Deep learning models challenge the prevailing assumption that face-like effects for objects of expertise support domain-general mechanisms

Galit Yovel[1,2], Idan Grosbard[1,2] & Naphtali Abudraham[1]

[1]School of Psychological Sciences

[2]Sagol School of Neuroscience

Tel Aviv University, Tel Aviv, Israel

Running head: Domain-specific inversion effect for object of expertise

Keywords: Expertise, Face Recognition, Computational modelling, Level of Categorization

Abstract

The question of whether task performance is best achieved by domain-specific, or domain-general processing mechanisms is prevalent in both artificial and biological systems. This question has generated a fierce debate in the study of expert object recognition. Because humans are experts in face recognition, face-like neural and cognitive effects for objects of expertise were considered support for domain-general mechanisms. However, effects of domain, experience, and level of categorization, are confounded in human studies, which may lead to erroneous inferences. To overcome these limitations, we trained deep learning algorithms on different domains (objects, faces, birds) and levels of categorization (basic, sub-ordinate, individual), matched for amount of experience. Like humans, the models generated a larger inversion effect for faces than for objects. Importantly, a face-like inversion effect was found for individual-based categorization of non-faces (birds) but only in a network specialized for that domain. Thus, contrary to prevalent assumptions, face-like effects for objects of expertise do not support domain-general mechanisms but may originate from domain-specific mechanisms. More generally, we show how deep learning algorithms can be used to dissociate factors that are inherently confounded in the natural environment of biological organisms to test hypotheses about their isolated contributions to cognition and behavior.

Keywords: Perceptual expertise, Deep learning, Face recognition, Categorization

1. Introduction

Perceptual expertise is an acquired skill to classify members of a homogenous category at the subordinate- or individual-level of categorization [1]–[4]. Faces are the only category for which most humans are experts. Humans' superb ability to classify faces of different individuals is acquired thorough the extensive social and perceptual experience that humans have with people. This ability for individual face recognition has also been shown in several non-human species [5]. Thus, a major question that has been hotly debated for over three and a half decades [6], is whether perceptual expertise is mediated by domain-specific or domain-general mechanisms [7]–[9]. According to the domain-specific hypothesis, expertise for faces is mediated by face-specific mechanisms that are not used for other domains, such as birds or cars [8], [10]–[13]. Conversely, according to the general-expertise hypothesis, classification of objects of expertise, including faces, is mediated by a general processing mechanism for subordinate or individual-level classification [14]–[16].

To decide between these two hypotheses, previous studies examined whether real-life or lab-based experts show the same behavioral and neural face-selective markers for their objects of expertise [11], [17], [18]. One well-established, face-selective marker is the face inversion effect, which was the focus of the first study that initiated the debate [6] and many other studies that followed [10], [19], [20][14]. The face inversion effect refers to the large drop in performance for inverted than upright stimuli in recognition or perceptual matching tasks. This effect is larger for faces than for any non-face objects [21]–[26], which led to the suggestion that faces are special [12], [13], [21]. Diamond and Carey (1986) were the first to report a face-sized inversion effect for dogs in dog experts, suggesting that the inversion effect is not a face-specific effect, but is found for all objects of expertise

that are individuated based on second order configurations. Later studies that examined the inversion effect with other objects of expertise such as birds or cars, in experts that could classify them at the sub-ordinate level of categorization, revealed mixed findings with most of them still reporting a larger inversion effect for faces than objects of expertise [10], [22], [27]–[29]. A recent study that did replicate Carey and Diamond's findings revealed a face-sized inversion effect in bird experts who can identify individual birds of one specific bird species [19]. This study concluded that faces and objects of expertise are processed by the same mechanism.

Here we propose that conclusions made by human studies of perceptual expertise are based on presumptions that are hard to evaluate in humans and may lead to the erroneous conclusions. First, a larger inversion effect for faces than non-face stimuli may not necessarily support a domain-specific account but may reflect the much greater experience humans have with faces than any other objects of expertise. Even extensive real-world expertise with dogs, birds or cars does not start on the first year of life, as does humans' expertise with faces. Recent studies indicate that during the first year of life, infant spend 25% of waking time looking at foveated, frontal faces [30], [31]. This gap is even greater when expertise is acquired in the lab for novel objects [2], [32]. Second, previous studies of perceptual expertise considered similar inversion effects for faces and objects of expertise as evidence for general-expert processing mechanisms [6], [7], [19], [20], [25]. This prevalent assumption was accepted by both sides of the debate. However, similar inversion effects for dogs, birds and faces may still originate from distinct expert systems, each specializes for its own domain, rather than a single, general-expert processing mechanism for all objects of expertise. These alternative accounts cannot be tested in

human experts, where effects of experience, domain, and level of categorization are often confounded.

In the current study, we used deep convolutional neural networks as computational models of perceptual expertise. Deep convolutional neural networks (DCNNs) are brain-inspired algorithms that reach human-level performance and generate human-like representations for objects and faces [33]–[38]. These models can be trained to classify images from different domains at different levels of categorization. This provides us with an unprecedented opportunity to directly compare between different expert and non-expert systems that are matched for the amount of training, in a way that cannot be achieved in humans. Accordingly, in the current experiment we compared the magnitude of the inversion effect in DCNNs that were trained to classify images at different levels of categorization including: objects (basic level), bird-species (subordinate level), faces (individual level) and individual birds (individual level) (Figure 1). Bird species [39]–[42] and individual birds [19] were both used to study perceptual expertise in humans. These models enable us to measure the magnitude of the inversion effect for different domains and levels of categorization in networks that were trained on the same number of classes/images across different domains. Importantly, we can examine whether similar inversion effects may originate from distinct,. z specialized, domain-specific mechanisms rather than a single general expert processing mechanism, as was presumed by all previous human studies of objects of expertise [6]–[8], [10], [19].

