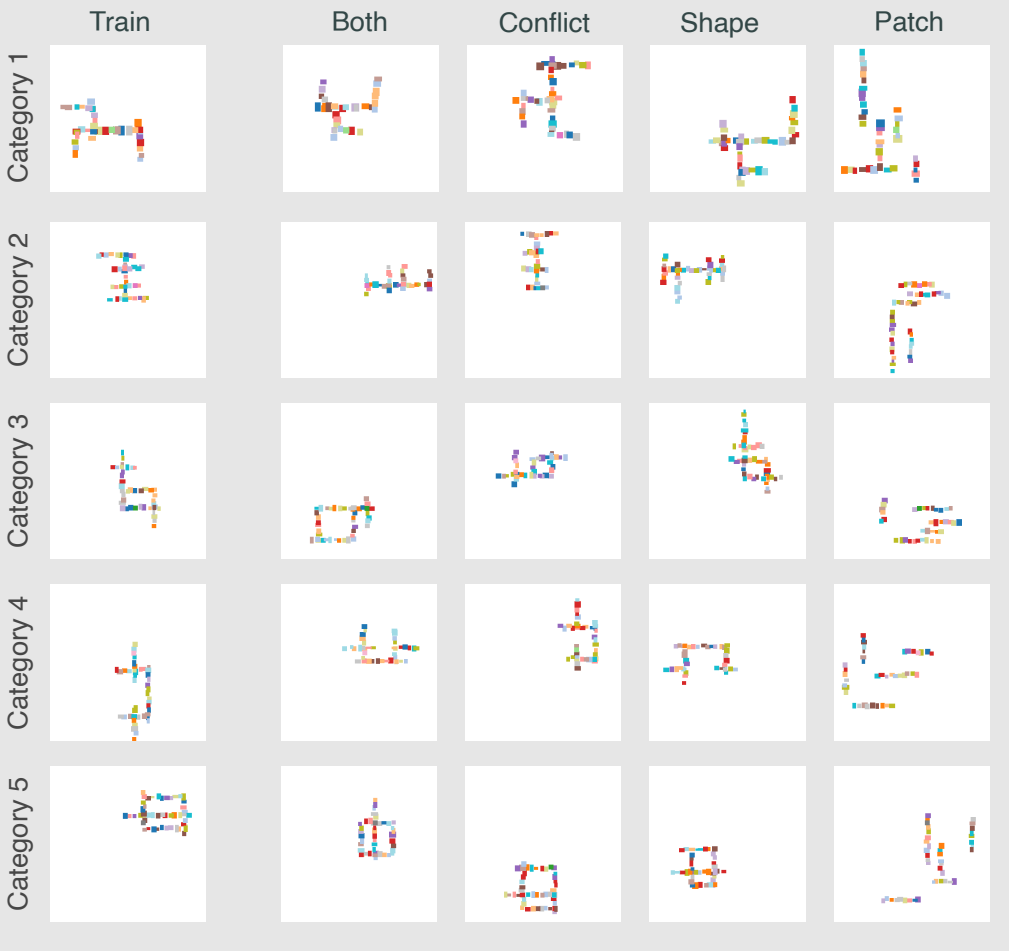


Fig S8. Examples of stimuli in Experiment 1 (patch). In each row we show (from left to right) an example image from the training set, `Both` condition, `Swap` condition, `Shape` condition and `Non-shape` (`Patch`) condition for a category. Each image in the training set contains a diagnostic patch of a certain colour that is present at a category-specific location. Additionally, all training images in Experiment 1a and 80% of images in Experiment 1b have a diagnostic shape. Images in the `Both` condition contain both these features. Images in the `Swap` condition contain the shape from one category but diagnostic patch from another category. Images in the `Shape` condition contain the shape feature but none of the diagnostic patches. Images in the `Patch` condition contain the diagnostic patch but none of the shapes from the training set.

