This electronic thesis or dissertation has been downloaded from Explore Bristol Research, http://research-information.bristol.ac.uk

Author:
Dujmovic, Marin

Title:
Examining visual representations in mind and machine

General rights
Access to the thesis is subject to the Creative Commons Attribution - NonCommercial-No Derivatives 4.0 International Public License. A copy of this may be found at https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode This license sets out your rights and the restrictions that apply to your access to the thesis so it is important you read this before proceeding.

Take down policy
Some pages of this thesis may have been removed for copyright restrictions prior to having it been deposited in Explore Bristol Research. However, if you have discovered material within the thesis that you consider to be unlawful e.g. breaches of copyright (either yours or that of a third party) or any other law, including but not limited to those relating to patent, trademark, confidentiality, data protection, obscenity, defamation, libel, then please contact collections-metadata@bristol.ac.uk and include the following information in your message:

•Your contact details
•Bibliographic details for the item, including a URL
•An outline nature of the complaint

Your claim will be investigated and, where appropriate, the item in question will be removed from public view as soon as possible.

University of Bristol Thesis

Examining visual representations in mind and machine

By

MARIN DUJMOVIĆ

School of Psychological Science
UNIVERSITY OF BRISTOL

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of DOCTOR OF PHILOSOPHY in the Faculty of Life Sciences.

MARCH 2023

Word count: 36345

Abstract

Deep learning has quickly become the dominant approach in machine learning and successes have led to increasing interest in modeling human cognition via deep learning models. Deep convolutional neural networks have been hailed as the best models of human visual processing based on state-of-the-art performance on large-scale benchmarks as well as impressive scores on neural predictivity and representational similarity analysis measures. In this thesis I test claims about the similarity of visual representations between deep neural networks and humans. First, through a series of experiments, I show that humans do not intuitively understand how neural networks classify adversarial images (stimuli designed to fool neural networks) and that these types of stimuli do not provide insight into human vision as has recently been claimed. Next, human and network inductive biases are explored by generating datasets which allow manipulation of both the type and statistics of predictive features. Findings show human shape bias remains robust in novel learning environments while networks (even ones pre-trained to develop shape bias) adapt to the statistics of the new environment (learning to classify based on the most predictive feature). Additionally, when shape was as predictive of category membership as more local features, an inductive bias towards more local features was observed in networks. Finally, a series of simulations demonstrate that high representational similarity analysis (RSA) scores can be achieved between systems that represent stimuli in qualitatively different ways. While high RSA scores will be a feature of models that truly capture human-like visual representations, they are not sufficient to claim a model does so. Overall, findings presented in this thesis highlight the importance of systematic experimental scrutiny in light of engineering developments outpacing scientific research. This is key if deep learning models are going to be properly evaluated in hopes of providing insight into human cognition.

DEDICATION AND ACKNOWLEDGEMENTS

First and foremost I would like to thank Jeffrey Bowers and Gaurav Malhotra, my supervisors and main collaborators. Jeff as the lead on the *Generalization in Mind and Machine* project has provided all the support for quality work and all the space for independent ideas as well as contradicting opinions. Gaurav has been invaluable to me and to the project generally with insightful feedback, novel ideas, support and attention to detail. I would also like to thank each and every member of the *Generalization in Mind and Machine* project for providing an enriching working environment and valuable feedback throughout. Finally, I would like to thank Peter Allen. Teaching alongside, and collaborating with, Peter has been a pleasure as well as providing me with a mental counterweight of sorts to work on the *Generalization in Mind and Machine* project. Without that experience these past few years would have been far less fulfilling and enjoyable.

Author's declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Marin Dujmović DATE: 10/12/2022

TABLE OF CONTENTS

LIST OF TABLES

List of Figures

Chapter 1

Introduction

The emergence of deep learning systems as the dominant approach to machine learning has propelled AI research exponentially in the past decade. Successes in areas such as object recognition and natural language processing have led to comparisons with human cognition. These developments have been the cause of much excitement both within and outside the scientific community.

1.1 Background

The `ImageNet` large scale visual recognition challenge evaluates algorithms for object classification on millions of naturalistic images. Machine learning systems take training images as inputs, learn to classify them into one of a thousand categories and are then tested on a held-out test set [154]. Top-5 accuracy is usually the choice measure of classification performance. This type of accuracy refers to whether the target class is amongst the top 5 choices a system makes for a given image. In 2012, the winning system boasted a Top-5 accuracy rate of 84.685%, beating out the runner up by a massive 10.857%. The system in question was a deep convolutional neural network - AlexNet [100]. AlexNet is an end-to-end deep neural network consisting of convolutional layers, max pooling operations and three fully connected layers (Figure 1.1).

Though AlexNet sparked the latest move towards deep learning, it was an evolution of models inspired by findings form neuroscience about visual systems of mammals. Hubel and Wiesel [75] examined responses of neurons in the primary visual system of cats to discover neurons broadly group into simple and complex cells. Simple cells responded optimally to light bars at specific locations and orientations. Complex cells shared some properties of simple cells such as being sensitive to orientation of the stimulus, but responded just as strongly to stimuli in different (though spatially adjacent) locations. They concluded that simple cells represent early stages

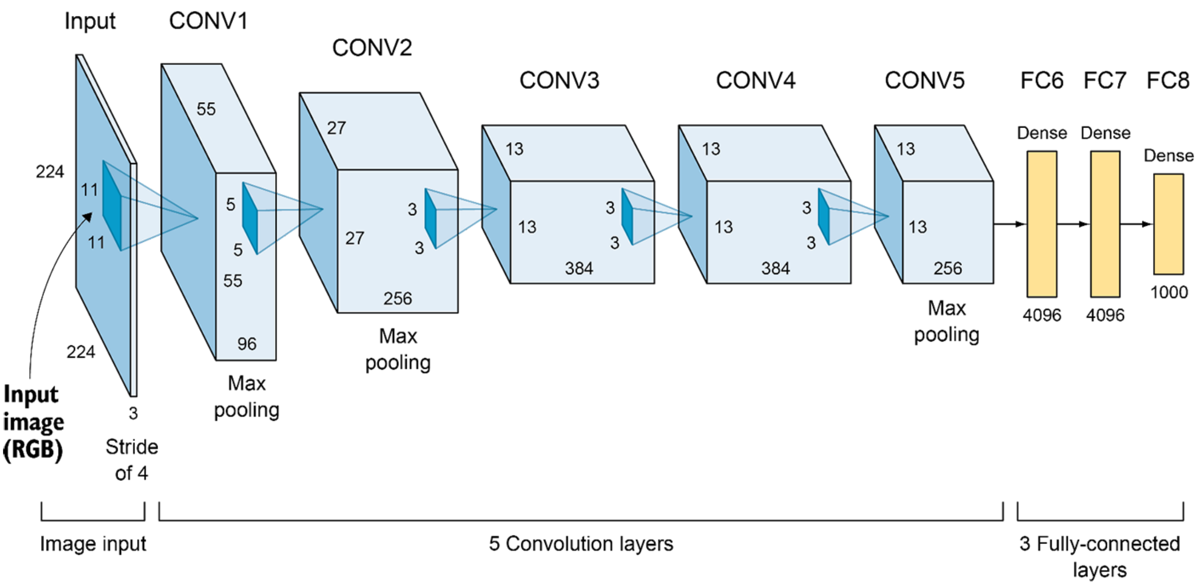


Figure 1.1: **Schematic of the AlexNet architecture.**

of organization with complex cells receiving inputs from multiple simple cells. These findings directly inspired Fukushima [46] to develop the Neocognitron. The model consists of multiple modules with each module containing S-layers and C-layers made up of cells corresponding to simple and complex neurons found by Hubel and Wiesel [75]. The first S-layer receives the 2D image as an input and responses of C-cells are nonlinear combinations of several S-cell projections. Receptive fields increase in size as the depth of the network increases (the number of modules). The final module C-cells have receptive fields encompassing the entire area of the input with each C-cell responding to a single input pattern. This network was found to do well at distinguishing between a set of letters (e.g., X, Y, Z, W) as well as a set of digits (0 to 9). Such hierarchical models became more common, including convolutional neural networks. Early convolutional networks were followed by the success of LeCun et al. [107] who applied the backpropagation algorithm [153] so that the convolutional kernels are learned rather than hand coding them. The network succeeded at recognizing handwritten digits (5% misclassification on the test set), performing better than unconstrained fully-connected networks.

While deep networks, convolutions, pooling operations, backpropagation and other concepts implemented in modern networks have been around for a long time, hardware advances and optimizations started a new deep learning revolution. Since AlexNet's success on the ImageNet challenge, companies like NVIDIA have spearheaded further hardware advances with every generation of GPU being more efficient at the kinds of computations that benefit deep learning. This has led to development of many different architectures beyond convolutional neural networks, both for machine vision and other areas like language processing. For example, transformer architectures, originally developed for text translation [177] have been applied to image recognition in the form of vision transformers, achieving state-of-the-art performance on naturalistic datasets like ImageNet [8, 31].

1.2 Similarities between human vision and deep convolutional networks

Since AlexNet won the ImageNet challenge, and particularly since the annual challenge was concluded in 2017 (with the top performing classification model achieving 97.749% accuracy), claims that these types of models are *the best* models of human vision have been made by many researchers and lab groups [102, 119, 171, 189].

There are a number of research paradigms which have been utilized in order to compare primate visual processing with convolutional neural networks. Roughly, these approaches can be grouped into behavioral-level comparisons and neural-level comparisons. The most basic comparison comes from performance on classification of naturalistic stimuli such as the images found in ImageNet dataset. Human-level performance has been matched and even surpassed using naturalistic stimuli of varying levels of difficulty [83]. Matching accuracy, however, is a rather crude metric, but further evidence for similarity was found by investigating error patterns. These misclassifications can be analyzed in a number of ways. For example, correspondence of confusion matrices can be evaluated. A confusion matrix quantifies how often stimuli from one category are misclassified as each of the remaining, non-target, categories. Rajalingham et al. [145] presented humans, monkeys and a set of deep neural networks with naturalistic stimuli from a number of categories (e.g., ‘camel’, ‘dog’, ‘elephant’, ‘rhino’, ‘pen’). They found good correspondence between deep networks and human category-level misclassifications. Humans were more likely to misclassify an ‘elephant’ as a ‘rhino’ than they were to misclassify an ‘elephant’ as a ‘pen’ etc. Deep convolutional networks mimicked the patterns of misclassiifications well, particularly when compared to a low-level V1 model and discriminability based on pixel values of the input images.

To further add to evidence from accuracy and confusion matrix analysis, researchers have studied human similarity judgments. Jozwik et al. [78] gathered similarity judgments from human participants for 92 images from a number of categories (e.g., ‘frog’, ‘lion’, ‘human face’) and compared them to similarities as measured via internal activations of deep neural networks (Euclidean distance between internal activations for each pair of stimuli of AlexNet and VGG-16). Results showed that most of the explainable variance (the percentage of variance one would expect could be explained from one human to another) of human similarity judgments can be explained by deep neural networks. Additionally, they showed that later layers of deep networks explain more variance in human similarity judgments when compared to early layers.

Similarly, human typicality judgments can be compared to network classification confidence. Lake et al. [104] presented both humans and a set of ImageNet pre-trained convolutional neural networks with 16 images from each of 8 categories (e.g., ‘banana’, ‘bathtub’, ‘coffee mug’). Participants had to judge how typical each image is of the category it comes from. Networks completed a classification task, but rather than simply recording classification decisions, confidence for each image was recorded (both raw class scores and normalized class probability from a softmax layer). Finally, a rank-order correlation between typicality scores and network classification confidence was computed to estimate the similarity. They found that network classification scores correlated well with human typicality judgments (average correlation of 0.71), especially compared to a baseline SIFT model (correlation of 0.28). When further analyzing internal layers of the network rather than the output, as was the case with similarity judgments, results showed better correspondence of deeper layers with typicality judgments.

There are quite a few other lines of research in which various degrees of similarity between neural networks and human behavior have been observed. For example, replicating Gestalt phenomena (such as the law of closure [87]) and investigating how classification accuracy decreases in networks for distortions that also reduce human performance [48, 179]. These studies find some similarity, for example, image degradation decreases accuracy for both humans and convolutional networks, but the effects are much more pronounced for networks both in the types of degradation leading to lower performance and the scale of performance loss.

The second, and perhaps more impressive, set of studies focuses on neural data from primates rather than behavioral outputs. There are two prominent approaches that attempt to measure the correspondence between neural activity (as measured by fMRI, EEG, MEG, single cell recordings and other measures) with internal representations (activations) from deep neural networks. One approach does so by attempting to predict neural activation patterns (e.g., from primates) from activation patterns of neural networks. This is usually done by fitting data from a training set using a multivariate regression model in which network activations are the predictors and neural activations the criterion. Once a model has been fitted to the data, it is evaluated on a held-out set of stimuli. These *neural predictivity* measures can vary in the exact steps by which they are computed but the most well-known of these measures is Brain Score [159]. Both networks and primates are shown a set of images. Activity patterns are recorded from layers of the neural network and from regions of interest of the primate brain (for example, IT cortex). Next, activity patterns from the deep nets are fitted to the activity patterns from brain areas - in the case of Brain-Score this is done via PLS regression. Studies utilizing such methods have shown performance-optimized deep networks explain more variance in neural activity when compared to competing models such as HMAX [159, 182]. These studies also highlight that early layers of deep networks better predict neural activity of early vision areas such as V1 and V2, whereas deeper layers do so for neural activity of later areas such as V4 and IT [35, 60, 162, 182].

The second approach is called representational similarity analysis (RSA) which compares *representational geometries* of two systems on a set of stimuli. The representational geometry is computed by measuring distances between activation patterns for all pairs of stimuli. The distance metric can vary (e.g., 1- Pearson correlation, Euclidean distance etc.) but all pair-wise distances are represented in a matrix (the representational dissimilarity matrix - RDM). RDMs capture the similarities of visual representations within a system. For example, neural activity patterns for images of `dogs` will be more similar to patterns for images of `cats` than they will be to patterns for images of `planes`. Such RDMs can be computed from internal activation patterns of deep neural networks at various depths and then a correspondence between human RDMs and network RDMs can be computed (most commonly as a Spearman rank-order correlation of the RDMs). This correspondence is then the RSA score between two systems (for more detail on the method see Chapter 4). Studies using this method have mirrored results from neural predictivity measures; deep networks achieve higher RSA scores with primates than competing models [25, 82, 84, 86, 182]. The scores can reach ceiling levels which are RSA scores one can expect when comparing an RDM from one human to an RDM of another. The hierarchical pattern of correspondence observed with neural predictivity measures (early layers corresponding to V1 and V2, deeper layers to V4 and IT) have been mirrored using RSA as well [25, 82]. An example of these patterns can be seen in Figure 1.2. These results come from Khaligh-Razavi and Kriegeskorte [82] in which humans and AlexNet were presented 92 images from various animate and inanimate categories. RSA scores were computed as the Kendall τ_a correlation between human RDMs and RDMs from different layers of AlexNet. The main findings are that best performing layers from AlexNet reach noise ceiling and that early layers correspond to early visual cortex better while deeper layers correspond better to IT.

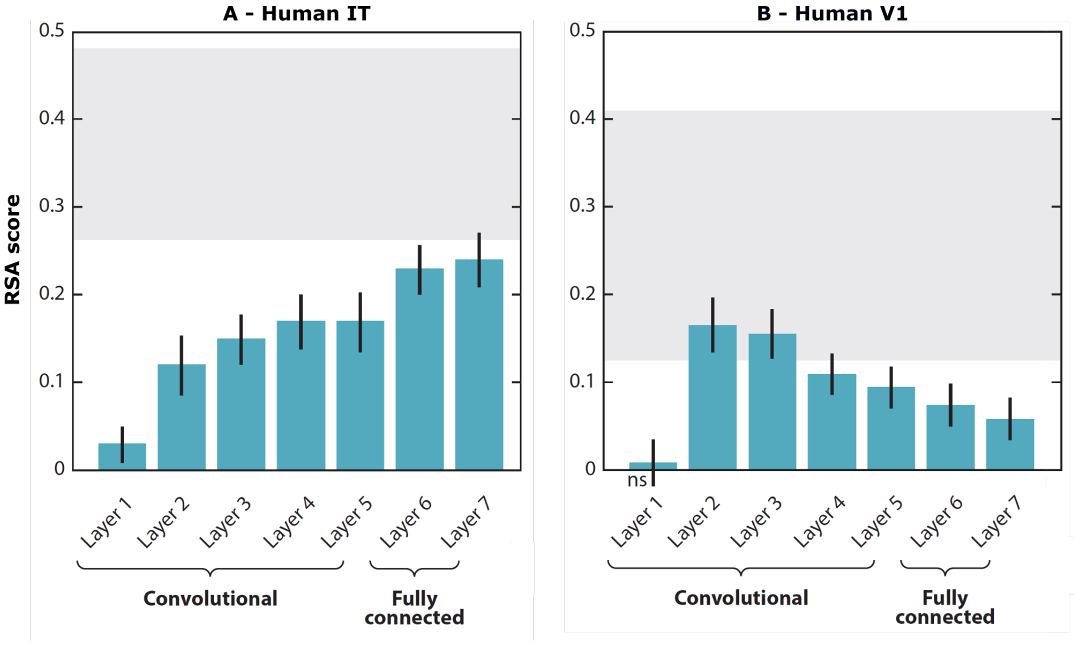


Figure 1.2: **RSA scores of AlexNet layers with neural activity from human IT (A) and V1 (B).** RSA scores between AlexNet layers and human neural fMRI patterns were computed as the Kendall τ_a between RDMs. The shaded region represents the estimated noise ceiling (expected human to human RSA scores). The figure was adopted from [82].

1.3 Dissimilarities between human vision and deep convolutional networks

Though the findings observing similarities between human vision and deep convolutional networks have been promising, they are far from unambiguous. As shown in the previous section, some studies find good correspondence between error patterns, not just accuracy, between neural networks and human classification [145]. However, Geirhos et al. [49] utilized a different measure to measure classification overlap. They introduce the measure of error consistency which corrects for accuracy (high accuracy by necessity means high overlap for many individual stimuli) and for overlap expected by chance and without averaging across human participants. They found that deep neural networks were quite consistent amongst themselves, but not with humans. This indicates that classification strategies are quite different rather than similar.

Similarly, results reported by Jozwik et al. [78] do show significant correspondence of similarity judgments from humans and deep neural networks (but the level of this correspondence is still well below ceiling - correspondence among human participants). Further, when utilizing a different dataset and method, Rosenfeld et al. [152] do not find that deep network representations account for human similarity judgments. They use the `Totally-Looks-Like` dataset (see examples in Figure 1.3). The dataset was downloaded from a popular website where visitors posted amusing pairs of images that look alike based on various types of features. In the actual study, participants were shown one of the images from a pairing alongside 5 candidate images. Distractor images were chosen differently in different experimental setups (e.g., random, based on generic features etc.) Their task was to choose which of the 5 images was most similar to the target image. The 'correct' response was the image which was in the original pair. Activation patterns from penultimate layers of a number of deep networks (AlexNet, VGG-16, ResNet-50 etc.) were extracted and distances between activation patterns for target and candidate images computed to check if the distance between the target and the matched pair was smaller than the distance between the target and distractor images. Results showed human selections corresponded well with the original dataset taken from the website (e.g., 82.5% when distractos chosen randomly), which was not the case for machine-human correspondence (25% in the same condition).

Beyond comparisons to human behavior seemingly depending on datasets and methodology, there are many other examples of dissimilar behavior. Deep learning models are susceptible to adversarial attacks (stimuli which are designed to fool the network) [56, 57]. Adversarial attacks consist of images which are classified as one category by humans but as a different category by neural networks. There is a wide range of different adversarial attack. Perhaps the most well-known attacks impose perturbations on otherwise normal images in order to induce misclassification. For example, a specifically generated high-frequency noise mask added to an image of a panda to make networks classify it as a gibbon [57]. Another class of adversarial

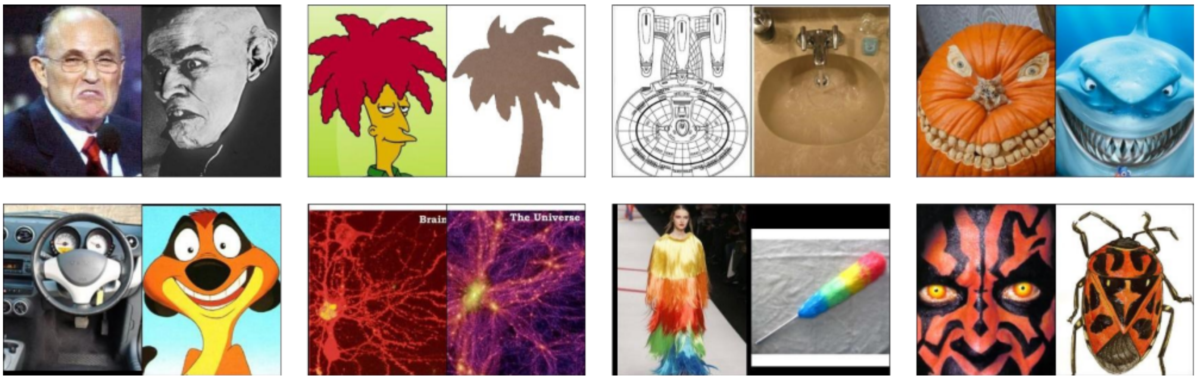


Figure 1.3: **Examples of image pairs from Rosenfeld et al. [152].** Image pairs generated by visitors to a popular website may be similar in a variety of ways (global shape, texture, colour etc.).

attack are so called *fooling* images which look like random patterns of features unrelated to the deep network classification [127] (examples of both types can be found in Chapter 2 - Figure 2.1).

Additionally, deep convolutional networks have been shown to exploit *shortcuts* which would lead to solving the task at hand [47]. Shortcuts can be defined as "decision rules that perform well on standard benchmarks but fail to transfer to more challenging testing conditions" ([47], p. 665). This includes learning to classify images based on any feature predictive of category membership during training. For example, Malhotra et al. [116] inserted a single pixel in images of the training dataset such that the location of this pixel was category-correlated in the CIFAR-10 dataset. For example, all 'cat' images had a white pixel at the same location, and all 'dog' images had a white pixel at a different location. The networks were then trained on this modified dataset and it was shown that they classified images by learning the location of the predictive pixel. Networks (VGG-16 and ResNet-101) performed nearly perfectly in a test set which retained the predictive pixels, but failed when the test set did not contain these predictive pixels. This confirms that the networks ignored other information present in the images during training in favour of the highly predictive pixels. In general, deep convolutional networks show high sensitivity to local features such as texture. Geirhos et al. [50] found that deep networks have a texture bias, as opposed to shape bias which humans exhibit. Similarly, Baker et al. [6] found that deep convolutional neural networks do not classify based on global shape. To counter this, Geirhos et al. [50] trained deep networks using the style-transfer approach. They created a style-transfer version of the ImageNet dataset (for an example see Figure 1.4). In each style-transfer image, global shape was retained while different textures were superimposed on the image. This meant that textures were no longer predictive of category membership, leaving global shape as the feature networks needed to learn during training.

Results indeed showed that networks trained using a style-transfer version of the dataset developed shape sensitivity. These networks performed well on both stylized images and standard ImageNet images (79% and 82.6% respectively for a ResNet-50). On the other hand, networks trained on the normal ImageNet images performed well on the standard test set (92.9%) but

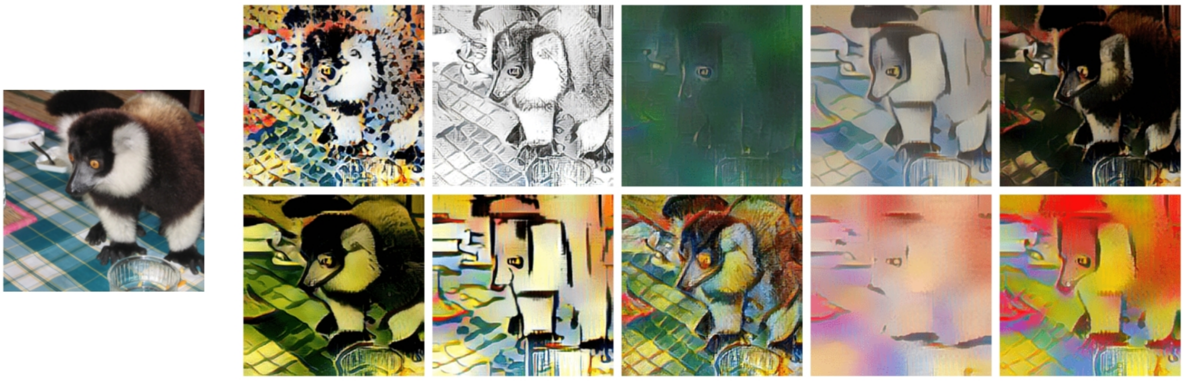


Figure 1.4: **Style-transfer training stimuli from Geirhos et al. [50].** An image from the `ImageNet` dataset (left) and 10 with the same shape/content but different texture/style (right).

much worse on the style-transfer test set in which texture was not predictive of category (16.4%). Style-transfer trained networks were also more robust to various types of image distortions such as adding uniform noise or changes in contrast which was more in line with human participants. Apart from style-transfer, data augmentation, initialization and hyper-parameter choices have been shown to increase shape bias [71, 73].

Even though studies have shown networks can develop shape sensitivity the nature of the shape representations seems to be dissimilar to that of humans. Malhotra et al. [114] tested whether networks trained to develop shape were sensitive to categorical changes of relations between object parts. Human participants represent shapes of objects based on relations between features. In contrast, CNNs in Malhotra et al. [114]showed no evidence of such a relational representation. They used the same type of stimuli as Stankiewicz and Hummel [170] which consisted of different objects and two modified versions of the objects (Figure 1.5). In one variant, a categorical relationship between object parts was changed. In the second variant, categorical relations between object parts were preserved, but coordinates of some parts were shifted. Humans have been shown to be more sensitive to relational changes when compared to coordinate changes [170]. In contrast, comparing internal representations of neural networks trained to have a shape bias showed that representations for both variants were equally similar to the representation of the original object. Malhotra et al. [114] concluded that this was likely because networks rely on features of the proximal stimulus (the image itself) which changed equally for both relational and non-relational deformations, while humans infer properties of the distal stimulus (like global shape).

There are a host of other dissimilarities of deep networks when compared to humans such as failure to exhibit uncrowding effects [30], having the capacity to learn random stimuli such as TV static [185], being much worse than humans at handling degraded images [50, 51], images with occluded objects [188], and objects in strange poses [2, 188], failing to generalize to novel objects in the same/different task [142], being poor at combinatorial generalization [122] etc. For a more detailed review see Bowers et al. [17].

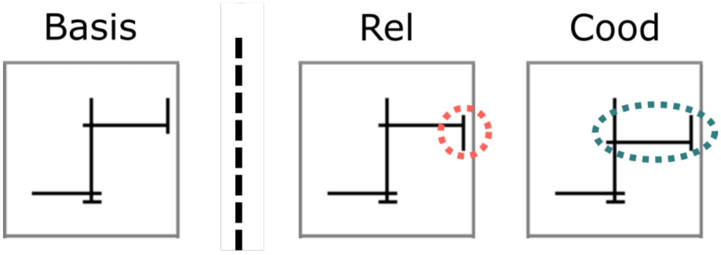


Figure 1.5: **Example of an object and modified variants from Malhotra et al. [114].** The basis object was modified to create two variants. (Rel) The first modification consisted of a categorical change of a relation between parts of the object. (Cood) The second modification preserved all relations but coordinates of some elements were shifted.

Evidence for similarities between internal representations of networks and humans / monkeys seems to be less controversial, but these findings have been based on a limited number of datasets and neural data. One set of findings that contradicts these similarities comes form Xu and Vaziri-Pashkam [180] who conducted an fMRI study utilizing a larger set of naturalistic and synthesized stimuli. Xu and Vaziri-Pashkam [180] identified potential issues with the approach taken in previous research of this type. First, the event-related fMRI approach produced data with a low signal to noise ratio and second, regions of interest were identified anatomically rather than functionally for each individual participant. They found that the previous pattern of early layers corresponding well to early visual processing and deeper layers to later visual processing could not be replicated across datasets when measuring correspondence using RSA. The correspondence was both dataset and network-dependent. For example, in AlexNet early layers did correspond better with early and deeper layers with later visual areas, but DenseNet-101 did not have the same pattern with all layers corresponding fairly poorly with all areas of the ventral visual stream. On te other and, DenseNet-101 does show a similar pattern to AlexNet for low-pass filtered images. However, the most important finding was that none of the deep networks tested reached beyond 60% of the total explainable variance for later regions of visual processing such as IT. This is in sharp contrast to previous findings as the ones shown in Figure 1.2 in which, at least, the fully connected layers reach ceiling RSA scores with human IT.

And while good prediction of neural activity will surely be a characteristic of the best model, it is not sufficient to infer a model is indeed mechanistically similar to the region of interest it is being compared with. A good example of why this is the case results from the rapid development of deep learning itself. Vision transformers were initially not comparable to convolutional neural network Brain-Score results, but have recently caught up [8]. This means we now have two different architectures of deep learning models which are both image-computable, trained end-to-end, achieve high object recognition performance and good neural predictivity. If one pursues these types of benchmarks in order to evaluate models, then we have already come to the point of them not being useful to adjudicate between models.

1.4 Aims and overview

The review presented in the sections above demonstrates there is a lively debate about how well deep convolutional neural networks model visual processing in the cortex. Further, it is plain that reliable inferences about similarities of processing and visual representations as well as adjudicating between models cannot be done based on one research paradigm. Therefore, during my PhD project, I chose to focus on, what are in my estimate, three key areas of the ongoing debate about the relation between convolutional networks and biological systems.

First, one key part of the debate has concerned adversarial images and has become more contentious recently. As discussed above, adversarial images have traditionally been seen as a striking example of dissimilarities between human visual processing and convolutional networks. However, recent studies conclude that adversarial images, rather than being striking examples of differences, could be used to better understand human vision [36, 61, 187]. Zhou and Firestone [187] claim humans can decipher adversarial images and posses a kind of 'machine theory of mind'. Therefore, in Chapter 2 I address these claims through a re-assessment of the methodology and data analysis that led to these claims. The re-assessment is followed by a series of experiments designed to test various factors contributing to correspondences between human and network classification of these images.

Second, when human-like classification behavior is replicated by networks, there is a debate whether the behavior is based on human-like visual representations. Specifically, it has long been established that humans primarily rely on shape for object recognition. In contrast, `ImageNet` trained convolutional networks rely on local features such as texture. As discussed above, altering the training regime (such as style-transfer augmentation as in [50]) can result in networks developing shape bias. Therefore, in Chapter 3 I tested whether shape bias is similar in humans and neural networks. Specifically, I asked the question: how robust are human and network inductive biases in new learning environments. This was done by generating a novel dataset in which predictive features can be manipulated. Humans, `ImageNet` and style-transfer trained networks were then trained and tested when varying the statistics of the training set.

Finally, perhaps the most impressive line of research comparing brains and convolutional networks looks at similarity of internal representations. Particularly, I focus on RSA as an approach of comparing visual representations between two systems. Recent findings by Xu and Vaziri-Pashkam [180] seem to contradict previous favorable results and indicate impressive similarity as measured by RSA might be a function of methodological decisions such as the choice of dataset. Therefore, in Chapter 4 I investigated whether obtaining a high RSA score necessarily indicates similarity of visual representations between networks and biological systems, and more generally, what types of inferences one can draw based on comparing representational geometries of two systems.

These three projects have in common that they tackle recent developments in long-standing areas of human-network comparisons. The choice of only investigating convolutional neural networks and not other classes of deep learning models is based on consequences of engineering solutions outpacing research evaluating these solutions as models of human vision. First, as mentioned before, architectures like vision transformers are becoming more popular at the expense of (by deep learning standards) old convolutional networks. What were hailed as the best models of human vision just a year or two ago might be abandoned for new architectures which achieve similar performance, neural predictivity and RSA. This is seemingly done without rigorous testing and adjudicating between the architectures. They cannot both be the best models of human vision, and if newer architectures are better then old ones should no longer be hailed as the best.

Second, successes and claims based on limited research are less likely to be scrutinized to an adequate degree. Certain claims become generally accepted without going through what we would usually consider enough scrutiny. Hopefully, the increase of interest in using deep learning models to study human cognition will result in a large enough body of work to provide the required scrutiny for claims going forward. However, rather than hoping sheer volume will lead to diverse, systematic experimentation - that should be the dominant approach from the outset.

Given these trends the work in this thesis has been limited to using feed-forward deep convolutional neural networks rather than exploring a large number of architectures in less detail. The aim was to test some of the strong claims being made about similarity between convolutional networks and human vision. Instead of focusing on one metric or methodological paradigm, the goal was to take a multifaceted approach, incorporating experimentation and stimulus manipulation at different levels of analysis (both performance measures and analysis of internal representations).

Chapter 2

Decoding adversarial images

A specific class of stimuli had caught the attention of researchers as soon as deep CNNs became the leading class of deep learning models. In object recognition, these stimuli are collectively called *adversarial images* [56, 57, 172]. Originally, adversarial images were defined as stimuli designed to cause the model to make a mistake (hence the term adversarial attacks also being prevalent). Usually, these images will include various perturbations (e.g., noise masks) to otherwise normal images which push the model to missclassify them. Since that time, the type of stimuli that can be labelled as adversarial has grown to include *fooling images* - which are not naturalistic stimuli with adversarial perturbations, but rather, images that do not contain what a human would classify as an object (of any class) but nevertheless result in a confident classification as a specific class by the model. Attacks can also be categorised depending on how much access one has to the targeted system. When the system is completely accessible (e.g., having access to model weights of a CNN), *white-box* attacks can be performed (e.g., gradient ascent), while *black-box* attacks are the ones in which the attacker only has access to the output of the system [1]. Attacks can also be classified as *targeted*, if they are designed to be classified as a specific target category, or *non-targeted*, if the goal is missclassification with no specific target category.

Adversarial attacks are clearly threats to robustness which can have costly consequences if encountered by security systems, autonomous vehicles and other systems. For example, Sharif et al. [165] demonstrate how adversarial masks can fool facial recognition systems, and Eykholt et al. [37] construct an adversarial street sign (e.g., a *stop* sign with perturbations that make deep learning models classify it as a 45 mph sign). Apart from being problematic in practical applications, adversarial attacks are problematic from the standpoint of using deep learning systems to model human visual processes. This is due to CNN behavior being completely counter-intuitive and contrasting with how humans perceive adversarial attacks. Recently, however, some researchers have made the claim that humans are susceptible to similar adversarial attacks as CNNs [36, 61, 187]. These researchers state that human classification performance is reduced when adversarial noise is superimposed on naturalistic stimuli [36], that adversarial noise can also influence neural firing rates in primate IT, and that humans generally agree with CNN classification of fooling images [187] (see Figure 2.1a for an example).

In this chapter, claims made by Zhou and Firestone [187] about general agreement between humans and CNNs on classification of fooling images are explored in detail. A re-assessment of the analysis from the original paper is followed by five experiments designed to ascertain the level of this agreement as well as potential causes.

This is a publication chapter - the chapter is a re-formatted and slightly edited version of the paper *What do adversarial images tell us about human vision* in eLife [33].

Contributions - this Zhou and Firestone [187] paper was brought to my attention by my co-authors (also serving as my supervisors) as a potential research project. As first co-author I initiated (and executed) the idea of re-assessing the findings from that paper. I identified factors to be experimentally explored and designed the experiments. I conducted all of the presented research and data analyses. Gaurav Malhotra generated stimuli for Experiment 4. All co-authors contributed to the write-up and revisions following reviews.

2.1 Introduction

Deep convolutional neural networks (DCNNs) have reached, and in some cases exceeded, human performance in many image classification benchmarks such as `ImageNet` [66]. In addition to having obvious commercial implications, these successes raise questions as to whether DCNNs identify objects in a similar way to the inferotemporal cortex (IT) that supports object recognition in humans and primates. If so, these models may provide important new insights into the underlying computations performed in IT. Consistent with this possibility, a number of researchers have highlighted various functional similarities between DCNNs and human vision [139] as well as similarities in patterns of activation of neurons in IT and units in DCNNs [181]. This has led some authors to make strong claims regarding the theoretical significance of DCNNs to neuroscience and psychology. For example, Kubilius et al. [102] write: “Deep artificial neural networks with spatially repeated processing (a.k.a., deep convolutional [Artificial Neural Networks]) have been established as the best class of candidate models of visual processing in primate ventral visual processing stream” (p.1).

One obvious problem in making this link is the existence of *adversarial* images. These are “inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake" [56]. Figure 2.1 shows examples of two types of adversarial images. On first impression, it seems inconceivable that these adversarial images would ever confuse humans. There is now a small industry of researchers creating adversarial attacks that produce images which DCNNs classify in bizarre ways [1]. The confident classification of these adversarial images by DCNNs suggests that humans and current architectures of DCNNs perform image classification in fundamentally different ways. If this is the case, the existence of adversarial images poses a challenge to research that considers DCNNs as models of human behaviour [e.g., 24, 29, 102, 138, 151], or as plausible models of neural firing patterns in primate and human visual cortex [e.g., 20, 25, 35, 82, 145, 182].

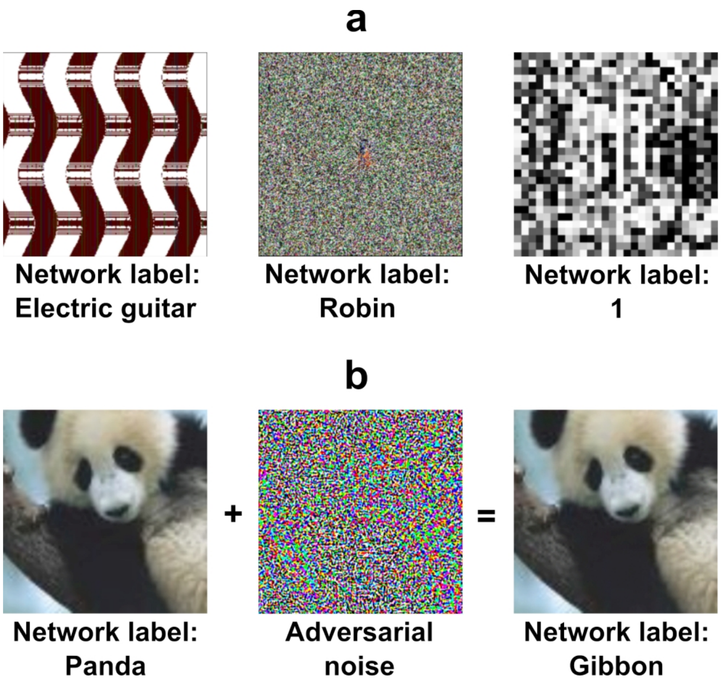


Figure 2.1: **Examples of two types of adversarial images.** (a) *fooling adversarial images* taken from Nguyen et al. [127] that do not look like any familiar object. The two images on the left (labelled 'Electric guitar' and 'Robin') have been generated using evolutionary algorithms using *indirect* and *direct* encoding, respectively, and classified confidently by a DCNN trained on `ImageNet`. The image on the right (labelled '1') is also generated using an evolutionary algorithm using *direct* encoding and it is classified confidently by a DCNN trained on `MNIST`. (b) An example of a *naturalistic adversarial image* taken from [57] that is generated by perturbing a naturalistic image on the left (classified as 'Panda') with a high-frequency noise mask (middle) and confidently (mis)classified by a DCNN (as a 'Gibbon').

However, some recent studies have suggested that there may, in fact, be theoretically relevant overlap between DCNNs and humans in how they process these adversarial images. Zhou and Firestone [187] (***Z&F*** from here on) recently reported that humans can reliably *decipher* fooling adversarial images that, on first viewing, look uninterpretable (as in Figure 2.1a). The authors took a range of published adversarial images that were claimed to be uninterpretable by humans and, in a series of experiments, they showed those images to human subjects next to the DCNN's preferred label and various foil labels. They reported that, over the course of an experimental session, a high percentage of participants (often close to 90%) chose the DCNN's preferred label at above-chance rates. Furthermore, they report evidence that humans appreciate subtler distinctions made by the machine rather than simply agreeing on the basis of some superficial features (such as predicting "bagel" rather than "pinwheel" when confronted with an image of a round and yellow blob). These results are important because they speak to an important theoretical question that ***Z&F*** pose in the first line of their abstract: "Does the human mind resemble the machine-learning systems that mirror its performance?" (p.1). The high level of agreement they report seems to suggest the answer is "yes".