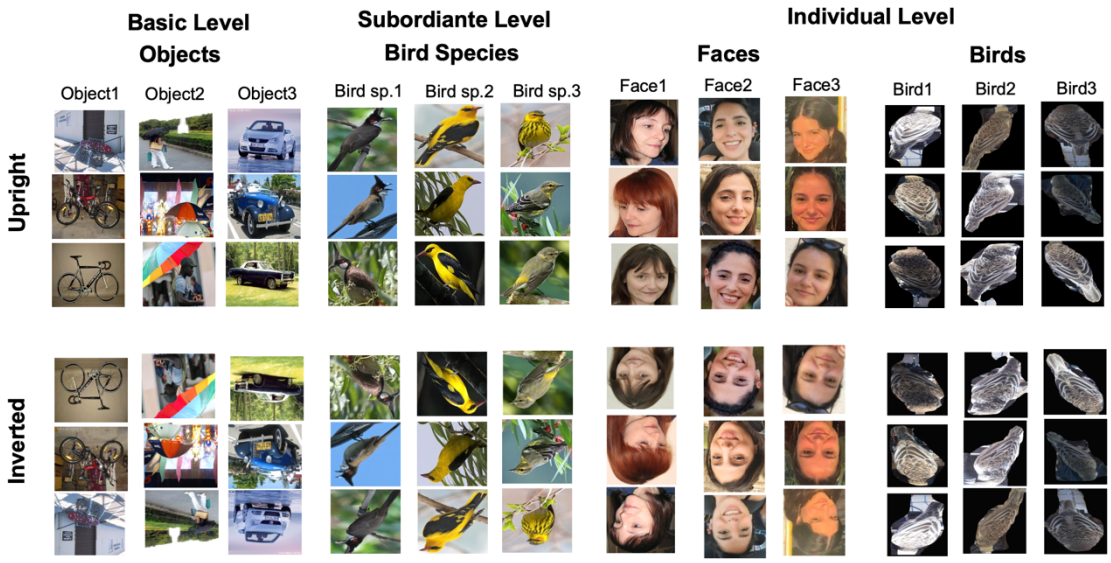

Figure 1: Examples of upright and inverted images for objects that were classified at the basic-level of categorization, bird species that were classified at the subordinate level of categorization and faces and birds that were classified at the individual-level categorization.

2. Methods

2.1 Stimuli: Training sets

We used four different data sets (Figure 1): Objects that are labeled at the basic-level of categorization (e.g., cars, chairs, bikes); Bird species that are labeled at the subordinate-level of categorization. Faces that are labeled at the individual-level of categorization and individual birds that are labeled at the individual level of categorization. To match the training of the different categories we used the number of classes of the category with the least number of classes. Because the bird species dataset included only 260 classes and the individual bird dataset included only 30 classes, we first compared the bird species DCNN to face and object DCNNs that are trained on 260 classes each. In this analysis, we used the maximum number of images per class in the bird species, which was still

smaller than the object and face datasets. We then ran a second analysis in which we added a DCNN trained on the 30 classes of individual birds and compared its performance to object, bird species and face DCNNs that were trained on 30 classes. In this training, the categories were fully matched for the number of images per class. The same stimulus categories were used to measure the inversion effect, with images that were not included in the training set. In a supplemental material we also report a face and object networks that were trained on 1000 classes and 300 images per class of faces and objects, for which we had enough images in the datasets to train from scratch on the same number of classes and images per class.

2.2 Training protocols

Because of the small number of classes in the tested domains, to create the different DCNNs we took a pre-trained object classification DCNN as a base-DCNN, and fine-tuned it from layer Conv1 and up (i.e. fine-tuning the whole network) to classify stimuli in the different domains. The object-trained base-DCNN was a randomly initialized VGG-16 DCNNs [43], that we trained to classify 500 classes of inanimate objects (basic-level categorization) from the ImageNet dataset [44], each class consisted of 300 training images, and 50 validation images. We then fine-tuned it from layer Conv1 and up (i.e. fine-tuning the whole network) to classify stimuli from the different categories at different levels of categorization.

To assure that fine-tuning a base-object model generates similar findings to a model trained from scratch, in supplementary material we report analysis for a model trained from scratch on 1000 classes and 300 images per class of objects and faces, for which we have enough images to train a network from scratch and examined their performance to upright and inverted faces and objects.

2.3 Fine-tuning protocol:

Fine-tuning was done using cross-entropy loss. The networks were optimized using Stochastic Gradient Descent with a learning rate of 0.01 and with PyTorch default parameters [45] for 60 full epochs, with batch size of 128. After 50 epochs the learning rate was reduced to 1e-3. Training images were normalized using the pixel means and standard-deviations of each training dataset, and we used the same image augmentations that were used for the pre-trained object DCNN, with an addition of random rotations of up to +/- 40 degrees (which was found to improve performance in [46]).

DCNNs trained on 260 classes of Objects, Bird Species & Faces

To create an object-DCNN we took the base-DCNN (pre-trained on 500 inanimate objects) and fine-tuned it to classify 260 inanimate objects, randomly selected from the training set of the base-DCNN. Each class included 300 images per class. To create a face-DCNN, we fine-tuned the base-DCNN on 260 face identities randomly selected from the VGGFace2 dataset. Each class included 300 images per class. For the bird-species DCNN, we fine-tuned the base-DCNN to classify 260 classes of bird-species. We used all the images that were included in each class of bird species: average 151 images per class (range: 105-310 images).

2.4 DCNNs trained on 30 classes of Objects, Bird Species, Faces & Individual birds

We used the same procedure to fine-tune the base-DCNN to create four DCNNs. The number of classes and number of images per class was fully matched across the different categories: an object-DCNN trained on 30 inanimate objects with 100 images per class, randomly selected from the 260 classes used above with 100 images per class, a face-DCNN trained on 30 face identities with 100 images per class, a bird-species-DCNN

trained on the 30 classes with 100 images per class. The individual-bird-DCNN was created by fine-tuning the base-DCNN on 30 classes of individual birds.

2.5 Measuring performance for upright and inverted images:

To test performance on the same/different classification task, we randomly selected 50 distinct image pairs from the validation images of each of the 30 classes, that were not included in the training set, making 1500 (30x50) same-class pairs and 1500 distinct different-class pairs. We randomly divided the 1500 same-class image pairs, and the 1500 different-class image pairs, to 30 batches, each with 50 same identity and 50 different identity pairs. Then, for each batch, we measured the cosine similarity between the embeddings of the images in the penultimate layer for each image-pair, and calculated based on these similarity scores the Receiver Operator Curves (ROCs) and Areas Under Curves (AUCs) for true and false same-class classification. The Python scikit-learn package was used to calculate ROC and AUC. We repeated these calculations for each one of the 30 batches and calculated the mean and std of AUCs. This procedure was used to measure performance for upright and inverted images in each of the fine-tuned networks.