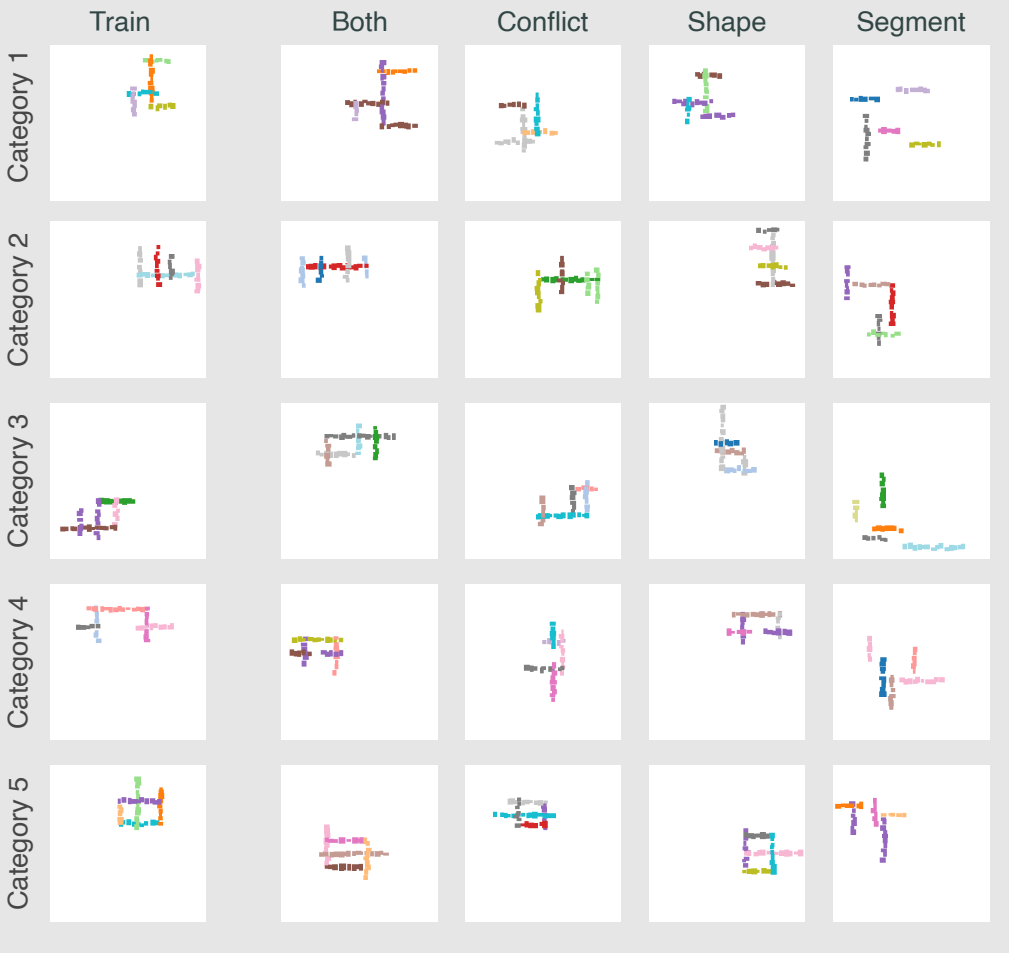


Fig S9. Examples of stimuli in Experiment 2 (segment). In each row we show (from left to right) an example image from the training set, `Both` condition, `Swap` condition, `Shape` condition and `Non-shape` (`Segment`) condition for a category. Each image in the training set contains a diagnostic segment of a category-specific colour. Only images of this category have a segment of this colour. Additionally, all training images in Experiment 2a and 80% of images in Experiment 2b have a diagnostic shape. Images in the `Both` condition contain both these features. Images in the `Swap` condition contain the shape from one category but diagnostic segment from another category. Images in the `Shape` condition contain the shape feature but none of the diagnostic segments. Images in the `Segment` condition contain the diagnostic segment but none of the shapes from the training set.