Here we show that the agreement between humans and DCNNs on adversarial images was weak and highly variable between participants and images. The remaining agreement appeared to reflect participants making educated guesses based on some superficial features (such as colour) within images and the limited response alternatives presented to them. We then carried out five experiments in which we systematically manipulated factors that can contribute to an observed agreement between humans and DCNNs in order to better understand how humans interpret adversarial images. The experiments demonstrate that the overlap between human and DCNN classification is contingent upon various details of the experimental design such as the selection of adversarial images used as stimuli, the response alternatives presented to participants during the experiment, the adversarial algorithm used to generate the images and the dataset on which the model was trained. When we controlled for these factors, we observed that the agreement between humans and DCNNs dropped to near chance levels. Even when adversarial images were selected such that multiple DCNNs confidently assigned the same label to these images, humans seldom agreed with the machine label, especially when they had to choose between response alternatives that contained superficial features present within these images. We also show that it is straightforward to generate adversarial images that fool networks trained on `ImageNet` but are truly meaningless to human participants, irrespective of how the stimuli are selected or response alternatives are presented to a participant. We take the findings to highlight a dramatic difference between human and DCNN object recognition.

2.2 Results

2.2.1 Reassessing the level of agreement in Zhou & Firestone

Our first step, in trying to understand the agreement between humans and DCNNs observed by ***Z&F***, was to assess how well their methods reflect the degree of agreement between humans and DCNNs. ***Z&F*** conducted seven experiments in which they measured agreement by computing the number of trials on which the participants matched the DCNN's classification and working out whether this number is numerically above or below chance level. So in Experiment 3, for example, a participant is shown an adversarial image on each trial and asked to choose one amongst 48 labels for that image. Each trial was independent, so a participant can choose any of the 48 labels for each image. Chance level is 1/48, so if the participant chooses the same label as the DCNN on two or more trials, they were labelled as agreeing with the DCNN. In addition, half of the participants who agreed with the DCNN on only 1/48 trials were also counted towards the number of participants who agreed with the DCNN. When computed in this manner, ***Z&F*** calculated that 142 out of 161 (88%) participants in Experiment 3a, and 156 out of 174 (90%) in Experiment 3b agreed with the DCNN at above chance levels.

This is a reasonable way of measuring agreement if the goal is to determine whether agreement between humans and DCNNs is statistically above chance levels. However, if the goal is to measure the *degree of agreement*, this method may be misleading and liable to misinterpretation. Firstly, the rates of agreement obtained using this method ignore inter-individual variability and assign the same importance to a participant that agrees on 2 out of 48 trials as a participant who agrees on all 48 trials with the DCNN. Secondly, this method obscures information about the number of trials on which humans and DCNNs disagree. So even if every participant disagreed with the network on 46 out of 48 trials, the rate of agreement, computed in this manner, would be 100% and even a sample of blindfolded participants would show 45% agreement (see Methods). In fact, not a single participant in Experiments 3a & 3b (from a total of 335 participants) agreed with the model on a majority (24 or more) of trials, yet the level of agreement computed using this method is nearly 90%.

A better way of measuring the degree of agreement is to simply report the average agreement. This can be calculated as the mean percentage of images (across participants) on which participants and DCNNs agree. This method overcomes the disadvantages mentioned above: it takes into consideration the level of agreement of each participant (a participant who agrees on 4/48 trials is not treated equivalently to a participant who agrees on 48/48 trials), and it reflects both the levels of agreement and disagreement observed (so a mean agreement of 100% would indeed mean that participants agreed with the DCNN classification on all the trials). ***Z&F*** reported mean agreement for the first of their seven experiments and in Table 2.1 we report mean agreement levels in all their experiments. Viewed in this manner, it is clear that the degree of agreement in the experiments carried out by ***Z&F*** is, in fact, fairly modest and far from "surprisingly universal" (p.2) or "general agreement" (p.4) the authors reported.

2.2.2 Reassessing the basis of the the agreement in Zhou & Firestone

Although the mean agreement highlights a much more modest degree of agreement, it is still the case that the agreement was above chance. Perhaps the most striking result is in ***Z&F***'s Experiment 3 where participants had to choose between 48 response alternatives and mean agreement was ~ 10% with chance being ~ 2%. Does this consistent, above chance agreement indicate that there are common underlying principles in the way humans and DCNNs perform object classification?

In order to clarify the basis of overall agreement we first assessed the level of agreement for

Table 2.1: **Mean DCNN-participant agreement in the experiments conducted by Zhou & Firestone.**

Exp.	Test type	Mean agreement	Chance
1	Fooling 2AFC ^{N15}	74.18% (35.61/48 images)	50%
2	Fooling 2AFC ^{N15}	61.59% (29.56/48 images)	50%
3a	Fooling 48AFC ^{N15}	10.12% (4.86/48 images)	2.08%
3b	Fooling 48AFC ^{N15}	9.96% (4.78/48 images)	2.08%
4	TV-static 8AFC ^{N15}	28.97% (2.32/8 images)	12.5%
5	Digits 9AFC ^{P16}	16% (1.44/9 images)	11.11%
6	Naturalistic 2AFC ^{K18}	73.49% (7.3/10 images)	50%
7	3D Objects 2AFC ^{A17}	59.55% (31.56/53 images)	50%

[a] To give the a sense of the levels of agreement observed in these experiments, we have also computed the average number of images in each experiment where humans and DCNNs agree as well as the level of agreement expected if participants were responding at chance.

[b] Stimuli sources: N15 - Nguyen et al. [127]; P16 - Papernot et al. [131]; K18 - Karmon et al. [80]; A17 - Athalye et al. [3]

each of the 48 images separately. As shown in Figure 2.2, the distribution of agreement levels was highly skewed and had a large variance. There was a small subset of images that looked like the target class (such as the Chainlink Fence, which can be seen in Appendix A - Figure A.1) and participants showed a high level of agreement with DCNNs on these images. Another subset of images with lower (but statistically significant) levels of agreement contained some features consistent with the target class, such as the Computer Keyboard which contains repeating rectangles. But agreement on many images (21/48) was at or below chance levels. This indicates that the agreement is largely driven by a subset of adversarial images, some of which (such as the Chainlink Fence) simply depict the target class.

We also observed that there was only a small subset of images on which participants showed a clear preference amongst response alternatives that matched the DCNN's label. For most adversarial images, the distribution of participant responses across response alternatives was fairly flat (see Appendix A - Figures A.3-A.5) and the most frequent human response did not match the machine label even when agreement between humans and DCNNs was above chance (see Appendix A - Figure A.1). In fact, the label assigned to the image by DCNNs was ranked 9th (Experiment 3a) or 10th (Experiment 3b) on average. 75% of the adversarial images in Experiment 3a and 79.2% in Experiment 3b were not assigned the label chosen by the DCNN with highest frequency (Appendix A - Figures A.1 and A.2). This indicates that most adversarial images do not contain features required by humans to uniquely identify an object category.

Collectively, these findings suggest that the above chance level of agreement was driven by two subsets of images. A very small subset of images have features that humans can perceive and are highly predictive of the target category (e.g., Chainlink Fence image that no one would call "uninterpretable"), and another subset of images that include visible features consistent with the target category as well as a number of other categories. These category-general features (such as colour or curvature) are what ***Z&F*** called "superficial commonalities" between images [187, p. 2].

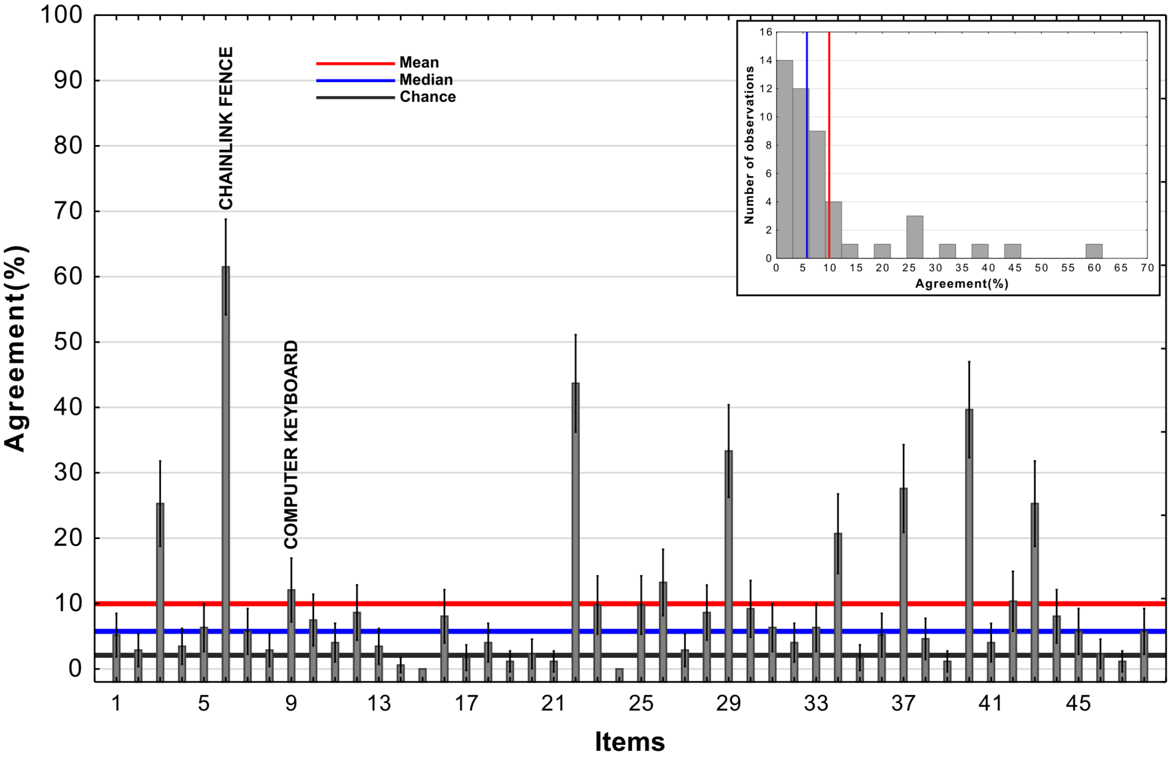


Figure 2.2: **Agreement across adversarial images from Experiment 3b in Zhou & Firestone.** The red line represents the mean, the blue line represents the median, and the black reference line represents chance agreement. The inset contains a histogram of agreement levels across the 48 images.

For this subset of images, the most frequent response chosen by participants does *not* usually match the label assigned by the DCNN. Participants in these cases seem to be making educated guesses using superficial features of the target images to hedge their bets. For the rest of the images agreement is at or below chance levels.

In order to more directly test how humans interpret adversarial images we carried out five experiments. First, if participants are making educated guesses based on superficial features, then agreement levels should decrease when presented with response alternatives that do not support this strategy. We test this in Experiment 1. Second, if a DCNN develops human-like representations for a subset of categories (e.g., the Chainlink Fence category for which human-DCNN agreement was high for a specific adversarial image of a chainlink fence), then it should not matter which adversarial image from these categories is used to evaluate agreement. We test this in Experiment 2. Third, if DCNNs are processing images in very different ways to humans, then it should be possible to find situations in which overall agreement levels are at absolute chance levels. In Experiment 3 we show that one class of adversarial images for the MNIST dataset generated overall chance level agreement. In Experiment 4 we show that it is straightforward to generate adversarial images for the ImageNet dataset that produce overall chance level agreement. Finally, in Experiment 5 we show that agreement levels between humans and DCNNs remain low and variable even for images that fool an ensemble of DCNNs. The findings further undermine any claim that DCCNs and humans categorize adversarial images in a similar way.

2.2.3 Experiment 1: Response alternatives

One critical difference between decisions made by DCNNs and human participants in an experiment is the number of response alternatives available. For example, DCNNs trained on ImageNet will choose a response from amongst 1000 alternatives while participants will usually choose from a much smaller cohort. In Experiment 1, we tested whether agreement levels are contingent on how these response alternatives are chosen during an experiment. We chose a subset of ten images from the 48 that were used by ***Z&F*** and identified four *competitive* response alternatives (from amongst the 1000 categories in ImageNet) for each of these images. One of these alternatives was always the category picked by the DCNN and the remaining three were subjectively established as categories which share some superficial visual features with the target adversarial image. For example, one of the adversarial images contains a *florescent orange curve* and is confidently classified by the DCNN as a Volcano. For this image, we chose the set of response alternatives {Lighter, Missile, Table lamp, Volcano}, all of which also contain this superficial visual feature. See Appendix A - Figure A.6 for the complete list of images and response alternatives. Participants were then shown each of these ten images and asked to choose one amongst these four competitive response alternatives. Note that if humans possess a "machine-theory-of-mind", it should not matter how one samples response alternatives as a DCNN classifies the fooling adversarial images with high confidence (> 99%) in the presence of *all* 999 alternative labels, including the competing alternatives we have selected. In the control condition an independent sample of participants completed the same task but the alternative labels were chosen at *random* from the 48 used by ***Z&F***.

We observed that agreement levels fell nearly to chance in the competitive condition while being well above chance in the random condition (see Figure 2.3). The mean agreement level in the competitive condition was at 28.5% (SD = 11.67) with chance being at 25%. A single sample t-test comparing the mean agreement level to the fixed value of 25% did show the difference was significant ($t(99) = 3.00, p = .0034, d = 0.30$). However, in the random condition mean agreement was 49.8% (SD = 16.02) which was both significantly above chance ($t(99) = 15.48, p < .0001, d = 1.54$) and well above agreement in the competitive condition ($t(198) = 10.75, p < .0001, d = 1.52$). Both conditions are in stark contrast to the DCNN which classified these images with a confidence > 99% even in the presence of these competing categories.

These results highlight a key contrast between human and DCNN image classification. While the features in each of these adversarial images are sufficient for a DCNN to uniquely identify one amongst a 1000 categories, for humans they are not. Instead features within these images only allow them to identify a cohort of categories. Thus, the observed decrease in agreement

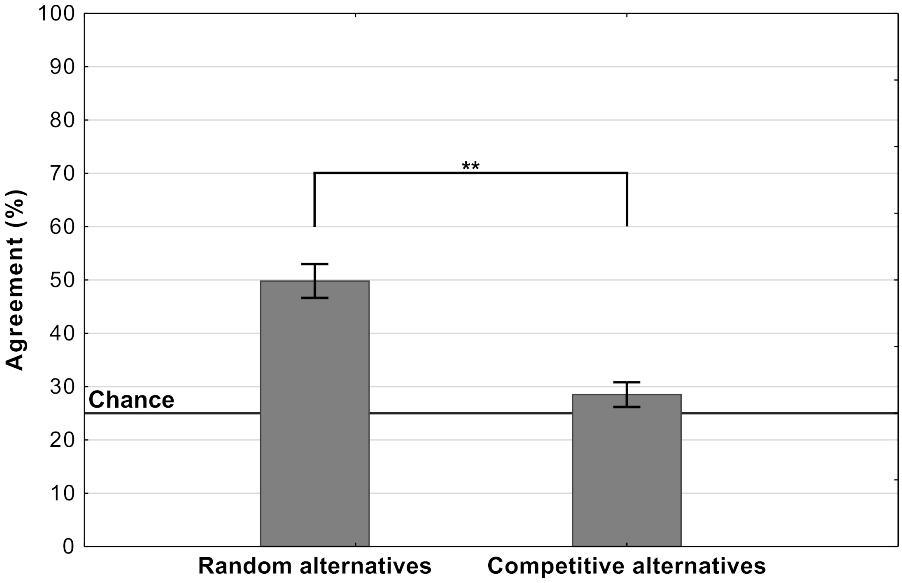


Figure 2.3: **Average levels of agreement in Experiment 1** (error bars denote 95% confidence intervals).

between the random and competitive conditions supports the hypothesis that participants are making plausible guesses in these experiments, using superficial features (shared amongst a cohort of categories) to eliminate response alternatives.

It should be noted that ***Z&F*** were themselves concerned about how the choice of response alternatives may have influenced their results. Therefore, they carried out another experiment where, instead of choosing between the DCNN's preferred label and another randomly selected label, the participants had to choose between the DCNN's 1^{st} and 2^{nd}-ranked labels. The problem with this approach is that the DCNN generally has a very high level of confidence ($> 99\%$) in it's 1^{st} choice. Accordingly, it is not at all clear that the 2^{nd} most confident choice made by the network provides the most challenging response alternative for humans. The results from Experiment 1 show that when the competing alternative is selected using a different criterion, the agreement between participants and DCNNs does indeed drop to near-chance levels.

2.2.4 Experiment 2: Target adversarial images

Our reanalysis above also showed that there was large variability in agreement between images. One possible explanation for this is that the DCNN learns to represent some categories (such as Chainlink Fence or Computer Keyboard) in a manner that closely relates to human object recognition while representations for other categories diverge. If there was meaningful overlap between human and DCNN representations for a category, we would expect participants to show a similar level of agreement on all adversarial images for this category as all adversarial images will capture these common features. So replacing an adversarial image from these categories with another image generated in the same manner should lead to little change in agreement. In Experiment 2 we directly tested this hypothesis by sampling two different images (amongst the

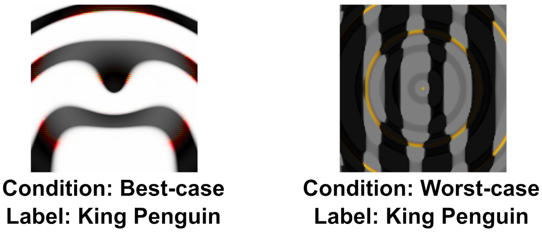


Figure 2.4: **Example of *best-case* and *worst-case* images for the same category ('king penguin') used in Experiment 2.**

five images for each category generated by Nguyen et al. [127]) for the same ten categories from Experiment 1. We chose the best and worst representative stimuli for each of the categories by running a pre-study (see the Methods section) and labelled the two conditions as *best-case* and *worst-case*. An example of each type of image is shown in Figure 2.4.

Figure 2.5 shows the mean agreement for participants viewing the *best-case* and *worst-case* adversarial images. The difference in agreement between the two conditions was highly significant ($t(198) = 22.28, p < .0001, d = 3.15$). Both groups showed agreement levels significantly different from chance (which was at 25%). The best-case group was significantly above chance ($t(99) = 20.12, p < .0001, d = 2.01$) while the worst-case was significantly below chance ($t(99) = 10.58, p < .0001, d = 0.99$).

Thus, we observed a large drop in agreement when we replaced one set of adversarial images with a different set, and there was no evidence for consistent above-chance agreement for all adversarial images from a subset of categories (see Appendix A - Figure A.7 for an item-wise breakdown). In other words, we did not observe any support for the hypothesis that DCNNs learn to represent even a subset of categories in a manner that closely relates to human object recognition.

2.2.5 Experiment 3: Different types of adversarial images

Although we can easily reduce DCNN-human agreement to chance by judiciously selecting the targets and foils, it remains the case that a random selection of targets and foils has led to above chance performance on this set of images. In the next experiment, we asked whether this effect is robust across different types of adversarial images. All the images in the experiments above were generated to fool a network that had been trained on `ImageNet` and belonged to the subclass of *regular* adversarial images generated by Nguyen et al. [127] using an *indirect encoding* evolutionary algorithm. In fact, Nguyen et al. [127] generated four different types of adversarial images by manipulating the type of encoding – *direct* or *indirect* – and the type of database the network was trained on – `ImageNet` or `MNIST` (see Figure 2.6). We noticed that ***Z&F*** used images designed to fool DCNNs trained on images from `ImageNet`, but did not consider the adversarial images designed to fool a network trained on `MNIST` dataset. To our eyes, these `MNIST` adversarial images looked completely uninterpretable and we wanted to test whether the above

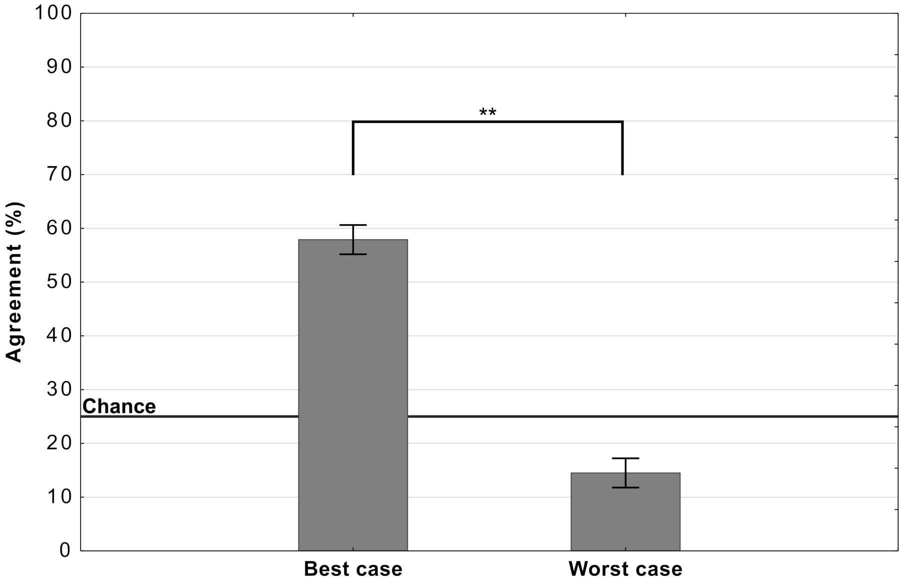


Figure 2.5: **Average levels of agreement in Experiment 2** (error bars denote 95% confidence intervals).

chance agreement was contingent on which set of images were used in the experiments.

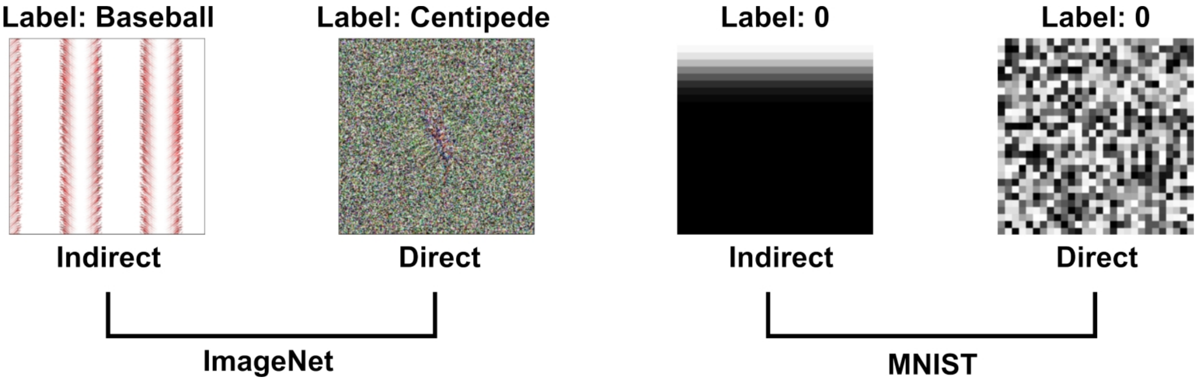


Figure 2.6: **Examples of images from Nguyen et al. [127] used in the four experimental conditions in Experiment 3.** Images are generated using an evolutionary algorithm either using the *direct* or *indirect* encoding and generated to fool a network trained on either `ImageNet` or `MNIST`

Accordingly, we designed a 2x2 experiment in which we tested participants on all four conditions corresponding to the four types of images (Figure 2.6). Since `MNIST` has ten response categories and we wanted to compare results for the `MNIST` images with `ImageNet` images, we used the same 10 categories from Experiments 1 and 2 for the two `ImageNet` conditions. On each trial, participants were shown an adversarial image and asked to choose one out of ten response alternatives that remained fixed for all trials.

Mean agreement levels in this experiment are shown in Figure 2.7. We observed a large difference in agreement levels depending on the types of adversarial images. Results of a two-way repeated measures ANOVA revealed a significant effect of dataset on agreement levels ($F(1,197) = 298.62, p < .0001, \eta_p^2 = 0.60$). Participants agreed with DCNN classification for images designed to fool ImageNet classifiers significantly more than for images designed to fool MNIST classifiers. Participants also showed significantly larger agreement for indirectly-encoded compared to directly-encoded images ($F(1,197) = 67.57, p < .0001, \eta_p^2 = 0.26$). The most striking observation was that agreement dropped from 26% for ImageNet images to near chance for MNIST images. Participants were slightly above chance for indirectly-encoded MNIST images ($t(197) > 6.30, p < .0001, d = 0.44$) and at chance agreement for directly-encoded MNIST images ($t(197) = 1.03, p = 0.31$).

In addition to the between-condition differences, we also found high within-condition variability for the ImageNet images.We observed that this was because agreement was driven by a subset of adversarial images (see Appendix A - Figure A.8 for a breakdown). Thus, even for these ImageNet images, DCNN representations do not consistently overlap with representations used by humans.

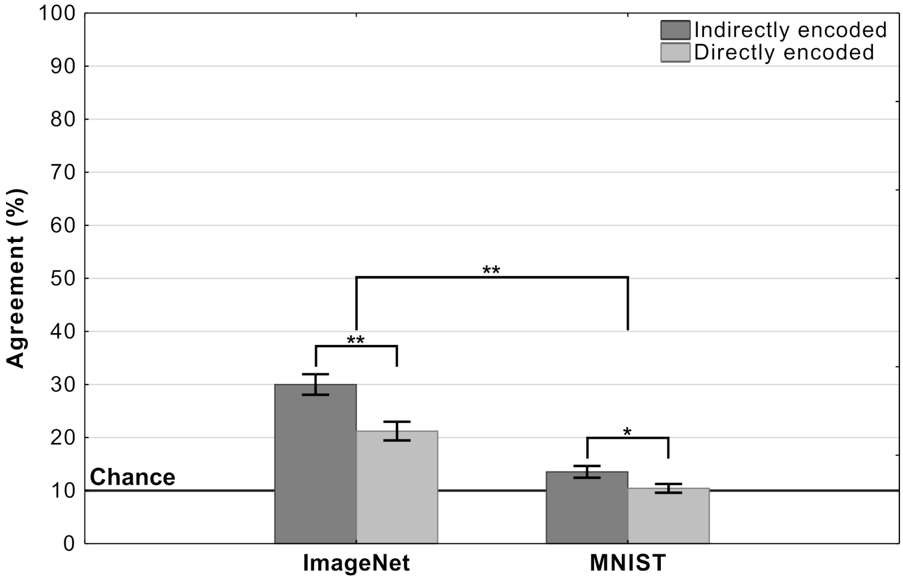


Figure 2.7: **Agreement (mean percentage of images on which a participant choices agree with the DCNN) as a function of experimental condition in Experiment 3** (error bars denote 95% confidence intervals).

2.2.6 Experiment 4: Generating fooling images for ImageNet

Experiment 3 showed that it is straightforward to obtain overall chance level performance on the MNIST images, and this raises the obvious question of whether it is also straightforward to observe chance performance for adversarial images designed to fool ImageNet classifiers? In order to test this we generated our own irregular (TV-static like) adversarial images using a standard method of generating adversarial images (see Methods section). Each of these images was confidently classified as one out of a 1000 categories by a network trained on ImageNet. Participants were presented three of these adversarial images and asked to choose the image that most closely matches the target category (see inset in Figure 2.8). In half of the trials participants were shown adversarial images that were generated to fool AlexNet while in the other half they were shown adversarial images generated to fool Resnet-18.

Results of the experiment are shown in Figure 2.8. For both types of images, the agreement between participants and DCNNs was at chance. Additionally, we ran binomial tests for each image in order to determine whether the number of participants which agreed with DCNN classification was significantly above chance and the results showed that not a single image showed agreement that was significantly above chance. Clearly, participants could not find meaningful features in any of these images, while networks were able to confidently classify each of these images.

2.2.7 Experiment 5: Transferable adversarial images

In the experiments above we observe that while DCNNs are vulnerable to adversarial attacks (they classify these images with extremely high confidence), participants do not show such a vulnerability or even a consistent agreement with the DCNN classification. But it does *not* necessarily follow that DCNNs are poor models of biological vision. In fact, there are many different methods of generating adversarial images [1] and some do not transfer even from one DCNN to another, and this does not merit the conclusion that the different the DCNNs function in fundamentally different ways (indeed, current DCNNs are highly similar to one another, by design). In a similar manner, the fact that adversarial images do not transfer between DCNNs and humans does not, by itself, support the conclusion that the human visual system and DCNNs are fundamentally different.

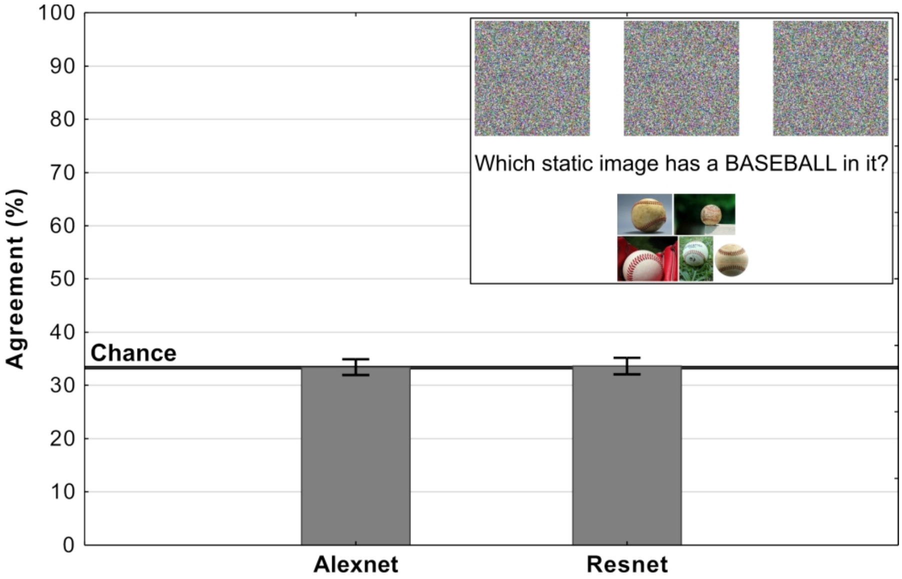


Figure 2.8: **Average levels of agreement in Experiment 4** (error bars denote 95% confidence intervals). The inset depicts a single trial in which participants were shown three fooling adversarial images and naturalistic examples from the target category. Their task was to choose the adversarial image which contained an object from the target category.

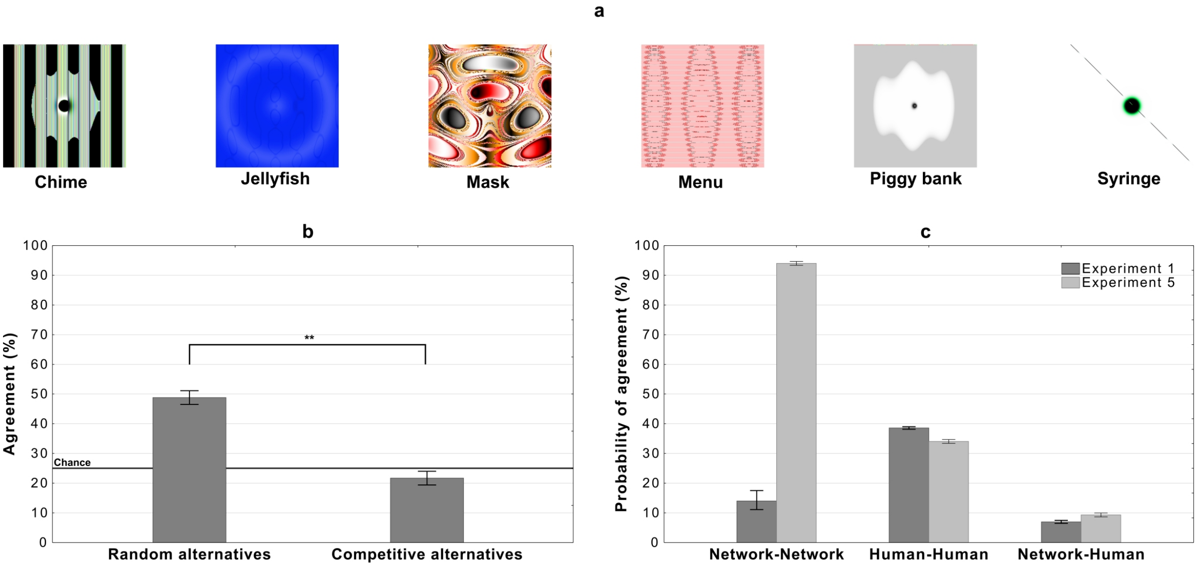


Figure 2.9: **Results for images that are confidently classified with high network-to-network agreement**. Networks: `Alexnet, Densenet-161, GoogLeNet, MNASNet 1.0, MobileNet v2, Resnet 18, Resnet 50, Shufflenet v2, Squeezenet 1.0`, and `VGG-16`. (a) Examples of images used in the experiment - for all the stimuli see Appendix A - Figures A.9 and A.10, (b) average levels of agreement between participants and DCNNs under the *random* and *competitive* alternatives conditions in Experiment 5, and (c) probability of network, human, and network to human agreement in the *competitive alternatives* condition of Experiment 1 and Experiment 5 (error bars denote 95% confidence intervals).

In order to provide an stronger test of the similarity of DCNNs and human vision, we asked whether adversarial images that fool multiple DCNNs are decipherable by humans. If indeed there are some underlying and reliable similarities in how stimuli are processed in DCNNs and humans, then it might be expected that highly transferable DCNN adversarial attacks should also lead to higher human to network agreement.

So in the next experiment, we chose 20 adversarial images that 10 DCNNs classify with high confidence and high between-network agreement (see the Methods section for details). The experiment then follows the same procedure as Experiment 1, where a participant is shown an adversarial image on each trial and asked to choose a label from four response alternatives. Like Experiment 1, participants are assigned to one of two conditions. In the *random alternatives* condition, participants were shown the network label and three other labels, which were randomly drawn from the remaining 19 labels. In the *competitive alternatives* condition, participants again had to choose from the network label and three alternative labels. However, in this condition the labels were chosen amongst the 999 remaining category labels in `ImageNet` such that they contain some superficial features contained within these images (see Methods for details). Note that all DCNNs classified these images with high confidence and with all 1000 `ImageNet` labels present as alternatives.

Results are depicted in Figure 2.9b. There was a significant difference between the two conditions $t(198) = 16.37, p < .0001, d = 2.32$. Additionally, both conditions were significantly different from chance. Agreement in the *random alternatives* was above chance ($t(99) = 18.66, p < .0001, d = 1.87$) and below chance in the *competitive labels* condition ($t(99) = 3.13, p < .01, d = 0.31$).

Thus we find very similar levels of agreement for these adversarial images, which fool multiple DCNNs, to the adversarial images from Experiment 1 (compare Figure 2.9b and Figure 2.3). To further examine how the DCNN-to-DCNN agreement compares to DCNN-to-human agreement, we computed the probability that two randomly sampled networks will agree on an image's label and compared it to the probability that a randomly sampled network will agree with a randomly sampled participant (see Methods for details). Figure 2.9c shows these probabilities for the competitive condition for both Experiment 1 and Experiment 5. We observed that: (i) even when the probability that two networks agree on an adversarial image is larger than 90% the probability of network-human agreement is low ($\sim 10\%$), and (ii) the increase in probability of network-network agreement (between Experiment 1 and Experiment 5) has very little impact on human classification as the probability of human-human and network-human agreement remains much the same in the two experiments. Thus, participants showed very little agreement with DCNNs even when DCNNs agreed with each other. Interestingly, humans showed more agreement amongst themselves, consistent with the hypothesis that participants represent these adversarial images in similar ways, even though we find no evidence that these representations overlap those of the networks. This again suggests that humans and current DCNNs process these images in fundamentally different ways.

2.3 Discussion

Zhou and Firestone [187] claim that humans can robustly decipher adversarial images which suggests that there are important similarities in how humans and DCNNs process these images, and objects more generally. However, when we examined their results using an alternative analysis, we found that the level of agreement was rather low, highly variable, and largely driven by a subset of images where participants could eliminate response alternatives based on superficial features present within these images. This was confirmed in a series of experiments that found that agreement between humans and DCNNs was contingent on the adversarial images chosen as stimuli (Experiments 2 and 3) and the response alternative presented to participants (Experiment 1). We also show that there are well-known methods for generating adversarial images that lead to overall chance level DCNN-human agreement (Experiments 3 and 4), again demonstrating that DCNNs confidently identify images on the basis of features that humans completely ignore. Furthermore, even when humans were presented with adversarial images that fooled at least 9 of 10 DCNNs, the level of agreement between humans and DCNNs remained low and variable (Experiments 5). Indeed, manipulating the level of agreement between DCNNs (by varying the adversarial images) had no impact on the level of agreement between

DCNNs and humans, or amongst humans, as highlighted in Figure 2.9. Taken together, these findings not only refute the claim that there is a robust and reliable similarity in processing these adversarial images, but also suggest that humans and current DCNNs categorize objects in fundamentally different ways.

A similar distinction between human and DCNN classification is made by Ilyas et al. [76], who argue that current architectures of DCNNs are vulnerable to adversarial attacks due to their tendency for relying on *non-robust features* present in databases. These are features that are predictive of a category but highly sensitive to small perturbations of the image. It is this propensity for relying on non-robust features that makes it easy to generate adversarial images that are completely uninterpretable by humans but classified confidently by the network (Experiment 4). A striking example of DCNNs picking up on non-robust features was recently reported by Malhotra and Bowers [113] who showed that DCNNs trained on a `CIFAR-10` dataset modified to contain a single diagnostic pixel per category, learn to categorize images based on single pixels ignoring everything else in the image. Humans, by contrast, tend to use robust features of objects, such as their shape, for classifying images [12].

We would like to note that we are not claiming that there is no role played by superficial and non-robust features in human object recognition. In a recent study, [36] asked human participants to classify naturalistic adversarial images (see Figure 2.1b) when these images were briefly flashed (for around 70ms) on the screen. They found that there is a small, but statistically significant, effect of the adversarial manipulation on choices made by participants (i.e., participants were slightly more likely to classify a 'cat' image as a 'dog' when the image was adversarially perturbed towards a 'dog'). Thus, these results seem to suggest that humans are sensitive to the same type of non-robust features that lead to adversarial attacks on DCNNs. However, it is important to note here that the size of these effects is small: while human accuracy drops by less than 10% when normal images are replaced by adversarially perturbed images, DCNNs (mis)classify these adversarially perturbed images with high confidence. These findings are consistent with our observation that some adversarial images capture some superficial features that can be used by participants to make classification decisions, leading to an overall above-chance agreement.

It should also be noted that we have only considered a small fraction of adversarial images here and, like Experiment 4, there are many other types of adversarial attacks that produce images that seem completely undecipherable for humans. It could be that humans find these images completely uninterpretable due to the difference in *acuity* of human and machine vision (a line taken by ***Z&F***). There are two reasons why we think a difference in acuity cannot be the primary explanation of the difference between human and machine perception of adversarial images. Firstly, we have shown above that the very same algorithm produced some images that supported above chance agreement and other images that supported no agreement (for example, Appendix A - Figure A.7). There is no reason to believe that the two sets of images are qualitatively different, with DCNNs selectively exploiting subliminal features only when overall agreement levels are chance. Secondly, a wide variety of adversarial attacks clearly do not rely on subtle visual features that are below human perceptual threshold. This includes semantic adversarial attacks that occur when the colour of an images is changed [72], or attacks that cause incorrect classification by simply changing the pose of an image [2], etc. These are all dramatic examples of differences between DCNNs and humans that cannot be attributed to the acuity of human perceptual front-end. Rather they reflect the fact that current architectures of DCNNs are often relying on visual features that humans can see but ignore.

Of course, it might be possible to modify DCNNs so that they perform more like humans in our adversarial tasks. For example, training similar models on data sets that are more representative of human visual experience might reduce their susceptibility to adversarial images and lead DCNNs to produce more variable responses in our tasks as a consequence of picking up on superficial visual features. In addition, modifying the architectures of DCNNs or introducing new ones may lead to a better DCNN-human agreement on these tasks (for example, capsule networks [155]). But researchers claiming that current DCNNs provide the best models of visual processing in primate ventral visual processing stream need to address this striking disconnect between the two systems.

There are two rhetorically appealing arguments against our conclusion that adversarial images still represent a striking example of dissimilarity between humans and deep nets. First, that adversarial images are akin to visual illusions humans experience [93, 112, 181]. And second, that if we could access the human brain similarly to how we can access connection weights in CNNs, we could generate adversarial stimuli for humans. Both of these arguments are appealing and thought provoking but do not further the argument about similarity between humans and CNNs. If we do accept the analogy that adversarial attacks are akin to visual illusions, then the fact that CNNs are susceptible to illusions so different to the ones experienced by humans is not a good argument for similarity. Visual illusions are a consequence of normal visual processing resulting with incorrect perceptions under specific circumstances. They provide insight into the process itself. Take, for example, the Ponzo illusion (Figure 2.10 left). We perceive the blue line as longer than the red line even though they are of the same length. We do so because the blue line is perceived to be further away from us. In order for two objects at different depths to be the same size in the retinal image, they need to be of different sizes in the real world. The farthest object has to be larger in order to be the same size on the retina from a farther distance. The illusion is therefore a consequence of a process which is normal and useful in navigating the world around us. If adversarial images are the same type of stimuli for CNNs then they are a consequence of processes which are highly dissimilar to human vision.