3. Results

Figure 2 shows the ROCs for upright (red) and inverted (blue) images for different types of images (columns) for the object, face and bird species DCNNs (rows), which were each trained on 260 classes, based on the maximal number of classes of bird species (see Methods). The descriptive statistics of the AUCs are reported in supplementary Table 1. A mixed ANOVA with DCNN domain (Objects, Faces, Bird Species) as a between factor and image domain (Objects, Faces, Bird Species) and orientation (Upright, Inverted) as

within factors revealed a main effect of orientation ($F(1,87) = 397.18$, $p < .001$, $\eta^2_p = 0.82$) indicating better performance for upright than inverted stimuli. An interaction between orientation, image domain and DCNN domain ($F(2,174) = 246.39$, $p < .001$, $\eta^2_p = 0.85$) indicates that the inversion effect was found for each stimulus domain only in the network that was trained for that domain. That is, performance was better for upright than inverted inanimate objects only in the object network, for bird species only in the bird species network and for faces only in the face network. Furthermore, the inversion effect for faces in the face network ($t(29) = 43.02$, $p < .001$, Cohen's $d = 5.94$) was much larger than the inversion effect for objects in the object network ($t(29) = 8.30$, $p < .001$, Cohen's $d = 2.53$) and for bird species in the bird species network ($t(29) = 7.13$, $p < .001$, Cohen's $d = 2.67$) (See Figure 2 diagonal & Figure 4A).

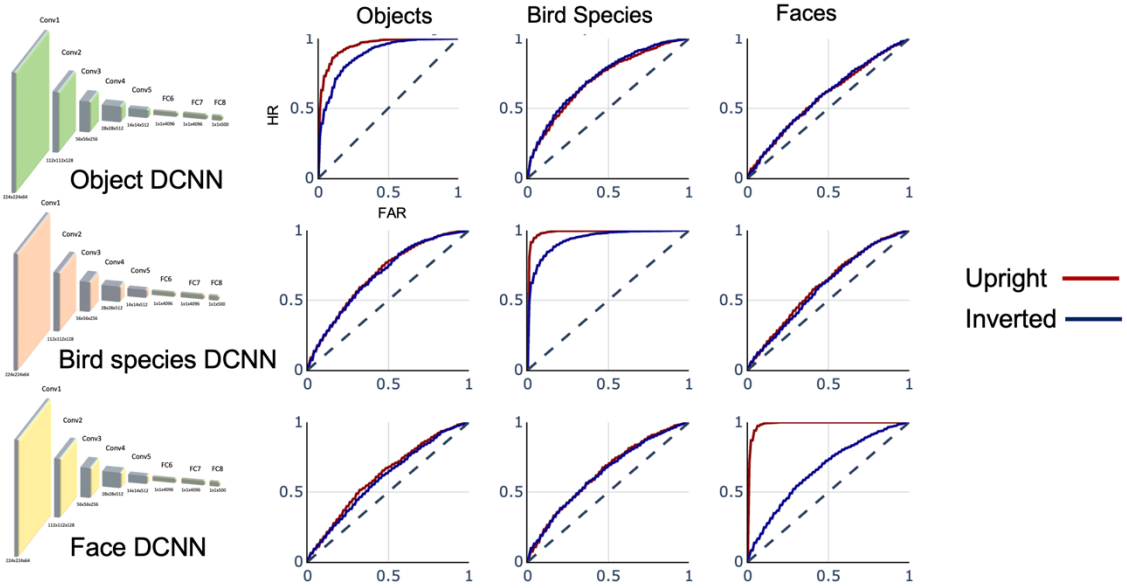

Figure 2: Performance (ROC) for upright (red) and inverted (blue) images from different domains (columns) across DCNNs optimized to classify 260 classes of objects (basic level), birds (sub-ordinate level) and faces (individual level) (rows). An inversion effect is found for each domain only in DCNNs optimized for that domain (diagonal). The inversion effect for faces is much larger than for objects and bird species.

Figure 3 shows the ROCs for upright and inverted stimuli for different types of stimuli (columns) for the four DCNNs that were trained on 30 classes (rows), which was the maximal number of classes of individual birds in the database. This enabled us to compare the face inversion effect with another category that is classified at the individual level of categorization – individual birds. Results show an inversion effect for faces in the face network and an inversion effect for individual birds in the individual bird network, which were larger than the inversion effect for objects in the object networks and for bird species in the bird species network (see Figure 3). The descriptive statistics is reported in supplementary Table 2. A mixed ANOVA with DCNN domain as a between factor (Objects, Faces, Bird Species, Bird individuals), image domain (Objects, Faces, Bird Species, Bird individuals) and orientation (Upright, Inverted) as within factors revealed a main effect of orientation ($F(1,116) = 634,72$, $p < .001$, $\eta^2_p = 0.85$) indicating better performance for upright than inverted stimuli. An interaction between orientation, image domain and DCNN domain ($F(1,348) = 234.06$ $p < .001$, $\eta^2_p = 0.86$) indicated that the inversion effect was found for each stimulus domain only in the network that was specialized for that domain. Post hoc analysis revealed that the inversion effect was much larger for faces in the face network ($t(29) = 24.95$, $p < .001$, Cohen's $d = 7.34$) and individual birds in the individual bird network ($t(29) = 17.13$, $p < .001$, Cohen's $d = 4.79$), and smaller for objects in the object network ($t(29) = 2.57$, $p = .07$, Cohen's $d = 1.84$) and for bird species in the bird species network ($t(29) = 5.88$, $p < .001$, Cohen's $d = 2.93$).