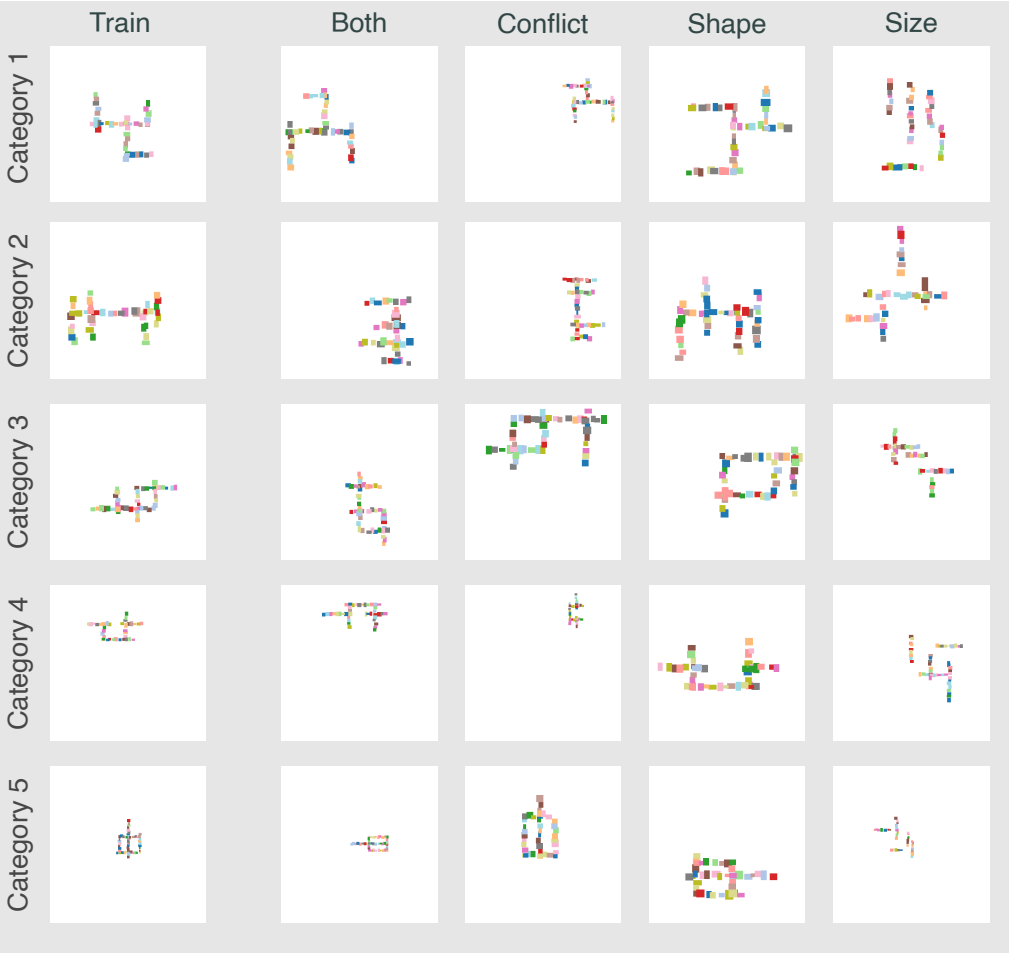


Fig S10. Examples of stimuli in Experiment 3 (size). In each row we show (from left to right) an example image from the training set, `Both` condition, `Swap` condition, `Shape` condition and `Non-shape` (`Size`) condition for a category. The average size of all images in the training set is diagnostic of the category. That is, different categories have images that have different average size of patches. Additionally, all training images in Experiment 3a and 80% of images in Experiment 3b have a diagnostic shape. Images in the `Both` condition contain both these features. Images in the `Swap` condition contain the shape from one category but diagnostic size from another category. Images in the `Shape` condition contain the shape feature and the average size of patches is larger than the diagnostic size of any category in the training set. Finally, the `Size` condition contains images where the average size of patches is diagnostic but shape is not.