The second argument, that we could generate human adversarial images if we had access to the brain the same way we can access deep nets, is uninformative on what these stimuli would look like. There is nothing to suggest that these human adversarial images would be similar to

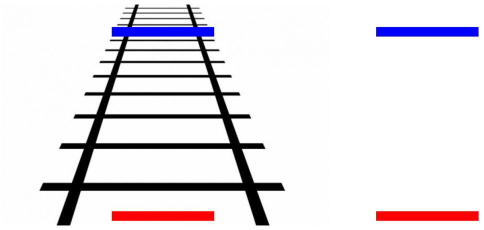


Figure 2.10: **The Ponzo ilussion.** The red and blue lines are of equal lengths but the blue line is perceived as being longer on the left-hand side. This is due to it being perceived as further away form the observer. Since it has the same length on the retinal image at a greater distance, the blue line has to be longer in 'reality' and is perceived as such. When depth information is not present, the lines are perceived as being the same length (right-hand side).

CNN adversarial images. Therefore, adversarial images indicate dissimilarity between humans and deep learning models, though they can be an interesting gateway into better understanding of deep learning models.

To conclude, our findings with fooling adversarial images pose a challenge for theorists using current DCNNs trained on data sets like `ImageNet` as psychological models of human object identification. An important goal for future research is to develop models that are sensitive to the visual features that humans rely on, but at the same time insensitive to other features that are diagnostic of object category but irrelevant to human vision. This involves identifying objects on the basis of shape rather than texture or color or other diagnostic features [6, 50], where vertices are the critical components of images [9], where Gestalt principles are used to organize features [141], where relations between parts are explicitly coded [170], where features and objects are coded independently of retinal position [15] size [11], left/right reflection [10], etc. When DCNNs rely on these set of features, we expect they will not be subject to adversarial attacks that seem so bizarre to humans, and will show the same set of of strengths and weakness (visual illusions) that characterize human vision.

2.4 Methods

Reassessing agreement: Blindfolded participants If a participant is blindfolded and chooses one of 48 options randomly on 48 trials, the probability of them making the same choice as the DCNN on k trials is given by the binomial distribution $\binom{48}{k}p^k(1-p)^{48-k}$, where $p = \frac{1}{48}$. Substituting different values of k, one can compute that 37.2% of these blindfolded participants will agree with the DCNN on 1 trial, 18.6% will agree on 2 trials, 6% will agree on 3 trials, and so on. To compute the proportion of participants who agree with the DCNN, Zhou and Firestone [187] count all participants who agree on 2 or more trials as agreeing with the DCNN (chance is 1 out of 48 trials) and half of the participants that agree on exactly 1 trial. Thus, summing up all the blindfolded participants that agree on 2 or more trials and half of those who agree on exactly 1 trial, this method will show $\sim 45\%$ agreement between participants and the DCNN.

Experiment 1. This experiment examined whether agreement between humans and DCNNs depended on the response alternatives presented to participants. We tested $N = 200$ participants and each participant completed 10 trials. During each trial, participants were presented a fooling adversarial image and four response alternatives underneath the image and asked to choose one of these alternatives. Participants indicated their response by moving their cursor to the response alternative and clicking. We selected 10 fooling adversarial images from amongst the 48 images used by ***Z&F*** in in their Experiments 1–3. Each of these images was classified with $> 99\%$ confidence by a DCNN which was trained to classify the `ImageNet` dataset. We selected these 10 adversarial images to minimise semantic and functional overlap in the labelled categories (for example, we avoided selecting both 'computer keyboard' and 'remote control'). The experiment consisted of two conditions, which differed in how the response alternatives were chosen on each trial. In the 'Competitive' condition ($N = 100$) we chose four response alternatives that subjectively seemed to contain one or more visual features that were present in the adversarial image. One of these response alternatives was always the label chosen by the DCNN. The other three were chosen from amongst 1000 `ImageNet` class labels. This was again done to minimise a semantic or functional overlap with the target class (e.g. an alternative for the 'baseball' class was *parallel bars* but not *basketball*). All ten images and the four competitive response alternatives for each image are shown in Appendix A - Figure A.6 In the 'Random' condition ($N = 100$) the three remaining alternative responses were drawn at random (on each trial) from the aforementioned 48 target classes from Experiment 3 in Zhou and Firestone [187]. We randomised the order of images, as well as the order of the response alternatives for each participant.

Experiment 2. This experiment was designed to examine whether all fooling adversarial images for a category show similar levels of agreement between humans and DCNNs. The experiment's design was the same as Experiment 1 above, except participants were now randomly assigned to the 'best-case' ($N = 100$) and 'worst-case' ($N = 100$) conditions. In each condition, participants again completed 10 trials and on each trial, they saw an adversarial image and four response alternatives. One of these alternatives was the category chosen by a DCNN with $> 99\%$ confidence and the other three were randomly drawn from amongst the 48 categories used by ***Z&F*** in their Experiments 1–3. The difference between the 'best-case' and 'worst-case' conditions was the adversarial image that was shown to the participants on each trial.

In order to choose the best and worst representative image for each of the categories we ran a pre-study. Each image used by ***Z&F*** in their Experiments 1–3 was chosen from a set consisting of five adversarial images for that category generated by Nguyen et al. [127]. In the pre-study, participants ($N = 100$) were presented all five fooling images and asked to choose an image that was most-like and least-like a member from that category (e.g. most like a computer keyboard). Then, during the study, participants in the 'best-case' condition were shown the image from each category that was given the *most-like* label with the highest frequency. Similarly, participants in the ‘worst-case’ condition were shown images that were labelled as *least-like* with the highest frequency. DCNNs showed the same confidence in classifying both sets of images. We again randomised the order of presentation of images.

Experiment 3. The experiment consisted of four experimental conditions in a 2x2 repeated measures design (every participant completed each condition). The first factor of variation was the database on which the DCNNs were trained – `ImageNet` or `MNIST` with one condition containing images designed to fool `ImageNet` and the other containing images designed to fool `MNIST` classifiers. The second factor of variation was the evolutionary algorithm used to generate the adversarial images – *direct* or *indirect*. The indirect encoding method leads to adversarial images which have regular features (e.g. edges) that often repeat, while the direct encoding method leads to noise-like adversarial images. All of the images were from the seminal Nguyen et al. [127] paper on fooling images. The `MNIST` dataset consists of ten categories (corresponding to handwritten numbers between 0 and 9), while `ImageNet` consists of 1000 categories. As we wanted to compare agreement across conditions, we selected ten images (from ten different categories) for both datasets. The indirectly-encoded `ImageNet` images were the same as the ones in Experiment 1 while the images for the other three conditions were randomly sampled from the images generated by [127]. Participants were shown one image at a time and asked to categorize it as one of ten categories (category labels were shown beneath the image). One of these ten categories was the label assigned to the image by a DCNN. Therefore, chance level agreement was 10%. The participants had to click on the label they thought represented what was in the image. The order of conditions was randomized for each participant and the order of images within each condition was randomized as well. A total of $N = 200$ participants completed the study. Two participants were excluded from analysis because their choices were made with average response times below 500 ms indicating random clicking rather than actually making decisions based on looking at the images themselves.

Experiment 4. In this experiment we used the Foolbox package ([148]) to generate images that fool DCNNs trained on `ImageNet`. The experiment consisted of two conditions, one with images designed to fool AlexNet [100] and the other with images designed to fool ResNet-18, both trained on `ImageNet`. We generated our own adversarial images by first generating an image in which each pixel was independently sampled and successively modifying this image using an Iterative Gradient Attack based on the fast gradient sign method ([57]) until a DCNN classified this image as a target category with a $> 99\%$ confidence. The single trial procedure mirrors Experiment 4 from ***Z&F***. Participants ($N = 200$) were shown three of the generated images and a set of five real-world example images of the target class (see Inset in Figure 2.8). They were asked to choose the adversarial image which contained an object from the target class. The example images were randomly chosen from the `ImageNet` dataset for each class. Each participant completed both experimental conditions. The order of trials was randomized for each participant.

Experiment 5. The experiment mirrors Experiment 1 in procedure and experimental conditions. Participants ($N = 200$) were sequentially shown 20 images and asked to choose one out of four response alternatives for each image. The order of presentation of these images was randomized. The stimuli were chosen from ten independent runs (a total of 10000 images) of the evolutionary algorithm used in [127] which were kindly provided to us by the first author. The images were selected such that at least 9 out of 10 networks classify the images as the same category with high confidence (median confidence of 92.61%). The DCNN models were pre-trained on `ImageNet` and are a part of the model zoo of the PyTorch framework. The models are: `Alexnet`, `Densenet-161`, `GoogLeNet`, `MNASNet 1.0`, `MobileNet v2`, `Resnet 18`, `Resnet 50`, `Shufflenet v2`, `Squeezenet 1.0`, and `VGG-16`. The input images were transformed in accordance with recommendations found in PyTorch documentation: 224x224 centre crop and normalization with $mean = [0.485, 0.456, 0.406]$ and $std = [0.229, 0.224, 0.225]$ prior to classification.

The two experimental conditions mirror Experiment 1. In the *random alternatives* condition, for each image, participants ($N = 100$) chose among labels which included the network classification and three alternatives chosen at random among the remaining 19 stimuli labels. In the *competitive alternatives* condition, participants ($N = 100$) chose among labels which included the network classification and three competitive labels. To determine these competitive labels, we conducted a pre-study, where participants ($N = 20$) were asked to generate three labels for each adversarial image. These labels were then used as a guide to select the three competitive categories from `ImageNet` while ensuring that these categories did *not* semantically overlap with the target category. Participants were assigned to one of the two conditions randomly, the order of images and label positioning on the screen were randomized for each participant. Stimuli and competitive labels can be seen in Appendix A - Figures A.9 and A.10.

Statistical analyses. All conducted statistical analyses were two-tailed with a p-value under 0.05 denoting a significant result. In Experiments 1, 2, 4, and 5 we conducted single sample t-tests to check if agreement levels were significantly above a fixed chance level (25% in Experiments 1, 2, and 5, 33.33% in Experiment 4). We additionally ran a between-subject t-test (Experiments 1, 2, and 5) and a within-subject t-test (Experiment 4) to determine whether the difference between experimental conditions was significant. We also conduct a Binomial test in Experiment 4 to determine for how many items was agreement level significantly above chance. In Experiment 3 we ran a two-way repeated measures analysis of variance. In Experiment 5 we ran a mixed two-way analysis of variance. We report effect size measures for all tests (Cohen's d for t-tests and partial eta squared for ANOVA effects). We calculate probability of network-network, human-human, and network-human agreement in the *competitive labels* condition of Experiment 1 and 5. This was done by calculating the percentage of agreements among all possible comparisons. For example, the total number of comparisons to calculate the probability of agreement between two networks, was: 20(images) $*$45(number of possible combinations of two networks). The number of such comparisons which resulted in agreement between two networks divided by the total number of comparisons gives the probability of two networks agreeing when classifying an adversarial image. We conducted the same calculation on data from Experiment 1, since those stimuli were not specifically chosen to be highly transferable between networks.

Power analysis. A sample size of $N = 200$ was chosen for each experiment which mirrors ***Z&F*** experiments 1-6 in order to detect similar effects. This allowed us to detect an effect size as low as $d = 0.18$ at $\alpha = .05$ with 0.80 power in within-subject and $d = 0.35$ in between-subject experiments.

Online recruitment. We conducted all four experiments online with recruitment through the Prolific platform. Each sample was recruited from a pool of registered participants which met the following criteria. Fluent English speakers living in the UK, USA, Canada or Australia of both genders between the ages of 18 and 50 with normal or corrected to normal vision and a high feedback rating on the Prolific platform (above 90). Participants were reimbursed for their time upon successful completion through the Prolific system.

Data availability. Data and stimuli form all our experiments are available via the Open Science Framework at `https://osf.io/a2sh5/`.
Stimuli from evolutionary runs producing fooling images by [127] can be found at `https://anhnguyen.me/project/fooling/`.

Ethics approval. Participants were informed about the nature of the study, and their right to withdraw during the study or to withdraw their data from analysis. The participants gave consent for anonymized data to be used for research and available publicly. The project has been approved by the IRB at the University of Bristol (application ID 76741).

Chapter 3

Feature blindness

Chapter 2 explored various datasets of adversarial stimuli to infer whether there is agreement between humans and deep neural networks. Adversarial images are edge-cases, however it is still possible that for most other stimuli network and human representations share more similarities. This chapter focuses on shape-bias in particular. As discussed in Chapter 1, studies have shown that CNNs trained on large naturalistic datasets do not exhibit a shape bias. In contrast, it is well known that humans exhibit a *shape bias* - we prefer to categorize objects based on shape if the task allows for it [12, 105, 108, 126]. Further, shape bias develops very early [26, 169] between 18 and 24 months of age [183].

In their seminal paper Landau et al. [105] presented children and adults with objects varying in size, shape and texture to test whether shape is weighed more heavily than the other two features. For example, in Experiment 1 they presented participants with a standard (Figure 3.1 top) stimulus and the name associated with the stimulus. Then, in a yes/no task they asked participants to simply respond whether a new stimulus is of the same category (e.g., "is the object a DAX"). They found that variations in size and texture did not affect the proportion of 'yes' responses. In other words, participants confirmed that a new object was of the same category as the standard if shape was not changed but size and texture were. However, for shape changes the proportion of these responses decreased, and decreased more for the more extreme shape distortion (the bottom row in Figure 3.1). Age was also a factor; 2-year olds rejected the greater shape distortion less often than 3-year olds (31% vs 50% of the trials) while adults did so in 94% of the trials. Thus, even though shape was more important than the remaining two features, the magnitude of shape bias increased with age.

Considering experience influences the magnitude of shape-bias this chapter investigated whether networks of different experiences (`ImageNet` and style-transfer trained) adapt to a new

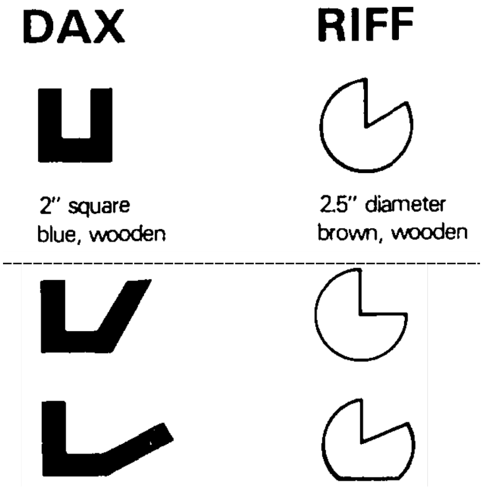


Figure 3.1: **Example stimuli from Landau et al. [105]**. Standard versions of the stimuli were shown to participants and named (top). In trials of a yes/no taks, participants were presented with objects that could vary from standards either in size, texture or in shape (shape variations can be seen in the bottom of the figure). The task was to answer whether the new objects were of the same kind as the standard object.

environment similarly to humans. We generated a new dataset in order to test how CNNs and humans learn when different features predict category membership. We designed datasets in which both global shape, and a secondary feature could be predictive of category membership (for example, the colour of a single part of the shape as shown in Figure 3.2b. We then conducted a number of simulations and behavioral studies by creating various experimental conditions using these new datasets. Our aim was to test how adaptable humans and CNNs are to statistical properties in novel learning environments and what impact previous biases (like the shape bias in humans) have on learning in these novel environments.

This is a publication chapter - this chapter is a re-formatted and slightly edited version of the paper *Feature blindness: A challenge for understanding and modelling visual object recognition* in PLOS Computational Biology [115].

Contributions - Gaurav Malhotra initiated and led this project. My role was consulting during initial stages of dataset generation, discussing various iterations and providing input. After the datasets had been generated Gaurav Malhotra conducted all of the simulations presented in this chapter while I designed and conducted all the behavioral studies and provided statistical analysis of the resulting data. All co-authors contributed to write-up and revisions during the review process.

3.1 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 [28]. A lot of effort in psychology, computational neuroscience and computer vision has gone into understanding how the brain constructs these invariant representations [27, 173].

One hypothesis is that the brain learns these invariant representations from the statistics of natural images [27, 52, 81]. But until 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 [106]. Like humans, these models show impressive generalisation to new images and to different translations, scales and viewpoints [83]. And like humans, this capacity to generalise seems to stem from the ability of Deep Networks to learn invariant internal representations [55]. It is also claimed that the learned representations in humans and networks are similar [83, 101, 151]. These results raise the exciting possibility that Deep Networks may finally provide a good model of human object recognition [24, 93, 137, 176] and provide important insights into visual information processing in the primate brain [20, 85, 119, 149, 181].

Many reasons could be, and are, given for why CNNs have succeeded where previous models have failed [7, 106]. 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 [107]. In addition, both systems are thought to implement optimisation frameworks, generating predictions by performing statistical inferences [149, 182]. Indeed, evidence suggests that humans perform some form of statistical optimisation for many cognitive tasks including language learning [156], spatial cognition [40], motor learning [90] and object perception [81]. 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 [33, 57, 127]. They often classify images of objects based on statistical regularities in the background [174], or even based on single diagnostic pixels present within images [116]. That is, CNNs are prone to overfitting, often relying on predictive features that are idiosyncratic to the training set [47].

To what extent do these findings reflect fundamental differences between CNNs and human vision? On the one hand, such differences could be simply down to differences in the learning environments of humans and artificial neural networks. Evidence supporting this hypothesis comes from studies that have compared features used by CNNs and humans to classify objects. Psychological experiments have repeatedly shown that humans rely primarily on global features, such as shape, for naming and recognising objects [12, 105, 126, 169]. By contrast, a number of studies have demonstrated that CNNs trained on standard datsets, such as `ImageNet`, rely on local textures [50, 70, 109]. But these studies have also shown that, when CNNs are trained on datsets with the right biases, their behaviour can be brought a lot closer to human behaviour [38, 50, 70, 71]. For example, Geirhos et al. [50] showed that CNNs trained on a modified version of `ImageNet` learn to show a shape-bias. Based on these results, Geirhos et al. conclude that "texture bias in current CNNs is not by design but induced by ImageNet training data". Similarly, Hermann et al. [70] showed that CNNs can learn to classify images based on global shape when they are trained with naturalistic data augmentations, leading them to conclude that "apparent differences in the way humans and ImageNet-trained CNNs process images may arise not primarily from differences in their internal workings, but from differences in the data that they see."

On the other hand, behavioural differences between humans and CNNs may arise out of more fundamental differences in resource constraints and mechanisms, rather than just differences in their training sets. While a CNN trained on a particular environment is able to mimic some aspects of shape-bias, the origin of shape-bias may be very different in the two systems. One way to distinguish between these two hypotheses is to check how the bias (of a network or human) is affected by moving to a new environment. If the origin of the bias is purely environmental, then a shift in the environment should also lead to a shift in the bias – that is, the system should start selecting features based on the statistical properties of the new environment. If, on the other hand, the bias is a reflection of a mechanistic principle or a resource constraint, it will be much more immune to a change in the statistical properties of the environment.

In this study, we explored this question by training models and humans to classify a set of novel objects. Each object contained multiple diagnostic features, all of which were clearly visible and could be used to perform the task. We manipulated the statistical bias for selecting these features, by manipulating the extent to which each feature type predicted the category labels. We wanted to explore the extent to which human adults and pre-trained CNNs were *adaptable* to the biases present within this task environment. At one extreme, people (and CNNs) could be completely adaptable, and select features solely based on the statistical properties of the new environment. At the other extreme, they could be completely inflexible and continue selecting features based on their prior biases. To gain a deeper insight into the role that prior biases play in learning new information, we compared the performance of both humans and CNNs to a statistical inference model that had no biases and learned to infer the category of a stimulus based on the sequence of samples observed in the task.

In a sequence of experiments that tested a range of different feature types and model settings, we observed that (i) the behaviour of human participants was in sharp contrast with the statistical inference model, with participants continuing to rely on global features, such as shape, and ignoring local features even when these features were better at predicting the target categories, (ii) when multiple global features were concurrently present (e.g. overall shape as well as colour), some participants chose to rely on one feature while others chose to rely on the second feature, but participants generally did not learn both features simultaneously, (iii) the behaviour of CNNs also contrasted with the statistical inference model, with the CNNs also preferring to rely on one feature, (iv) however, unlike human participants, CNNs frequently relied on diagnostic local features and, crucially, this dependence on local features increased when the features were made more predictive, (v) CNNs were highly adaptable in the feature they used for learning – even when they were trained to have a shape-bias, this bias was lost as soon as they were trained on a new dataset with a different bias.

In two follow-up studies, we investigated whether human participants can overcome their bias for global features by (a) learning in an environment where there is no concurrent shape at all, or (b) being told what type of local feature to look for. In both cases, we observed that participants still failed to learn these tasks based on local features. Thus, the reason why participants ignore some clearly visible features is not simply due to the competition from shape, or to the difficulty in discovering these types of features. Rather, participants seem to struggle with the computational demands of learning the task based on certain features.

These results highlight important differences in how human participants and CNNs learn to extract features from objects and the role that existing biases play in adapting to novel learning environments. In general, CNNs are highly adaptable in learning new information, with the statistical structure of their learning environment driving their learning. While performing statistical learning is also clearly important for humans, their behaviour is much more strongly constrained by prior biases. Models of visual object recognition need to explain how such strong biases can be acquired and how they constrain learning in order to adequately capture human object recognition. The training and test sets developed in this study can be used to constrain and falsify models towards this end.

3.2 Results

3.2.1 Behavioural tasks and Simulations

The behavioural tasks mimicked the process of learning object categorisation through supervised learning. In each experiment, participants were trained to perform a 5-way classification task, where they had 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,

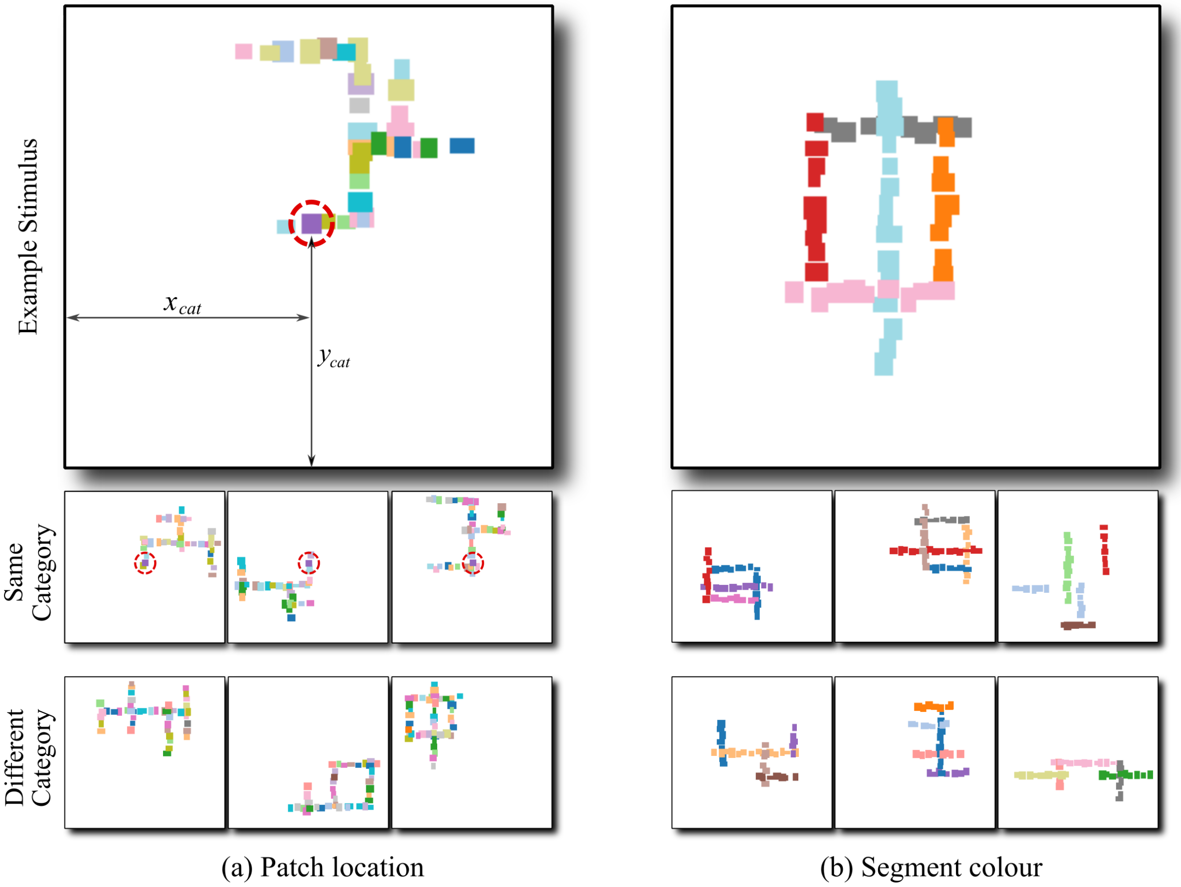


Figure 3.2: **Example training images from Experiments 1 and 2.** (a) Two features predict stimulus category: global shape and location (x_{cat}, y_{cat}) of one of the patches. For illustration, the predictive patch is circled. Stimuli in the same category (middle row, reduced size) have a patch with 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 B - Figures B.1–B.5.

organised so that they appeared to form a solid structure. Within each figure, the relative location, size and colour of patches as well as segments was perturbed (within some bounds) from image to image, making each stimulus unique and avoiding any unintentional diagnostic features, such as local features where segments intersect. In order to successfully perform the task, the participants and CNNs had to generalise over all these variables and discover the invariant shape or non-shape feature. See Figure 3.2 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 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

Table 3.1: **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.

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 3.1 summarises the different combinations of features that were examined in the behavioural tasks in this study. Examples for all Experiments are shown in Appendix B - Figures B.1–B.5.

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

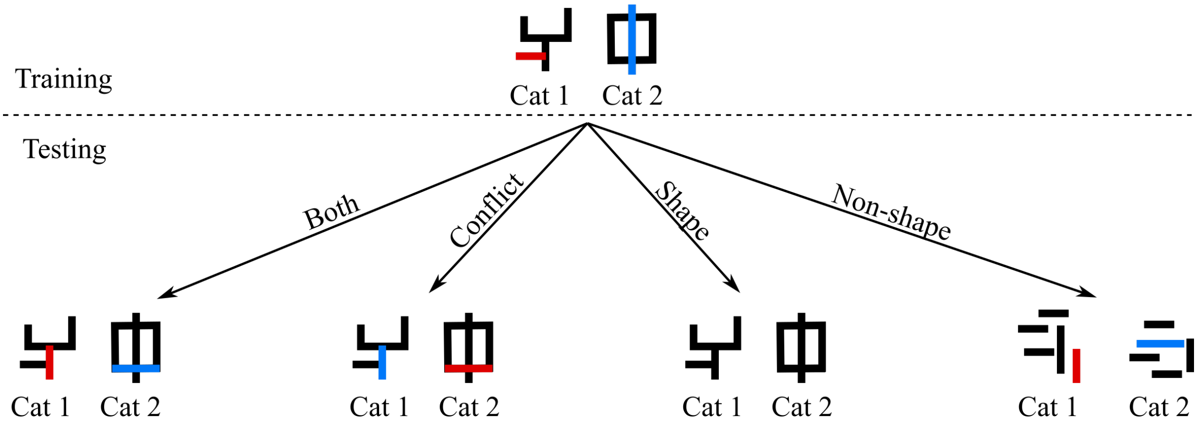


Figure 3.3: **An illustration of the four types of test conditions.** Each category has two diagnostic features: here, the overall shape and the colour of one of the segments. In the training images, features are mapped to categories using the following mapping: {(Shape A, Red) → Category 1; (Shape B, Blue) → Category 2}, where Shape A and Shape B are the shapes on the left and right, respectively. In the `Both` test condition, both types of features (shape and colour) have the same mapping as training. In the `Conflict` condition the mapping of the non-shape feature is swapped – i.e., the new mapping is {(Shape A , Blue) → Category 1; (Shape B, Red) → Category 2}. In the `Shape` condition, images have only one diagnostic feature – the overall shape – which has the same mapping as training: {Shape A → Category 1; Shape B → Category 2}. In the `Non-shape` condition, images have no coherent shape, but contain the same diagnostic colours as the training images: {Red → Category 1; Blue → Category 2}.

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 (see Appendix B - Figures B.1–B.5 for some examples). An illustration of the four test conditions is shown in Figure 3.3. We measured accuracy in the `Both`, `Conflict` and `Shape` test trials based on the category predicted by the shape feature and accuracy in the `Non-shape` trials based on the category predicted by the non-shape feature.

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 [54].

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 had no prior biases and wanted to be statistically as efficient as possible, using all the information available during training trials. This model 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 [167], (pp. 85–89), but additionally has constraints built in through the choice of its architectural properties, such as performing convolutions and pooling. Both models are described in Materials and Methods below.

3.2.2 Both features equally predictive

Figure 3.4A 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 and significantly lower in the `Conflict` condition. 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.

Note that the ideal-inference model predicts that the accuracy for the `Conflict` condition is different for Experiment 1a from the other experiments. The shape and non-shape cues are equally competitive in Experiments 2a, 3a and 4a. Consequently the probability of choosing the correct (shape-based) category is around 0.50 in the `Conflict` condition in these experiments. However, the results in Figure 3.4A show that, in Experiment 1a, the non-shape cue dominates the shape cue in the `Conflict` condition. This is because an image with a diagnostic patch at one of the diagnostic locations contains two types of information: (i) a diagnostic colour at one of the diagnostic locations, and (ii) white (background) patches at all the other diagnostic locations. These two signals together dominate the shape signal in `Conflict` trials in Experiment 1.

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

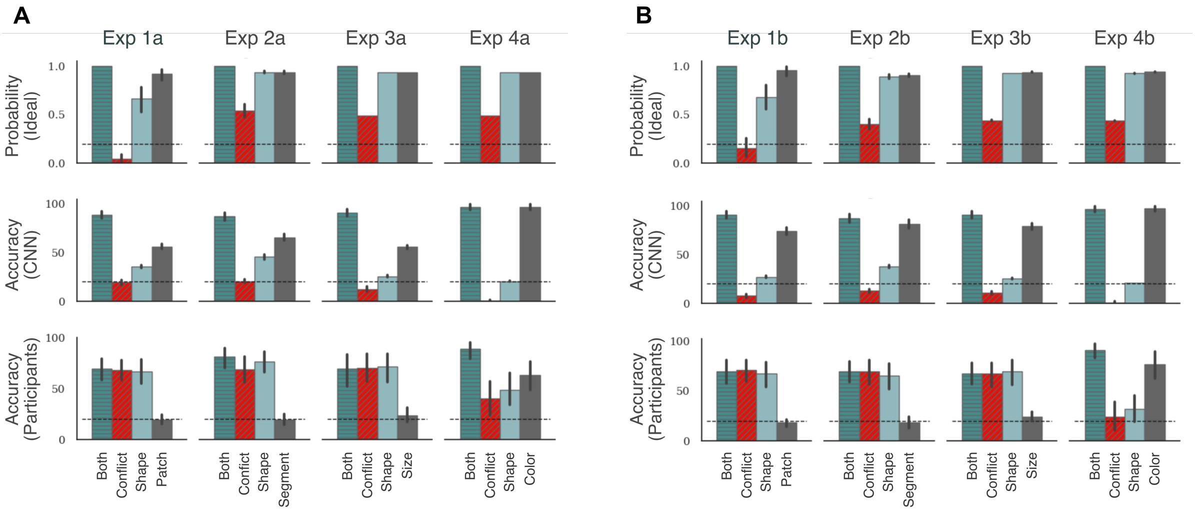


Figure 3.4: **Results in Experiments 1–4**. Each column corresponds to an experiment and each row corresponds to the type of learner (ideal inference model, CNN or human participants). The top row shows the posterior probability of choosing the labelled class for a test trial given the training data. The bottom two rows show categorisation accuracy for this labelled class. Panel A shows results for experiments where both features are equally predictive (1a, 2a, 3a and 4a), while Panel B shows results for experiments where the non-shape feature is more predictive (1b, 2b, 3b and 4b). Each plot 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 plot, 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.

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 (from an information-theoretic perspective) to learn both features in parallel. A similar result was observed by Hermann et al.[71], 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.

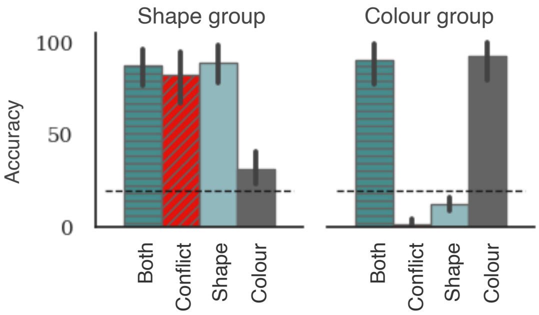


Figure 3.5: **Two groups in Experiment 4a.** Each panel shows the accuracy under the four test conditions for a subgroup of participants. Participants were split based on whether they performed better in the shape or colour conditions. The first group contained N=12 participants and the second group contained N=13 participants.

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 3.5 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 3.4A.

3.2.3 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 the nature of shape-bias is similar in humans and CNNs, we expect both systems will adapt in a similar way to a new statistical environment, which favours a non-shape feature. 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 3.2 and B.5 for some examples of training images that do not contain a shape feature but contain a non-shape feature. Figure 3.4B 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.

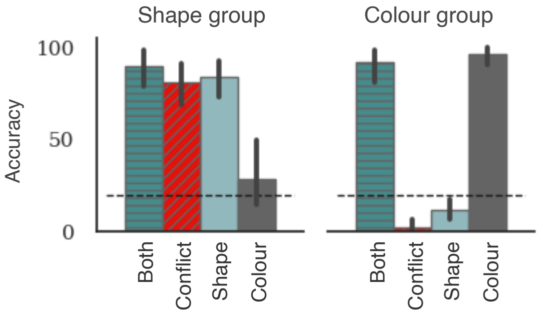


Figure 3.6: **Two groups in Experiment 4b.** Each panel again shows accuracy under the four test conditions for the subgroups of participants who prefer to rely on shape and colour, respectively. In this case, the first group consisted of N=7 participants and the second group consisted of N=18 participants.

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 3.6).

3.2.4 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 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 [50, 70]. 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 [50] 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.

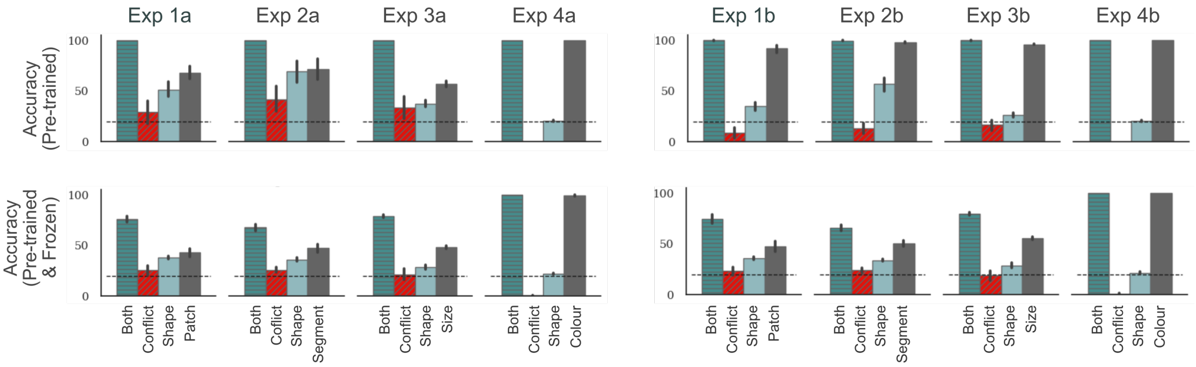


Figure 3.7: **Results for pre-training on a dataset with a shape-bias.** The first row shows results when the CNN was pre-trained on the Style-transfer ImageNet [50] and allowed to learn throughout the network. The second row shows results of the same network when weights for all convolution layers are 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.

The results under these two settings are summarised in Figure 3.7. In line with previous results [50, 70], 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 3.4). 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 3.7, 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 3.7, 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.

3.2.5 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 3.8A 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 Figure B.6). 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.

3.2.6 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

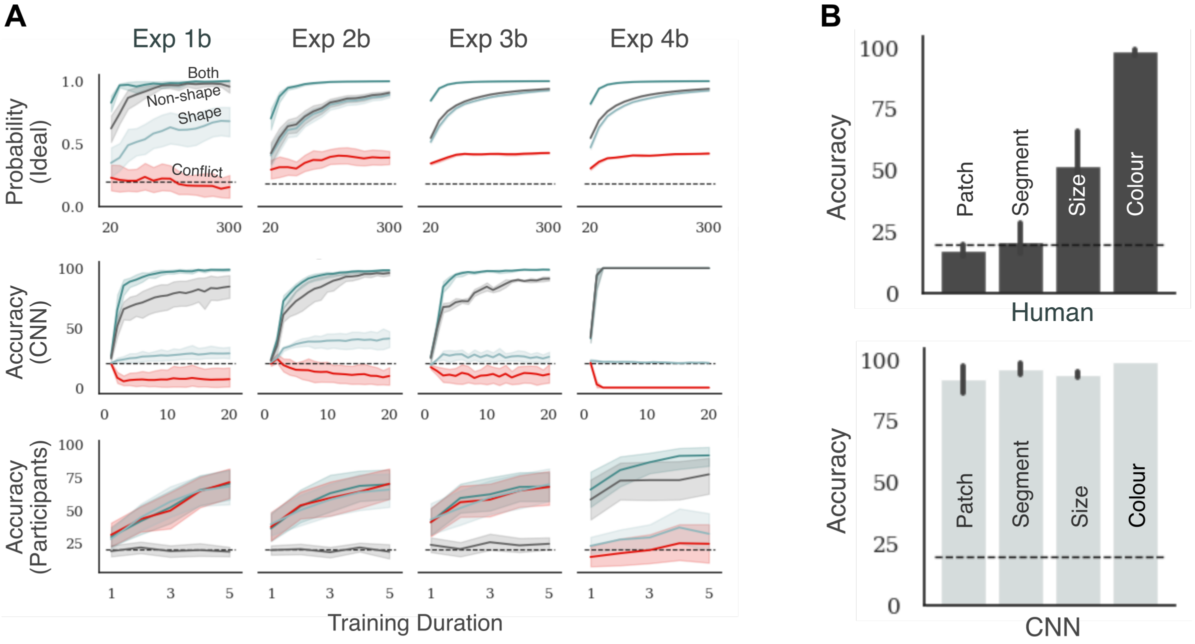


Figure 3.8: **Change in test performance with training in Experiments 1a, 2a, 3a, and 4a.** Each plot in Panel A shows how accuracy on the four types of test trials changes with experience. The top, middle and bottom row correspond to ideal inference model, CNN and human participants respectively. Columns 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 plots in Panel 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.

and segments, but the segments were not consistently organised in a spatial structure (see Figure B.5 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 3.8B. 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 3.8B 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.

3.2.7 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 3.8B 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.

The results of Experiment 6 are shown in Figure 3.9. Like Experiment 5, mean performance across participants was above chance in the Size condition but at chance in the Patch and

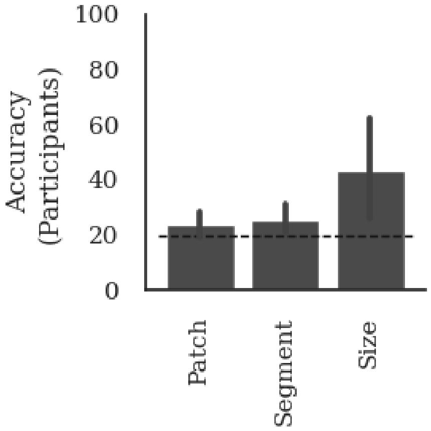


Figure 3.9: **Results of telling participants the diagnostic cue.** Each bar shows mean accuracy across 10 participants in the test block. Participants were divided into three groups based on the diagnostic cue – patch location, segment colour, or average size – used to train the participants.

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.

3.3 Discussion

In a series of experiments we repeatedly observed that participants learned to classify a set of novel objects on the basis of global features such as overall shape and colour even when local non-shape features were more predictive of category (Figure 3.4B). This behaviour is in keeping with psychological studies which show that humans prefer to categorise objects based on shape [12, 105, 108, 126] but, additionally, shows that this shape-bias is retained in novel learning environments where the statistics favoured learning based on a different feature. This observation is consistent with category-learning studies which show that participants overlook salient cues when multiple cues can be used to solve the task [160] and especially in high-dimensional classification tasks [132].

We found that one cannot explain human behaviour using a simple statistical model that infers the category of a test stimulus based solely on the evidence observed in the training trials and no prior biases. We also found that human behaviour was inconsistent with the behaviour of CNNs as the the predictive value of features play a key role in how CNNs learns to classify novel objects. Unlike human participants, previous biases of the network (either learnt through training or built-in through architectural constraints) were not sufficient to overcome this reliance on predictive features. If humans indeed learn in novel environments through a process of statistical learning, these results motivate an exploration of why humans do not quickly adapt to the novel environment in the same way the statistical models presented in this study do. Note that this may be a challenging problem to solve for CNNs and statistical inference models as in Experiment 5 and 6 participants struggle to learn some features even when there is no concurrent shape feature.