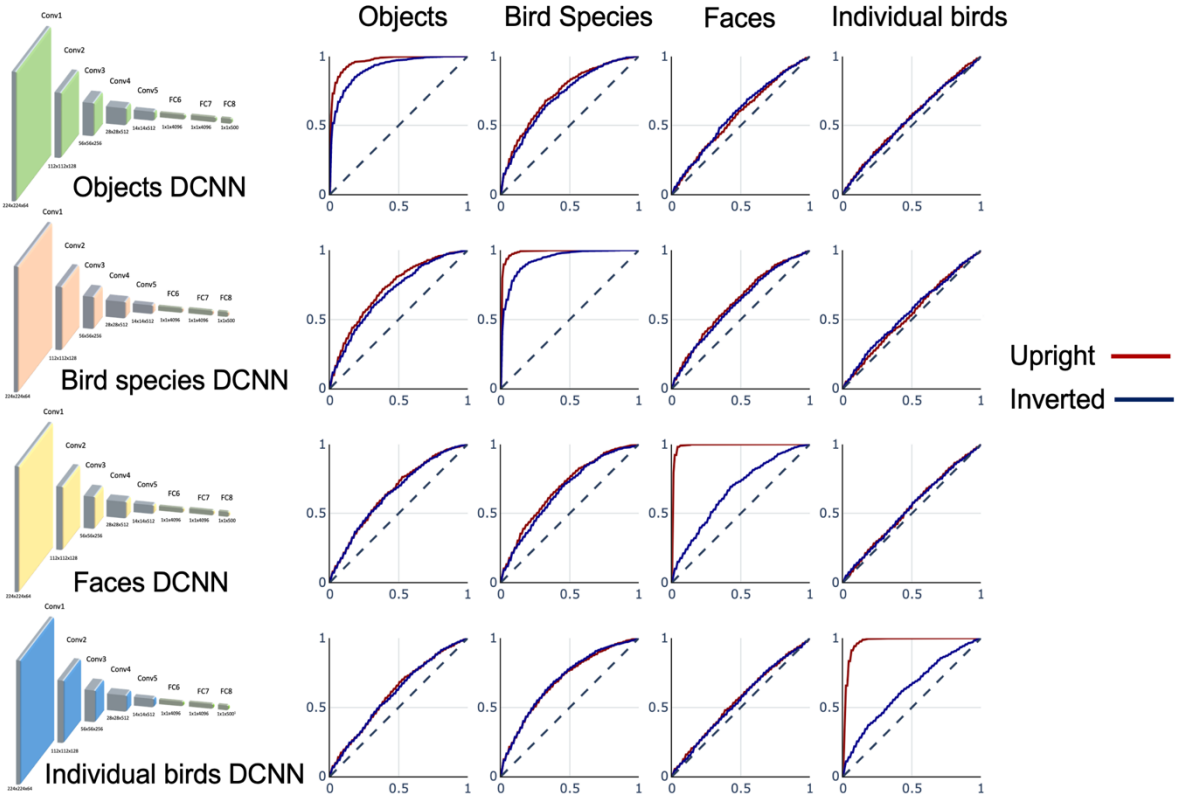

Figure 3: Performance (ROC) for upright (blue) and inverted (red) stimuli across DCNNs optimized to classify 30 classes of objects, bird species, faces and individual birds at different levels of classification. An inversion effect is found for each domain only in DCNNs optimized for that domain. An inversion effect is found for each domain only in DCNNs optimized for that domain (diagonal). The inversion effects for faces and individual birds are much larger than for objects and bird species.

Figure 4 displays performance for upright and inverted stimuli of each stimulus domain only in the network that was trained for that domain. We first examined the networks that were trained on 260 classes of objects, faces and bird species (Figure 4A). An ANOVA with image domain as a between factor and orientation as a within factor revealed a significant effect of orientation ($F(1,87) = 1460.38$ $p < .001$, $\eta^2_p = 0.94$) and a significant interaction of image domain and orientation ($F(2,87) = 533.44$, $p < .001$, $\eta^2_p = 0.93$), indicating a much larger inversion effect for faces than for bird species and for objects. We confirmed this observation with an interaction for each pair of image domains. The orientation by image domain interaction was significant for faces and objects $F(1,58) =$

1252.53, $p < .001$, $\eta^2_p = .96$ and for faces and bird species $F(1,58) = 1268.40$, $p < .001$, $\eta^2_p = .96$, but not for bird species and objects, which showed a significant inversion effect ($F(1,58) = 401.35$, $p < .001$, $\eta^2_p = .87$), but no significant interaction of domain and orientation ($F(1,58) = 2.29$, $p = .13$, $\eta^2_p = .04$).

We then performed a similar analysis for the four image domains including the individual birds (Figure 4B). An ANOVA with image domain (Objects, Faces, Bird Species, Individual birds) as a between factor and orientation as a within factor revealed a significant effect of orientation ($F(1,116) = 2293.22$ $p< .001$, $\eta^2_p = 0.95$) and a significant interaction of image domain and orientation ($F(3,116) = 460.19$, $p< .001$, $\eta^2_p = 0.92$, indicating a much larger inversion effect for faces and individual birds than for bird species and objects. A similar ANOVA which included only faces and individual birds revealed a main effect of orientation ($F(1,58) = 1977.8$, $p< .001$, $\eta^2_p = 0.97$) but no significant interaction between orientation and stimulus domain ($F(1,58) = 1.39$, $p= .24$, $\eta^2_p = 0.02$). An ANOVA which included only objects and bird species revealed a main effect of orientation ($F(1,58) = 317.59$, $p < .001$, $\eta^2_p = 0.85$) and no significant interaction between orientation and stimulus domain ($F(1,58) = 3.29$, $p= .08$, $\eta^2_p = 0.05$), replicating results we found in the networks that were trained on 260 classes. In supplementary results we report similar findings of a larger inversion effect for faces than objects in a face-trained and an object-trained DCNN, respectively that were trained from scratch on these categories. Thus, our findings are not limited to object DCNNs that are fine-tuned for these different categories.