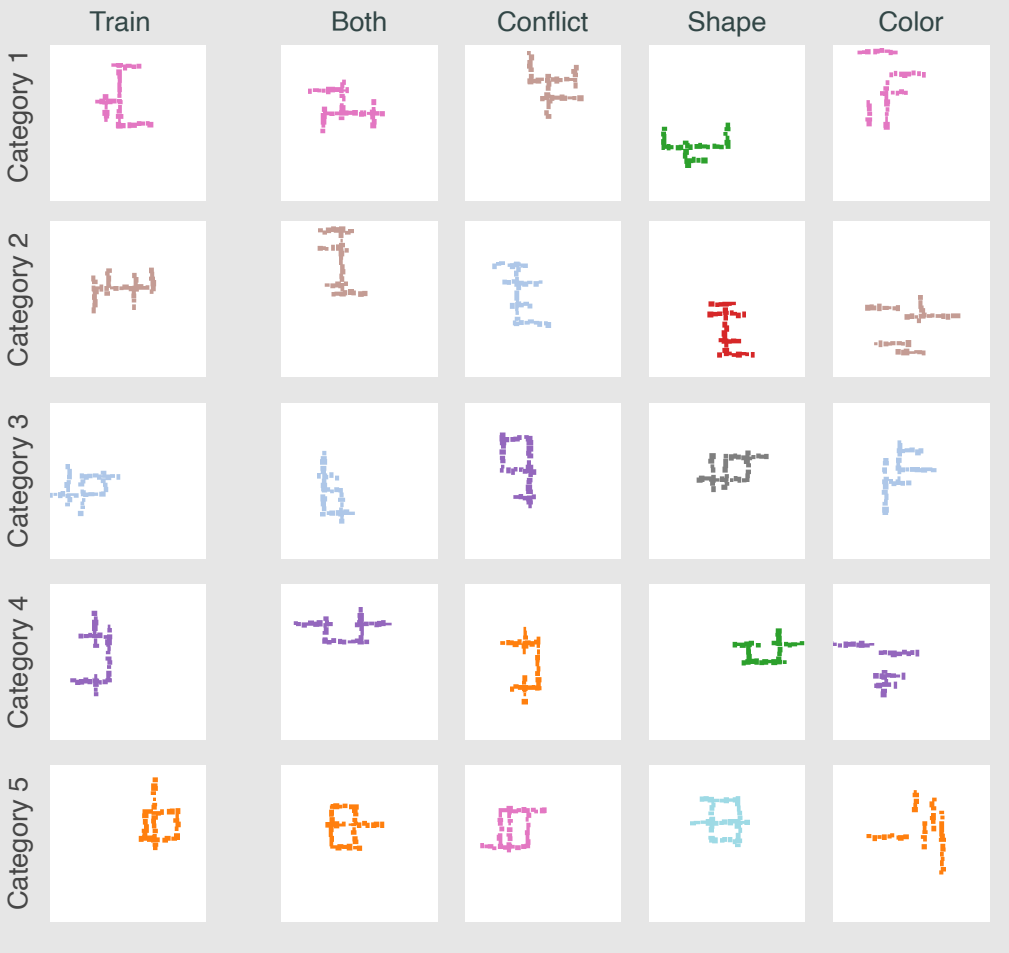


Fig S11. Examples of stimuli in Experiment 4 (colour). In each row we show (from left to right) an example image from the training set, `Both` condition, `Swap` condition, `Shape` condition and `Non-shape` (`Size`) condition for a category. All patches in an image have the same colour. This colour is diagnostic of an image's category in the training set. Additionally, all training images in Experiment 4a and 80% of images in Experiment 4b have a diagnostic shape. Images in the `Both` condition contain both these features. Images in the `Swap` condition contain the shape from one category but diagnostic colour from another category. Images in the `Shape` condition contain the shape feature and a colour that is not diagnostic of any category in the training set. Finally, the `Colour` condition contains images with no coherent shape but where the colour of segments is diagnostic of the category.

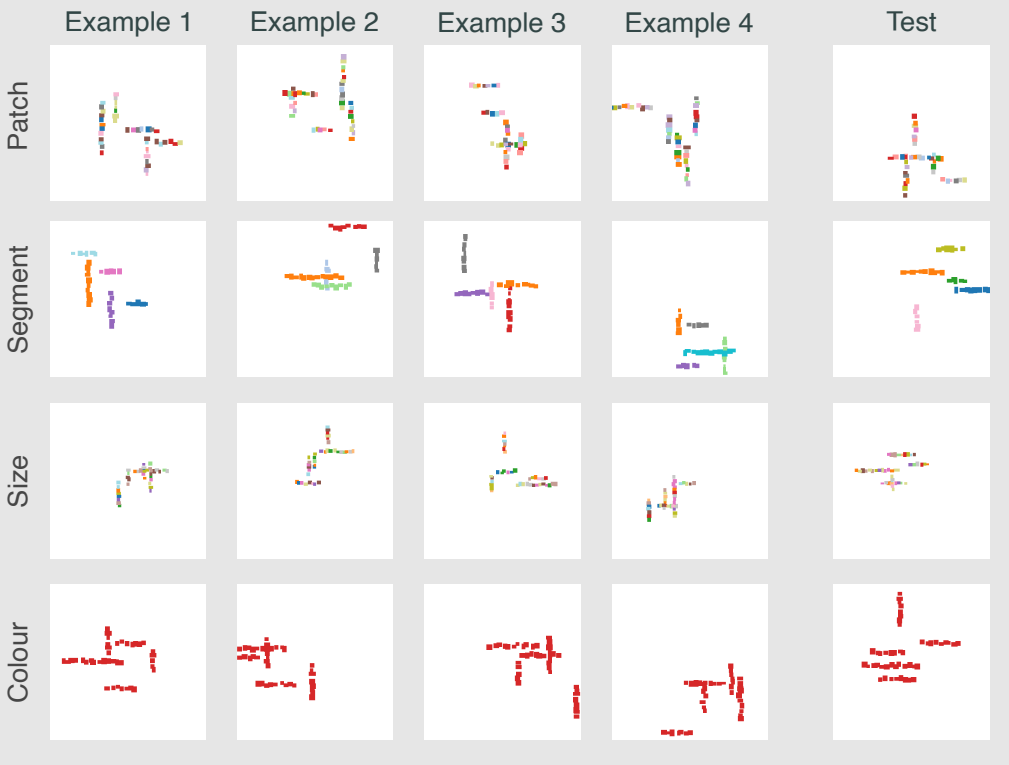


Fig S12. Examples of stimuli in Experiment 5 and 6 (no shape). Each row shows four examples from the training set that have the same category label as well as one example from the test set with the same label. The four rows correspond to the four conditions. In row 1, the predictive feature is patch location. In row 2, the predictive feature is colour of one of the segments. In row 3, the predictive feature is average size of patches. And in row 4, the predictive feature is colour of all patches.