Our results were robust across a range of experimental and simulation conditions. Of course, this does not mean that we have controlled for all differences between human experiments and CNN simulations, but our study shows that our findings are robust across multiple CNN architectures, a range of hyper-parameters, different types of pre-training and different types of predictive features (patch location, segment colour, patch size). While we believe that there is unlikely to be a set of experiment conditions that will make humans behave like CNNs, we acknowledge that we do not control for all possible differences in the experimental setup.

Another difference between human participants and CNNs is that participants in our studies had a life-time of exposure to a natural world where shape may be the most diagnostic feature. Indeed, some longitudinal studies have shown that it is possible to create a shape-bias in very young children by intensively teaching them new categories but keeping the statistical properties of their linguistic environment [158]. Accordingly, it is possible that our participants had acquired a shape-bias early on in life [26, 169] that constrained how the new objects in our experiments were learned. But we observed that CNNs did not *retain* a shape-bias 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 if one assumes that CNNs adequately capture why humans learn to classify objects based on their shape, they do not capture why humans continue to look for certain features (like shape) and are agnostic to other features when learning about new objects.

Note, we do not want to claim that humans could never learn to use features other than shape. In Experiment 4, many participants learn to rely on another global feature – the overall colour of objects. And in some of our experiments (for example, where the size of the patches predicted category membership) it is possible that if participants were given a lot more training, some participants may switch to using the more predictive feature. If this were the case, the pattern would be that participants prefer relying on global features early in learning, then switch to more predictive features. The dynamics of the ideal inference model and the CNN (Figure 3.8A) show that neither of the models predict this behaviour.

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 [105, 169]. 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 functional blindness towards certain features cannot necessarily be explained as a shape-bias as there is no competing shape feature to learn.

These findings are consistent with a recent study conducted by Shah et al.[164], who found that CNNs learn to classify images on the basis of simple diagnostic features and ignore more complex features. The focus of Shah et al.[164] was not on comparing CNNs to humans, but rather, showing how a simplicity bias limits generalisation in CNNs. Nevertheless, their study may shed light on another key difference between human and CNN vision, namely, humans are much better at generalising to out-of-distribution image datasets compared to CNNs, such as identifying degraded and distorted images [47, 51]. It may be that that the shape bias we observed in humans but lacking in CNNs plays a role in more robust human visual generalisation.

An important outstanding question is *why* participants in our study relied on global features such as shape or overall colour and struggled to learn salient features that were highly diagnostic. Some of the observations made in our experiments provide clues to the reasons underlying participant behaviour. In Experiment 6, we observed that even when the relevant features are pointed out, participants still could not learn to classify objects based on patch location and segment colour. This shows that the inability to learn these local features is not limited to the difficulty of discovering the type of feature, but may be due to the computational demand of learning how features map to categories. For example, consider the Segment condition in Experiment 6, where the colour of a segment predicted the object category. One strategy to learn this task is to simultaneously store colours of all five segments in memory during each trial and compare these colours across trials of the same category, eliminating colours that do not overlap. This type of strategy will have strained or exceeded the visual capacity of humans, leading them to ignore this predictive cue and focus on shape, even though it is less diagnostic.

Similarly, we also observed that participants 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 in the task). Learning multiple features may lead to better prediction in certain circumstances, however it also requires using more cognitive resources. The fact that participants generally rely on only one feature suggests that participants trade off their performance in the task with the mental effort [89, 166] required to learn how each feature maps to the object category.

By contrast, CNNs 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 [185], something far beyond the capacity of humans [175]. 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 [5], visual crowding [117, 178] and selective attention [110, 186].

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 [33, 47, 163].

To counter these conclusions, there is an argument to be made that a strong shape bias in humans has been demonstrated primarily in behavioural studies. We used performance metrics based on classification behavior. However, researchers claiming similarity of mid-level visual representations with representations in layers of deep networks would argue that this is unfair. There are a number of neural systems between areas such as IT in the ventral stream and the final behevioral output, as well as recurrent and top down processes which are not present in models. Additionally, the fact that networks which were trained to have shape bias still contain non-shape information is not unreasonable. After all, representations in the human visual system certainly encode information about many other features. They would argue that mid-level representations could still be similar and that shape bias emerges downstream. To this end, Ayzenberg and Behrmann [4] discuss whether the ventral stream processes global shape. From their point of view, findings form neuroscience seem to be ambiguous and the contribution of the dorsal stream largely ignored. In related work, Jagadeesh and Gardner [77] conducted an MVPA-style study in which neural patterns from various areas in the ventral stream were used in an 'odd-one out' task (see Figure 3.10). Neural patterns for images with preserved spatial structure and without preserved spatial structure but matched in texture were presented to participants in an attempt to correctly identify the odd stimulus from the set of three. Results revealed that neural patterns can be used to identify the odd-one out if done at category level. When the task is to identify the odd image in a set with two images from one category and a single image from another - neural patterns can be successfully used to do so. However, when the task is to identify a naturalistic (spatial structure preserved) image among synthesized (spatial structure not preserved) images - then it was not possible to do so from BOLD signals in regions of interest from the ventral stream which was also true for activity patterns from layers of CNNs.

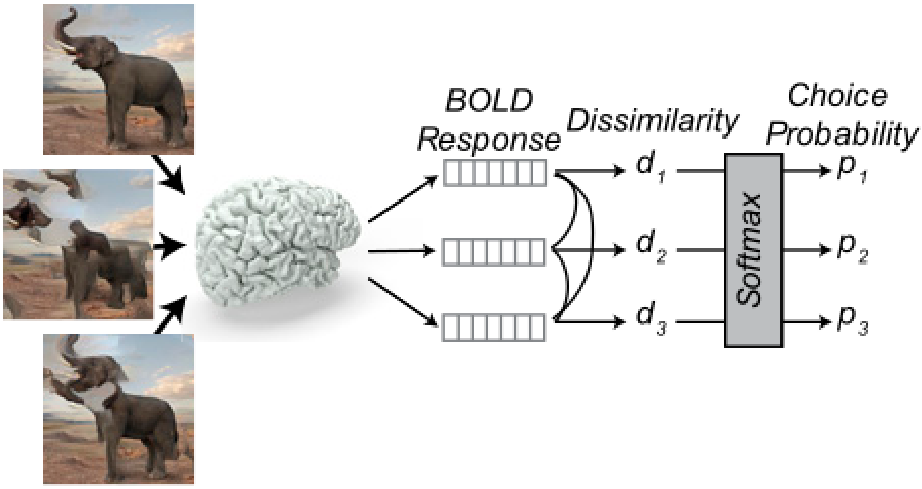


Figure 3.10: **Decoding representations from the ventral stream.** Three images are shown to participant during an fMRI imaging session. Two of the images are synthesized to match the third naturalistic image in local features but do not have spatially preserved structure. Activity patterns are then used in order to idenrify the odd-one out. In this example, the naturalistic image. Figure adapted from Figure 4 in [77].

The researchers conclude that the ventral stream encodes local shape and other local features such as texture which are then combined downstream depending on task demands rather than the ventral stream being particularly sensitive to spatially structured stimuli. It is not clear how these findings reconcile with a number of previous studies showing areas in the ventral stream are indeed sensitive to shape [118, 133, 134] but this approach is the type of research advocated for in this thesis. Further research is clearly required to reconcile the contradicting findings at different levels of analysis as well as demonstrate robustness of these findings across different datasets.

3.4 Methods

Ethics Statement All studies adhered to the University of Bristol ethics guidelines and obtained an ethics approval from the School of Psychological Science Research Ethics Committee (approval code 10350). For all behavioural experiments, we obtained formal consent from participants to use their anonymised data for research.

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*. These shapes were chosen such that the five shapes were clearly distinct from one another. A pilot experiment showed that most participants could learn to categorise based on the chosen shapes within 300 trials. 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, S1 and S2 there were N = 25 participants per experiment (total N = 250), and in Experiments 5, 6, S3 and S4 there were N = 10 participants per experimental condition (total N = 100). In Experiments 1–5 as well as S1-S3 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 and S4 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 7.62 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 a fixed duration and were asked to predict its category label (see Figure 3.11). In Experiments 1–5, this duration was 1000 ms, but we experimented with both longer durations (Experiments 6 and S4) and shorter durations (Experiments S1–S3, see below) and obtained a similar pattern of results. After each training trial, participants were told whether their choice was correct and received feedback on the correct label if their choice was

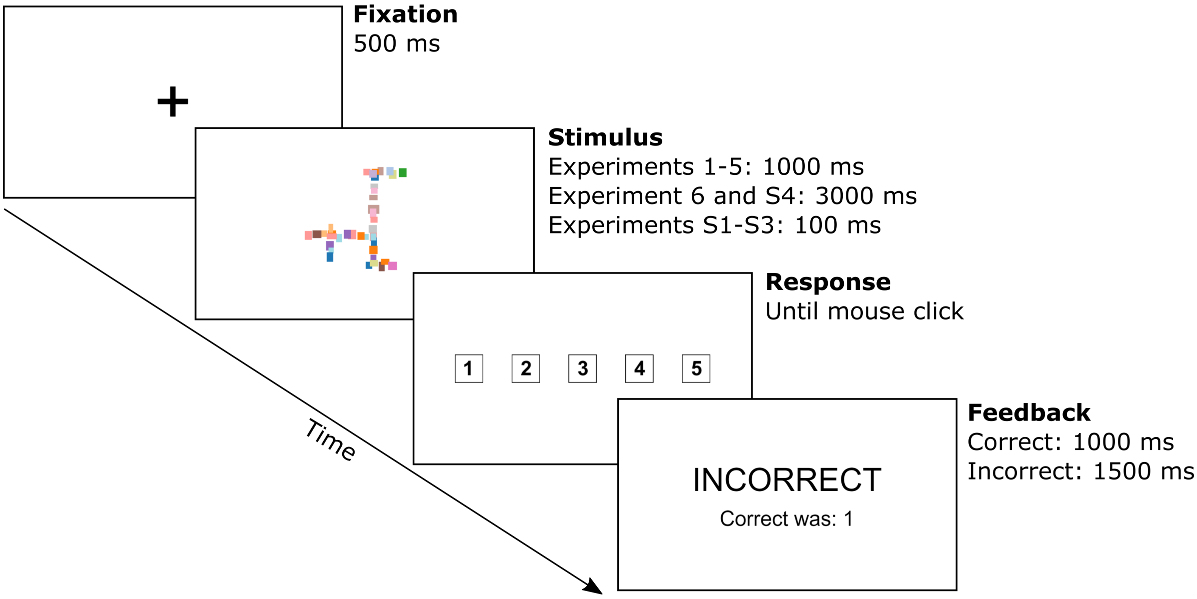


Figure 3.11: **Procedure for human experiments.** Time course for a single training trial in human experiments. The test trials followed an identical procedure, except participants were not given any feedback on their choices.

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 Figure B.7).

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.

In order to ensure that our results are not affected by the presentation time or field of view, we conducted three control experiments. The results of our main experiments (see Figure 3.4) showed that be biggest contrasts between participants and humans were in Experiment 1 and 2, where the diagnostic non-shape feature was the location of a patch or the colour of a segment. Therefore, we conducted three control experiments, reproducing the setup of Experiments 1, 2 and 5 (`Patch` and `Segment` conditions). All details of these control experiments were the same as above, except (i) presentation time of stimulus was reduced to 100ms, (ii) the stimulus was re-scaled so that it was always within 10° visual angle, (iii) instead of testing participants in between every training block, we tested participants only at the end, and (iv) in order to ensure that participants are able to learn the task despite the shorter presentation time, we increased the number of training trials from 300 to 450. We re-scaled the stimulus by using the `ScreenScale` script `https://pavlovia.org/Wake/screenscale`, which has been shown to give good estimates of visual angle in online experiments [19]. Participants were asked to adjust the size of a displayed rectangle to the size of a credit card. To ensure that participants did this correctly, we asked participants to measure the size of a second rectangle and rejected all participants whose measurements did not match the correct size. To compute the visual angle, we asked participants to sit at an arm's length from the screen and asked them to measure the distance between their eyebrow and a fixation cross on the screen. Based on this measurement and how participants re-scaled the displayed rectangle, we re-scaled the stimulus so that the entire image subtended a visual angle of 10°. Participants were reminded to sit at an arm's length at the end of every training block. The results of these control experiments are shown in Appendinx B - Figure B.8.

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.

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 [67], VGG-16 [168] and AlexNet [100] which performs image classification on some image datasets to a near-human standard. We obtained the same pattern of 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 B.9–B.11 Figs. Since evolution and learning both play a role in how the human visual system classifies natural objects, we used a network that was pre-trained on naturalistic images (ImageNet) rather than trained from scratch. However, we observed the same pattern of results for a network that was trained from scratch. 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 [184] 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 [88] 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 3.7), we used a network that was pre-trained on a variation of ImageNet that induces a shape bias [50] 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 [135] and we used torchvision implementation of all models.

Ideal inference model In order to understand how prior biases affect human and CNN classification in the new task environment, we compared their classification to a statistical inference model that computes the ideal category label for a stimulus based solely on the information observed in a sequence of trials. In all our experiments, a trial presented a mapping between a group of features and a category label. The goal of the Ideal Inference model was to accumulate this information over a sequence of trials to predict the mapping in a future trial. It does this by creating a generative model that predicts the probability of observing each feature, given a category label. For example, in Experiment 2, each trial presents a shape, five segment colours and a category label. Based on this information, we can update the generative model, assigning a higher probability for observing the shape and segments observed in the trial, given the class label. Over a sequence of trials, a participant will observe more colours, shapes and category labels and in each trial we can keep adjusting the generative model predicting shapes and colours given the class labels. In a test trial, we can then use the generative model and Bayes' rule to infer the probability of all category labels given the observed shape and colours. We now describe this sequential Bayesian updating procedure formally.

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 $p(Y = c|\mathbf{X} = \mathbf{x_{test}}, \mathscr{D})$, the probability distribution over class labels given the training data, $\mathscr{D}$, and a test image, $\mathbf{x_{test}}$. Using Bayes' law, we have:

$$p(Y = c|\mathbf{X} = \mathbf{x_{test}}, \mathscr{D}) \propto p(\mathbf{X} = \mathbf{x_{test}}|Y = c, \mathscr{D})\;\; p(Y = c) \tag{3.1}$$

where $p(Y = c)$ is the class prior and $p(\mathbf{X} = \mathbf{x_{test}}|Y = c, \mathscr{D})$ 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:

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

Our approach is to estimate these class-conditional densities by constructing a generative model $p(X^f = x_{test}^f|Y = c, \boldsymbol{\theta})$. 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, $Mult(x_{test}^f|1, \boldsymbol{\theta})$. The Bayesian method of estimating these parameters is to start with the prior distribution $p(\boldsymbol{\theta})$ and update it based on training data, $\mathscr{D}$, to obtain the posterior $p(\boldsymbol{\theta}|\mathscr{D})$. 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. Here we assume the flat prior, $\boldsymbol{\alpha} = \mathbf{1}$, which corresponds to Laplace smoothing. 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 [13].

Once we have a posterior distribution on the model parameters, $p(\boldsymbol{\theta}|\mathscr{D})$, we can obtain the required class-conditional densities, $p(X^f = x_{test}^f|Y = c, \mathscr{D})$ by integrating over these parameters. This leads to the following expression (see [124]):

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

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 3.1 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.

In all experiments, the class label, Y can take one of five possible values, that is $Y \in \{1,\ldots,5\}$. In Experiment 1, where the location and colour of a single patch is diagnostic, the feature vector on any trial, $\mathbf{x_{trial}}$, is $(x_{shape}, x^1_{loc}, \ldots, x^F_{loc})$, where $x_{shape} \in \{1,\ldots,5\}$ is a multinomial random variable for the shape feature that can take one of five values, and each of the $x^f_{loc} \in \{1,\ldots 20\}$ is multinomial random variable for a location that can take one of twenty possible colour values (we restricted the number of colours to 20 to make sure colours are clearly discernible by human participants). In Experiment 2, where the colour of one of the segments is diagnostic, the feature vector on a trial, $\mathbf{x_{trial}}$ is $(x_{shape}, x^1_{colour}, \ldots, x^{20}_{colour})$, where x_{shape} is again a multinomial variable for the shape feature that can take one of five values and each $x^f_{colour} \in \{0,\ldots,5\}$ is a count variables that represents the number of segments of colour, f, in the image. In Experiment 3, where the average size of patches is diagnostic, $\mathbf{x_{trial}} = (x_{shape}, x_{size})$, where $x_{size} \in \{1,\ldots,5\}$ is a multinomial random variable for the average size of patches in the image. In Experiment 4, where the global colour of the figure is diagnostic, $\mathbf{x_{trial}} = (x_{shape}, x_{colour})$, where $x_{colour} \in \{1,\ldots,5\}$ is a multinomial random variable representing the global colour of the figure.

Chapter 4

Representational similarity analysis

In previous chapters I used performance measures as our main dependent variable when assessing how humans and models categorize fooling images and a dataset specifically designed to manipulate the statistics of learning environments. While performance metrics can be used very effectively when combined with rigorous experimental design, delving into internal representations provides further opportunities to gain insight into how information is represented throughout networks. Methods derived to do so have heavily borrowed from neuroscience. The family of multivariate pattern analysis methods (MVPA) has in particular been of use to researchers and provides interesting possibilities for future research.

For example, multivariate pattern classification usually takes neural patterns of activation as inputs to a decoder (which could be as simple as a regression model and as complex as a deep neural network) which is then trained to classify stimuli based on the input. The approach has been heavily used to study category selectivity along the ventral visual stream (for an excellent review see Bracci et al. [18]). Similar approaches have been adopted when studying representations of deep neural networks. By placing decoders at various layers of the network, a better understanding of information being represented at various levels can be achieved. For example Doerig et al. [30] train decoders to investigate whether a vernier offset can be decoded at different levels of a pre-trained deep learning model in a study about uncrowding in humans and machines.

More recently, representation similarity analysis (RSA) has become a popular addition to the MVPA family of methods [96]. RSA abstracts away neural tuning information and operates as a second-order measure. Rather than decoding information from activity patterns, it is used to compare representational geometries (how activity patterns between stimuli relate to each other in representational space). The major benefit of this method is that any type of data can be used to compute similarity - e.g., human fMRI patterns and activation vectors from deep networks. However, the flexibility this provides also represents a potential source of confusion regarding the inferences that can be made about similarity of representations using this method. As a second-order metric, it measures similarity between representational similarities - and similar representational geometries (theoretically) could be achieved between systems that represent the external world in quite dissimilar ways. Therefore, inferring that two systems have similar representations because RSA scores are high is a logical fallacy.

This major issue has been hinted at on a theoretical level by Kriegeskorte and Diedrichsen [94], Kriegeskorte and Wei [98] and more concretely by Haxby et al. [65]. However, the method itself has not been used with adequate caution. This is either because both the possibility and plausibility of high RSA scores coming from drastically dissimilar systems is dismissed, or is not taken into account at all. In this chapter a series of simulations demonstrate that it is both possible and plausible that confounds in stimulus datasets can drive RSA scores. Apart from demonstrating these pitfalls, we attempt to revive what is now a neglected debate on the nature of representations.

This is a publication chapter - the chapter is a re-formatted and slightly edited version of the paper *Some pitfalls of measuring representational similarity using representational similarity analysis* which is at the time of writing under reviews, a pre-print can be found at [32].

Contributions - The project was originally initiated and conceptualized by myself. After developing a number of research ideas I (with valuable feedback and commentary from my co-authors) settled on the simulations presented in this chapter. Gaurav Malhotra conducted Study 1 while I developed and conducted all other simulations and accompanying data analyses. All co-authors contributed to write-up and revisions during the review process.

4.1 Introduction

How do other animals see the world? Do different species represent the world in a similar manner? How do the internal representations of AI systems compare with humans and animals? The traditional scientific method of probing internal representations of humans and animals (popular in both psychology and neuroscience) relates them to properties of the external world. By moving a line across the visual field of a cat, Hubel & Wisel [74] found out that neurons in the visual cortex represent edges moving in specific directions. In another Nobel-prize winning work, O'Keefe, Moser & Moser [62, 129] discovered that neurons in the hippocampus and entorhinal cortex represent the location of an animal in the external world. Despite these successes it has proved difficult to relate internal representations to more complex properties of the world. Moreover, relating representations across individuals and species is challenging due to the differences in experience across individuals and differences of neural architectures across species.

These challenges have led to recent excitement around multivariate analyses methods, such as Multi-Voxel Pattern (MVP) Classification, which uses machine learning algorithms to decode neural activity [65]. MVP classification assesses whether a brain region codes for a stimulus feature by examining whether the feature can be easily decoded from neural response patterns in the region. However, there are at least two issues with using MVP classification for comparing mental representations across individuals. Firstly, just because a stimulus feature can be easily decoded from neural response patterns in a region does not imply that downstream regions in the brain actually decode this information [150]. Different individuals (or species) may use this information in different ways and MVP classification does not provide a way of capturing this. Secondly, there are methodological limitations on mapping brain regions and neural activity patterns between individuals and species. Therefore, even if two individuals represent a visual stimulus in the same manner, a decoder trained on one individual will show a significant performance drop when applied across individuals [63].

A newer addition to multivariate analysis, Representation Similarity Analysis (RSA), is specifically designed to compare representations between different systems and overcomes some of these obstacles. RSA usually takes patterns of activity from two systems and computes how the distances between activations in one system correlate with the distances between corresponding activations in the second system (see Figure 4.1). Rather than compare each pattern of activation in the first system directly to the corresponding pattern of activation in the second system, it computes representational distance matrices (RDMs), a *second-order* measure of similarity that compares systems based on the relative distances between neural response patterns. This arrangement of neural response patterns in a representational space has been called a system's *representational geometry* [95]. The advantage of looking at representational geometries is that one no longer needs to match the architecture of two systems, or even the feature space of the two activity patterns (see section 4.2 for a brief history of RSA and its philosophical origins). One could compare, for example, fMRI signals with single cell recordings, EEG traces with behavioural data, or vectors in a computer algorithm with spiking activity of neurons [96]. RSA is now ubiquitous in computational psychology and neuroscience and has been applied to compare object representations in humans and primates [97], representations of visual scenes by different individuals [63, 130], representations of visual scenes in different parts of the brain [120], to study specific processes such as cognitive control [45] or the dynamics of object processing [79], and most recently, to relate neuronal activations in human (and primate) visual cortex with activations of units in Deep Neural Networks [25, 82, 84, 86, 182].

However, this flexibility in the application of RSA comes at the cost of inferences one can draw from this analysis. If the goal of the neuroscientist, psychologist or AI researcher is to establish whether two systems are similar in mechanism, feature representation or information processing, then RSA may not be the correct analytical method to use. This is because RSA is a second-order

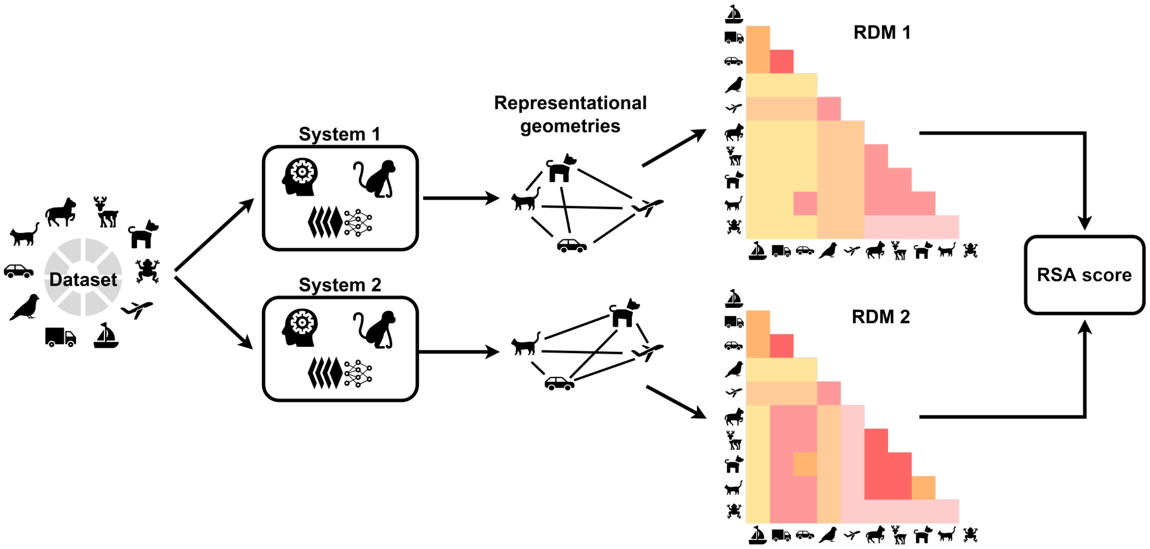


Figure 4.1: **RSA calculation.** Stimuli from a set of categories (or conditions) are used as inputs to two different systems (for example, a human brain and a primate brain). Activity from regions of interest is recorded for each stimulus. Pair-wise distances in activity patterns are calculated to get the representational geometry of each system. This representational geometry is expressed as a representational dissimilarity matrix (RDM) for each system. Finally, an RSA score is determined by computing the correlation between the two RDMs. It is up to the resercher to make a number of choices during this process including the choice of distance measure (e.g., 1-Pearson's r, Euclidean distance etc.) and a measure for comparing RDMs (e.g., Pearson's r, Spearman's ρ, Kendall's τ, etc.).

measure – it looks at the similarity of similarities – that abstracts over mechanism, feature representations and information processing. This point has been made before. For example, Haxby et al. [65] write that the disadvantage of using RSA is that:

> ...one cannot investigate whether the spaces in different subjects share the same feature tuning functions or how these tuning function codes differ for different brain regions. One cannot predict the response to a new stimulus in a subject on the basis of the responses to that stimulus in other subjects. One cannot predict the tuning function for individual neural features in terms of stimulus features, precluding investigators from predicting the response pattern vector for a new stimulus on the basis of its features. (p. 446)

Despite these warnings, RSA continues to be used to infer that different individuals or brain regions or computational models have similar mechanism (that is, they are similar in nested functions and algorithms that transform inputs into neural response vectors). One area where these conclusions are frequently made is the comparison between the hierarchical representations in the visual cortex and Deep Neural Networks (DNNs). For example, Cichy et al. [25] observed a correspondence in the RDMs of DNNs performing object categorization and neural responses in human visual cortex recorded using MEG and fMRI. Based on this correspondence, the authors concluded that:

> ...hierarchical systems of visual representations emerge in both the human ventral and dorsal visual stream as the result of task constraints of object categorization posed in everyday life, and provide strong evidence for object representations in the dorsal stream independent of attention or motor intention. (p. 5)

Thus, the correspondence in RDMs is used to infer the mechanism of emergence of visual representations. Based on a similar comparison, Kriegeskorte [93] concluded that:

> Deep convolutional feedforward networks for object recognition are not biologically detailed and rely on nonlinearities and learning algorithms that may differ from those of biological brains. Nevertheless they learn internal representations that are highly similar to representations in human and nonhuman primate IT cortex. (p. 441)

While authors are sometimes careful in stating that the term 'similarity in representations' is used as a shorthand for a 'similarity in representational geometries', they nevertheless also invite the reader to accept that different systems show similar representational geometries because it is likely that they also use similar mechanisms to transform sensory information into latent representations, or they use similar (downstream) mechanisms to decode these latent representations. But how safe are these assumptions?

The main goal of our paper is to show that high RSA scores should not be used to infer two systems have similar mechanisms. In Study 1, in a bare-bones setup, we show that it is possible for two systems to transform input stimuli through known functions that are vastly different but end up with similar representational geometries. In particular, the study shows that 1) the presence of second-order confounds in the training data can lead systems to mimic each other's representational geometry even in the absence of mechanistic similarity, and 2) the intrinsic structure of datasets rather than mechanistic alignment can lead to artifactual modulation of RSA scores. Then in Studies 2 and 3 we show these problems extend to more complex datasets directly relevant to artificial intelligence and computational neuroscience by making comparisons within and between sets of artificial and biological systems. Finally, in Study 4, we show that not only are misleadingly high RSA scores possible in practice but they are also highly plausible given the hierarchical structure of categories in datasets that are routinely used.

Our demonstrations provide an explanation of how these phenomena, which arise ubiquitously, can lead to incorrect inferences and contradictory or paradoxical findings. For example, it has been recently observed that correlations in representational geometries between human visual cortex and DNNs can vary from being close to the noise ceiling to being uncorrelated based on the visual stimuli used in the experiments [180]. Since our results have considerable generality with respect to current practices across multiple fields, we discuss the implications for published results, including a discussion of two alternative philosophical perspectives on the nature of mental representations that our findings speak to. We conclude by providing some general recommendations regarding how to best use RSA going forward.

4.2 The nature of representation and a brief history of RSA

In the 1990s there was an important debate taking place on how to compare the mental representations of two individuals. On one side of this debate was Paul Churchland. Inspired by the success of connectionist models, Churchland argued that the brain represents reality as a pattern of activations over its network of neurons [22]. This pattern of activation can be seen as a position in the brain's (high-dimensional) state-space. So, Churchland argued that one could compare how two individuals represent an object by comparing the corresponding positions in each individual's state-space. On the other side of the debate were Jerry Fodor and Ernie Lepore [43]. They pointed out that a problem with Churchland's proposal was that it 'offers no robust account of content identity' (p. 147). On Churchland's account, they argued, two mental representations have the same meaning only if they are embedded in identical state-spaces. This condition was highly unlikely to be satisfied in practice, given that no two brains have either the same number or connectivity of neurons and no two individuals have exactly the same experiences.

A possible solution to this problem of comparing representations across state-spaces of different dimensions was proposed by Laasko and Cottrell [103], who were investigating whether different neural networks, trained on the same data, represented an input stimulus in a similar manner. A direct comparison of activations across networks was not possible due to the difference in the number of units. To overcome this problem, [103] devised a method that compared encodings based on their *relative* positions in state-space. That is, based on a second-order isomorphism. They argued that two networks could be said to represent a concept in a similar manner if both networks partitioned their activation space (amongst concepts) in a similar manner – that is, if the activation spaces in both systems had a similar *geometry*. [103] conducted a series of experiments with neural networks, showing that neural networks with different sensory encodings and different number of hidden units nevertheless partitioned their activation space in a similar manner, leading them to conclude that these networks learned similar internal representations.

Churchland [23] saw Laasko and Cottrell's method as a decisive response to Fodor & Lepore's scepticism. He argued that, using Laasko and Cottrell's method, one could use the state-space approach to compare representations across individuals, even individuals that had different dimensions of their representational spaces. All one needed to do was to replace the requirement of 'content identity' with 'content similarity'. That is, instead of comparing absolute positions of representations, one could simply compare how representations were organised *relative* to each other within each representational space.

However, Fodor & Lepore [44] argued that Churchland's reply was, in fact, 'an egregious *ignoratio elenchi*' (p. 382). The problem was *not*, they argued, that one couldn't find the right metric to measure similarity across vector spaces of different dimensions. Rather, it was the fact that Churchland (and Laasko & Cottrell [103]) were interested in a *semantic* similarity – i.e., they wanted to compare whether representations had the same meaning in the two systems. Fodor & Lepore [44] argued that this problem of semantic similarity was intractable because similarity of concepts across systems of different dimensions is undefined. Consider the concept of a ’dog’. Let’s say one person’s representational space has a dimension of ’loyalty’, while the other person’s representational space does not. There is no principled answer for how similar the representation of ‘dog’ should be for these two individuals as it depends on how the dimension of ‘loyalty’ is weighted in the concept of ‘dog’. And the relative weight of dimensions can differ for different concepts and circumstances. Moreover, Fodor & Lepore [44] argued that even identical representational geometries could *mean* very different things. For example, one individual may represent a dog along the dimensions of ‘size’ and ‘speed’ as being small (sized) and medium (speed). Another individual may represent a dog along the dimensions of ‘usefulness’ and ‘furriness’ as being of small (usefulness) and medium (furriness). Even if the concept of a dog occupies a similar position in both state-spaces (small, medium) the two individuals clearly represent dogs differently.

Representation Similarity Analysis is an evolution of Laasko and Cottrell’s method for comparing representations across systems. It retains its core principle of comparing representations based on their relative locations within each system’s state-space. In addition, it formalises the ideas of similarity of representations within and across systems [96]. Like Laasko and Cottrell’s method, a representation is usually coded as a vector of activation over some units (in a neural network or the brain). However, it could also be a behavioural measure, such as similarity judgments or even measures like accuracy or response times. We believe that many of the objections levelled by Fodor & Lepore against Churchland’s idea of comparing systems based on relative positions in state-space also hold for representation similarity analysis. For example, Fodor & Lepore’s point that similar state-space representations could mean different things can also be extended to RSA and in the main text we show how different systems with same representational geometries can, in fact, be encoding very different properties of sensory stimuli. From an externalist’s perspective, activations within these systems *mean* very different things and yet have very comparable state-space representations (i.e. geometries). The only way to argue that concepts have a similar meaning in systems with similar representational geometries is to adopt a holistic perspective on representations. And as Fodor & Lepore [44] argued, and we discuss in the main text, adopting this perspective comes with its own set of problems.

4.3 Results

It may be tempting to infer that two systems which have similar representational geometries for a set of concepts do so because they encode similar properties of sensory data and transform sensory data through a similar set of functions. In section 4.3.1, we show that it is possible, at least in principle, for qualitatively different systems to end up with very similar representational geometries even though they (i) transform their inputs through very different functions, and (ii) select different features of inputs.

4.3.1 Study 1: Demonstrably different transformations of inputs can lead to low or high RSA-scores

We start by considering a simple two-dimensional dataset and two systems where we know the closed-form functions that project this data into two representational spaces. This simple setup helps us gain a theoretical understanding of the circumstances under which it is possible for qualitatively different projections to show similar representational geometries.

Consider a population of animate and inanimate objects that consist of four categories of objects – birds, dogs, airplanes and bicycles. Each object in this population will have a set of stimulus features, using which one can map each exemplar from all four categories into a feature space. In Figure 4.2A (left), we show a hypothetical 2D feature space where exemplars from each category cluster together. Futhermore, we consider two datasets sampled from this population – Dataset A (Figure 4.2A, middle) which consists of birds and bicycles and Dataset B (Figure 4.2A, right) which consists of dogs and airplanes. Both datasets consist of animate and inanimate objects, but they differ in how items in each category are represented in the input space.

Now, consider two information-processing systems that re-represent Dataset A into two different latent spaces (Figure 4.2B). These could be two recognition systems designed to distinguish animate and inanimate categories. We assume that we can observe the representational geometry of the latent representations of each system and we are interested in understanding whether observing a strong correlation between these geometries implies whether the two systems have a similar *representational space* – that is, they project inputs into the latent space using similar functions. To examine this question, we consider a setup where we know the functions, Φ_1 and Φ_2, that map the inputs to the latent space in each system. We will now demonstrate that even when these functions are qualitatively different from each other, the geometry of latent representations can nevertheless be highly correlated. We will also show that the difference in representational spaces becomes more clear when one considers a different dataset (Dataset B), where inputs projected using the same functions now lead to a low correlation in representational geometries.

We can compute the geometry of a set of representations by establishing the pair-wise distance between all vectors in each representational space Φ. There are many different methods of computing this representational distance between any pair of vectors, all deriving from the dot product between vectors (see, for example, Figure 1 in Bobadilla-Suarez et al. [16]). Previous research has shown that the choice of the distance metric itself can influence the inferences one can draw from one's analysis [16, 147]. However, here our focus is not the distance metric itself, but the fundamental nature of RSA. Therefore, we use the same generic distance metric – the dot product – to compute the pair-wise distance between all vectors in both representational spaces. In other words, the representational distance $d[\Phi(\boldsymbol{x_i}), \Phi(\boldsymbol{x_j})]$, between the projections of any pair of input stimuli, $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ into a feature space Φ, is proportional to the inner product between the projections in the feature space:

$$d[\Phi(\boldsymbol{x_i}),\Phi(\boldsymbol{x_j})] \propto \Phi(\boldsymbol{x_i})\cdot\Phi(\boldsymbol{x_j}) \tag{4.1}$$

And we can obtain the representational geometry of the input stimuli $\{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}\}$ in any representational space Φ by computing the pairwise distances, $d[\Phi(\boldsymbol{x_i}),\Phi(\boldsymbol{x_j})]$ for all pairs of data points, (i,j). Here, we assume that the projections Φ_1 and Φ_2 are such that these pairwise distances are given by two positive semi-definite kernel functions $\kappa_1(\boldsymbol{x_i},\boldsymbol{x_j})$ and $\kappa_2(\boldsymbol{x_i},\boldsymbol{x_j})$, respectively:

$$d[\Phi_1(\boldsymbol{x_i}),\Phi_1(\boldsymbol{x_j})] \propto \Phi_1(\boldsymbol{x_i})\cdot\Phi_1(\boldsymbol{x_j}) = \kappa_1(\boldsymbol{x_i},\boldsymbol{x_j}) \tag{4.2}$$

$$d[\Phi_2(\boldsymbol{x_i}),\Phi_2(\boldsymbol{x_j})] \propto \Phi_2(\boldsymbol{x_i})\cdot\Phi_2(\boldsymbol{x_j}) = \kappa_2(\boldsymbol{x_i},\boldsymbol{x_j}) \tag{4.3}$$

Now, let us consider two qualitatively different kernel functions: $\kappa_1(\boldsymbol{x_i},\boldsymbol{x_j}) = e^{\frac{||\boldsymbol{x_i}-\boldsymbol{x_j}||^2}{2\sigma^2}}$ is a radial-basis kernel (where σ^2 is the bandwidth parameter of the kernel), while $\kappa_2(\boldsymbol{x_i},\boldsymbol{x_j}) = \frac{\boldsymbol{x_i^T x_j}}{||\boldsymbol{x_i}||||\boldsymbol{x_j}||}$ is a cosine kernel. In other words, Φ_1 and Φ_2 are two fundamentally different projections of the inputs $\{\boldsymbol{x_1},\ldots,\boldsymbol{x_n}\}$ – while Φ_2 maps a 2D input $\boldsymbol{x_i}$ into a 2D feature space, Φ_1 maps the same 2D input into an infinite-dimensional space. Nevertheless, since cosine and RBF kernels are Mercer kernels, we can compute the distances (as measured by the dot product) between each pair of projected vectors using the kernel trick [157, 161]. That is, we can find the distance between any pair of points in the representational space by applying the kernel function to those points in the input space. These pairwise distances are shown by the kernel matrices in Figure 4.2B.

Next, we can determine how the geometry of these projections in the two systems relate to each other by computing the correlation between the kernel matrices, shown on the right-hand-side of Figure 4.2B. We can see from these results that the kernel matrices are highly correlated – i.e., the input stimuli are projected to very similar geometries in the two representational spaces.

If one did not know the input transformations and simply observed the correlation between kernel matrices, it would be tempting to infer that the two systems Φ_1 and Φ_2 transform an unknown input stimulus $\boldsymbol{x}$ through a similar set of functions – for example functions that belong to the same class or project inputs to similar representational spaces. However, this would be an error. The projections $\Phi_1(\boldsymbol{x})$ and $\Phi_2(\boldsymbol{x})$ are fundamentally different – Φ_1 (radial basis kernel) projects an input vector into an infinite dimensional space, while Φ_2 (cosine kernel) projects it onto a unit sphere. The difference between these functions becomes apparent if one considers how this correlation changes if one considers a different set of input stimuli. For example, the set of data points from Dataset B (sampled fromt the same population) are projected to very different geometries, leading to a low correlation between the two kernel matrices (Figure 4.2C).