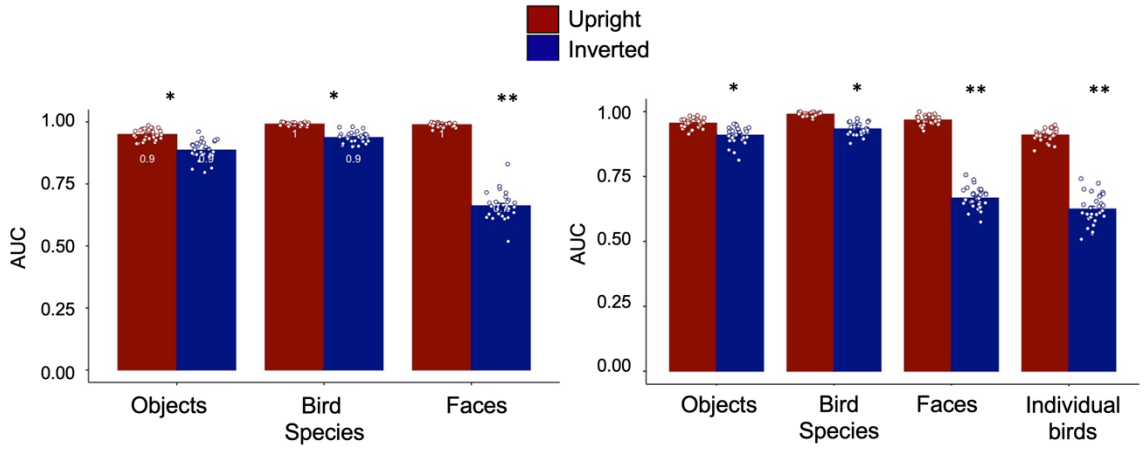

Figure 4: Performance (AUC) for upright and inverted stimuli in DCNNs that are optimized for their domain. The inversion effect is substantially larger for faces and birds that are classified at the individual level of categorization than bird species and objects. A: DCNNs that were trained on 260 classes (diagonal of Fig. 2). B: DCNNs that were trained on 30 classes (diagonal of Fig. 3). * < .01, ** < .001

Taken together, results of both sets of networks trained on different number of classes between experiments (30, 260 or 1000), show a significantly larger inversion effect for networks that are optimized for classification at the individual level than the basic or subordinate level of categorization, but only for the images that the network was optimized to classify. This pattern of results were similar across different sizes of training sets, which indicates that they are found for different amounts of experience.

4. Discussion

The purpose of the present study was to reevaluate the theoretical assumptions of the domain-specific and domain-general hypotheses of expert object recognition. Because domain, experience and level of categorization cannot be dissociated in humans, we designed computational models of perceptual expertise, by training DCNNs to classify images at different levels of categorization, matched for amount of experience. Our findings reveal that, similar to humans, DCNNs showed a larger inversion effect for faces that are classified at the individual-level of categorization than objects that are classified

at the basic-level of categorization. Interestingly, we found a face-like inversion effect for individual-level classification of birds. Importantly, this face-like inversion effect for individual birds was found only in a system that was trained for classification of individual birds. We therefore conclude that face-like effects in objects of expertise may not necessarily reflect a general-expert processing mechanism, as was presumed in all previous human studies of perceptual expertise. Instead, they may originate from separate distinct systems that are specialized for their own domain, consistent with a domain-specific account of perceptual expertise.

Results of our computational models are consistent with human studies that revealed a face-sized inversion effects for individual dogs in dog experts [6],but see [10] and for individual birds in bird experts [19]. Most other studies of perceptual expertise examined expertise at the subordinate-level of categorization (e.g., car types, bird species) and similar to our models, revealed an inversion effect that was smaller for objects of expertise than faces [10], [22], [27]–[29]. One inherent difference between individual-level and sub-ordinate level of categorization is the degree of variance between classes, which is much smaller in the former. We therefore suggest that a disproportionally large inversion effect reflects poor generalization to untrained inverted images in a system that is highly specialized for fine-grained discrimination of the upright orientation of a specific domain. This suggestion is in line with recent findings that show that individual or subordinate-level classification required deeper retraining of additional layers in a face than an object DCNN to reach similar levels of performance [47], [48]. Thus, expertise in one domain does not transfer well to other domains.

It is noteworthy that the amount of experience (i., number of classes/images per category in the training sets) did not influence the magnitude of the inversion effect. We found the

same pattern of findings for models that were trained to classify 260 classes or only 30 classes (see Figure 4 and supplementary data for 1000 classes). This finding is important for the interpretation of results of human studies of perceptual expertise where experience with objects of expertise can be never matched with faces. It indicates that the smaller inversion effect for objects of expertise than faces that were reported in previous studies may not be due to the greater experience that human experts have with faces than any other object of expertise. Thus, it is the level of categorization rather than the amount of experience that determines the magnitude of the inversion effect.

The expert processing mechanisms proposed by our computational models may in fact offer a possible resolution to the debate between the domain-specific and domain-general accounts of perceptual expertise. Consistent with the domain-specific account, we propose that the fine-grained discrimination that is required for classification of the within-category perceptually similar images is best achieved by a domain-specific mechanism that is optimized for its specific critical features[49]. These domain-specific mechanisms are over-fitted to a certain category and therefore show poor generalization to untrained images. This overfitting may generate similar patterns of behavior in different expert systems, such as a disproportionally large inversion effect. However, these similar patterns of behavior may not necessarily reflect similar types of processing (e.g. configural processing), but the outcome of specialization to within-category discrimination which is similarly required for different experts system.

In the seminal study that first proposed the domain-general expertise hypothesis, Carey and Diamond suggested that because within-category members share first order configuration, objects of expertise can be discriminated only by their second order configuration (i.e., distance between features)[6]. Because configural processing is found

for upright but not inverted faces [24], [50], this led many studies to treat face processing, the face inversion effect and configural processing almost as synonyms [6], [16], [19], [51]. Later studies have questioned this prevalent assumption [52] showing that a face inversion effect is found also for faces that differ in face parts not only face configuration [53], [54], that face recognition is intact despite configural distortions [55], [56] but impaired for non-configural distortions [57], that effects of configural processing are not correlated with face recognition [58] and that distance between facial features are not useful for generalization across head-views of the same identity [59]. Our findings also suggest that perceptual expertise may not necessarily rely on configural processing. Inspection of the individual bird images (see Figure 1) indicates that even stimuli that do not have a clear first or second order configuration of their parts may generate a face-sized inversion effect. Thus, perceptual expertise and inversion effects may simply reflect fine-grained discrimination between perceptually similar classes of stimuli.