In fact, the reason for highly correlated kernel matrices in Figure 4.2B is not a similarity in the transformations Φ_1 and Φ_2 but the structure of the dataset. The representational distance

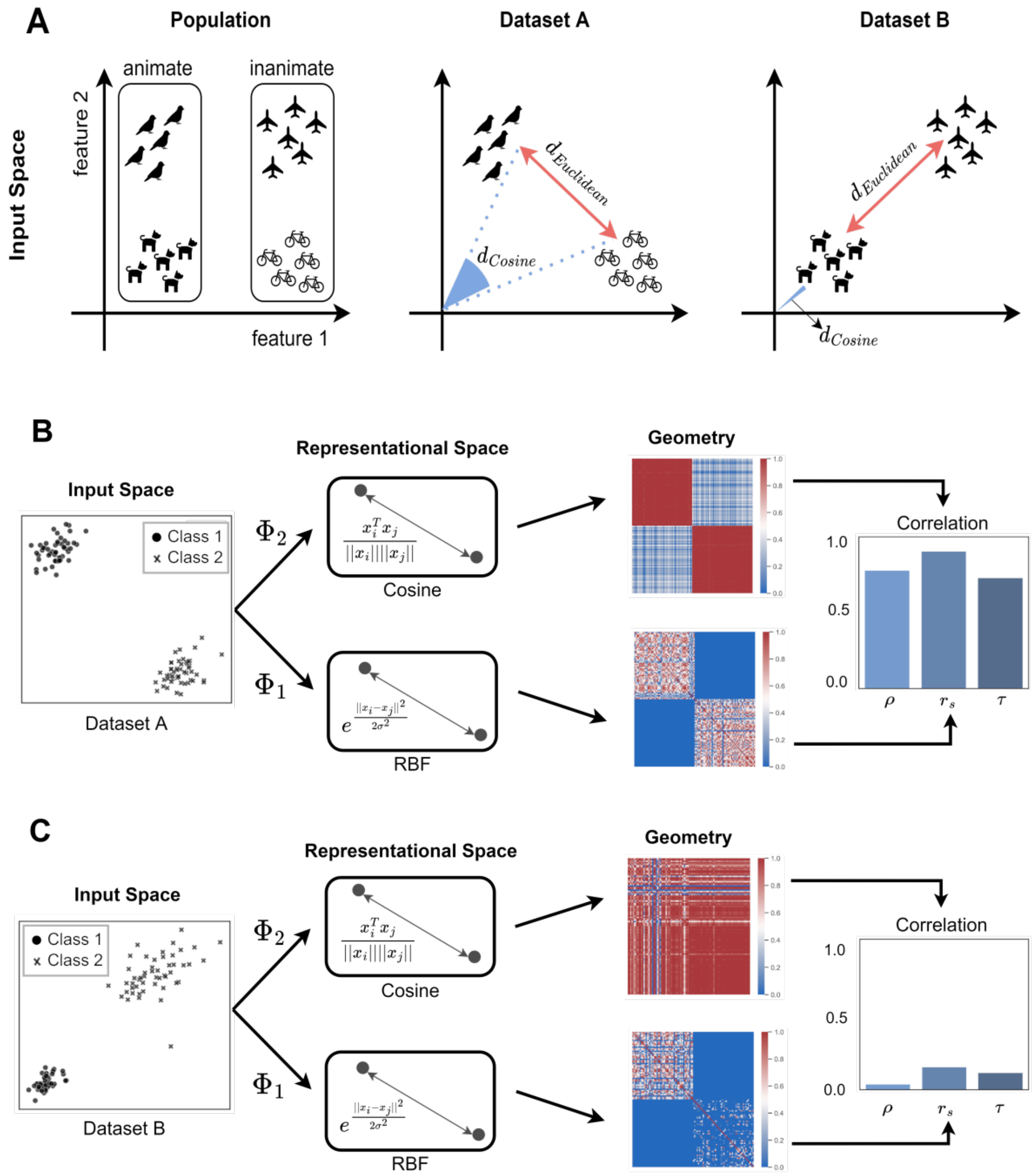


Figure 4.2: **Mimic and modulation effect in representational geometries.** (A) An example of a population of animate (birds, dogs) and inanimate (planes, bikes) objects, plotted in a hypothetical 2D stimulus feature space. Two datasets are sampled from this population: In Dataset A (middle), the Euclidean distance (in input space) between categories mirrors the Cosine distance, while in Dataset B (right) it does not. (B) Simulation where two systems transform stimuli in Dataset A into latent representations such that the (dot product) distance between latent vectors is given by RBF and Cosine kernels, respectively. As Euclidean and Cosine distances in the input space mirror each other, the representational geometries (visualised here using kernel matrices) end up being highly correlated (shown using Pearson (ρ), Spearman (r_s) and Kendall's (τ) correlation coefficients on the right). We call this strong correlation in representational geometries despite a difference in input transformation a *mimic effect*. (C) Simulation where objects in Dataset B are projected using same transformations as (B). The (dot product) distance is still given by the same (RBF and Cosine) kernels. However, for this dataset, the Euclidean and Cosine distances in input space do *not* mirror each other and as a consequence, the representational geometries show low correlation. Thus the correlation in representational geometries depends on how the datasets are sampled from the population. We call this change in correlation a *modulation effect*.

between any two points in the first representational space, $d[\Phi_1(\boldsymbol{x_i}), \Phi_1(\boldsymbol{x_j})]$, is $e^{\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{2\sigma^2}}$. That is, the representational distance in Φ_1 is a function of their Euclidean distance $||\boldsymbol{x_i} - \boldsymbol{x_j}||$ in the input space. On the other hand, the representational distance between any two points in the second representational space, $d[\Phi_2(\boldsymbol{x_i}), \Phi_2(\boldsymbol{x_j})]$, is, $\frac{\boldsymbol{x_i^T x_j}}{||\boldsymbol{x_i}||||\boldsymbol{x_j}||}$. That is, the representational distance in Φ_2 is a function of their cosine distance in the input space. These two stimulus features – Euclidean distance and cosine distance – are *confounds* that lead to the same representational geometries for certain datasets. In Dataset A, the stimuli is clustered such that the Euclidean distance between any two stimuli is correlated with their cosine distance (see Figure 4.2A, middle). However, for Dataset B, the Euclidean distance is no longer correlated with the angle (see Figure 4.2A, right) and the confounds lead to different representational geometries, as can be seen in Figure 4.2C. Thus, this example illustrates two effects: (i) a *mimic* effect, where two systems that transform sensory input through very different functions end up with similar representational geometries (Figure 4.2B), and (ii) a *modulation* effect, where two systems that are non-identical have similar representational geometries for one set of inputs, but dissimilar geometries for a second set (compare Figure 4.2B and 4.2C).

4.3.2 Study 2: Complex systems encoding different features of inputs can show a high RSA-score

Study 1 made a number of simplifying assumptions – the dataset was two-dimensional, clustered into two categories and we intentionally chose functions Φ_1 and Φ_2 such that the kernel matrices were correlated in one case and not correlated in the other. It could be argued that, even though the above results hold in principle, they are unlikely in practice when the transformations and data structure are more complex. For example, it might be tempting to assume that accidental similarity in representational geometries becomes less likely as one increases the number of categories (i.e., clusters or conditions) being considered. However, In Figure 4.3 we illustrate how complex systems transforming high-dimensional input from a number of categories may achieve high RSA scores. Even though one system extracts surface reflectance and the other extracts global shape, they can end up with very similar representational geometries. This would occur if objects similar in their reflectance properties were also similar in shape (e.g., glossy balloons and light bulbs) and if objects dissimilar according to reflectance properties were also dissimilar in shape (e.g., dogs and light bulbs). This is the mimic effect, where representational geometries of these two systems end up being similar because reflectance and shape are second-order confounds in this dataset. Conducting RSA on this dataset will show a high correlation in RDMs, even though the latent representations in these systems are related to very different stimulus features.

To demonstrate this empirically, we now consider a more complex setup, where the transformations Φ_1 and Φ_2 are modelled as feedforward deep neural networks (DNNs), trained to classify a high-dimensional dataset into multiple categories. Many studies that use RSA compare systems using naturalistic images as visual inputs [97, 182]. While using naturalistic images

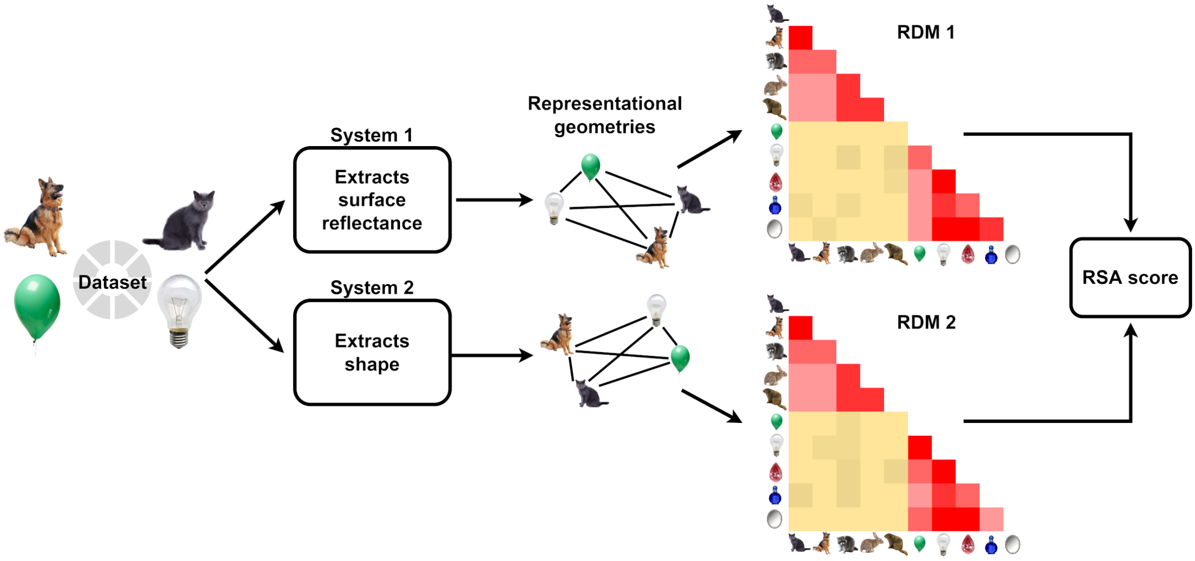


Figure 4.3: **Example of a second-order confound.** Two systems, one forming representations based on surface reflectance of objects (while ignoring all other features such as colour or texture) and the other based on global shape (while ignoring other features), can have very similar representational geometries. This similarity would lead to a high RSA score but would not justify an inference about the representations being similar.

brings research closer to the real-world, it is also well-known that datasets of naturalistic images frequently contain confounds – independent features that can predict image categories [174]. We will now show how the simplest of such confounds, a single pixel, can lead to a high RSA score between two DNNs that encode qualitatively different features of inputs.

Consider the same setup as above, where an input stimulus, $\boldsymbol{x}$, is transformed to a representation space by two systems, Φ_1 and Φ_2. Instead of a two-dimensional input space, $\boldsymbol{x}$ now exists in a high-dimensional image space and Φ_1 and Φ_2 are two versions of a DNN – `VGG-16` – trained to classify input images into different categories. We ensured that Φ_1 and Φ_2 were qualitatively different transformations of input stimuli by making the networks sensitive to different predictive features within the stimuli. The first network was trained on an unperturbed dataset, while the second network was trained on a modified version of the dataset, where each image was modified to contain a confound – a single pixel in a location that was diagnostic of the category (see Figure 4.4 for the general approach).

The locations of these diagnostic pixels were chosen such that they were correlated to the corresponding representational distances between classes in Φ_1. Our hypothesis was that if the representational distances in Φ_2 preserve the physical distances of diagnostic pixels in input space, then this confound will end up mimicking the representational geometry of Φ_1, even though the two systems use qualitatively different features for classification. Furthermore, we trained two more networks, Φ_3 and Φ_4, which were identical to Φ_2, except these networks were trained on datasets where the location of the confound was uncorrelated (Φ_3) or negatively correlated (Φ_4) with the representational distances in Φ_1 (see Figure 4.5 and Methods for details).

Classification accuracy (Figure 4.6 (left)) revealed that the network Φ_1, trained on the

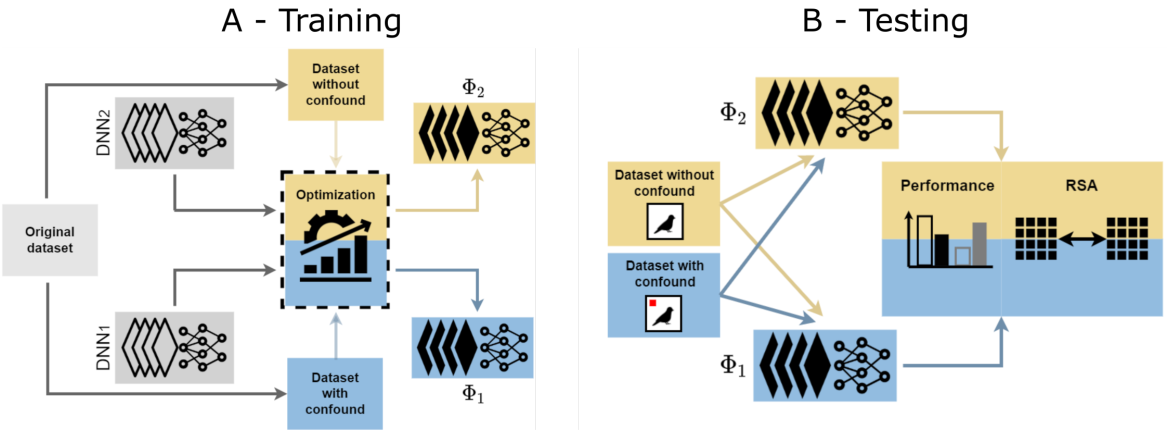


Figure 4.4: **Training and testing DNNs with different feature encodings.** Panel A shows the training procedure for Studies 2–4, where we created two versions of the original dataset (gray), one containing a confound (blue) and the other left unperturbed (yellow). These two datasets were used to train two networks (gray) on a categorisation task, resulting in two networks that learn to categorise images either based on the confound (projection Φ_2) or based on statistical properties of the unperturbed image (projection Φ_1). Panel B shows the testing procedure where each network was tested on stimuli from each dataset – leading to a 2x2 design. Performance on these datasets was used to infer the features that each network encoded and their internal response patterns were used to calculate RSA-scores between the two networks.

unperturbed images, learned to classify these images and ignored the diagnostic pixel – that is, it's performance was identical for the unperturbed and modified images. In contrast, networks Φ_2 (positive), Φ_3 (uncorrelated) and Φ_4(negative) failed to classify the unperturbed images (performance was near chance) but learned to perfectly classify the modified images, showing that these networks develop qualitatively different representations compared to normally trained networks.

Next we computed pairwise RSA scores between the representations at the last convolution layer of Φ_1 and each of Φ_2,Φ_3 and Φ_4 (Figure 4.6 (right)). When presented unperturbed test images, the Φ_2,Φ_3 and Φ_4 networks all showed low RSA scores with the normally trained Φ_1 network. However, when networks were presented with test images that included the predictive pixels, RSA varied depending on the geometry of pixel locations in the input space. When the geometry of pixel locations was positively correlated to the normally trained network, RSA scores approached ceiling (i.e., comparable to RSA scores between two normally trained networks). Networks trained on uncorrelated and negatively correlated pixel placements scored much lower.

These results mirror Study 1: we observed that it is possible for two networks (Φ_1 and Φ_2) to show highly correlated representational geometries even though these networks learn to classify images based on very different features. One may argue that this could be because the two networks could have learned similar representations at the final convolution layer of the DNN and it is the classifier that sits on top of this representation that leads to the behavioural differences between these networks. But if this was true, it would not explain why RSA scores diminish for the two other comparisons (with Φ_3 and Φ_4). This modulation of RSA-scores for different datasets suggests that, like in Study 1, the correlation in representational geometry is

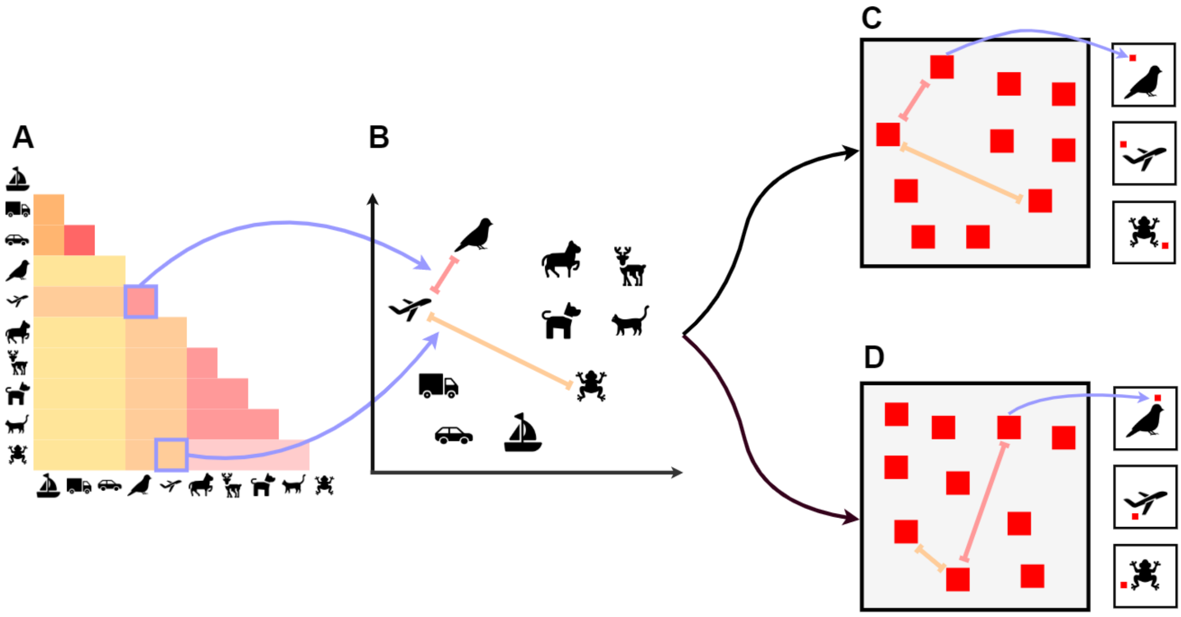


Figure 4.5: **Study 2 confound placement.** The representational geometry (Panel A and B) from the network trained on the unperturbed `CIFAR-10` images is used to determine the location of the single pixel confound (shown as a red patch here) for each category. In the 'Positive' condition (Panel C), we determined 10 locations in a 2D plane such that the distances between these locations were positively correlated to the representational geometry – illustrated here as the red patches in Panel C being in similar locations to category locations in Panel B. These 10 locations were then used to insert a single diagnostic – i.e., category-dependent – pixel in each image (Insets in Panel C). A similar procedure was also used to generate datasets where the confound was uncorrelated (Panel D) or negatively correlated (not shown here) with the representational geometry of the network.

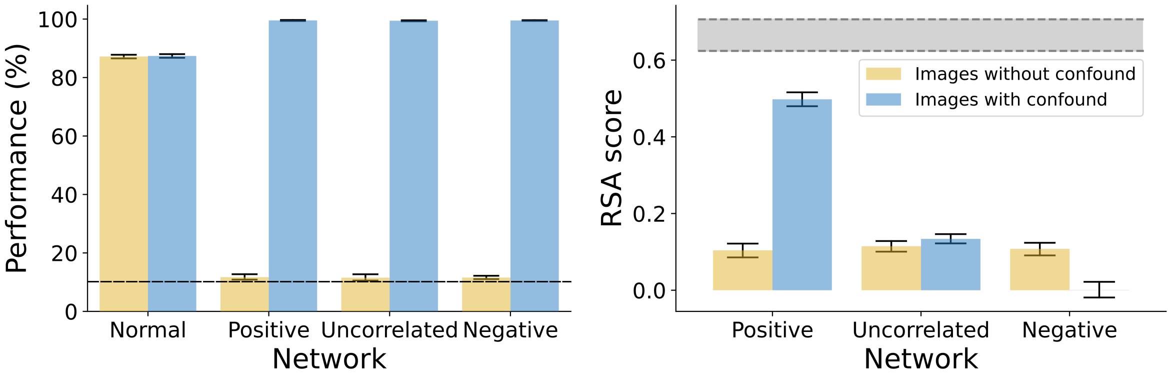


Figure 4.6: **Study 2 results.** *Left:* Performance of normally trained networks did not depend on whether classification was done on unperturbed `CIFAR-10` images or images with a single pixel confound (error bars represent 95% CI, the dashed line represents chance performance). All three networks trained on datasets with confounds could perfectly categorise the test images when they contained the confound (blue bars), but failed to achieve above-chance performance if the predictive pixel was not present (yellow bars). *Right:* The RSA score between the network trained on the unperturbed dataset and each of the networks trained on datasets with confounds. The three networks showed similar scores when tested on images without confounds, but vastly different RSA scores when tested on images with confounds. Networks in the Positive condition showed near ceiling scores (the shaded area represents noise ceiling) while networks in the Uncorrelated and Negative conditions showed much lower RSA.

not because the two systems encode similar features of inputs, but because different features mimic each other in their representational geometries.

In Studies 1 and 2, we showed that it is possible for qualitatively different systems to end up with similar representational geometries. However, it may be argued that while this is possible in principle, it is unlikely in practice in real-world scenarios. In the following two studies, we consider real-world data from some recent influential experiments, recorded from both primate and human cortex. We show how RSA-scores can be driven by confounds in these real-world settings and how properties of training and test data may contribute to observed RSA-scores.

4.3.3 Study 3: Neural activations in monkey IT cortex can show a high RSA-score with DNNs despite different encoding of input data

In our next study, we consider data from experiments comparing representational geometries between computational models and macaque visual cortex [159, 182]. The experimental setup was similar to Study 2, though note that unlike Study 2, where both systems used the same architecture and learning algorithm, this study considered two very different systems – one artificial (DNN) and the other biological (macaque IT cortex). We used the same set of images that were shown to macaques by Majaj et al. [111] and modified this dataset to superimpose a small diagnostic patch on each image. In the same manner as in Study 2 above, we constructed three different datasets, where the locations of these diagnostic patches were either positively correlated, uncorrelated or negatively correlated with the RDM of macaque activations. We then trained four CNNs. The first CNN was pre-trained on `ImageNet` and then fine-tuned on the unmodified dataset of images shown to the macaques. Previous research has shown that CNNs trained in this manner develop representations that mirror the representational geometry of neurons in primate inferior temporal (IT) cortex [182]. The other three networks were trained on the three modified datasets and learned to entirely rely on the diagnostic patches (accuracy on images without the diagnostic patches was around chance).

Figure 4.7 (right) shows the correlation in representational geometry between the macaque IT activations and activations at the final convolution layer for each of these networks. The correlation with networks trained on the unmodified images is our baseline and shown as the gray band in Figure 4.7. Our first observation was that a CNN trained to rely on the diagnostic patch can indeed achieve a high RSA score with macaque IT activations. In fact, the networks trained on patch locations that were positively correlated to the macaque RDM matched the RSA score of the CNNs trained on `ImageNet` and the unmodified dataset. This shows how two systems having very different architectures, encoding fundamentally different features of inputs (single patch vs naturalistic features) can show a high correspondence in their representational geometries. We also observed that, like in Study 2, the RSA score depended on the clustering of data in the input space – when patches were placed in other locations (uncorrelated or negatively correlated to macaque RDMs) the RSA score became significantly lower.

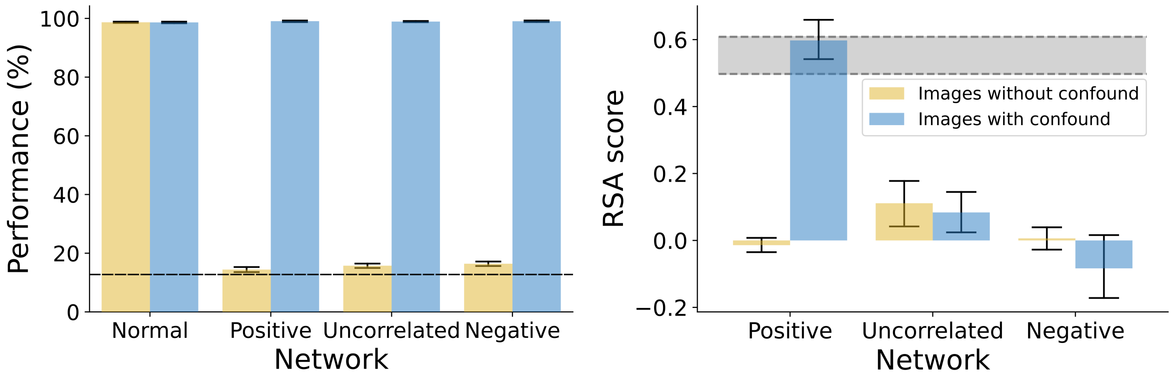


Figure 4.7: **Study 3 results.** *Left:* Classification Performance of the network trained on unperturbed images (Normal condition) did not depend on the presence or absence of the confound, while performance of networks trained with the confound (Positive, Uncorrelated and Negative conditions) highly depended on whether the confound was present (dashed line represents chance performance). *Right:* RSA-scores with macaque IT activations were low for all three conditions when images did not contain a confound (yellow bars). When images contained a confound (blue bars), the RSA-scores depended on the condition, matching the RSA-score of the normally trained network (grey band) in the Positive condition, but decreasing significantly in the Uncorrelated and Negative conditions. The grey band represents a 95% CI for the RSA-score between normally trained networks and macaque IT activations.

It is quite usual to compute RSA scores on category-level RDMs like the score in Figure 4.7 (right panel) or on a smaller number of stimuli per category (as was the case in Study 2). It is rare to see RSA scores being computed using a large number of items per category. Therefore, in Figure 4.8 we show the pattern of RSA scores when computed on RDMs of size 3200×3200 (all of the stimuli in the dataset). The overall pattern of results remains the similar - RSA scores in the Positive condition are high (even above that of normally trained networks) with lower scores in the Uncorrelated and Negative conditions. The difference being that RSA scores were overall lower than in the category-level analysis and that in the Uncorrelated and Negative conditions were higher relative to the Positive condition (when the confound was present). The lower overall scores can be explained by the increase of noise when computing over so many stimuli, this would be even more pronounced in a dataset such as the full CIFAR-10 or large subsets of stimuli from ImageNet since compared to this, more curated, dataset. This is because within-category similarity in naturalistic datasets tends to be rather low. The relatively higher scores, particularly in the Negative condition, can be explained by the increased number of stimuli per category. RSA scores with multiple items per category (and without averaging) will depend on both within-category similarity and on whether relations between categories is similar between two systems. As long as compared systems represent stimuli within a category as more similar to each other than to stimuli from other categories - the larger the number of items per category, the larger the RSA score between the two systems. Thus, even though networks in the Negative condition were specifically trained to achieve low scores (and on a category level they do so as can be seen in Figure 4.7), when there are 400 images per category the RSA score gets inflated even if the between-category structure is dissimilar to macaque IT. Therefore, the selection of dataset and number of stimuli/conditions can impact scores in a significant way. For a simulation of how completely unrelated systems can achieve significant RSA scores due to a larger number of within-category stimuli see Appendix C.

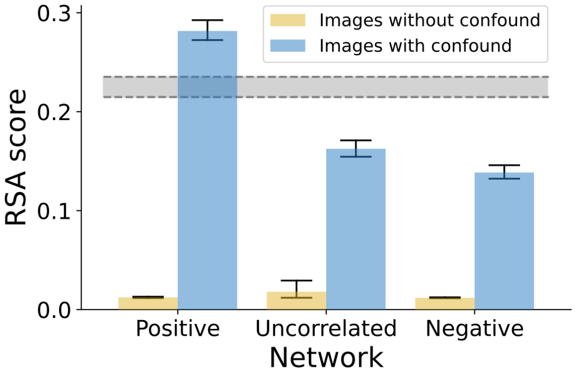


Figure 4.8: **Image-level RSA scores from Study 3.** RSA-scores with macaque IT activations were low for all three conditions when images did not contain a confound (yellow bars). When images contained a confound (blue bars), the RSA-scores depended on the condition, even exceeding the RSA-score of the normally trained network (grey band) in the Positive condition, but decreasing significantly in the Uncorrelated and Negative conditions. The grey band represents a 95% CI for the RSA-score between normally trained networks and macaque IT activations.

4.3.4 Study 4: High RSA-scores may be driven by the structure of testing data

All the studies so far have used the same method to construct datasets with confounds – we established the representational geometry of one system (Φ_1) and constructed datasets where the clustering of features (pixels) mirrored this geometry. However, it could be argued that confounds which cluster in this manner are unlikely in practice. For example, even if texture and shape exist as confounds in a dataset, the inter-category distances between textures are not necessarily similar to the inter-category distances between shape.

However, categories in real-world datasets are usually hierarchically clustered into higher-level and lower-level categories. For example, in the CIFAR-10 dataset, the Dogs and Cats (lower-level categories) are both animate (members of a common higher-level category) and Airplanes and Ships (lower-level categories) are both inanimate (members of a higher-level category). Due to this hierarchical structure, Dog and Cat images are likely to be closer to each other not only in their shape, but also their colour and texture (amongst other features) than they are to Airplane and Ship images. In our next simulation, we explore whether this hierarchical structure of categories can lead to a correlation in representational geometries between two systems that learn different feature encodings.

For this study, we selected a popular dataset used for comparing representational geometries in humans, macaques and deep learning models [82, 92]. This dataset consists of six categories which can be organised into a hierarchical structure shown in Figure 4.9. Kriegeskorte et al. [97] showed a striking match in RDMs for response patterns elicited by these stimuli in human and

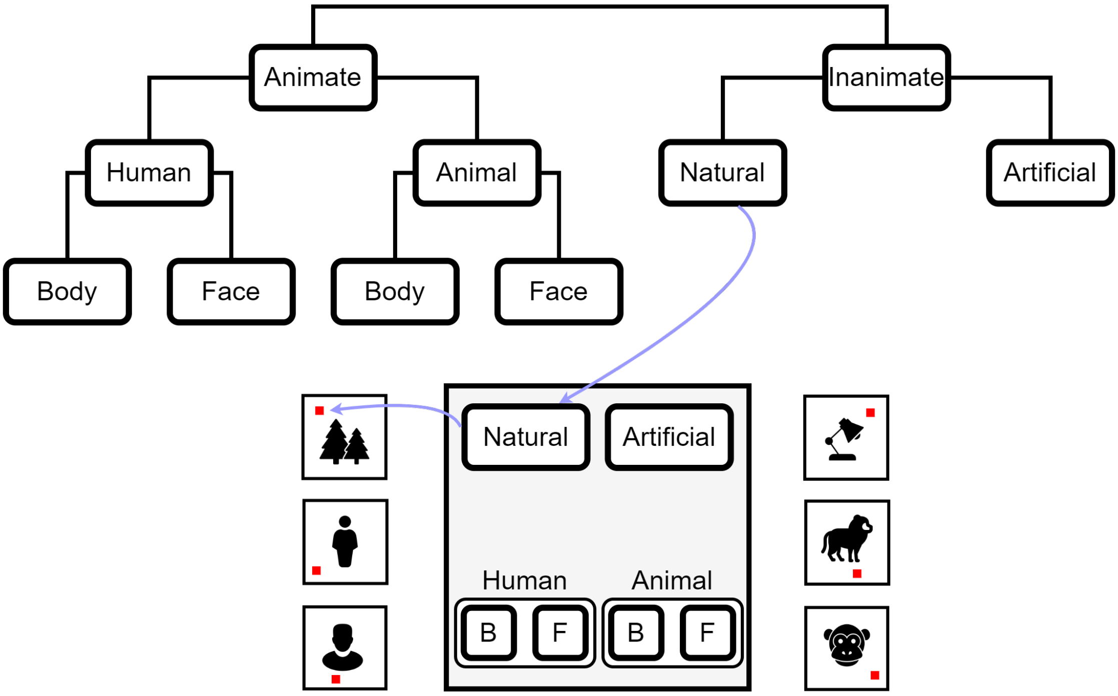


Figure 4.9: **Exploiting intrinsic dataset hierarchy in order to place confounds.** The top panel shows the hierarchical structure of categories in the dataset, which was used to place the single pixel confounds. The example at the bottom (middle) shows one such hierarchical placement scheme where the pixels for Inanimate images were closer to the top of the canvas while Animate images were closer to the bottom. Within the Animate images, the pixels for Humans and Animals were placed at the left and right, respectively, and the pixels for bodies (B) and faces (F) were clustered as shown.

macaque IT. For both humans and macaques, distances in response patterns were larger between the higher-level categories (animate and inanimate) than between the lower-level categories (e.g., between human bodies and human faces).

We used a similar experimental paradigm to the above studies, where we trained networks to classify stimuli which included a single predictive pixel. But instead of using an RDM to compute the location of a diagnostic pixel, we used the hierarchical categorical structure. In the first modified version of the dataset, the location of the pixel was based on the hierarchical structure of categories in Figure 4.9 – predictive pixels for animate kinds were closer to each other than to inanimate kinds, and pixels for faces were closer to each other than to bodies, etc. One such configuration can be seen in Figure 4.9. In the second version, the predictive pixel was placed at a random location for each category (but, of course, at the same location for all images within each category). We call these conditions 'Hierarchical' and 'Random'. Khaligh-Razavi and Kriegeskorte [82] showed that the RDM of average response patterns elicited in the human IT cortex (Φ_1) correlated with the RDM of a DNN trained on naturalistic images (Φ_2). We explored how this compared to the correlation with the RDM of a network trained on the Hierarchical pixel placement (Φ_3) and Random pixel placement (Φ_4).

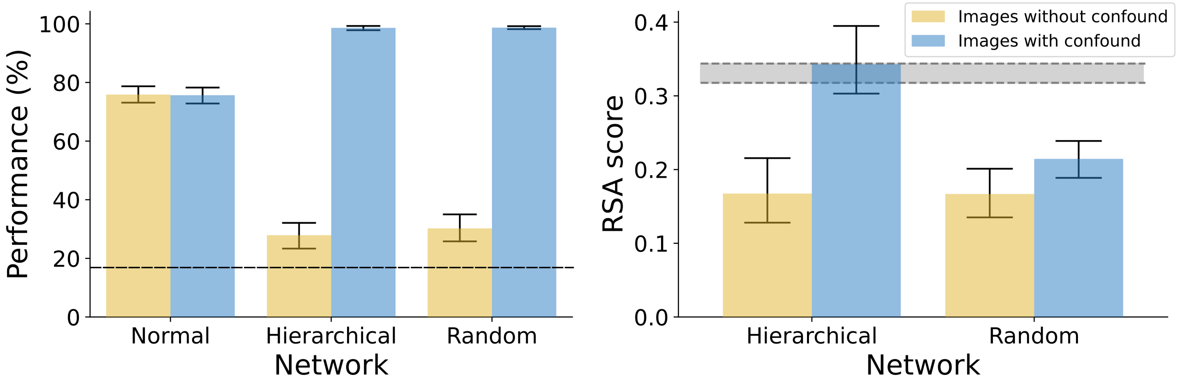


Figure 4.10: **Study 4 results.** *Left:* Performance of normally trained networks did not depend on whether the confound was present. Networks trained with the confound failed to classify stimuli without the confound (yellow bars) while achieving near perfect classification of stimuli with the confound present (blue bars, dashed line represents chance performance). *Right:* RSA with human IT activations reveals that, when the confound was present, the RSA-score for networks in the Hierarchical condition matched the RSA-score of normally trained network (gray band), while the RSA-score of the network in the Random condition was significantly lower. The grey band represents 95% CI for the RSA score between normally trained networks and human IT.

Results for this study are shown in Figure 4.10. We observed that representational geometry of a network trained on Hierarchically placed pixels (Φ_3) was just as correlated to the representational geometry of human IT responses (Φ_1) as a network trained on naturalistic images (Φ_2). However, when the pixel locations for each category were randomly chosen, this correlation decreased significantly. These results suggest that any confound in the dataset (including texture, colour or low-level visual information) that has distances governed by the hierarchical clustering structure of the data could underlie the observed similarity in representational geometries between CNNs and human IT. More generally, these results show how it is plausible that many confounds present in popular datasets may underlie the observed similarity in representational geometries between two systems. The error of inferring a similarity in mechanism based on a high RSA score is not just possible but also probable.

4.4 Discussion

In four studies, we have illustrated a number of conditions under which it can be problematic to infer a similarity of representations between two systems based on a correlation in their representational geometries. In particular, we showed that two systems may transform their inputs through very different functions and encode very different features of inputs and yet have highly correlated representational geometries. Of course, one may acknowledge that a second-order isomorphism of activity patterns does *not* strictly imply that two systems are similar mechanistically but still assume that it is highly likely to be the case. That is, as a practical matter, a researcher may assume that RSA is a reliable method to compare systems. However, our findings challenge this assumption. We show how a high RSA score between different systems can not only occur in a bare-bones simulation (Study 1), but also in practice, in high-dimensional systems operating on high-dimensional data (Studies 2–3). Furthermore, we show that the hierarchical structure of datasets frequently used to test similarity of representations lends itself to a high RSA score arising because of second-order confounds present in the dataset (Study 4). Therefore, second-order confounds driving high RSA scores is not only possible but plausible.

One limitation of our method is that we manually insert a confound in input stimuli (in Studies 2–4) and train a network based on this confound. Even though our findings demonstrate that second-order confounds are plausible, they do not allow us to infer whether such confounds *are* present in existing datasets and driving the observed similarity in existing studies. In our view, there are two methods one could use to check whether confounds are driving results of RSA. The best way would be to identify the stimulus features in a dataset that mimic each other in representational space (e.g. shape and reflectance in Figure 4.3). This is not straightforward to do in high-dimensional stimuli, such as naturalistic images, which consist of millions of features. However, another approach is more tractable: conduct controlled experiments to establish whether the two systems are representing information in similar ways. We have argued for this approach in relation to making inferences about mechanistic similarity between DNNs and humans [17]. In fact, research relating DNNs to human vision provides a striking case of a disconnect between RSA and behavioural findings from psychology [17, 33, 163]. The findings here may explain contradictory RSA scores between DNNs and human visual processing as pointed out by Xu and Vaziri-Pashkam [180]. At the very least, a researcher claiming that two systems are mechanistically similar to one another based on high RSA scores should have an explanation for this discrepancy.

A related point has been made by Kriegeskorte and Diedrichson [94] and Kriegeskorte and Wei [98], who point out that two systems may have the same representational geometry, even if they have a different activity profile over neurons. In this sense, the geometry abstracts away the information about how information was distributed over a set of neurons. Kriegeskorte and Diedrichson [94] equate this loss in information to 'peeling a layer of an onion' – downstream decoders that are sensitive to the representational geometry rather than activity profiles over neuron populations can focus on difference in information as reflected by a change in geometry and be agnostic to how this information is distributed over a set of neurons. We agree that this invariance over activity profiles is indeed a useful property of representational geometries for downstream decoders. However, we are not aware of any studies that highlight how representational geometries also abstract over behaviourally relevant stimulus properties. While abstracting over activity profiles may be useful, abstracting over stimulus properties loses an important piece of information when comparing representations across brain regions, individuals, species and between brains and computational models. Our studies show how two systems may appear similar based on their representational geometries in one circumstance (e.g. Figure 4.2B) but drastically different in another circumstance (Figure 4.2C).

It is important to note how our results differ from previous studies exploring limitations of RSA. A number of studies have focused on the importance of how neural data is pre-processed and how the distance between neural patterns is computed. For example, Ramirez [146] found that pre-processing steps, such as centering (de-meaning) activation vectors may lead to incorrect inference about the representational geometry of activations. He demonstrated that subtracting the mean from activations could change the rank order of similarity between conditions. In turn, this could lead to clearly distinct RDMs becoming highly correlated and vice-versa. While this is an important methodological point, it is clearly distinct from the point we are making in this study. Indeed, the results here are agnostic of the data pre-processing steps and hold whether or not activations are centered.

Some previous studies have also explored how confounds present in data can influence the results of RSA. For example, Henriksson et al. [69] and Cai et al. [21] demonstrated that RDMs measured based on fMRI data can be severely biased because of temporal and spatial correlations in neural activity. These authors have pointed out that if activity patterns from different brain regions are recorded during the same trial, the similarity estimates will be exaggerated due to correlated neural fluctuations in these regions. Similarly, neural activity is correlated over time, which means estimated similarity based on activity patterns from the same imaging run also introduces a strong bias in RDMs. These sources of bias are important to understand, but they can also be addressed by a more careful task design and analysis [21]. In contrast, the confounds that are highlighted in this study exist in the stimulus itself. Therefore, even if one were to completely mitigate the bias in estimating RDMs, the types of confounds we highlight in our work would still pose problems when drawing inferences from correlation in RDMs.

Similarly, previous research has also highlighted the importance of choosing the correct distance metric when using RSA. For example, Ramirez [147] compared Euclidean distance with an angular metric (such as cosine similarity) and showed that the choice of distance metric can reveal different aspects of the same fMRI data. They argued that the Euclidean distance is particularly sensitive to the mean activity over a recorded voxel. Based on this analysis, Ramirez [147] suggested using an angular distance metric, especially when neural signal is aggregated over large number of neurons. Similarly, in another exhaustive study over distance measures, Bobadilla-Suarez et al. [16], evaluated neural similarity using various distance measures, including angle-based measures (cosine, Pearson, Spearman) and magnitude-based measures (Euclidean, Mahalanobis, Minkowski) and found that the choice of metric significantly influenced the measured similarity. They also found that there was no one metric that outperformed all others – rather, the preferred metric varied across different studies, but was consistent across brain regions within a study. The choice of distance metric is again a related but orthogonal issue to the one we highlight in this study. Our results show that representational geometry loses information about stimulus features and different stimulus features (and indeed transformations of input stimulus) can lead to the same geometry. This is fundamental to the nature of representational geometries, rather than a consequence of the distance metric used.

Of course, the problem of confounds in stimuli is not unique to RSA and will affect other statistical analyses, including multivariate regression methods such as MVP classification. Indeed, the problem of confounds in stimuli is well appreciated in many different contexts [47, 121, 174], but there has been no consideration of whether these confounds are contributing to RSA findings. Perhaps this is because, unlike for MVP classification, a confound for RSA needs to not only help decode category membership, but also lead to a second-order isomorphism. Nevertheless, as we illustrate in Figure 4.3, there could be confounds with a second-order similarity structure in many datasets that are the product of unexpected properties of the world or the product of how these datasets are curated or hierarchically organized. This is problematic as we have clearly shown that these second order confounds can drive high RSA scores.

A reader could ask why these results matter. Couldn't a researcher take the view that representational geometry *is* representation and therefore, a strong correlation in representational geometries between two systems is sufficient to infer that the systems are representing the world in a similar manner? This question goes to the heart of an existing debate in philosophy, where philosophers distinguish between the *externalist* and *holistic* views on mental representations. According to the first view, the content of representations is determined by their relationship to entities in the external world. This perspective is implicitly taken by most neuroscientists and psychologists, who are interested in comparing mechanisms underlying cognitive processes – that is, they are interested in the set of nested functions and algorithms responsible for transforming sensory input into a set of activations in the brain. From this perspective, our finding that high RSA scores can be obtained between systems that work in qualitatively different ways poses a challenge to researchers using RSA to compare systems.

Alternatively, a researcher may reject an externalist view and adopt the perspective that representations obtain their meaning based on how they are related to each other within each system, rather than based on their relationship to entities in the external world. That is, 'representation *is* the representation of similarities' [34]. From this perspective, as long as the two systems share the same relational distances between internal activations, one can validly infer that the two systems have similar representations. That is, a second-order isomorphism implies a similarity of representations, by definition. This view has been called *holism* in the philosophy of mind [14, 42] and is related to a similar idea of *meaning holism* in language, which is the idea that the meaning of a linguistic expression is determined by its relation to other expressions within a language [68, 143]. For example, Firth [39] (p. 11) writes: "you shall know a word by the company it keeps". Similarly, Griffiths and Steyvers [58], and Griffiths, Steyvers, and Tenenbaum [59] have adopted meaning holism accounts of semantic representations in neural networks. More recently, Piantadosi and Hill [140] have argued that large language models capture important aspects of meaning and approximate human cognition because they represent relations between concepts and their roles within a representational geometry. Even if a researcher was to adopt this holistic perspective on representations, our results should still be of interest to them as they show that the similarity between representational geometries can vary based on the visual stimulus that is used to compare them (the modulation effect). Additionally, our results show that adopting this view misses the information about differences in mechanistic processes that a psychologist or neuroscientist is frequently interested in, for instance, whether the visual system processes surface reflectance or shape (or the location of diagnostic pixels) in order to identify objects. Fodor and Lepore long ago criticized this philosophical stance [41, 42], and interestingly, this philosophical debate played an important part in the development of RSA (see S1 Appendix). Unfortunately, this debate has largely been ignored by researchers who use RSA as a method to compare similarity of systems.

In closing, we describe our recommendations for practitioners who find RSA to be useful for their research goals. These will be especially relevant to the large majority of researchers in computational, cognitive, and systems neuroscience, cognitive scientists and AI practitioners, who are interested in mechanistic similarities (i.e., they adopt an externalist position). But they should also be relevant to adopters of the holistic view who are interested in how observed representational geometries depend on the stimulus used to extract them.

First, since the intrinsic structure of datasets can artificially modulate RSA scores, researchers should compare systems on a wider variety of datasets and sampling schemes than currently done. Second, given that confounding features can lead to mimicked representational geometries, researchers should consider running additional controlled experiments to rule out this possibility when inferences hinge crucially on it. This point has recently been made by Bowers et al [17] in relation to testing DNNs. Similarly, the 'controversial stimuli' designed by Kriegeskorte and Golan [53] should also enable researchers to test representational geometries for stimuli where different models make contrasting predictions. Third, when studies are conducted to search for evidence of mechanistic similarity between two or more systems, researchers should use a wider range of complementary methods, each addressing the others' blindspots (e.g., RSA combined with neural predictivity [182], MVPC [63, 64], CCA [123], SVCCA [144], CKA [91]).

Lastly, perhaps the most important general recommendation we make is that researchers should acknowledge, procedurally and in writing, which inferences are afforded by the use of RSA, and what dissimilarities remain possible despite having observed a given pattern of RSA scores. To this end, we believe that general statements of similarity tend to obfuscate rather than accurately summarize any set of RSA-based results. Instead, we urge researchers using RSA (1) to justify the use of this method by theoretically motivated interest in representational geometry or otherwise consider other tools that best fit their goals, and (2) to state in precise terms that RSA scores reflect the similarity of representational geometries in particular, and generally avoid underspecified claims of similarity.

4.5 Methods

Dataset generation and training

All DNN simulations (Studies 2–4) were carried out using the `Pytorch` framework [136]. The model implementations were downloaded from the `torchvision` library. Networks trained on unperturbed datasets in all studies were pre-trained on `ImageNet` as were networks trained on modified datasets in Study 2. Networks trained on modified datasets in Studies 3 and 4 were randomly initialised. For the pre-trained models, their pre-trained weights were downloaded from `torchvision.models` subpackage.

Study 1 Each dataset in Study 1 consists of 100 samples (50 in each cluster) drawn from two multivariate Gaussians, $\mathcal{N}(x|\mu,\Sigma)$, where μ is a 2-dimensional vector and Σ is a 2×2 covariance matrix. In Figure 4.2A, the two Gaussians have means $\mu_1 = (1,8)$ and $\mu_2 = (8,1)$ and a covariance matrices $\Sigma_1 = \Sigma_2 = \frac{1}{2}I$, while in Figure 4.2B the Gaussians have means $\mu_1 = (1,1)$ and $\mu_2 = (8,8)$ and a covariance matrices $\Sigma_1 = I$, $\Sigma_2 = 8I$. All kernel matrices were computed using the `sklearn.metrics.pairwise` module of the `scikit-learn` Python package.

Study 2 First, a VGG-16 deep convolutional neural network [168], pre-trained on the `ImageNet` dataset of naturalistic images, was trained to classify stimuli from the `CIFAR-10` dataset [99]. The `CIFAR-10` dataset includes 10 categories with 5000 training, and 1000 test images per category. The network was fine-tuned on `CIFAR-10` by replacing the classifier so that the final fully-connected layer reflected the correct number of target classes in `CIFAR-10` (10 for `CIFAR-10` as opposed to 1000 for `ImageNet`). Images were rescaled to a size of 224×224px and then the model learnt to minimise the cross-entropy error using the RMSprop optimizer with a mini-batch size of 64, learning rate of 10^{-5}, and momentum of 0.9. All models were trained for 10 epochs, which were sufficient for convergence across all datasets.

Second, 100 random images from the test set for each category were sampled as input for the network and activations at the final convolutional layer extracted using the `THINGSVision` Python toolkit [125]. The same toolkit was used to generate a representational dissimilarity matrix (RDM) from the pattern of activations using `1-Pearson's r` as the distance metric. The RDM was then averaged by calculating the median distance between each instance of one category with each instance of the others (e.g., the median distance between `Airplane` and `Ship` was the median of all pair-wise distances between activity patterns for airplane and ship stimuli). This resulted in a 10×10, category-level, RDM which reflected median between-category distances.

Third, three modified versions of the `CIFAR-10` datasets were created for the 'Positive', 'Uncorrelated' and 'Negative' conditions, respectively. In each dataset, we added one diagnostic pixel to each image, where the location of the pixel depended on the category (See Figure 4.5). The locations of these pixels were determined using the averaged RDM from the previous step. We call this the target RDM. In the 'Positive' condition, we wanted the distances between pixel placements to be positively correlated to the distances between categories in the target RDM. We achieved this by using an iterative algorithm that sampled pixel placements at random, calculated an RDM based on distances between the pixel placements and computed an RSA score (Spearman correlation) with the target RDM. Placements with a score above 0.70 were retained and further optimized (using small perturbations) to achieve an RSA-score over 0.90. The same procedure was also used to determine placements in the Uncorrelated (optimizing for a score close to 0) and Negatively correlated (optimizing for a negative score) conditions.

Finally, datasets were created using 10 different placements in each of the three conditions. Networks were trained for classification on these modified `CIFAR-10` datasets in the same manner as the VGG-16 network trained on the unperturbed version of the dataset (See Figure 4.4).

Study 3 The procedure mirrored Study 2 with the main difference being that the target system was the macaque inferior temporal cortex. Neural data from two macaques, as well as the dataset were obtained from the Brain Score repository [159]. This dataset consists of 3200 images from 8 categories (animals, boats, cars, chairs, faces, fruits, planes, and tables), we computed an 8×8 averaged RDM based on macaque IT response patterns for stimuli in each category.

This averaged RDM was then used as the target RDM in the optimization procedure to determine locations of the confound (here, a white predictive patch of size 5×5 pixels) for each category. Using a patch instead of a single pixel was required in this dataset because of the structure and smaller size of the dataset (3200 images, rather than 50,000 images for `CIFAR-10`). In this smaller dataset, the networks struggle to learn based on a single pixel. However, increasing the size of the patch makes these patches more predictive and the networks are able to again learn entirely based on this confound (see results in Figure 4.6). In a manner similar to Study 2, this optimisation procedure was used to construct three datasets, where the confound's placement was positively correlated, uncorrelated or negatively correlated with the category distances in the target RDM.

Finally, each dataset was split into 75% training (2432 images) and 25% test sets (768 images) before VGG-16 networks were trained on the unperturbed and modified datasets in the same manner as in Study 2. One difference between Studies 2 and 3 was that here the networks in the Positive, Uncorrelated and Negative conditions were trained from scratch, i.e., not pre-trained on `ImageNet`. This was done because we wanted to make sure that the network in the Normal condition (trained on `ImageNet`) and the networks in the Positive, Uncorrelated and Negative conditions encoded fundamentally different features of their inputs – i.e., there were no `ImageNet`-related features encoded by representations Φ_2, Φ_3 and Φ_4 that were responsible for the similarity in representational geometries between these representations and the representations in macaque IT cortex.

Study 4 The target system in this study was human IT cortex. The human RDM and dataset were obtained from [97]. Rather than calculating pixel placements based on the human RDM, the hierarchical structure of the dataset was used to place the pixels manually. The dataset consists of 910 images from 6 categories: human bodies, human faces, animal bodies, animal faces, artificial inanimate objects and natural inanimate objects. These low-level categories can be organised into the hierarchical structure shown in Figure 4.9. Predictive pixels were manually placed so that the distance between pixels for Animate kinds were closer together than they were to Inanimate kinds and that faces were closer together than bodies. This can be done in many different ways, so we created five different datasets, with five possible arrangements of predictive pixels. Results in the Hieararchical condition (Figure 4.10) are averaged over these five datasets. Placements for the Random condition were done similarly, except that the locations were selected randomly.

Networks were then trained on a 6-way classification task (818 training images and 92 test images) in a similar manner to the previous studies. As in Study 3, networks trained on the modified datasets (both Hierarchical and Random conditions) were not pre-trained on `ImageNet`.

RDM and RSA computation

For Studies 2-4 all image-level RDMs were calculated using $1-r$ as the distance measure. RSA scores were computed as the Spearman rank correlation between RDMs.

In Study 2, a curated set of test images was selected due to the extreme heterogeneity of the `CIFAR-10` dataset (low activation pattern similarity between instances of the same category). This was done by selecting 5 images per category which maximally correlated with the averaged activation pattern for the category. Since `CIFAR-10` consists of 10 categories, the RSA-scores in Study 2 were computed using RDMs of size 50×50.

In Study 3, the dataset consisted of 3200 images belonging to 8 categories. We first calculated a full 3200×3200 RDM using the entire set of stimuli. An averaged, category-level, 8×8 RDM was then calculated using median distances between categories (in a manner similar to that described for Study 2 in the Section 'Dataset generation and training'). This 8×8 RDM was used to determine the RSA-scores. We also obtained qualitatively similar results using the full 3200×3200 RDMs. These results can be found in the S2 Appendix.

In Study 4, the dataset consisted of 818 training images and 92 test images. Kriegeskorte et al. [97] used these images to obtain a 92×92 RDM to compare representations between human and macaque IT cortex. Here we computed a similar 92×92 RDM for networks trained in the Normal, Hierarchical and Random training conditions, which were then compared with the 92×92 RDM from human IT cortex to obtain RSA-scores for each condition.

Testing

In Study 2, we used a 4×2 design to measure classification performance for networks in all four conditions (Normal, Postive, Uncorrelated and Negative) on both unperturbed images and modified images. We computed six RSA-scores: three pairs of networks – Normal-Positive, Normal-

Uncorrelated and Normal-Negative – and two types of inputs – unperturbed and modified test images. The noise ceiling (grey band in Figure 4.6) was determined in the standard way as described in [128] and represents the expected range of the highest possible RSA score with the target system (network trained on the unperturbed dataset).

In Study 3, performance was estimated in the same manner as in Study 2 (using a 4×2 design), but RSA-scores were computed between RDMs from macaque IT activations and the four types of networks – i.e. for the pairs Macaque-Normal, Macaque-Positive, Macaque-Uncorrelated and Macaque-Negative. And like in Study 2, we determined each of these RSA-scores for both unperturbed and modified test images as inputs to the networks.

In Study 4, performance and RSA were computed in the same manner as in Studyn 3, except that the target RDM for RSA computation came from activations in human IT cortex and the networks were trained in one of three conditions: Normal, Hierarchical and Random.

Data analysis

Performance and RSA scores were compared by running analyses of variance and Tukey HSD post-hoc tests. In Study 2 and 3, performance differences were tested by running a 4 (type of training) by 2 (type of dataset) mixed ANOVAs. In, Study 4, the differences were tested by running a 3x2 mixed ANOVA.

RSA scores with the target system between networks in various conditions were compared by running 3x2 ANOVAs in Studies 2 and 3, and a 2x2 ANOVA in Study 4. We observed that RSA-scores were highly dependent on both the way the networks were trained and also the test images used to elicit response activations.

In this section we provide more detailed statistical analyses for Studies 2-4.

Study 2

In order to test for differences in performance (Figure 4.6, left panel), a 4 (normally trained/positive/uncorrelated/negat by 2 (dataset with/without confound) mixed analysis of variance (ANOVA) was conducted. The finding was a significant interaction effect ($F(3,36) = 12256.10, p < .001, \eta_p^2 = .99$). Tukey HSD post-hoc comparisons revealed that performance in the positive, uncorrelated and negative conditions was significantly better on datasets which included the confounds (all $p < .001$) while the normally trained networks performed equally well on both datasets with and without the confound ($p = .99$). This shows that networks trained on datasets with confounds learned to classify based on the predictive confounding feature (single pixel) and ignored other features in the dataset (failing to classify if the confound is not present) while the normally trained networks remain unaffected by the presence or absence of the confounding feature.

Differences in RSA scores (Figure 4.6, right panel) were tested by conducting a 3 (positive/uncorrelated/negative) by 2 (dataset with/without the confound) mixed ANOVA. The key findings was a significant interaction effect ($F(2,297) = 289.27, p < .001, \eta_p^2 = .66$). Post-hoc comparisons revealed that there were no differences between the networks in RSA scores with normally trained networks when images without the confound were used as input (all $p > .954$). On the other hand, for images which contained the confound, networks in the positive condition achieved a significantly higher RSA score than both networks in the uncorrelated and negative conditions ($p < .001$), at the same time, networks in the uncorrelated condition achieved significantly higher RSA scores than networks in the negative condition ($p < .001$). This indicates a very strong modulation effect of RSA scores - depending on the relation between the representational geometry of the confounding feature exploited by these networks, RSA scores with normally trained networks can vary from high to low when the confound is present, but are consistently low when there is no confound in the test stimuli.

Study 3

The same analytical approach was taken as in Study 2, performance (Figure 4.7, left panel) was analyzed by conducting a 4 (normal/positive/uncorrelated/negative) by 2 (dataset with/without confound) mixed ANOVA. Again, the key finding was an interaction effect ($F(3,51) = 8086.60, p < .001, \eta_p^2 = .99$). Post-hoc comparisons revealed that performance in the positive, uncorrelated and negative conditions was significantly better on datasets which included the confounds (all $p < .001$) while the normally trained networks performed equally well on both datasets with and without the confound ($p > .99$).

RSA scores (Figure 4.7, right panel) were analyzed by conducting a 3 (positive/uncorrelated/negative) by 2 (dataset with/without confound) mixed ANOVA. The key result being a significant interaction effect ($F(2,42) = 122.46, p < .001, \eta_p^2 = .85$). Post-hoc comparisons revealed that there were no differences between the networks in RSA scores with normally trained networks when images without the confound were used as input (all $p > .071$). However, for images with the confound present, networks in the positive condition achieve a significantly higher RSA score with macaque IT when compared to networks in the uncorrelated and negative conditions (all $p < .001$). Networks in the uncorrelated condition achieve higher RSA scores than networks in the negative condition ($p = .005$). Finally, it is worth emphasizing that networks in the positive condition match RSA scores with macaque IT achieved by networks pretrained on naturalistic images and then finetuned on the dataset without confounds ($t(23) = 0.89$, $p = .384$) when the confound is present in the dataset.

Study 4

For this simulation, performance differences between conditions (Figure 4.10, left panel) were tested by conducting a 3 (normal/hierarchical/random) by 2 (dataset with/without confound) mixed ANOVA. As in previous studies, the eky result was a significant interaction effect ($F(2,42) = 407.61, p < .001, \eta_p^2 = .95$). Post-hoc comparisons revealed that performance in the hierarchical and random conditions was significantly better on datasets which included the confounds (all $p < .001$) while the normally trained networks performed equally well on both datasets with and without the confound ($p > .99$).

RSA scores with human IT (Figure 4.10, right panel) were analyzed by conducting a 2 (hierarchical/random) by 2 (dataset with/without) mixed ANOVA. The interaction effect was significant ($F(1,28) = 8.46, p = .007, \eta_p^2 = .23$). Follow-up comparisons show that there was no difference between networks in the hierarchical and radnom conditions when the dataset did not contain the confound ($p > 99$), but networks in the hierarchical condition achieved significantly higher RSA scores when the dataset did contain the confound ($p < .001$). Again, it is worth emphasizing that networks in the hierarchical condition match RSA scores with human IT achieved by networks pretrained on naturalistic images and then finetuned on the dataset without confounds ($t(28) = 0.46$, $p = .647$) when the confound is present in the dataset.

Data Availability

Confound placement coordinates (Studies 2-4), unperturbed datasets (Studies 3 and 4), macaque activation patterns and RDMs (Study 3) and human RDM (Study 4) are available at OSF.

Chapter 5

Summary, contributions and conclusions

Throughout these projects, the aim was to take a multifaceted approach to investigating visual representations in humans and deep learning models with the hope to better understand both. Beyond the main goals set for each of the studies, the findings also resulted in lessons learned about how this research can, and should, be approached in the future.

5.1 Summary

My main aims were to provide a rigorous examination of the similarities and dissimilarities of visual representations in humans and deep learning models. Throughout this process I maintained that taking a multifaceted approach is key. The idea was to provide converging evidence via multiple different methodological approaches and rigorous experimentation within each approach. Reliance on one measure or a set of measures using one or a constrained set of stimuli is likely to lead to inaccurate inferences. Studies can, of course, be focused on one specific phenomenon but a multifaceted approach would benefit this type of research as well.

In Chapter 2 I show how a different combinations of stimulus and data analysis choices can lead to completely different inferences. The original Zhou and Firestone [187] paper dealt with a curated set of adversarial stimuli and examined human-CNN agreement in an unorthodox manner, concluding that humans reliably decipher adversarial images. This line of reasoning ignores the fact that there are nearly endless adversarial examples for which agreement would be exactly at chance. It also ignores that under certain setups agreement will be below chance (see Figure 2.4) when humans agree amongst themselves on a label different to the one assigned to the image by a CNN. By and large, findings concerning adversarial images show that for a small subset of them, humans can have some intuitive insight into why and how they were misclassified. A subset of these intuitions is trivial; they actually depict an exemplar from the target class (e.g., chain-link fence) and should not be considered adversarial at all. For a subset of images humans do indeed pick up on features which allow them to correctly identify the label a deep learning model assigned to the image at an above chance level (e.g., black and yellow color scheme for the school bus image in Figure A.2). Note that even in that case, it is not clear whether humans and networks transform these inputs into visual representations in a mechanistically similar manner.

In Chapter 3 I demonstrate how resilient human shape bias is to a novel environment, even when environment statistics were manipulated so that shape was less predictive of category membership than local features. Comparisons with an ideal inference model and CNNs reveal that CNNs also have an inductive bias. While the ideal inference model learns to use both global shape and the secondary feature, CNNs prefer local features. Perhaps most importantly, we show that human-like shape bias is not simply a matter of better training. When we froze all of the convolutional layers of a network trained to develop shape bias (style-transfer trained networks from [50]) we found that bias is not robust in a new environment. Classifiers on top of the frozen convolutional layers readily learn to classify based on local features, particularly when predictivity of such features is higher than that of global shape.

In Chapter 4 I demonstrate how good RSA scores do not necessarily imply similarity of representations. Systems processing stimuli in qualitatively different ways can achieve high similarity as measured via RSA scores. This is important from the perspective of neuroscientists, computer scientists and cognitive scientists given that the method has become increasingly popular in all three of these fields and at their intersections. It provides high flexibility, as dissimilarity matrices can be constructed from any type of data and provide a summary of how stimulus representations relate to each other. However, it does not allow for reliable inferences about which features and by which processes were transformed into those representations. High RSA scores and good neural predictivity achieved by deep learning models has become the leading argument for these models being *the best* models of human vision. These claims remain strong in spite evidence to the contrary (for a detailed review see Bowers et al. [17]. Findings in Chapter 4 provide a plausible explanation for the reasons of some of these impressive results despite the many dissimilarities between human and network behavior. As more researchers start using these methods there will be better understanding of what exactly they imply. For example, Xu and Vaziri-Pashkam [180] found RSA scores do not follow the same patterns as previously reported and do not generalize across different stimulus sets - which could be explained by our findings of dataset dependent scores in Chapter 4. The overall conclusion is that, as most measures, RSA is an indirect measure of representations and processes (even more so, being a second-order metric) and does not allow for inferences without being embedded in rigorous experimentation and/or combined with other measures to provide converging evidence.

5.2 Contributions

Apart from the specific contributions and furthering of understanding of human and CNN visual representations, one would hope work written up in this thesis highlights the importance of experimental rigour, adequate choice of methods, the importance of converging evidence and nuance in research. While the vast majority of work surrounding deep learning has been done in the engineering domain, there is a part of deep learning which intersects with neuroscience and cognitive science with an ever increasing number of researchers. The potential for interdisciplinary approaches, cross-pollinating of ideas and research paradigms is exciting, however the deep learning revolution comes at a cost as well. The pressure to publish at high volume and highlight successes is increasing daily - competition for careers and funding, are all threats to nuanced, rigorous science. Even cautious researchers are encouraged to make strong, flashy claims for the sake of promotion. This oftentimes results in nuanced research being interpreted thorough the lens of that attention-grabbing claim. By the hundredth citation, a lot of the nuance and understanding may be gone from how the work is perceived and interpreted in the community. As researchers we need to be cognizant of how the work is going to be consumed.

Furthermore, the work presented here has resulted in very useful tools and approaches. The 'patchwork' datasets generated for studies in Chapter 3 is a prime example of such a contribution. The dataset allows for independent manipulation of different features and their level of predictivity. This, in turn, allows for testing specific hypotheses about which the models and which features are being prioritized.

The approach taken in Chapter 4 to study RSA demonstrates how deep learning models can be useful test-beds for evaluating properties of statistical methods conducted on complex, multi-dimensional data. In this instance, CNNs were easily trained to develop qualitatively different visual representations in order to study factors influencing RSA scores. This type of manipulation is not possible with human participants. They are complex systems, transforming high-dimensional inputs into representations useful for solving a specific task. Even if they are not modeling the ventral visual stream (or any other region of interest) particularly well, they can prove to be useful because they provide similarly complex data as human studies would. At the same time they are easy to experimentally manipulate and probe. In my mind, this is certainly a major contribution of the deep learning revolution, beyond attempts to use them as models of human cognition. Neural networks are used in MVPA as decoders of neural activity and the manner in which I used them in Chapter 4 could lead to more such research. There are many methodological decisions to be made when using RSA and similar methods - deep learning models can provide an excellent tool for evaluating the impact of those decisions. For example, one could investigate how curating datasets, decisions on distance metrics, deciding on the level of analysis and other decisions impact whether models developing qualitatively different representations can be distinguished based on RSA. Therefore, the hope is that researchers not interested in deep learning models as models of human cognition may find them to be valuable auxiliary tools in their work as well.

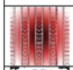
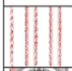
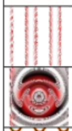
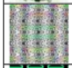
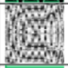
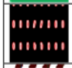
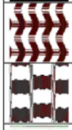
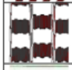
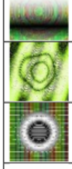
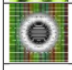
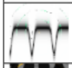
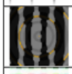
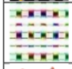
5.3 Conclusions

Through a number of simulations and behavioural studies from different perspectives we have found the following.

- Humans do not poses a robust ‘machine theory of mind’ as claimed by Zhou and Firestone [187] when controlling for important methodological factors
- The numerous types of adversarial stimuli and the lack of agreement between human and CNN classification on a large range of these stimuli provides evidence for dissimilarity between human and machine object recognition.
- Robust human-like shape bias cannot be induced purely through training on style-transfer stimuli which are designed to eliminate category predictivity of textures.
- Human shape bias is robust to changing statistics of the learning environment - humans rely on shape when compared to local features even when it is less predictive of category membership.
- Humans have inherent resource (attentions and working memory) limitations which contribute to the development of shape bias.
- Convolutional neural networks pre-trained on large-scale naturalistic stimuli datasets show a bias towards learning local features.
- Representational similarity analysis as a second order measure can produce high similarity scores between systems which represent stimuli qualitatively differently in the presence of confounds.

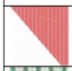
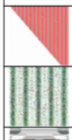
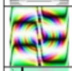
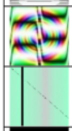
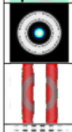
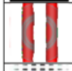
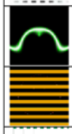
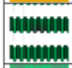
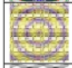
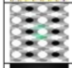
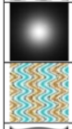
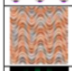
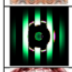
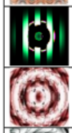
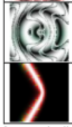
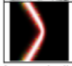
APPENDIX A

DECODING ADVERSARIAL IMAGES - APPENDIX

Stimulus	DCNN label	1	2	3	4	5	6	7	8		
	1. Accordion**	Traffic light 10.92%	Strawberry 8.05%	Tile roof 6.32%	**Accordion 5.17%**	Spotlight 5.17%	Electric Guitar 4.60%	Pole 4.60%	Chainlink fence 4.02%		
	2. Assault rifle*	Chainlink fence 21.26%	Crossword puzzle 8.04%	Computer keyboard 6.32%	Accordion 5.17%	Soccer ball 4.60%	Ski mask 3.45%	Tile roof 3.45%	**Assault rifle 2.87%**		
	3. Bagel**	**Bagel 25.28%**	Spotlight 12.07%	Traffic light 5.74%	Volcano 5.17%	Pinwheel 4.02%	Car wheel 2.87%	Chameleon 2.87%	Digital clock 2.87%		
	4. Baseball*	Pole 12.64%	Four-poster bed 5.74%	Sea snake 5.74%	Chainlink fence 4.59%	Hair Clip 4.02%	Panpipe 4.02%	Roundworm 4.02%	**Baseball 3.45%**		
	5. Car Wheel**	Traffic light 10.92%	Volcano 6.90%	**Car wheel 6.32%**	Obelisk 5.75%	Stethoscope 5.75%	Electric guitar 4.02%	Projector 4.02%	Digital clock 3.45%		
	6. Chainlink fence**	**Chainlink fence 61.49%**	Accordion 2.87%	Hair clip 2.30%	Sea snake 2.30%	Ski mask 2.30%	Tile roof 2.30%	Computer keyboard 1.72%	Digital clock 1.72%		
	7. Chameleon**	Pinwheel 6.90%	**Chameleon 5.75%**	Starfish 5.75%	Green snake 5.17%	Stethoscope 5.17%	Peacock 4.60%	Projector 4.60%	Traffic light 4.60%		
	8. Comic book*	Peacock 8.05%	Projector 6.90%	Chainlink fence 5.75%	Chameleon 5.75%	Green snake 4.60%	Remote control 4.60%	Accordion 4.02%	Stethoscope 4.02%	...	**Comic book T11 2.87%**
	9. Computer keyboard**	Chainlink fence 14.37%	**Computer keyboard 12.07%**	Green snake 8.05%	Tile roof 4.60%	Paddle 4.02%	Chameleon 3.45%	Digital clock 3.45%	Slot machine 3.45%		
	10. Crossword puzzle**	Chainlink fence 8.62%	**Crossword puzzle 7.45%**	Tile roof 6.90%	Computer keyboard 6.32%	Accordion 4.60%	Electric guitar 4.60%	Photocopier 4.60%	Digital clock 4.02%		
	11. Dial telephone*	Traffic light 9.77%	Soccer ball 8.05%	Bagel 6.32%	Car wheel 6.32%	Chameleon 6.32%	Projector 5.75%	Baseball 5.17%	Roundworm 4.60%	...	**Dial telephone T10 4.02%**
	12. Digital clock**	Strawberry 14.37%	Roundworm 12.07%	**Digital clock 8.62%**	Accordion 4.60%	Remote control 4.60%	Slot machine 4.60%	Spotlight 4.02%	Computer keyboard 3.45%		
	13. Electric guitar*	Sea snake 12.07%	Chainlink fence 6.32%	Hair clip 5.75%	Roundworm 5.75%	Tile roof 5.75%	Accordion 4.02%	Hand blower 4.02%	**Electric guitar 3.45%**		
	14. Four-poster bed	Accordion 8.62%	Tile roof 7.47%	Freight car 5.17%	Chainlink fence 4.60%	Soccer ball 4.60%	Electric guitar 4.02%	Paddle 4.02%	Comic book 3.45%	...	**Four poster bed T35 0.57%**
	15. Freight car	Projector 8.05%	Peacock 7.47%	Digital clock 6.90%	Electric guitar 6.90%	Stethoscope 6.90%	Photocopier 5.75%	Slot machine 5.17%	Chameleon 4.60%	...	**Freight car T39 0%**
	16. Green snake**	**Green snake 8.05%**	Roundworm 8.05%	Spotlight 8.05%	Chameleon 6.90%	Bagel 4.02%	Digital clock 4.02%	Soccer ball 4.02%	Traffic light 4.02%		
	17. Grey parrot	Bagel 14.37%	Stethoscope 8.05%	Car wheel 5.75%	Soccer ball 5.17%	Spotlight 5.17%	Vacuum 5.17%	Baseball 4.60%	Projector 4.60%	...	**Grey parrot T15 1.72%**
	18. Hair clip*	Monarch butterfly 32.76%	Peacock 6.32%	Chameleon 5.75%	Ski mask 5.75%	Obelisk 4.60%	**Hair clip 4.02%**	Green snake 3.45%	Paddle 2.87%		
	19. Hand blower	Computer keyboard 9.20%	Digital clock 6.90%	Stethoscope 5.17%	Volcano 4.60%	Accordion 4.02%	Hair clip 4.02%	Chameleon 3.45%	Dial telephone 3.45%	...	**Hand blower T25 1.15%**
	20. King penguin*	Four-poster bed 7.47%	Car wheel 5.75%	Chainlink fence 4.60%	Stethoscope 4.60%	Projector 4.02%	Grey parrot 3.45%	Punching bag 3.45%	Slot machine 3.45%	...	**King penguin T12 2.30%**
	21. Medicine chest	Slot machine 8.62%	Computer keyboard 8.05%	Chainlink fence 6.32%	Photocopier 6.32%	Electric guitar 5.17%	Panpipe 5.17%	Chameleon 4.60%	Tile roof 4.60%	...	**Medicine chest T20 1.15%**
	22. Monarch butterfly**	**Monarch butterfly 43.68%**	Volcano 8.05%	Punching bag 4.02%	Hair clip 3.45%	Paddle 3.45%	Starfish 3.45%	Stethoscope 3.45%	Traffic light 3.45%		
	23. Obelisk**	**Obelisk 9.77%**	Pole 6.90%	Four-poster bed 4.60%	Projector 4.60%	Tile roof 4.60%	Chainlink fence 4.02%	Crossword puzzle 4.02%	Photocopier 4.02%		
	24. Paddle	Green snake 25.86%	Sea snake 6.32%	Accordion 4.02%	Obelisk 4.02%	Chainlink fence 3.45%	Chameleon 3.45%	Electric guitar 3.45%	Hand blower 3.45%	...	**Paddle T41 0%**

* numerically above chance agreement; ** statistically above chance agreement

Figure A.1: **Participant responses ranked by frequency (Experiment 3b).** Each row contains the adversarial image, the DCNN label for that image, the top 8 participant responses. Shaded cells contain the DCNN choice, when not ranked in the top 8, it is shown at the end of the row along with the rank in brackets.

Timulus	DCNN label	1	2	3	4	5	6	7	8	
	25. Panpipe**	Tile roof 12.07%	**Panpipe 9.77%**	Accordion 4.60%	Obelisk 4.02%	Hair clip 3.45%	Hand blower 3.45%	Pinwheel 3.45%	Pole 3.45%	
	26. Peacock**	Green snake 13.22%	**Peacock 13.22%**	Sea snake 7.47%	Accordion 5.75%	Chainlink fence 5.75%	Monarch butterfly 4.02%	Chameleon 3.45%	Crossword puzzle 3.45%	
	27. Photocopier*	Car wheel 16.67%	Spotlight 6.32%	Pinwheel 4.60%	Accordion 4.02%	Projector 4.02%	Traffic light 4.02%	Assault rifle 3.45%	Digital clock 3.45%	… **Photocopier T11 2.87%**
	28. Pinwheel**	**Pinwheel 8.62%**	Computer keyboard 6.90%	Projector 6.32%	Chameleon 5.75%	Peacock 5.17%	Slot machine 4.60%	Spotlight 4.60%	Monarch butterfly 4.02%	
	29. Pole**	**Pole 33.33%**	Traffic light 5.17%	Comic book 4.02%	Dial telephone 4.02%	Obelisk 4.02%	Photocopier 4.02%	Chainlink fence 3.45%	Accordion 2.87%	
	30. Projector**	Spotlight 19.54%	Car wheel 9.20%	**Projector 9.20%**	Stethoscope 5.75%	Soccer ball 5.17%	Traffic light 4.02%	Assault rifle 3.45%	Digital clock 3.45%	
	31. Punching bag**	Pole 20.69%	**Punching bag 6.32%**	Spotlight 5.75%	Remote control 4.02%	Roundworm 4.02%	Traffic light 4.02%	Four-poster bed 3.45%	Accordion 2.87%	
	32. Remote control*	Computer keyboard 10.92%	Chainlink fence 8.05%	Tile roof 5.75%	Stethoscope 5.17%	Digital clock 4.60%	Projector 4.60%	Obelisk 4.02%	**Remote control 4.02%**	
	33. Roundworm**	Green snake 27.01%	**Roundworm 6.32%**	Sea snake 5.75%	Hair clip 5.75%	Electric guitar 3.45%	Digital clock 3.45%	Ski mask 3.45%	Slot machine 3.45%	
	34. School bus**	**School bus 20.69%**	Accordion 5.17%	Photocopier 5.17%	Chainlink fence 4.02%	Monarch butterfly 4.02%	Medicine chest 3.45%	Tile roof 3.45%	Chameleon 2.87%	
	35. Screwdriver	Accordion 12.64%	Panpipe 9.77%	Green snake 8.62%	Chainlink fence 5.75%	Computer keyboard 5.75%	Digital clock 4.02%	Slot machine 4.02%	Stethoscope 3.45%	… **Screwdriver T17 1.72%**
	36. Sea snake**	Traffic light 14.37%	Green snake 9.20%	Pole 6.90%	Spotlight 5.75%	Chameleon 5.17%	**Sea snake 5.17%**	Digital clock 4.60%	Photocopier 4.60%	
	37. Ski mask**	**Ski mask 27.59%**	King penguin 8.05%	Monarch butterfly 8.05%	Peacock 4.60%	Electric guitar 3.45%	Pinwheel 3.45%	Chameleon 2.87%	Comic book 2.87%	
	38. Slot machine*	Spotlight 10.35%	Pinwheel 9.20%	Crossword puzzle 5.17%	Projector 5.17%	Bagel 4.60%	**Slot machine 4.60%**	Traffic light 4.60%	Computer keyboard 4.02%	
	39. Soccer ball	Chainlink fence 17.24%	Crossword puzzle 6.90%	Photocopier 6.32%	Electric guitar 5.75%	Accordion 5.17%	Traffic light 4.60%	Digital clock 2.87%	Projector 2.87%	… **Soccer ball T26 1.15%**
	40. Spotlight**	**Spotlight 39.66%**	Projector 15.52%	Traffic light 5.75%	Vacuum 4.60%	Obelisk 3.45%	Stethoscope 3.45%	Soccer ball 2.87%	Freight car 2.30%	
	41. Starfish*	Sea snake 11.49%	Electric guitar 6.90%	Peacock 6.90%	Accordion 6.32%	Chainlink fence 5.17%	Slot machine 4.60%	Pinwheel 4.02%	**Starfish 4.02%**	
	42. Stethoscope**	Car wheel 12.06%	**Stethoscope 10.34%**	Chainlink fence 6.90%	Spotlight 6.32%	Obelisk 4.02%	Photocopier 4.02%	Projector 4.02%	Soccer ball 4.02%	
	43. Strawberry**	**Strawberry 25.29%**	Traffic light 16.67%	Slot machine 6.90%	Baseball 4.60%	Monarch butterfly 3.45%	Stethoscope 3.45%	Soccer ball 2.87%	Medicine chest 2.30%	
	44. Tile roof**	Accordion 8.62%	Volcano 8.62%	**Tile roof 8.05%**	Monarch butterfly 6.90%	Chainlink fence 5.17%	Starfish 5.17%	Dial telephone 4.02%	Hand blower 4.02%	
	45. Traffic light**	Projector 7.47%	Pinwheel 6.90%	Spotlight 6.32%	Car wheel 5.75%	**Traffic light 5.75%**	Stethoscope 4.02%	Obelisk 3.45%	Photocopier 3.45%	
	46. Trifle*	Strawberry 9.77%	Pinwheel 7.47%	Roundworm 6.90%	Traffic light 6.90%	Spotlight 6.32%	Volcano 5.75%	Car wheel 4.60%	Stethoscope 4.60%	… **Trifle T13 2.30%**
	47. Vacuum	Green snake 4.60%	Chameleon 4.02%	Monarch butterfly 4.02%	Obelisk 4.02%	Peacock 4.02%	Pinwheel 4.02%	Projector 4.02%	Starfish 4.02%	… **Vacuum T31 1.15%**
	48. Volcano**	Pole 9.77%	Hair clip 6.90%	Spotlight 6.90%	Traffic light 6.32%	**Volcano 5.75%**	Roundworm 5.17%	Panpipe 4.60%	Punching bag 4.60%	

* numerically above chance agreement; ** statistically above chance agreement

Figure A.2: **Participant responses ranked by frequency (Experiment 3b).** Continued.