Unlike behavioral studies that cannot indicate whether an inversion effect or a holistic effect, originates from a face-selective or other domain-selective mechanisms, neuroimaging studies do enable to separately examine the response of face-selective mechanisms to objects of expertise. Whereas many studies of objects of expertise only focused on the response of the FFA [60], other studies examined the response of additional areas within and outside the occipital temporal lobe [17], [61]–[63]. These studies suggest that attentional effects may account for increased response to objects of expertise in multiple areas including the FFA. The current findings are consistent with studies that show that expertise for non-face objects is mediated by brain regions in the occipital-temporal cortex that do not overlap with face-selective regions [62]. Such a brain region was recently reported in Pokemon experts [63]. The word-form area [64] may be

another example for such a dedicated processing mechanism for a specific domain. Accordingly, we predict that human expertise for individual bird stimuli will be mediated by dedicated a category-selective brain region for individual birds.

Our models were specifically trained on different domains at different levels of categorization, which is different from the biological system that is exposed to all categories. This enabled us to dissociate effects of domain and experience that are inherently confounded in humans, but generated an artificial system that cannot be directly compared to humans. Importantly, however, a recent study that trained a DCNN on both faces and objects, revealed that the dual-trained network spontaneously segregated to face and object units each specialized for its own domain [65]. Despite dissimilarities between the architecture and computations of the artificial and biological systems, the generation of human-like behavior in these DCNNs provides us with a powerful tool to ask about the origins and possible mechanisms of this behavior in ways that cannot be tested in humans [66], [67].

In summary, the present study used computational models of expert and non-expert recognition to reevaluate long-standing assumptions of previous studies that tested the general and domain-specific accounts of perceptual expertise. Notably, computational models cannot provide direct evidence for the type of operation of a biological/cognitive system, which may still use a general expert processing mechanism for objects of expertise. Importantly, our study does show the computational plausibility of an alternative account, which was not considered by the many previous human studies that compared the magnitude of face-like effects for objects of expertise. This approach can be used to dissociate the effects of factors that are inherently confounded in the natural environment

of biological organisms to reevaluate prevalent assumptions on animal cognition and behavior.

Acknowledgment

We would like to thank Noam Avidor, Amit Bardos, Koby Boyango and Danielle Chason who performed initial data-base searches and training of DCNNs for bird species in earlier stages of this research as part of their undergraduate research seminar. This research was supported by a grant from the Israeli Science Foundation (ISF 917/21) to GY.

Conceptualization – GY and NA; Methodology – GY, NA, IG; Formal Analysis – IG; Investigation – GY, IG, NA; Writing – Original Draft Preparation – GY; Writing – Review & Editing – GY,NA, IG

References

[1] A. Harel, D. Kravitz, C. I. Baker, G. Campitelli, E. Cowan University, and A. Elinor McKone, “Beyond perceptual expertise: revisiting the neural substrates of expert object recognition WHAT IS EXPERTISE AND WHY IS IT IMPORTANT TO STUDY IT?,” *Front. Hum. Neurosci.*, 2013, doi: 10.3389/fnhum.2013.00885.

[2] I. Gauthier and M. J. Tarr, “Becoming a ‘Greeble’ expert: Exploring mechanisms for face recognition,” *Vision Res.*, vol. 37, no. 12, pp. 1673–1682, 1997, doi: 10.1016/S0042-6989(96)00286-6.

[3] J. W. Tanaka, “The entry point of face recognition : Evidence for face expertise,” *Gen. Exp. Psychol.*, vol. 130, no. 3, pp. 534–543, 2001, doi: 10.1037/0096-3445.130.3.534.

[4] J. W. Tanaka and M. Taylor, “Object categories and expertise: Is the basic level in the eye of the beholder?,” *Cogn. Psychol.*, vol. 23, no. 3, pp. 457–482, 1991, doi: 10.1016/0010-0285(91)90016-H.

[5] M. Behrmann and G. Avidan, “Face perception: computational insights from phylogeny,” *Trends Cogn. Sci.*, vol. 26, no. 4, pp. 350–363, 2022, doi: 10.1016/j.tics.2022.01.006.

[6] R. Diamond and S. Carey, “Why Faces Are and Are Not Special. An Effect of Expertise,” *J. Exp. Psychol. Gen.*, vol. 115, no. 2, pp. 107–117, 1986, doi: 10.1037/0096-3445.115.2.107.

[7] I. Gauthier and C. Bukach, “Should we reject the expertise hypothesis?,” *Cognition*, vol. 103, no. 2, pp. 322–330, 2007, doi: 10.1016/j.cognition.2006.05.003.

[8] E. McKone, N. Kanwisher, and B. C. Duchaine, “Can generic expertise explain special processing for faces?,” *Trends Cogn. Sci.*, vol. 11, no. 1, pp. 8–15, 2007, doi: 10.1016/j.tics.2006.11.002.

[9] A. Harel, D. Kravitz, and C. I. Baker, “Beyond perceptual expertise: Revisiting the neural substrates of expert object recognition,” *Front. Hum. Neurosci.*, vol. 7, no. DEC, pp. 1–12, 2013, doi: 10.3389/fnhum.2013.00885.

[10] R. Robbins and E. McKone, “No face-like processing for objects-of-expertise in three behavioural tasks,” *Cognition*, vol. 103, no. 1, pp. 34–79, 2007, doi: 10.1016/j.cognition.2006.02.008.

[11] K. Grill-Spector, N. Knouf, and N. Kanwisher, “The fusiform face area subserves

face perception, not generic within-category identification," *Nat. Neurosci.*, 2004, doi: 10.1038/nn1224.

[12] N. Kanwisher, "Domain specificity in face perception," *Nat. Neurosci.*, vol. 3, no. 8, pp. 759–763, 2000, doi: 10.1038/77664.

[13] N. Kanwisher and G. Yovel, "The fusiform face area: A cortical region specialized for the perception of faces," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 361, no. 1476, 2006, doi: 10.1098/rstb.2006.1934.

[14] M. J. Tarr and I. Gauthier, "FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise," *Nature Neuroscience*. 2000. doi: 10.1038/77666.

[15] C. M. Bukach, I. Gauthier, and M. J. Tarr, "Beyond faces and modularity: the power of an expertise framework," *Trends Cogn. Sci.*, vol. 10, no. 4, pp. 159–166, 2006, doi: 10.1016/j.tics.2006.02.004.

[16] I. Gauthier and M. J. Tarr, "Becoming a 'Greeble' Expert: Exploring Mechanisms for Face Recognition," *Vision Res.*, vol. 37, no. 12, pp. 1673–1682, Jun. 1997, doi: 10.1016/S0042-6989(96)00286-6.