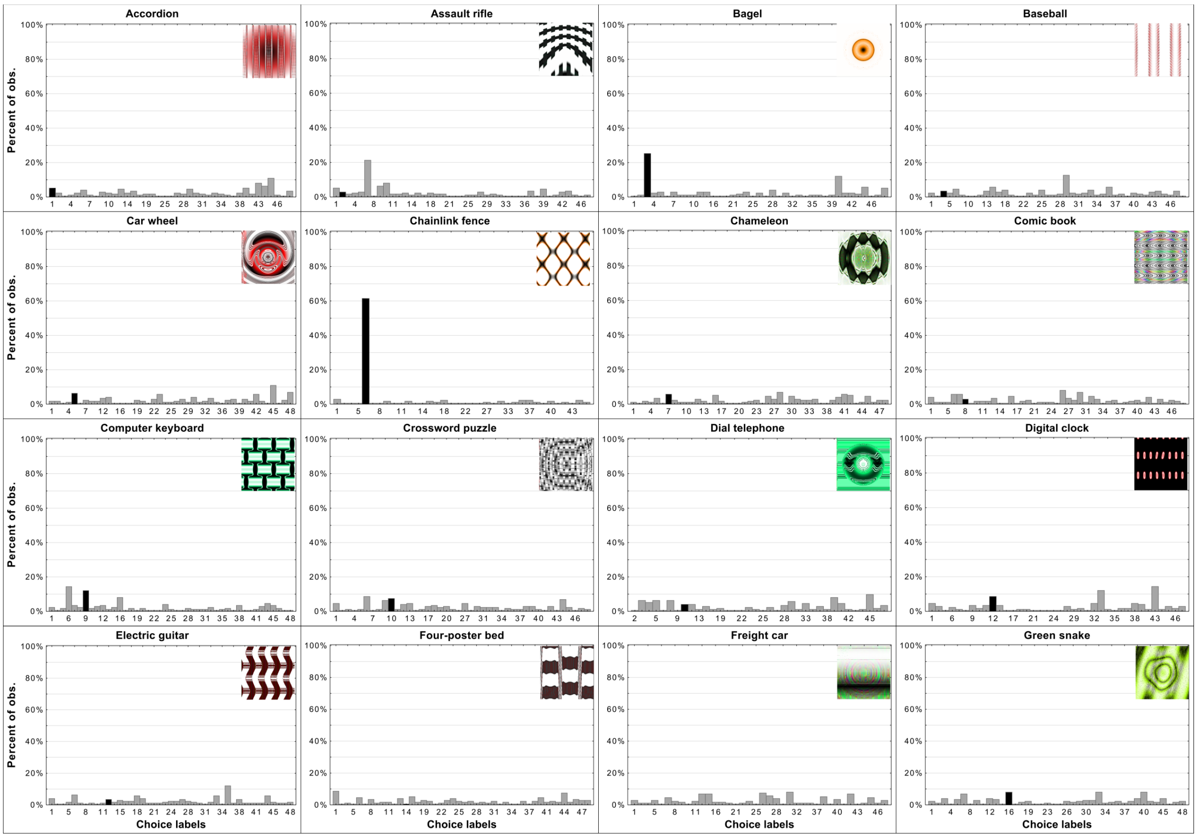


Figure A.3: **Per-item histograms of response choices from Experiment 3b in Zhou and Firestone.** Each histogram contains the adversarial stimuli and shows the percentage of responses per each choice (y-axis). The choice labels (x-axis) are ordered the same way as in Figure 2.2 from 1 to 48. Black bars indicate the DCNN choice for a particular adversarial image.

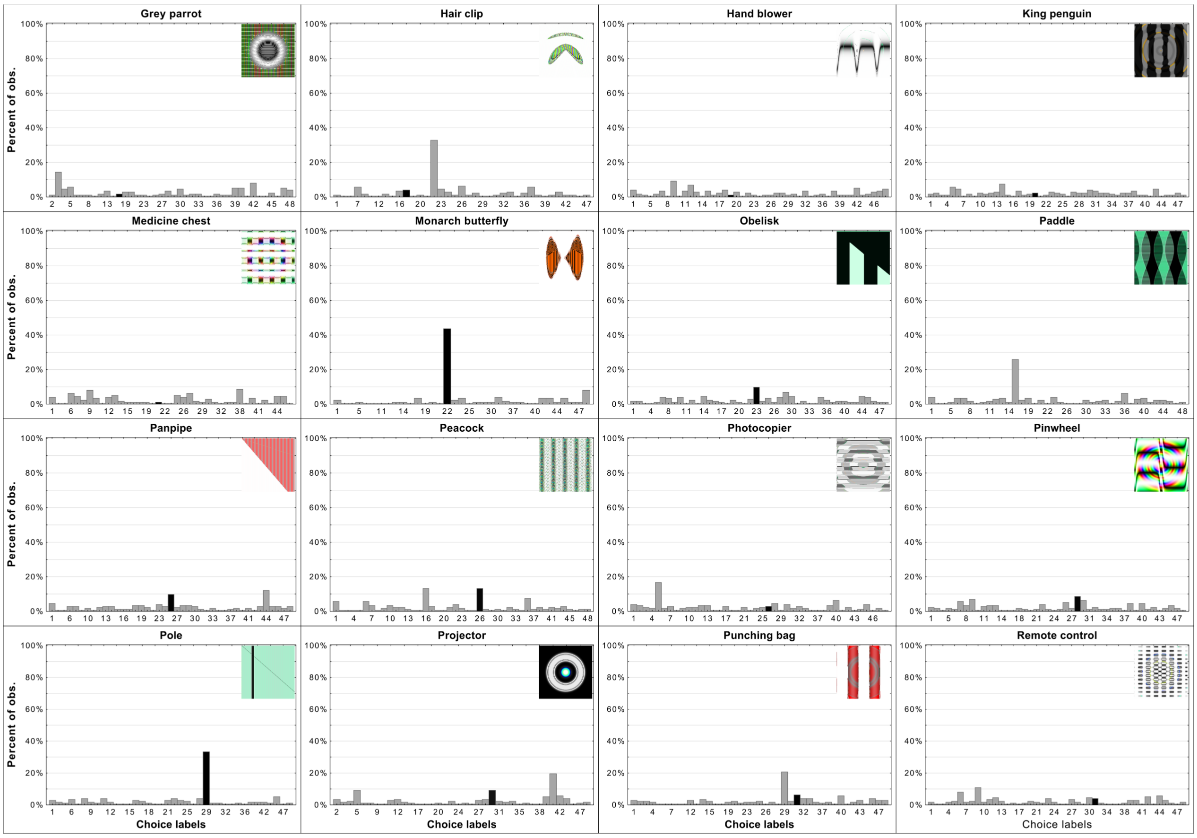


Figure A.4: **Per-item histograms of response choices from Experiment 3b in Zhou and Firestone.** Continued.

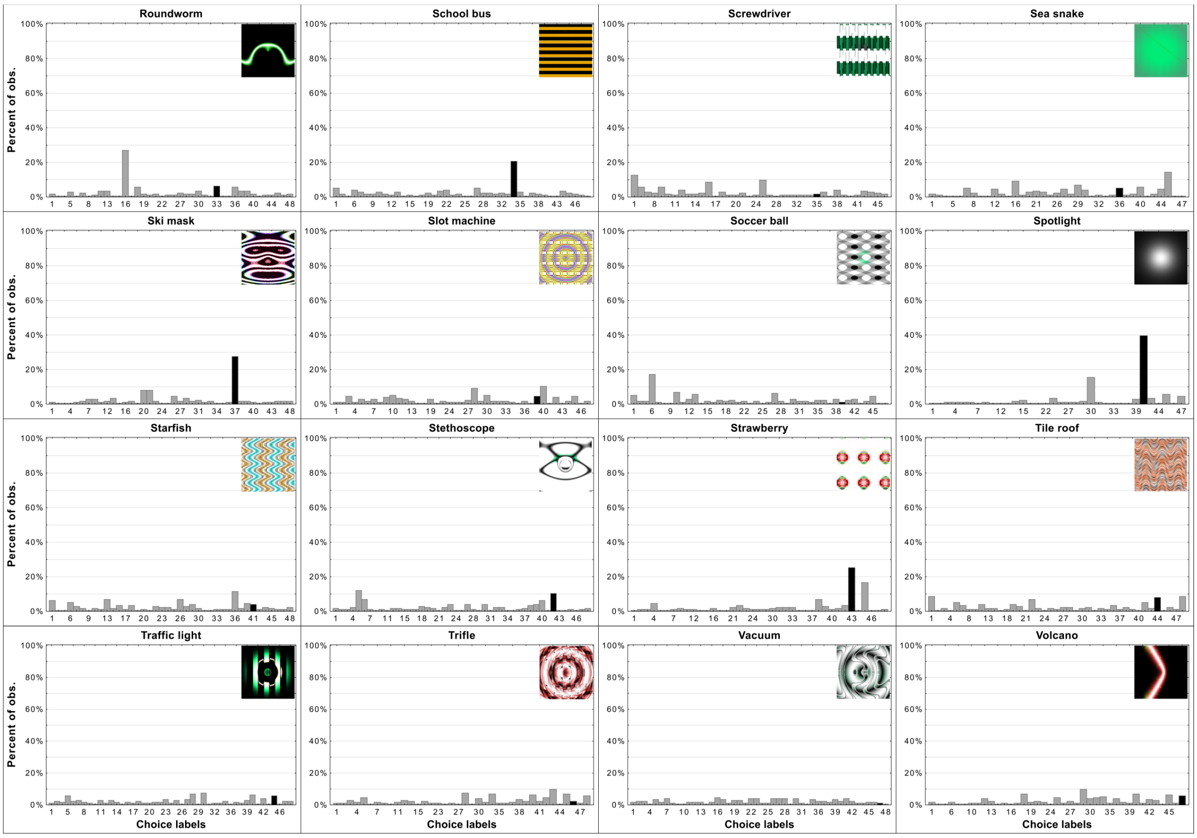


Figure A.5: **Per-item histograms of response choices from Experiment 3b in Zhou and Firestone.** Continued.

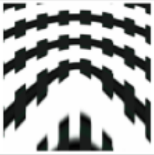
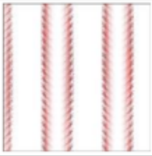
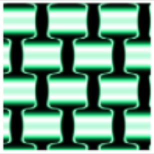
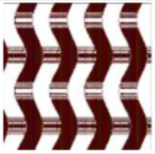
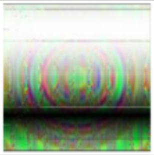
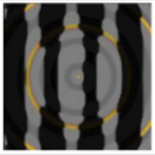
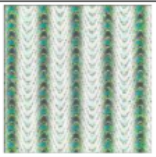
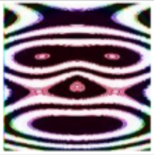
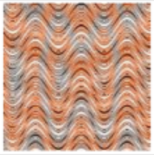
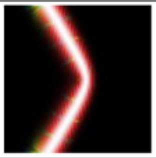
Stimulus	DCNN label	Alternative 1	Alternative 2	Alternative 3
	Assault rifle	Accordion	Chainsaw	Piano
	Baseball	Bath towel	Parallel bars	Wool
	Computer keyboard	Chainlink fence	Crossword puzzle	Honeycomb
	Electric guitar	Harmonica	Picket fence	Radiator grille
	Freight car	Bubble	Car wheel	Traffic light
	King penguin	CD player	Plate	Wall clock
	Peacock	Broccoli	Grass snake	Valley
	Ski mask	Baboon	Loudspeaker	Totem pole
	Tile roof	Rock crab	Shower curtain	Wooden spoon
	Volcano	Lighter	Missile	Table lamp

Figure A.6: **Experiment 1 stimuli and competitive alternative labels.**

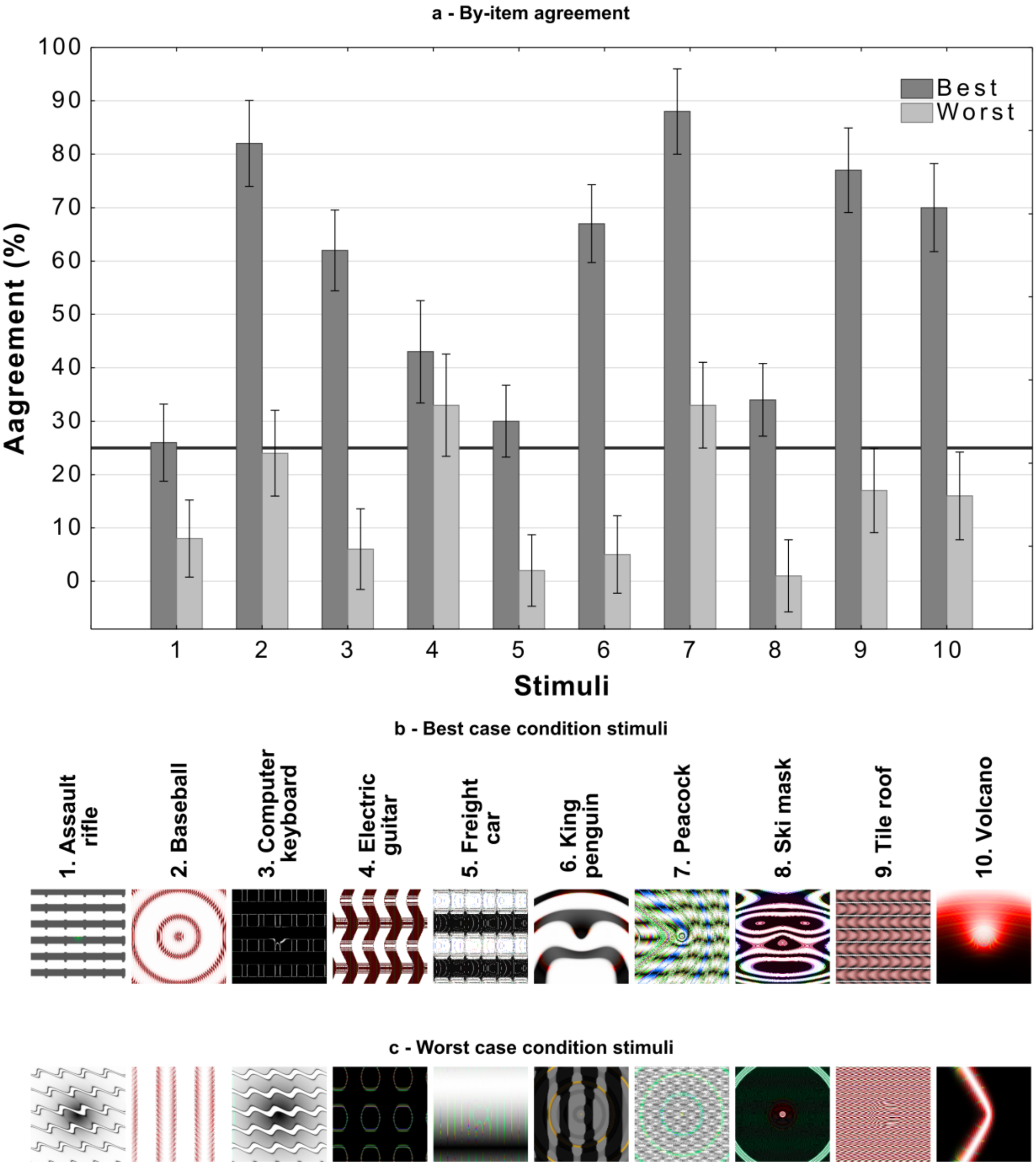


Figure A.7: **An item-wise breakdown of agreement levels in Experiment 2 as a function of experimental condition and category.** Average agreement levels for each category in each condition with 95% CI are presented in (a) with the black line referring to chance agreement. The best case stimuli are presented in (b), these stimuli were judged as containing the most features in common with the target category (out of 5 generated by Nguyen et al. [127]). The worst case stimuli are presented in (c), these were judged to contain the least number of features in common with the target category.

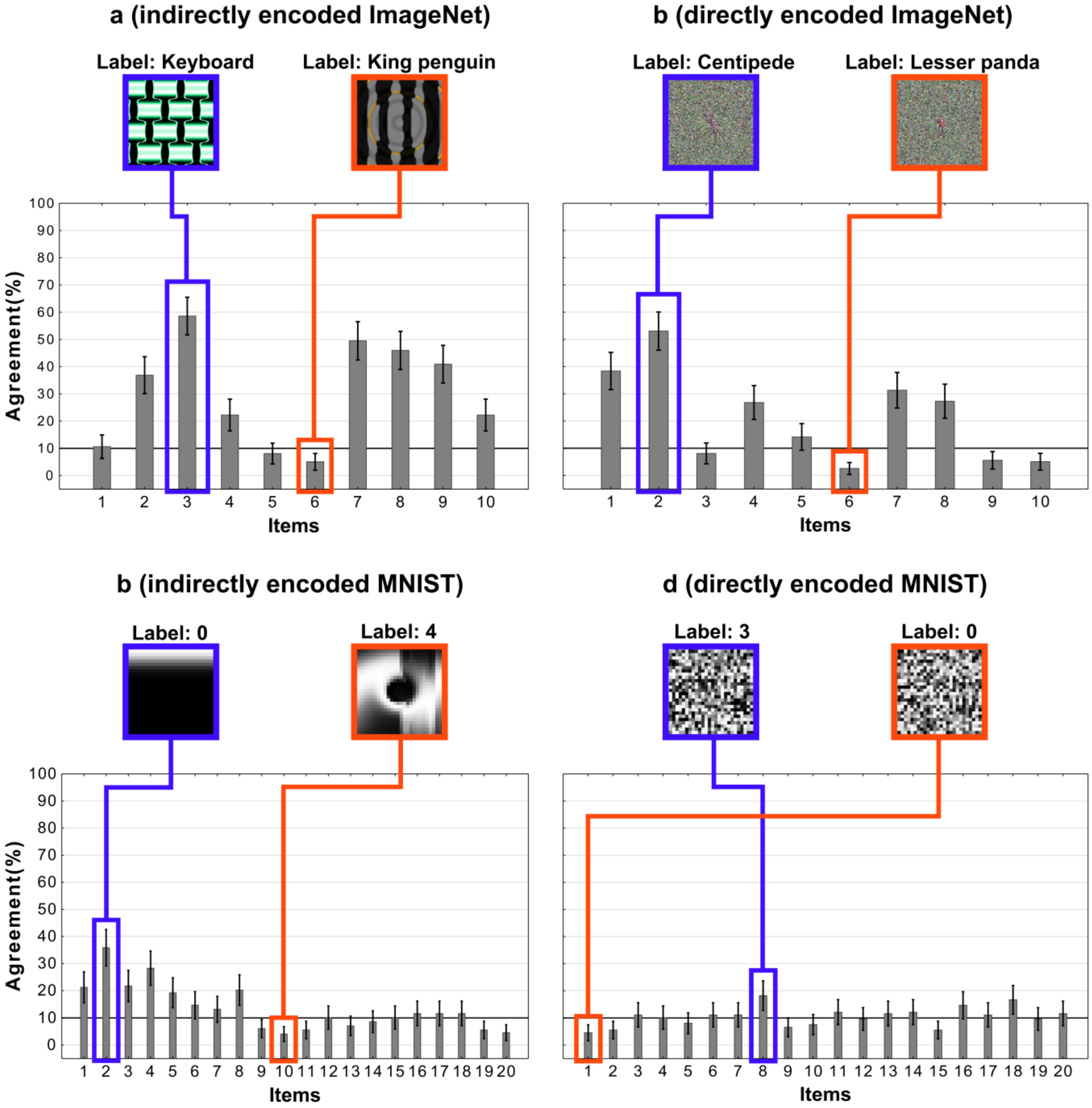


Figure A.8: **An item-wise breakdown of agreement levels for the four conditions in Experiment 3.** Each bar shows the agreement level for a particular image, that is, the percentage of participants that agreed with DCNN classification for that image. Each sub-figure also shows the images that correspond to the highest (blue) and lowest (red) levels of agreement under that condition.

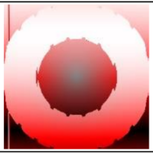
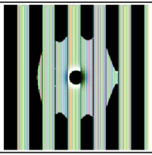
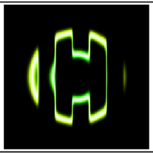
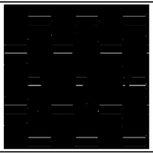
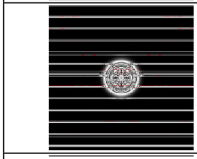
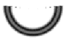
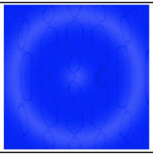
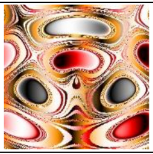
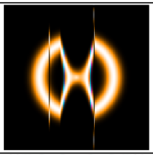
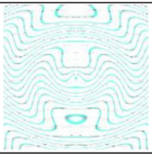
Stimulus	DCNN label	Alternative 1	Alternative 2	Alternative 3
	Balloon	Camera	Car wheel	Missile
	Chime	Prison	Sliding door	Vault
	Digital clock	Mask	Traffic light	Speaker
	File cabinet	Computer keyboard	Computer screen	Window shade
	Radiator grille	Bottle cap	Manhole cover	Safe
	Hook	Cup	Soup bowl	Washbasin
	Jellyfish	Bubble	Seashore	Sea snake
	Mask	Butterfly	Oil filter	Vase
	Matchstick	Bagel	Lighthouse	Orange
	Maze	Bubble	King crab	Spider web

Figure A.9: **Experiment 5 stimuli and competitive alternative labels.**

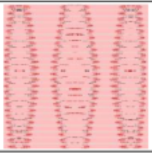
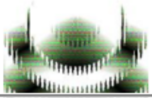
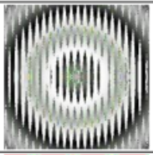
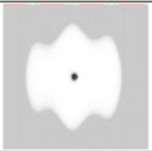
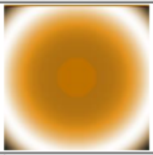
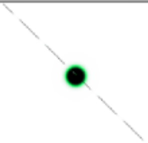
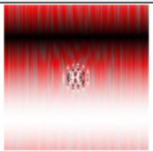
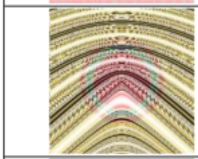
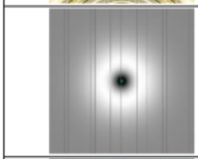
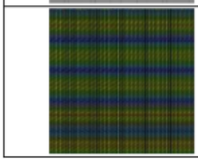
Stimulus	DCNN label	Alternative 1	Alternative 2	Alternative 3
	Menu	Crate	Flatworm	Helix
	Mosque	Chainsaw	Computer screen	Missile
	Picket fence	Camera	Car wheel	Zebra
	Piggy bank	Band aid	Butterfly	Pencil eraser
	Soup	Gong	Orange	Potpie
	Syringe	Ruler	Safety pin	Street sign
	Theatre curtain	Daisy	Microphone	Soccer ball
	Vault	Rock crab	Seashore	Volcano
	Wardrobe	Camera	Shower curtain	Spotlight
	Window screen	Dishrag	Prayer rug	Theatre curtain

Figure A.10: **Experiment 5 stimuli and competitive alternative labels.** Continued.

APPENDIX B

FEATURE BLINDNESS - APPENDIX

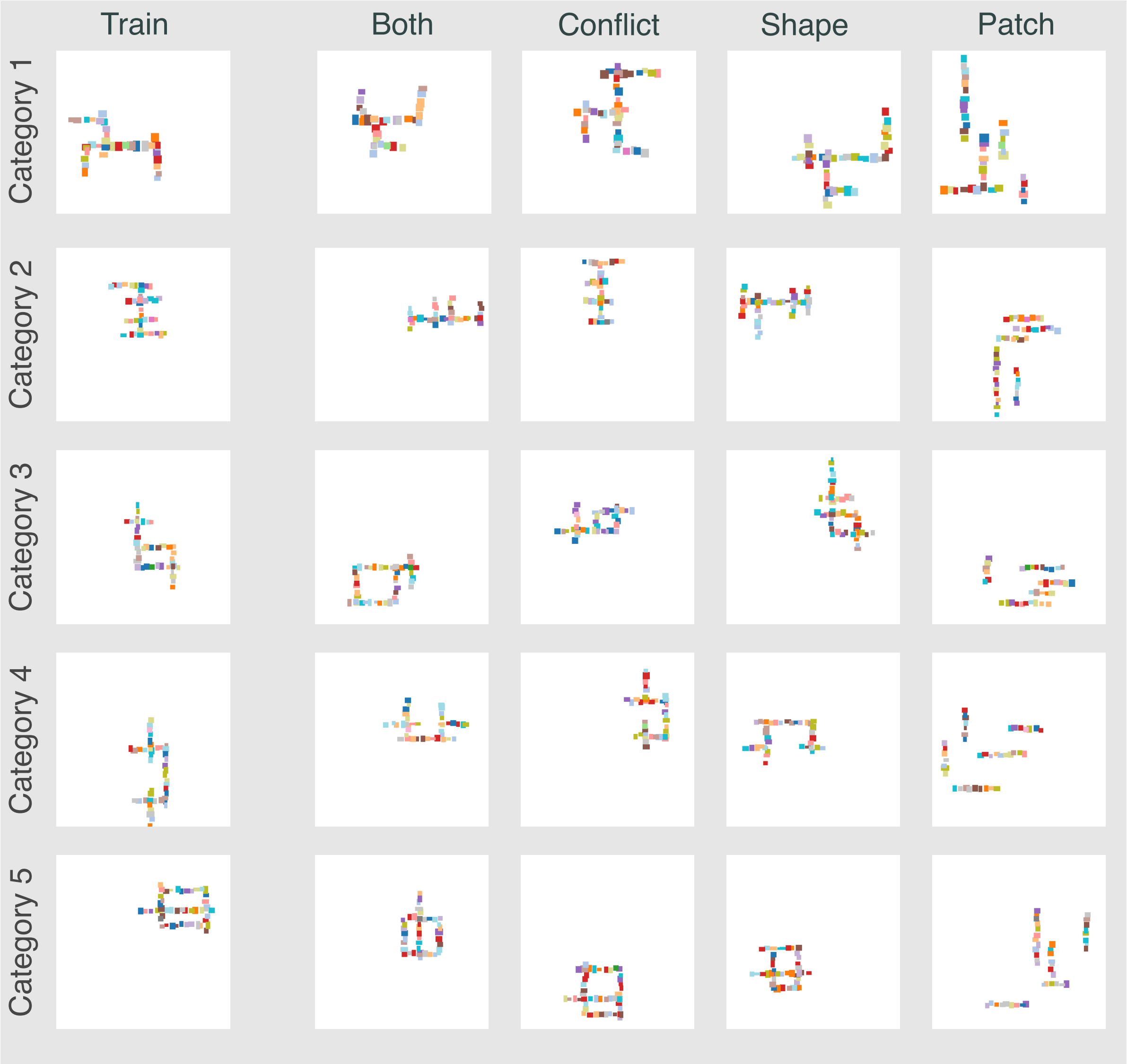


Figure B.1: **Examples of stimuli in Experiment 1 (patch).** In each row we show (from left to right) an example image from the training set, `Both` condition, `Conflict` condition, `Shape` condition and `Non-shape` (`Patch`) condition for a category. Each image in the training set contains a diagnostic 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 `Conflict` 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.

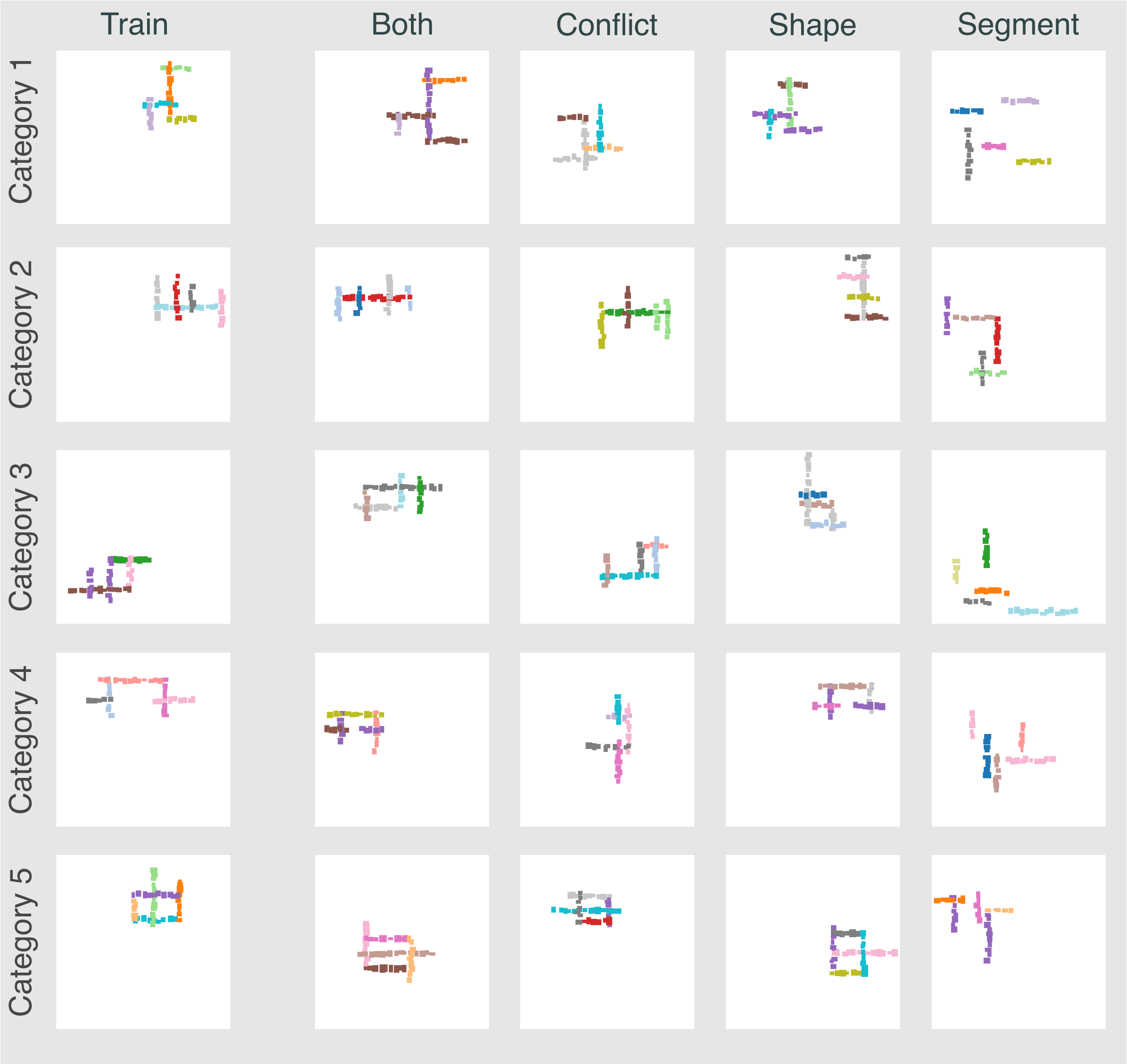


Figure B.2: **Examples of stimuli in Experiment 2 (segment).** In each row we show (from left to right) an example image from the training set, `Both` condition, `Conflict` condition, `Shape` condition and `Non-shape` (`Segment`) condition for a category. Each image in the training set contains a diagnostic 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 `Conflict` 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.

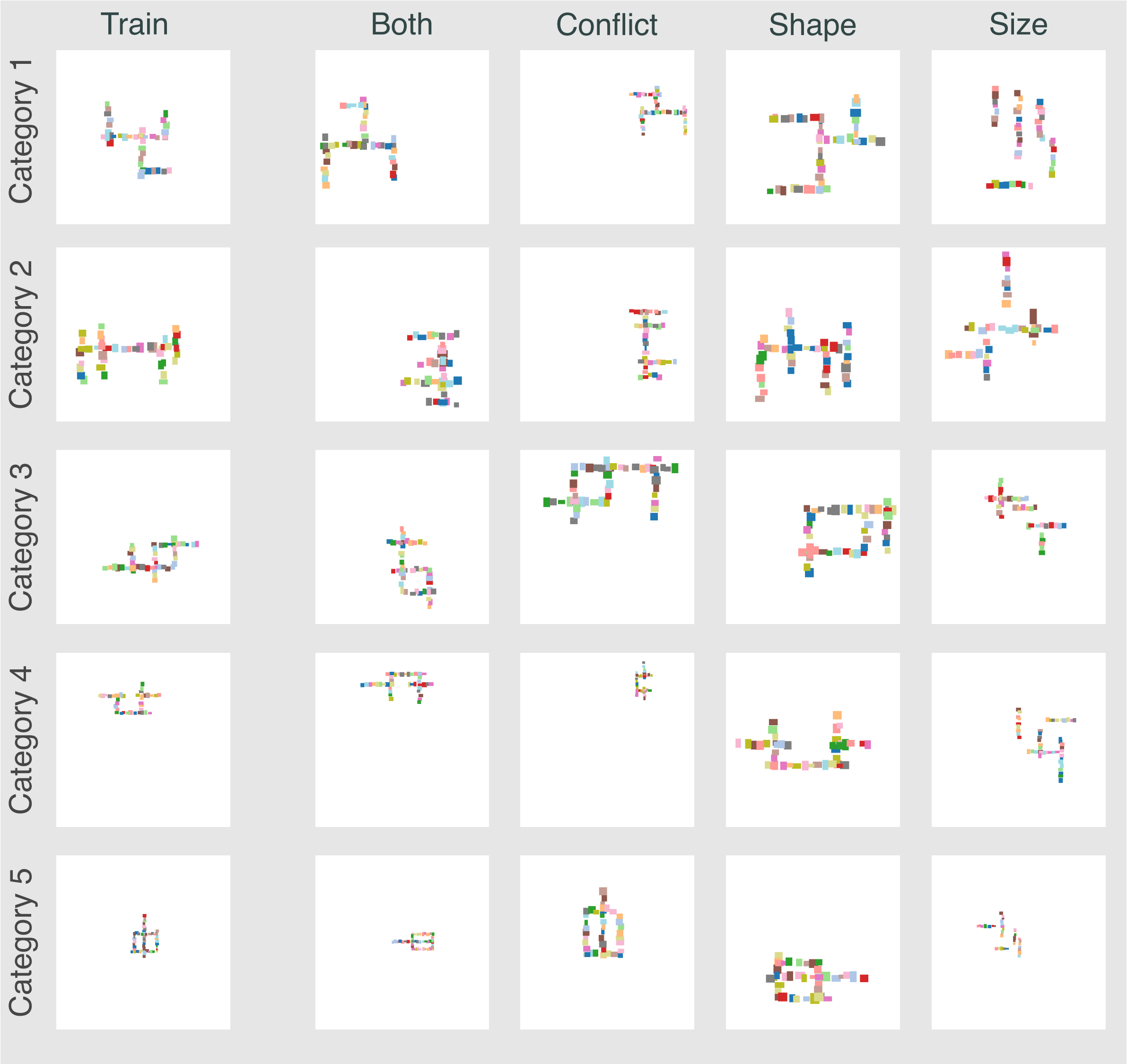


Figure B.3: **Examples of stimuli in Experiment 3 (size).** In each row we show (from left to right) an example image from the training set, `Both` condition, `Conflict` condition, `Shape` condition and `Non-shape` (`Size`) condition for a category. The average size of all images in the training set 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 `Conflict` 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.

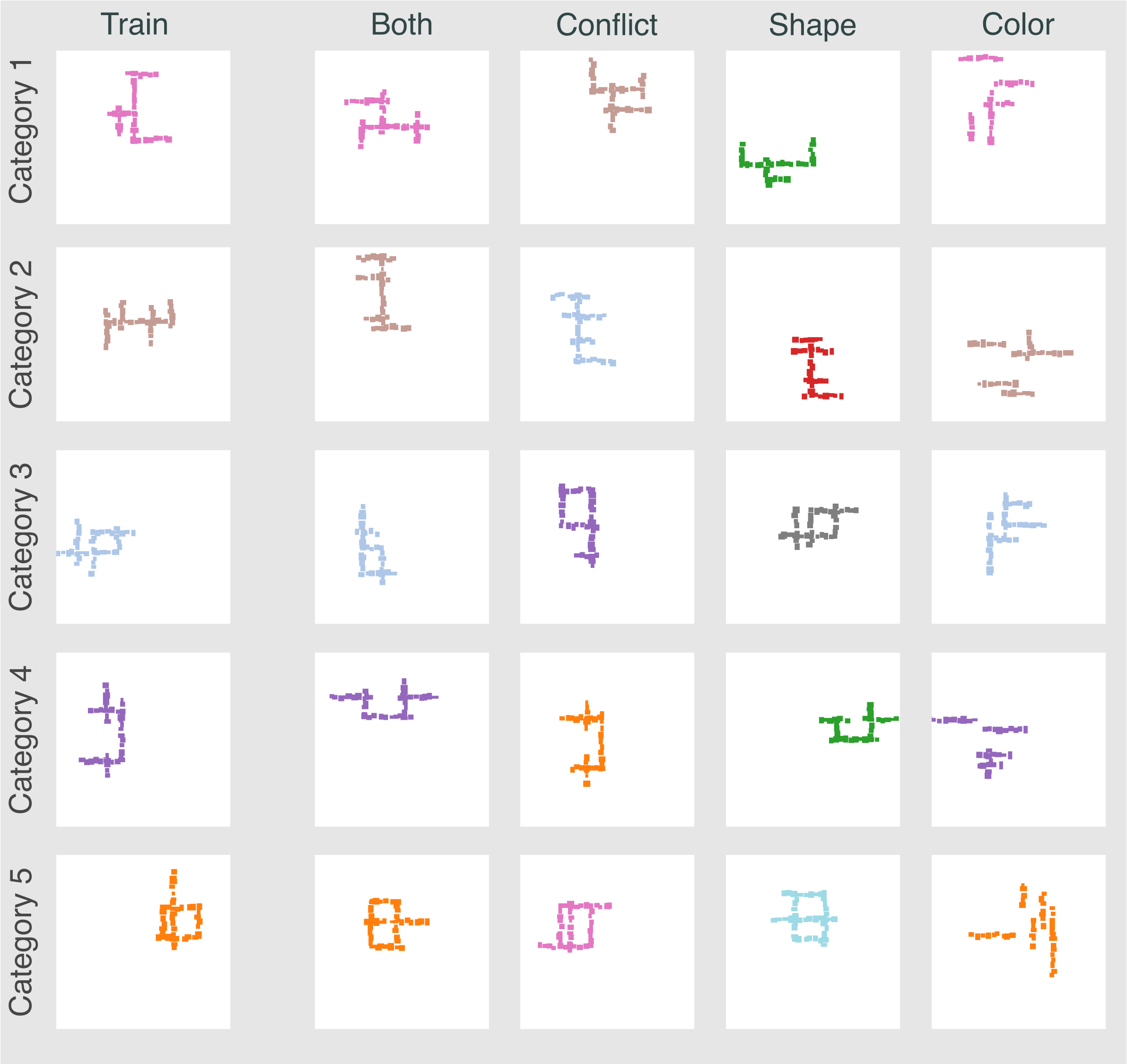


Figure B.4: **Examples of stimuli in Experiment 4 (colour).** In each row we show (from left to right) an example image from the training set, `Both` condition, `Conflict` condition, `Shape` condition and `Non-shape (Size)` condition for a category. All patches in an image have the same colour. This 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 `Conflict` 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.

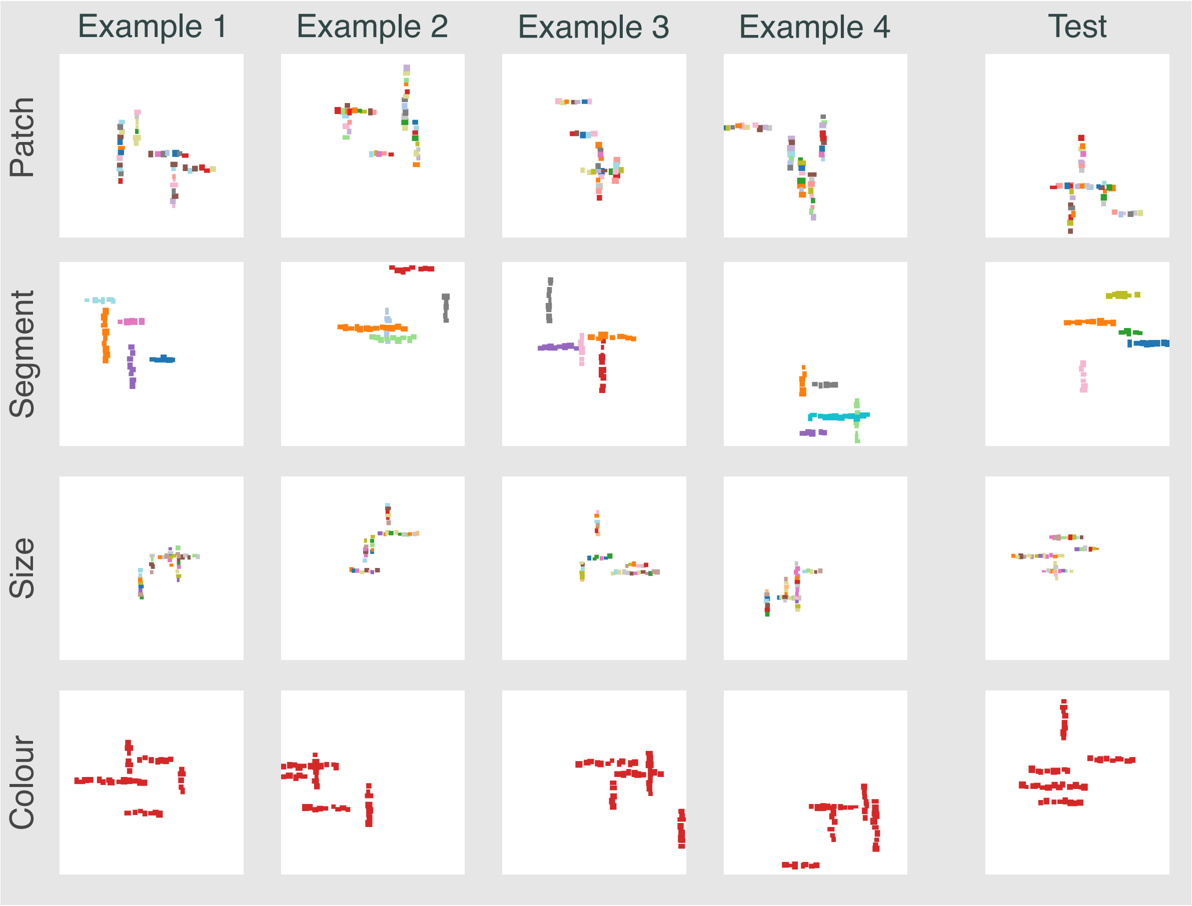


Figure B.5: **Examples of stimuli in Experiment 5 and 6 (no shape).** Each row shows four examples from the training set that have the same category label as well as one example from the test set with the same label. The four rows correspond to the four conditions. In row 1, the predictive feature is 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.

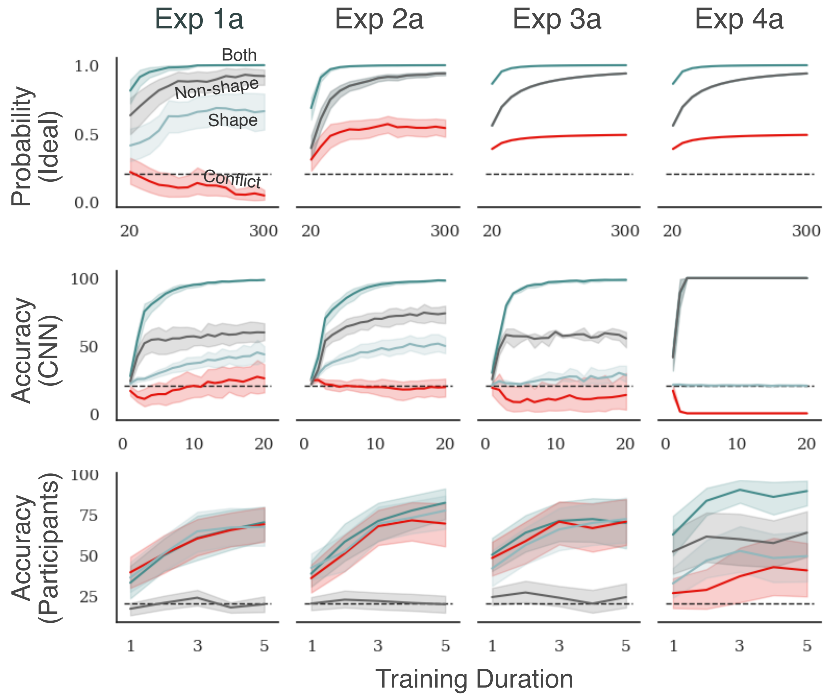


Figure B.6: **Change in test performance with training in Experiments 1a, 2a, 3a, and 4a.** Figure 3.8A in the main text shows the change in performance under the four test conditions in Experiment 1b, 2b, 3b and 4b, where the non-shape feature and more predictive than shape features in training. Here we have plotted how performance changes in Experiments 1a, 2a, 3a and 4a, where both features are equally likely. 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. A comparison of Figsure B.6 and 3.8A 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 `Conflict` 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.

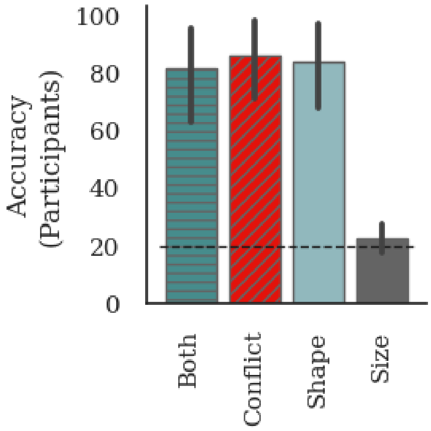
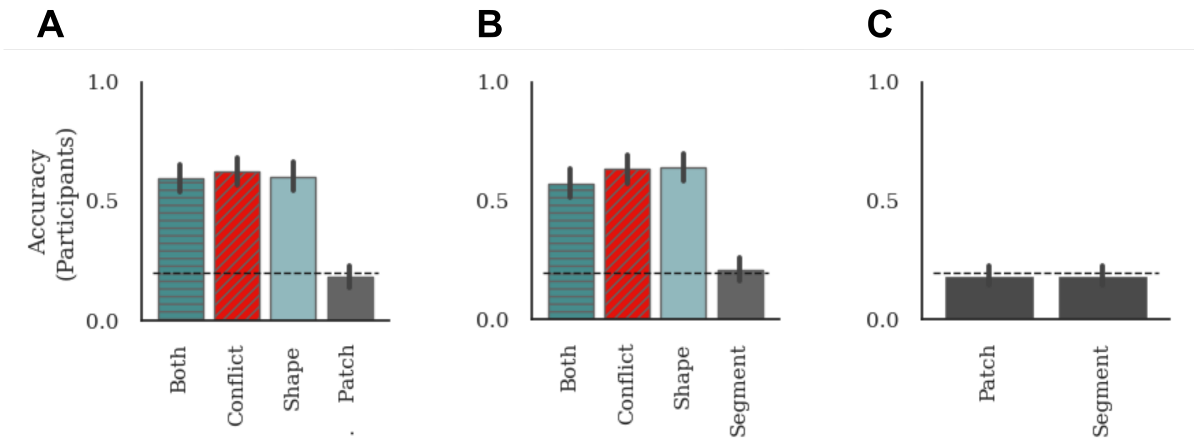


Figure B.7: **Results for Experiment S4.** Accuracy in the four conditions when participants are shown the stimuli for 3s instead of 1s. In this experiment, every trial has two diagnostic features – global shape and average size. Despite the increase in the duration of the stimulus, participants performed 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.



(a) **Exp S1.** Patch (b) **Exp S2.** Segment (c) **Exp S3.** No shape

Figure B.8: **Results for Experiments S1, S2 and S3.** In three experiments, we tested how participant behaviour changed when we presented the stimuli for a shorter duration (100ms) and restricted the field of view, such that the stimulus was always presented within 10° of fixation (see Materials and Methods in main text). (a) Accuracy of N=25 participants in the four test conditions in an experiment that mirrors Experiment 1b – i.e., all training images contain a diagnostic patch and 80% images contain a diagnostic shape, (b) Accuracy of N=25 participants in the four test conditions in an experiment mirroring Experiment 2b – i.e., all training images contain a diagnostic segment and 80% images contain a diagnostic shape, (c) Accuracy of two groups of N=10 participants in test block where the training images contained only non-shape cues. Performance of participants in all experiments was consistent with their performance observed in other experiments. Though the overall accuracy of participants in this control experiments was slightly lower (mean accuracy in the Both condition was $M = 59.60\%$ in (a) and $M = 57.20\%$ in (b)), which is understandable given the faster presentation time, there was statistically no difference in their performance in the Both, Conflict and Shape conditions and their performance in the Non-shape condition was at chance. In Experiment (c), where there was no shape features in the training set, performance of both the Patch and Segment groups was statistically at chance. That is, participants consistently learned based on shape cues; when a diagnostic shape was not present during training, no participant managed to learn the task. (Compare results with Figure 3.4B and 3.8B)

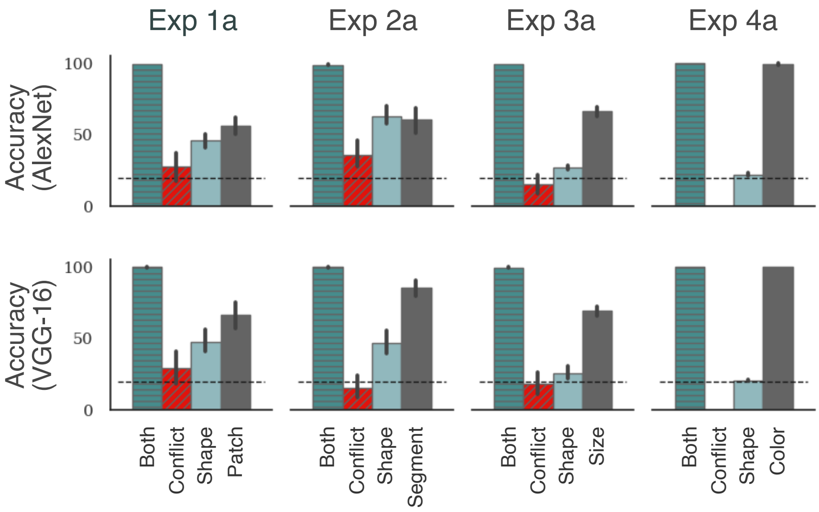


Figure B.9: **Results when both features are equally predictive.** Each panel shows the accuracy under the four test conditions for AlexNet (top row) or VGG-16 (bottom row). Each column corresponds to a different experiment. Both models were pre-trained on `ImageNet` and fine-tuned by reshaping the final layer to reflect the number of target classes in each experiment and trained on 2000 images from the training set (see Materials and Methods for details). A comparison with Figure 3.4A 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 Conflict condition).

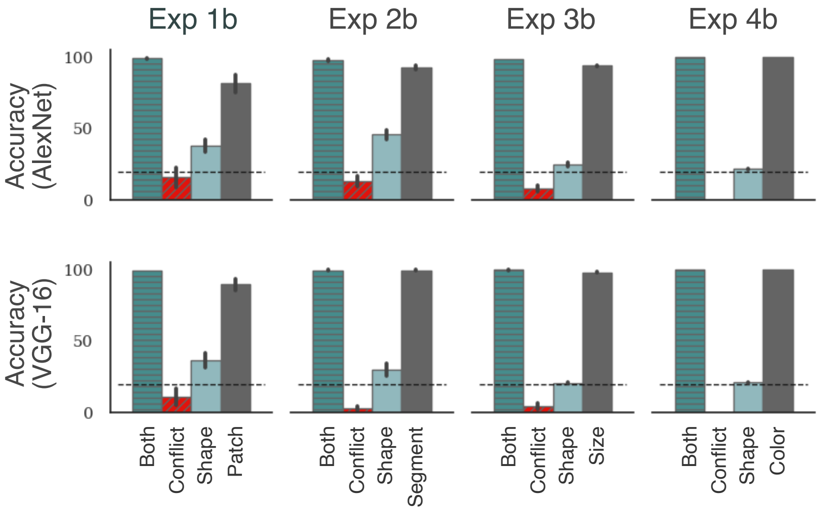


Figure B.10: **Results when non-shape feature is more predictive.** Each panel again shows the accuracy under the four test conditions for AlexNet (top row) or VGG-16 (bottom row). Each column corresponds to a different experiment. A comparison with Figure 3.4B shows that both architectures showed the 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-Conflict-Shape-Non-shape conditions) and this preference became larger than the experiments where both features were equally predictive (compare with Figure B.9).

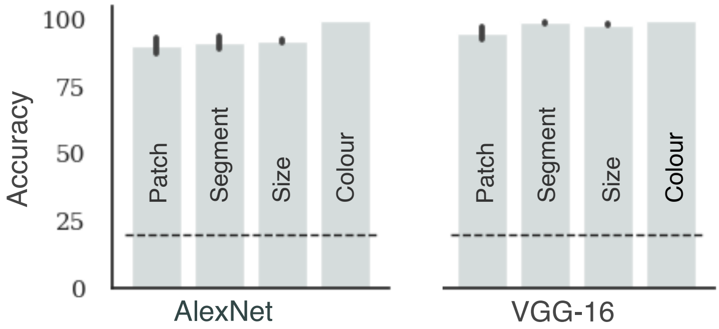


Figure B.11: **Results for learning without shape feature.** The two panels show accuracy in test blocks for AlexNet and VGG-16, respectively, when these models were trained on images that lack any coherent shape (Experiment 5). Each bar corresponds to the type of non-shape feature used in training. Like ResNet50, but unlike human participants (compare with Figure 3.8B), both models were able to learn all types of non-shape features.

APPENDIX C

REPRESENTATION SIMILARITY ANALYSIS - APPENDIX

In Study 3 we found that, relative to the highest RSA scores, networks from the Negative condition obtained higher scores when RSA was calculated using image-level RDMs when compared to averaging and calculating on category-level RDMs. Here we present a simple set of Monte Carlo-like simulations which explain how an increase of stimuli per category can inflate RSA scores between systems.

Imagine we are computing RSA using a dataset of 100 images for each of 10 categories. After extracting activation patterns for each stimulus and computing distances for each pair of stimuli, we would be left with RDMs of size 1000×1000. Since RSA is computed after excluding distances on the diagonal (which have a value of 0) and values above the diagonal (or below, the RDM is mirrored) - this leaves us with a total of 4950 pairwise distances from each RDM. A correlation (usually rank-order) is then computed between sets of 4950 values from different systems in order to obtain the RSA score between them.

From the 4950 pair-wise distances - 4500 are distances between stimuli from different categories and 450 are within-category distances. Let us further assume that both systems being compared are representing within-category stimuli as more similar to each other than to stimuli from other categories. This means that the 450 within-category distances should be the smallest distances for each system.

In a set of simple simulations distances were sampled randomly in the following way. First, 4500 distances were sampled randomly from a Gaussian distribution with a mean of 1 and a standard deviation of 0.1 (distances in real RDMs - if computed as in Chapter 4 using the 1-Pearson's r metric have a range of 0-2). These represent the between-category distances. The distances were sampled independently for each system. Since the sampling was done in this manner, the resulting distances do not correlate between two systems. In effect, this would mean that two systems do not share the between-category representational structure. Second, 450 distances were sampled in the same manner as the initial 4500 but the mean of the Gaussian distribution was 0.5 rather than 1. These are within-category distances, and if the two systems are representing stimuli within categories as more similar to each other than to stimuli from different categories - then these 450 distances should be smaller (similarity should be higher) than the first set of sampled distances. Again, due to the manner of sampling, these sets of distances do not correlate meaning that within-category structures are not similar between systems. Finally, we combine the two sets of distances into the full set of 4950 for each system to compute the RSA score between pairs of systems.

The simulation was conducted by generating distances for a 100 systems. This meant that 4950 pairs of systems were compared. The median RSA score (Spearman rank-order correlation) between these systems was 0.248 with a range of 0.206-0.285. These results show that even systems which do not share between-category nor within-category representational geometry can achieve significant RSA scores simply due to the fact that they represent within-category stimuli as generally more similar to each other than to stimuli from different categories.

BIBLIOGRAPHY

[1] Akhtar, N. and Mian, A. (2018).
Threat of adversarial attacks on deep learning in computer vision: A survey.
IEEE Access, 6:14410–14430.

[2] Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. (2019).
Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4845–4854.

[3] Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. (2017).
Synthesizing robust adversarial examples.
arXiv preprint arXiv:1707.07397.

[4] Ayzenberg, V. and Behrmann, M. (2022).
Does the brain's ventral visual pathway compute object shape?
Trends in Cognitive Sciences.

[5] Baddeley, A. (2010).
Working memory.
Current biology, 20(4):R136–R140.

[6] Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018).
Deep convolutional networks do not classify based on global object shape.
PLoS computational biology, 14(12):e1006613.

[7] Bengio, Y., Courville, A., and Vincent, P. (2013).
Representation learning: A review and new perspectives.
IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828.

[8] Berrios, W. and Deza, A. (2022).
Joint rotational invariance and adversarial training of a dual-stream transformer yields state of the art brain-score for area v4.
arXiv preprint arXiv:2203.06649.

[9] Biederman, I. (1987).
Recognition-by-components: a theory of human image understanding.
Psychological review, 94(2):115.

[10] Biederman, I. and Cooper, E. E. (1991).
Evidence for complete translational and reflectional invariance in visual object priming.
Perception, 20(5):585–593.

[11] Biederman, I. and Cooper, E. E. (1992).
Size invariance in visual object priming.
Journal of Experimental Psychology: Human Perception and Performance, 18(1):121.

[12] Biederman, I. and Ju, G. (1988).
Surface versus edge-based determinants of visual recognition.
Cognitive psychology, 20(1):38–64.

[13] Bishop, C. M. (2006).
Pattern recognition and machine learning.
springer.

[14] Block, N. (1986).
Advertisement for a semantics for psychology.
Midwest Studies in Philosophy, 10(1):615–678.

[15] Blything, R., Vankov, I., Ludwig, C., and Bowers, J. (2019).
Extreme translation tolerance in humans and machines.
In *Conference on Cognitive Computational Neuroscience*.

[16] Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., and Love, B. C. (2020).
Measures of neural similarity.
Computational Brain & Behavior, 3(4):369–383.

[17] Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., and et al. (2022).
Deep problems with neural network models of human vision.
Behavioral and Brain Sciences, page 1–74.

[18] Bracci, S., Ritchie, J. B., and de Beeck, H. O. (2017).
On the partnership between neural representations of object categories and visual features in the ventral visual pathway.
Neuropsychologia, 105:153–164.
Special Issue: Concepts, Actions and Objects: Functional and Neural Perspectives.

[19] Brascamp, J. W. (2021).
Controlling the spatial dimensions of visual stimuli in online experiments.
Journal of Vision, 21(8):19–19.

[20] Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014).
Deep neural networks rival the representation of primate it cortex for core visual object recognition.
PLoS computational biology, 10(12):e1003963.

[21] Cai, M. B., Schuck, N. W., Pillow, J. W., and Niv, Y. (2019).
Representational structure or task structure? bias in neural representational similarity analysis and a bayesian method for reducing bias.
PLOS Computational Biology, 15(5):1–30.

[22] Churchland, P. M. (1989).
Some reductive strategies in cognitive neurobiology.
In Silvers, S., editor, *Rerepresentation: Readings in the Philosophy of Mental Representation*, pages 223–253. Springer Netherlands, Dordrecht.

[23] Churchland, P. M. (1998).
Conceptual similarity across sensory and neural diversity: The fodor/lepore challenge answered.
Journal of Philosophy, 95(1):5–32.

[24] Cichy, R. M. and Kaiser, D. (2019).
Deep neural networks as scientific models.
Trends in cognitive sciences, 23:305–317.

[25] Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016).
Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence.
Sci Rep, 6:27755.

[26] Colunga, E. and Smith, L. B. (2008).
Knowledge embedded in process: the self-organization of skilled noun learning.
Developmental Science, 11(2):195–203.

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

[28] DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012).
How does the brain solve visual object recognition?
Neuron, 73(3):415–434.

[29] Dodge, S. and Karam, L. (2017).
A study and comparison of human and deep learning recognition performance under visual distortions.
In *2017 26th international conference on computer communication and networks (ICCCN)*, pages 1–7. IEEE.

[30] Doerig, A., Bornet, A., Choung, O., and Herzog, M. (2020).
Crowding reveals fundamental differences in local vs. global processing in humans and machines.
Vision Research, 167:39–45.

[31] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020).
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv:2010.11929.

[32] Dujmović, M., Bowers, J. S., Adolfi, F., and Malhotra, G. (2022).
Some pitfalls of measuring representational similarity using representational similarity analysis.
bioRxiv.

[33] Dujmović, M., Malhotra, G., and Bowers, J. S. (2020).
What do adversarial images tell us about human vision?
eLife, 9:e55978.

[34] Edelman, S. (1998).
Representation is representation of similarities.
Behavioral and Brain Sciences, 21(4):449–467.

[35] Eickenberg, M., Gramfort, A., Varoquaux, G., and Thirion, B. (2017).
Seeing it all: Convolutional network layers map the function of the human visual system.
NeuroImage, 152:184–194.

[36] Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., and Sohl-Dickstein, J. (2018).
Adversarial examples that fool both computer vision and time-limited humans.
In *Advances in Neural Information Processing Systems*, pages 3910–3920.

[37] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. (2018).
Robust physical-world attacks on deep learning visual classification.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Feinman, R. and Lake, B. M. (2018).
Learning inductive biases with simple neural networks.
arXiv preprint arXiv:1802.02745.

[39] Firth, J. R. (1957).
A synopsis of linguistic theory, 1930-1955.
Studies in linguistic analysis, pages 1–31.

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

[41] Fodor, J. (1987).
Psychosemantics: The Problem of Meaning in the Philosophy of Mind.
MIT Press, Cambridge.

[42] Fodor, J. and Lepore, E. (1992).
Holism: A Shoppers Guide.
Blackwell, Cambridge.

[43] Fodor, J. and Lepore, E. (1996).
Churchland on state space semantics.
In McCauley, R. N., editor, *The Churchlands and Their Critics*, pages 145–158. Blackwell.

[44] Fodor, J. and Lepore, E. (1999).
All at sea in semantic space: Churchland on meaning similarity.
Journal of Philosophy, 96(8):381–403.

[45] Freund, M. C., Etzel, J. A., and Braver, T. S. (2021).
Neural coding of cognitive control: The representational similarity analysis approach.
Trends in Cognitive Sciences, 25(7):622–638.

[46] Fukushima, K. (1980).
Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.
Biological Cybernetics, 36(4):193–202.

[47] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020a).
Shortcut learning in deep neural networks.
Nature Machine Intelligence, 2(11):665–673.

[48] Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2017).
Comparing deep neural networks against humans: object recognition when the signal gets weaker.
arXiv preprint arXiv:1706.06969.

[49] Geirhos, R., Meding, K., and Wichmann, F. A. (2020b).
Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency.
Advances in Neural Information Processing Systems, 33:13890–13902.

[50] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2018a).
Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness.
arXiv preprint arXiv:1811.12231.

[51] Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2018b).
Generalisation in humans and deep neural networks.
Advances in neural information processing systems, 31.

[52] Geisler, W. S. (2011).
Contributions of ideal observer theory to vision research.
Vision research, 51(7):771–781.

[53] Golan, T., Raju, P. C., and Kriegeskorte, N. (2020).
Controversial stimuli: Pitting neural networks against each other as models of human cognition.
Proceedings of the National Academy of Sciences, 117(47):29330–29337.

[54] Goldwater, M. B., Don, H. J., Krusche, M. J., and Livesey, E. J. (2018).
Relational discovery in category learning.
Journal of Experimental Psychology: General, 147(1):1.

[55] Goodfellow, I., Lee, H., Le, Q., Saxe, A., and Ng, A. (2009).
Measuring invariances in deep networks.
Advances in neural information processing systems, 22:646–654.

[56] Goodfellow, I. J., Papernot, N., Huang, S., Duan, R., Abbeel, P., and Clark, J. (2017).
Attacking machine learning with adversarial examples.
last accessed on 18/06/2020.

[57] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014).
Explaining and harnessing adversarial examples.
arXiv preprint arXiv:1412.6572.

[58] Griffiths, T. L. and Steyvers, M. (2002).
A probabilistic approach to semantic representation.
In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ. Erlbaum.

[59] Griffiths, T. L., Steyvers, M., and Tenenbaum, J. (2007).
A probabilistic approach to semantic representation.
Psychological Review, 114(2):211–244.

[60] Güçlü, U. and van Gerven, M. A. (2015).
Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream.
Journal of Neuroscience, 35(27):10005–10014.

[61] Guo, C., Lee, M., Leclerc, G., Dapello, J., Rao, Y., Madry, A., and Dicarlo, J. (2022).
Adversarially trained neural representations are already as robust as biological neural representations.
In *International Conference on Machine Learning*, pages 8072–8081. PMLR.

[62] Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., and Moser, E. I. (2005).
Microstructure of a spatial map in the entorhinal cortex.
Nature, 436(7052):801–806.

[63] Haxby, J., Guntupalli, J., Connolly, A., Halchenko, Y., Conroy, B., Gobbini, M., Hanke, M., and Ramadge, P. (2011).
A common, high-dimensional model of the representational space in human ventral temporal cortex.
Neuron, 72(2):404–416.

[64] Haxby, J. V. (2012).
Multivariate pattern analysis of fMRI: the early beginnings.
Neuroimage, 62(2):852–855.

[65] Haxby, J. V., Connolly, A. C., and Guntupalli, J. S. (2014).
Decoding neural representational spaces using multivariate pattern analysis.

Annu Rev Neurosci, 37:435–456.

[66] He, K., Zhang, X., Ren, S., and Sun, J. (2015).
Delving deep into rectifiers: Surpassing human-level performance on imagenet classification.
In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

[67] He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep residual learning for image recognition.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[68] Hempel, C. G. (1950).
Problems and changes in the empiricist criterion of meaning.
Revue Internationale de Philosophie, 4(11):41–63.

[69] Henriksson, L., Khaligh-Razavi, S.-M., Kay, K., and Kriegeskorte, N. (2015).
Visual representations are dominated by intrinsic fluctuations correlated between areas.
NeuroImage, 114:275–286.

[70] Hermann, K., Chen, T., and Kornblith, S. (2020).
The origins and prevalence of texture bias in convolutional neural networks.
Advances in Neural Information Processing Systems, 33.

[71] Hermann, K. and Lampinen, A. (2020).
What shapes feature representations? exploring datasets, architectures, and training.
Advances in Neural Information Processing Systems, 33.

[72] Hosseini, H. and Poovendran, R. (2018).
Semantic adversarial examples.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619.

[73] Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. (2018).
Assessing shape bias property of convolutional neural networks.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1923–1931.

[74] Hubel, D. H. and Wiesel, T. N. (1959).
Receptive fields of single neurones in the cat's striate cortex.
The Journal of Physiology, 148(3):574–591.

[75] Hubel, D. H. and Wiesel, T. N. (1962).
Receptive fields, binocular interaction and functional architecture in the cat's visual cortex.
The Journal of Physiology, 160(1):106–154.

[76] Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. (2019).
Adversarial examples are not bugs, they are features.
arXiv preprint arXiv:1905.02175.

[77] Jagadeesh, A. V. and Gardner, J. L. (2022).
Texture-like representation of objects in human visual cortex.
Proceedings of the National Academy of Sciences, 119(17):e2115302119.

[78] Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., and Mur, M. (2017).
Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments.
Frontiers in Psychology, 8.

[79] Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., and Suppes, P. (2015).
A representational similarity analysis of the dynamics of object processing using single-trial eeg classification.
PLOS ONE, 10(8):1–27.

[80] Karmon, D., Zoran, D., and Goldberg, Y. (2018).
Lavan: Localized and visible adversarial noise.
arXiv preprint arXiv:1801.02608.

[81] Kersten, D., Mamassian, P., and Yuille, A. (2004).
Object perception as bayesian inference.
Annu. Rev. Psychol., 55:271–304.

[82] Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014).
Deep supervised, but not unsupervised, models may explain it cortical representation.
PLOS Computational Biology, 10:1–29.

[83] Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016).
Deep networks can resemble human feed-forward vision in invariant object recognition.
Scientific reports, 6(1):1–24.

[84] Kiat, J. E., Luck, S. J., Beckner, A. G., Hayes, T. R., Pomaranski, K. I., Henderson, J. M., and Oakes, L. M. (2022).
Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis.
Developmental Science, 25(1):e13155.

[85] Kietzmann, T. C., McClure, P., and Kriegeskorte, N. (2018).
Deep neural networks in computational neuroscience.
BioRxiv, page 133504.

[86] Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019).
Recurrence is required to capture the representational dynamics of the human visual system.
Proceedings of the National Academy of Sciences, 116(43):21854–21863.

[87] Kim, B., Reif, E., Wattenberg, M., Bengio, S., and Mozer, M. C. (2021).
Neural networks trained on natural scenes exhibit gestalt closure.
Computational Brain & Behavior, 4(3):251–263.

[88] Kingma, D. P. and Ba, J. (2014).
Adam: A method for stochastic optimization.
arXiv preprint arXiv:1412.6980.

[89] Kool, W. and Botvinick, M. (2018).
Mental labour.
Nature human behaviour, 2(12):899–908.

[90] Körding, K. P. and Wolpert, D. M. (2004).
Bayesian integration in sensorimotor learning.
Nature, 427(6971):244–247.

[91] Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019).
Similarity of neural network representations revisited.
In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

[92] Kriegeskorte, N. (2009).
Relating population-code representations between man, monkey, and computational models.
Frontiers in Neuroscience, 3(3):363–373.

[93] Kriegeskorte, N. (2015).
Deep neural networks: A new framework for modeling biological vision and brain information processing.
Annual Review of Vision Science, 1(1):417–446.
PMID: 28532370.

[94] Kriegeskorte, N. and Diedrichsen, J. (2019).
Peeling the onion of brain representations.
Annual Review of Neuroscience, 42(1):407–432.

[95] Kriegeskorte, N. and Kievit, R. A. (2013).
Representational geometry: integrating cognition, computation, and the brain.
Trends in Cognitive Sciences, 17(8):401–412.

[96] Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a).
Representational similarity analysis - connecting the branches of systems neuroscience.
Frontiers in Systems Neuroscience, 2.

[97] Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008b).
Matching categorical object representations in inferior temporal cortex of man and monkey.
Neuron, 60(6):1126–1141.

[98] Kriegeskorte, N. and Wei, X.-X. (2021).
Neural tuning and representational geometry.
Nature Reviews Neuroscience, 22(11):703–718.

[99] Krizhevsky, A. and Hinton, G. (2009).
Learning multiple layers of features from tiny images.
Technical report, University of Toronto, Toronto, Ontario.

[100] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
Advances in neural information processing systems, 25:1097–1105.

[101] Kubilius, J., Bracci, S., and Op de Beeck, H. P. (2016).
Deep neural networks as a computational model for human shape sensitivity.
PLoS computational biology, 12(4):e1004896.

[102] Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., and DiCarlo, J. J. (2018).
Cornet: modeling the neural mechanisms of core object recognition.
BioRxiv, page 408385.

[103] Laakso, A. and Cottrell, G. (2000).
Content and cluster analysis: Assessing representational similarity in neural systems.
Philosophical Psychology, 13(1):47–76.

[104] Lake, B. M., Zaremba, W., Fergus, R., and Gureckis, T. M. (2015).
Deep neural networks predict category typicality ratings for images.
In *CogSci*.

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

[106] LeCun, Y., Bengio, Y., and Hinton, G. (2015).
Deep learning.
nature, 521(7553):436–444.

[107] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989).
Backpropagation applied to handwritten zip code recognition.
Neural computation, 1(4):541–551.

[108] Leek, E. C., Roberts, M., Oliver, Z. J., Cristino, F., and Pegna, A. J. (2016).
Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects.
Neuropsychologia, 89:495–509.

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

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

[111] Majaj, N. J., Hong, H., Solomon, E. A., and DiCarlo, J. J. (2015).
Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance.
Journal of Neuroscience, 35(39):13402–13418.

[112] Majaj, N. J. and Pelli, D. G. (2018).
Deep learning—Using machine learning to study biological vision.
Journal of Vision, 18(13):2–2.

[113] Malhotra, G. and Bowers, J. (2019).
The contrasting roles of shape in human vision and convolutional neural networks.
In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 2261–2267.

[114] Malhotra, G., Dujmovic, M., Hummel, J., and Bowers, J. S. (2021).
The contrasting shape representations that support object recognition in humans and cnns.
bioRxiv.

[115] Malhotra, G., Dujmović, M., and Bowers, J. S. (2022).
Feature blindness: A challenge for understanding and modelling visual object recognition.
PLOS Computational Biology, 18(5):1–27.

[116] Malhotra, G., Evans, B. D., and Bowers, J. S. (2020).
Hiding a plane with a pixel: examining shape-bias in cnns and the benefit of building in biological constraints.
Vision Research, 174:57–68.

[117] Manassi, M., Sayim, B., and Herzog, M. H. (2012).
Grouping, pooling, and when bigger is better in visual crowding.
Journal of Vision, 12(10):13–13.

[118] Margalit, E., Biederman, I., Tjan, B. S., and Shah, M. P. (2017).
What Is Actually Affected by the Scrambling of Objects When Localizing the Lateral Occipital Complex?
Journal of Cognitive Neuroscience, 29(9):1595–1604.

[119] Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021).
An ecologically motivated image dataset for deep learning yields better models of human vision.
Proceedings of the National Academy of Sciences, 118(8).

[120] Michael L. Mack, Alison R. Preston, B. L. (2013).
Decoding the brain's algorithm for categorization from its neural implementation.
Current Biology, 23(20):2023–2027.

[121] Mitchell, M. and Krakauer, D. C. (2022).
The Debate Over Understanding in AI's Large Language Models.

[122] Montero, M. L., Bowers, J. S., Costa, R. P., Ludwig, C. J., and Malhotra, G. (2022).
Lost in latent space: Disentangled models and the challenge of combinatorial generalisation.
arXiv preprint arXiv:2204.02283.

[123] Morcos, A., Raghu, M., and Bengio, S. (2018).
Insights on representational similarity in neural networks with canonical correlation.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

[124] Murphy, K. P. (2012).
Machine learning: a probabilistic perspective.
MIT press.

[125] Muttenthaler, L. and Hebart, M. N. (2021).
Thingsvision: A python toolbox for streamlining the extraction of activations from deep neural networks.
Frontiers in Neuroinformatics, 15:679838.

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

[127] Nguyen, A., Yosinski, J., and Clune, J. (2015).
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.

[128] Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014).
A toolbox for representational similarity analysis.
PLOS Computational Biology, 10:1–11.

[129] O'Keefe, J. (1976).
Place units in the hippocampus of the freely moving rat.
Experimental Neurology, 51(1):78–109.

[130] O'Hearn, K., Larsen, B., Fedor, J., Luna, B., and Lynn, A. (2020).
Representational similarity analysis reveals atypical age-related changes in brain regions supporting face and car recognition in autism.
NeuroImage, 209:116322.

[131] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., and Swami, A. (2016).
The limitations of deep learning in adversarial settings.
In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE.

[132] Pashler, H. and Mozer, M. C. (2013).
When does fading enhance perceptual category learning?
Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(4):1162.

[133] Pasupathy, A. and Connor, C. E. (2002).
Population coding of shape in area V4.
Nat Neurosci, 5(12):1332–1338.

[134] Pasupathy, A., Kim, T., and Popovkina, D. V. (2019).
Object shape and surface properties are jointly encoded in mid-level ventral visual cortex.
Current Opinion in Neurobiology, 58:199–208.
Computational Neuroscience.

[135] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017).
Automatic differentiation in pytorch.
31st Conference on Neural Information Processing Systems.

[136] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019).

Pytorch: An imperative style, high-performance deep learning library.
In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[137] Perconti, P. and Plebe, A. (2020).
Deep learning and cognitive science.
Cognition, 203:104365.

[138] Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2017).
Adapting deep network features to capture psychological representations: An abridged report.
In *IJCAI*, pages 4934–4938.

[139] Peterson, J. C., Abbott, J. T., and Griffiths, T. L. (2018).
Evaluating (and improving) the correspondence between deep neural networks and human representations.
Cognitive science, 42(8):2648–2669.

[140] Piantadosi, S. and Hill, F. (2022).
Meaning without reference in large language models.
In *NeurIPS 2022 Workshop on Neuro Causal and Symbolic AI (nCSI)*.

[141] Pomerantz, J. R. and Portillo, M. C. (2011).
Grouping and emergent features in vision: Toward a theory of basic gestalts.
Journal of Experimental Psychology: Human Perception and Performance, 37(5):1331.

[142] Puebla, G. and Bowers, J. S. (2022).
Can deep convolutional neural networks support relational reasoning in the same-different task?
Journal of Vision, 22(10):11–11.

[143] Quine, W. V. (1951).
Main trends in recent philosophy: Two dogmas of empiricism.
The Philosophical Review, 60(1):20–43.

[144] Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017).
Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[145] Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. (2018).

Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.
Journal of Neuroscience, 38(33):7255–7269.

[146] Ramírez, F. M. (2017).
Representational confusion: the plausible consequence of demeaning your data.
bioRxiv.

[147] Ramírez, F. M. (2018).
Orientation encoding and viewpoint invariance in face recognition: Inferring neural properties from large-scale signals.
The Neuroscientist, 24(6):582–608.
PMID: 29855217.

[148] Rauber, J., Brendel, W., and Bethge, M. (2017).
Foolbox: A python toolbox to benchmark the robustness of machine learning models.
arXiv preprint arXiv:1707.04131.

[149] Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019).
A deep learning framework for neuroscience.
Nature neuroscience, 22(11):1761–1770.

[150] Ritchie, J. B., Kaplan, D. M., and Klein, C. (2019).
Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience.
The British Journal for the Philosophy of Science, 70(2):581–607.

[151] Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. (2017).
Cognitive psychology for deep neural networks: A shape bias case study.
In *International conference on machine learning*, pages 2940–2949. PMLR.

[152] Rosenfeld, A., Solbach, M. D., and Tsotsos, J. K. (2018).
Totally looks like-how humans compare, compared to machines.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1961–1964.

[153] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986).
Learning representations by back-propagating errors.
Nature, 323(6088):533–536.

[154] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015).

ImageNet Large Scale Visual Recognition Challenge.
International Journal of Computer Vision (IJCV), 115(3):211–252.

[155] Sabour, S., Frosst, N., and Hinton, G. E. (2017).
Dynamic routing between capsules.
In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3859–3869, Red Hook, NY, USA. Curran Associates Inc.

[156] Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996).
Statistical learning by 8-month-old infants.
Science, 274(5294):1926–1928.

[157] Sahami, M. and Heilman, T. D. (2006).
A web-based kernel function for measuring the similarity of short text snippets.
In *Proceedings of the 15th International Conference on World Wide Web*, WWW '06, page 377–386, New York, NY, USA. Association for Computing Machinery.

[158] Samuelson, L. K. (2002).
Statistical regularities in vocabulary guide language acquisition in connectionist models and 15-20-month-olds.
Developmental psychology, 38(6):1016.

[159] Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and DiCarlo, J. J. (2018).
Brain-score: Which artificial neural network for object recognition is most brain-like?
bioRxiv preprint: 407007.

[160] Schuck, N. W., Gaschler, R., Wenke, D., Heinzle, J., Frensch, P. A., Haynes, J.-D., and Reverberi, C. (2015).
Medial prefrontal cortex predicts internally driven strategy shifts.
Neuron, 86(1):331–340.

[161] Schölkopf, B. and Smola, A. J.and Bach, F. (2002).
Learning with kernels: support vector machines, regularization, optimization, and beyond.
MIT Press.

[162] Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J.-M., Bosch, S., and van Gerven, M. (2018).
Convolutional neural network-based encoding and decoding of visual object recognition in space and time.
NeuroImage, 180:253–266.
New advances in encoding and decoding of brain signals.

[163] Serre, T. (2019).
Deep learning: The good, the bad, and the ugly.
Annual Review of Vision Science, 5(1):399–426.

[164] Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020).
The pitfalls of simplicity bias in neural networks.
arXiv preprint arXiv:2006.07710.

[165] Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. (2016).
Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition.
In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540.

[166] Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., and Botvinick, M. M. (2017).
Toward a rational and mechanistic account of mental effort.
Annual review of neuroscience, 40:99–124.

[167] Simon, H. (1999).
Neural networks: a comprehensive foundation.
Prentice hall.

[168] Simonyan, K. and Zisserman, A. (2014).
Very deep convolutional networks for large-scale image recognition.
arXiv preprint arXiv:1409.1556.

[169] Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., and Samuelson, L. (2002).
Object name learning provides on-the-job training for attention.
Psychological science, 13(1):13–19.

[170] Stankiewicz, B. J. and Hummel, J. E. (1996).
Categorical relations in shape perception.
Spatial Vision, 10(3):201–236.

[171] Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., and Kriegeskorte, N. (2021).
Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting.
Journal of Cognitive Neuroscience, 33(10):2044–2064.

[172] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013).
Intriguing properties of neural networks.
arXiv preprint arXiv:1312.6199.

[173] Tacchetti, A., Isik, L., and Poggio, T. A. (2018).
Invariant recognition shapes neural representations of visual input.
Annual review of vision science, 4:403–422.

[174] Torralba, A. and Efros, A. A. (2011).
Unbiased look at dataset bias.
In *CVPR 2011*, pages 1521–1528.

[175] Tsvetkov, C., Malhotra, G., Evans, B. D., and Bowers, J. S. (2020).
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*.

[176] VanRullen, R. (2017).
Perception science in the age of deep neural networks.
Frontiers in psychology, 8:142.

[177] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is all you need.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

[179] Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., and Bethge, M. (2017).
Methods and measurements to compare men against machines.
Electronic Imaging, 2017(14):36–45.

[180] Xu, Y. and Vaziri-Pashkam, M. (2021).
Limits to visual representational correspondence between convolutional neural networks and the human brain.
Nature Communications, 12:2065.

[181] Yamins, D. L. and DiCarlo, J. J. (2016).
Using goal-driven deep learning models to understand sensory cortex.
Nature neuroscience, 19(3):356–365.

[182] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014).
Performance-optimized hierarchical models predict neural responses in higher visual cortex.
Proceedings of the National Academy of Sciences, 111(23):8619–8624.

[183] Yee, M., Jones, S., and Smith, L. (2012).
Changes in visual object recognition precede the shape bias in early noun learning.
Frontiers in Psychology, 3.

[184] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014).
How transferable are features in deep neural networks?
arXiv preprint arXiv:1411.1792.

[185] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016).
Understanding deep learning requires rethinking generalization.
arXiv preprint arXiv:1611.03530.

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

[187] Zhou, Z. and Firestone, C. (2019).
Humans can decipher adversarial images.
Nature communications, 10(1):1334.

[188] Zhu, H., Tang, P., Park, J., Park, S., and Yuille, A. (2019).
Robustness of object recognition under extreme occlusion in humans and computational models.
arXiv preprint arXiv:1905.04598.

[189] Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. K. (2021).
Unsupervised neural network models of the ventral visual stream.
Proceedings of the National Academy of Sciences, 118(3):e2014196118.