[17] A. Harel, S. Gilaie-Dotan, R. Malach, and S. Bentin, "Top-Down engagement modulates the neural expressions of visual expertise," *Cereb. Cortex*, vol. 20, no. 10, pp. 2304–2318, 2010, doi: 10.1093/cercor/bhp316.

[18] I. Gauthier, P. Williams, M. J. Tarr, and J. Tanaka, "Training 'greeble' experts: A framework for studying expert object recognition processes," *Vision Res.*, vol. 38, no. 15–16, pp. 2401–2428, 1998, doi: 10.1016/S0042-6989(97)00442-2.

[19] A. Campbell and J. W. Tanaka, “Inversion Impairs Expert Budgerigar Identity Recognition: A Face-Like Effect for a Nonface Object of Expertise,” *Perception*, vol. 47, no. 6, pp. 647–659, 2018, doi: 10.1177/0301006618771806.

[20] B. Rossion and T. Curran, “Visual expertise with pictures of cars correlates with rt magnitude of the car inversion effect,” *Perception*, vol. 39, no. 2, pp. 173–183, 2010, doi: 10.1068/p6270.

[21] R. K. Yin, “Looking at upside-down faces.,” *J. Exp. Psychol.*, vol. 81, no. 1, p. 141, 1969.

[22] N. Weiss, E. Mardo, and G. Avidan, “Visual expertise for horses in a case of congenital prosopagnosia,” *Neuropsychologia*, vol. 83, pp. 63–75, 2016, doi: 10.1016/j.neuropsychologia.2015.07.028.

[23] G. Yovel, T. Pelc, and I. Lubetzky, “It’s all in your head: Why is the body inversion effect abolished for headless bodies?,” *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 36, no. 3, 2010, doi: 10.1037/a0017451.

[24] M. J. Farah, J. W. Tanaka, and H. M. Drain, “What Causes the Face Inversion Effect?,” *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 21, no. 3, pp. 628–634, 1995, doi: 10.1037/0096-1523.21.3.628.

[25] J. P. McCleery *et al.*, “The roles of visual expertise and visual input in the face inversion effect: Behavioral and neurocomputational evidence,” *Vision Res.*, vol. 48, no. 5, pp. 703–715, 2008, doi: 10.1016/j.visres.2007.11.025.

[26] G. Yovel and N. Kanwisher, “Face perception: Domain specific, not process specific,” *Neuron*, vol. 44, no. 5, 2004, doi: 10.1016/j.neuron.2004.11.018.

[27] R. Bruyer and G. Crispeels, “Expertise in person recognition,” *Bull. Psychon. Soc.*, vol. 30, no. 6, pp. 501–504, 1992, doi: 10.3758/BF03334112.

[28] B. Rossion, I. Gauthier, V. Goffaux, M. J. Tarr, and M. Crommelinck, “Expertise training with novel objects leads to left-lateralized facelike electrophysiological responses,” *Psychol. Sci.*, vol. 13, no. 3, pp. 250–257, 2002, doi: 10.1111/1467-9280.00446.

[29] T. A. Busey and J. R. Vanderkolk, “Behavioral and electrophysiological evidence for configural processing in fingerprint experts,” *Vision Res.*, vol. 45, no. 4, pp. 431–448, 2005, doi: 10.1016/j.visres.2004.08.021.

[30] S. Jayaraman, C. M. Fausey, and L. B. Smith, “The Faces in Infant-Perspective Scenes Change over the First Year of Life,” pp. 13–15, 2015, doi: 10.1371/journal.pone.0123780.

[31] C. M. Fausey, S. Jayaraman, and L. B. Smith, “From faces to hands: Changing visual input in the first two years,” *Cognition*, vol. 152, pp. 101–107, 2016, doi: 10.1016/j.cognition.2016.03.005.

[32] A. C. N. Wong, T. J. Palmeri, and I. Gauthier, “Conditions for facelike expertise with objects: Becoming a ziggerin expert - but which type?,” *Psychol. Sci.*, vol. 20, no. 9, pp. 1108–1117, 2009, doi: 10.1111/j.1467-9280.2009.02430.x.

[33] N. Abudarham, I. Grosbard, and G. Yovel, “Face Recognition Depends on Specialized Mechanisms Tuned to View-Invariant Facial Features: Insights from Deep Neural Networks Optimized for Face or Object Recognition,” *Cogn. Sci.*, vol. 45, no. 9, 2021, doi: 10.1111/cogs.13031.

[34] K. Dobs, J. Martinez, A. J. E. Kell, and N. Kanwisher, “Brain-like functional specialization emerges spontaneously in deep neural networks,” *Sci. Adv.*, vol. 8, no. 11, p. eabl8913, 2022.

[35] G. Jacob, R. T. Pramod, H. Katti, and S. P. Arun, “Qualitative similarities and differences in visual object representations between brains and deep networks,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–14, 2021, doi: 10.1038/s41467-021-22078-3.

[36] P. J. Phillips *et al.*, “Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms,” *Proc. Natl. Acad. Sci.*, vol. 115, no. 24, pp. 6171–6176, 2018.

[37] F. Tian, H. Xie, Y. Song, S. Hu, and J. Liu, “The Face Inversion Effect in Deep Convolutional Neural Networks,” *Front. Comput. Neurosci.*, vol. 16, no. May, pp. 1–8, 2022, doi: 10.3389/fncom.2022.854218.

[38] K. Dobs, J. Martinez, K. Yuhan, and N. Kanwisher, “Using deep convolutional neural networks to test why human face recognition works the way it does,” *bioRxiv*, pp. 1–26, 2022.

[39] I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson, “Expertise for cars and birds recruits brain areas involved in face recognition,” *Nat. Neurosci.*, 2000, doi: 10.1038/72140.

[40] J. W. Tanaka, T. Curran, and D. L. Sheinberg, “The training and transfer of real-world perceptual expertise,” *Psychol. Sci.*, vol. 16, no. 2, pp. 145–151, 2005, doi: 10.1111/j.0956-7976.2005.00795.x.

[41] F. Martens, J. Bulthé, C. van Vliet, and H. Op de Beeck, “Domain-general and

domain-specific neural changes underlying visual expertise," *Neuroimage*, vol. 169, no. July 2017, pp. 80–93, 2018, doi: 10.1016/j.neuroimage.2017.12.013.

[42] S. Duyck, F. Martens, C. Y. Chen, and H. Op de Beeck, "How visual expertise changes representational geometry: A behavioral and neural perspective," *J. Cogn. Neurosci.*, vol. 33, no. 12, pp. 2461–2476, 2021, doi: 10.1162/jocn_a_01778.

[43] K. Simonyan and A. Zisserman, "VGG-16," *arXiv Prepr.*, 2014, doi: 10.1016/j.infsof.2008.09.005.

[44] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015, doi: 10.1007/s11263-015-0816-y.

[45] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst.*, vol. 32, no. NeurIPS, 2019.

[46] A. C. Ferreira *et al.*, "Deep learning-based methods for individual recognition in small birds," *Methods Ecol. Evol.*, vol. 11, no. 9, pp. 1072–1085, 2020, doi: 10.1111/2041-210X.13436.

[47] G. Yovel, I. Grosbard, and N. Abudarham, "Deep learning models of perceptual expertise support a domain specific account," *bioRxiv*, 2022.

[48] N. Kanwisher, P. Gupta, and K. Dobs, "CNNs Reveal the Computational Implausibility of the Expertise Hypothesis," *iScience*, vol. 26, no. 2, p. 105976, 2023, doi: 10.1016/j.isci.2023.105976.

[49] N. Abudarham, L. Shkiller, and G. Yovel, "Critical features for face recognition,"

Cognition, vol. 182, pp. 73–83, 2019.

[50] A. W. Young, D. Hellawell, and D. C. Hay, “Configurational Information in Face Perception,” *Perception*, vol. 42, no. 11, pp. 1166–1178, 2013, doi: 10.1068/p160747n.

[51] D. Maurer, R. Le Grand, and C. J. Mondloch, “The many faces of configural processing,” *Trends in Cognitive Sciences*. 2002. doi: 10.1016/S1364-6613(02)01903-4.

[52] A. M. Burton, S. R. Schweinberger, R. Jenkins, and J. M. Kaufmann, “Arguments against a configural processing account of familiar face recognition,” *Perspect. Psychol. Sci.*, vol. 10, no. 4, pp. 482–496, 2015, doi: 10.1177/1745691615583129.

[53] G. Yovel and B. Duchaine, “Specialized face perception mechanisms extract both part and spacing information: Evidence from developmental prosopagnosia,” *J. Cogn. Neurosci.*, vol. 18, no. 4, 2006, doi: 10.1162/jocn.2006.18.4.580.

[54] E. Mckone and G. Yovel, “Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? toward a new theory of holistic processing,” *Psychon. Bull. Rev.*, vol. 16, no. 5, 2009, doi: 10.3758/PBR.16.5.778.

[55] S. Gilad-Gutnick, E. S. E. S. Harmatz, K. Tsourides, G. Yovel, and P. Sinha, “Recognizing facial slivers,” *J. Cogn. Neurosci.*, vol. 30, no. 7, 2018, doi: 10.1162/jocn_a_01265.

[56] G. J. Hole, P. A. George, K. Eaves, and A. Rasek, “Effects of geometric

distortions on face-recognition performance," *Perception*, vol. 31, no. 10, pp. 1221–1240, 2002, doi: 10.1068/p3252.

[57] R. E. Galper, "Recognition of faces in photographic negative," *Psychon. Sci.*, vol. 19, no. 4, pp. 207–208, 1970, doi: 10.3758/BF03328777.

[58] C. Rezlescu, J. B. Wilmer, and A. Caramazza, "The Inversion, Part-Whole, and Composite Effects Reflect Distinct Perceptual Mechanisms With Varied Relationships to Face Recognition," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 43, no. 12, pp. 1961–1973, 2017, doi: 10.1037/xhp0000400.supp.

[59] N. Abudarham and G. Yovel, "Reverse engineering the face space: Discovering the critical features for face identification," *J. Vis.*, vol. 16, no. 3, 2016, doi: 10.1167/16.3.40.

[60] E. J. Burns, T. Arnold, and C. M. Bukach, "P-curving the fusiform face area: Meta-analyses support the expertise hypothesis," *Neurosci. Biobehav. Rev.*, vol. 104, no. July, pp. 209–221, 2019, doi: 10.1016/j.neubiorev.2019.07.003.

[61] H. P. Op de Beeck and C. I. Baker, "The neural basis of visual object learning," *Trends Cogn. Sci.*, vol. 14, no. 1, pp. 22–30, 2010, doi: 10.1016/j.tics.2009.11.002.

[62] H. P. Op De Beeck, C. I. Baker, J. J. DiCarlo, and N. G. Kanwisher, "Discrimination training alters object representations in human extrastriate cortex," *J. Neurosci.*, vol. 26, no. 50, pp. 13025–13036, 2006, doi: 10.1523/JNEUROSCI.2481-06.2006.

[63] J. Gomez, M. Barnett, and K. Grill-Spector, "Extensive childhood experience with

Pokémon suggests eccentricity drives organization of visual cortex," *Nat. Hum. Behav.*, vol. 3, no. 6, pp. 611–624, 2019, doi: 10.1038/s41562-019-0592-8.

[64] B. D. McCandliss, L. Cohen, and S. Dehaene, "The visual word form area: Expertise for reading in the fusiform gyrus," *Trends Cogn. Sci.*, vol. 7, no. 7, pp. 293–299, 2003, doi: 10.1016/S1364-6613(03)00134-7.

[65] K. Dobs, J. Martinez, A. J. E. Kell, and N. Kanwisher, "Brain-like functional specialization emerges spontaneously in deep neural networks," *bioRxiv*, vol. 8913, no. March, p. 2021.07.05.451192, 2021, [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.07.05.451192v1%0Ahttps://www.biorxiv.org/content/10.1101/2021.07.05.451192v1.abstract

[66] N. Kanwisher, M. Khosla, and K. Dobs, "Using artificial neural networks to ask 'why' questions of minds and brains," *Trends Neurosci.*, vol. 46, no. 3, pp. 240–254, 2023, doi: 10.1016/j.tins.2022.12.008.

[67] W. J. Ma and B. Peters, "A neural network walks into a lab: towards using deep nets as models for human behavior," pp. 1–39, 2020, [Online]. Available: http://arxiv.org/abs/2005.02181