A Model of Categorical Perception in the Human Visual System

Solim LeGris

NEUR503: Computational Neuroscience

McGill University

In this paper, a broadly biologically inspired model of categorical perception in human vision is explored and results from the originators of the model are replicated. Categorical perception is a cognitive phenomenon that occurs as a result of category learning. The model in this paper provides computational evidence that categorical perception may occur at later stages of processing without the need for feedback mechanisms to earlier areas as has previously been suggested. Evidence of the model's limitations through additional simulations is provided showing that the CP effect may not be specific to the later stage of processing but to how and where the signal indicating category membership is provided to the model. Lastly, evidence that the model generalizes well to novel variations of the stimuli are compared to human data.

Keywords: categorical perception; computational modelling; categorization; competitive learning

1. Introduction

A key property of the human brain is its ability to make sense of an otherwise noisy and chaotic world. One of the fundamental mechanisms underlying this ability is categorization. The categories that an individual possesses guide their behaviour such that they behave differently towards different objects, events, etc. It has been noted that categorization is essential to cognition itself and may even be the basis for the vast majority of cognitive tasks [1]. Categorization allows us to deal with the infinite variation in the world effectively and is also the process by which continuous analog sensory perception becomes discrete symbolic processing (e.g. categorical). Identifying the category to which a stimulus belongs allows to rapidly access information about the potential properties of the stimulus and how we might respond to it.

Although some of our categories may be innate [2], most categories are arguably acquired throughout a lifetime. An interesting question that arises with respect to category learning concerns the

effect that this acquired knowledge may have on functional areas involved in perception. Categorical perception results from having learnt to categorize and appears to be a phenomenon revealing the influence of high-level category learning on feature-based perception [3]. In order to successfully categorize stimuli, the relevant features of a given stimulus must be identified and learnt whereas the irrelevant ones must be discarded or ignored [4]. Although some categories may naturally have physical properties such that members of different categories do not overlap and members of the same category are similar enough to be immediately recognized as such, this is not the case for all categories [4]. Often, stimuli that ought to be treated as being of the same type are not superficially similar enough for this behavior to be immediate or, furthermore, stimuli that ought to be treated differently overlap. CP's function becomes evident here as it is known to underlie transformations of relatively linear sensory signals into relatively non-linear representations [5]. Through CP, our nervous systems learn to place sensory information into different categories by warping perception in accord with the perceptual features that are relevant to categorization [6]. It is important to note that the CP discussed here occurs at the level of perceptual categories as opposed to semantic categories (e.g. those defined by abstract relations). Perceptual categories are defined by the perceptual relationships between stimuli that change with respect to the physical properties through experience[6].

Experimentally, CP is revealed when a subject's ability to make perceptual discriminations between stimuli is more accurate when they belong to different categories rather than the same category [5]. Moreover, these differences in perceptual discrimination may occur regardless of the physical differences between stimuli. In other words, CP may occur even when the difference (e.g. based on some physical metric such as wavelength) between stimuli of the same category is the same as that of stimuli of different categories. Typically, the phenomenon has been described as causing an expansion of perceived between-category differences and a compression of within-category differences [4][5]. Participants' results in experiments provided evidence that, following category learning, members of the same categories seemed perceptually more alike than before having learnt the categories whereas members of different categories seemed perceptually more different than before having learnt the categories. Moreover, it has been shown that changes in the neural correlates associated with categorization occur as well [7].

The neural mechanisms of the complex interplay evidenced by CP between low-level and high-level processing remain poorly understood as demonstrated by the numerous theories and models of

the phenomenon [5]. Understanding how early in the information processing stream the influences of category learning occur to generate CP is a core question of CP research. There is evidence for various candidate functional areas being involved in CP. For instance, the prefrontal cortex of monkeys show strong categorical representations [8], patterns of activity in the inferotemporal cortex code for different object categories [9] and cells in this area become tuned along characteristic dimensions [10]. It seems that many sites involved in vision (if not all) from precortical areas through occipital lobes and to the ventral stream may be involved in CP [5].

Previous studies have attempted to determine the possible neural loci of the CP effect. In one study, human participants were trained to categorize Gabor patch stimuli (Figure 1) that varied in spatial phase and their ability to discriminate within and between category difference before and after training was assessed [11]. It was found that CP was induced by category learning but that training restricted to a specific orientation did not generalize to variations in the orientation of the Gabor patch stimuli. According to the authors, this lack of transfer from one orientation to another was evidence that the perceptual changes resulting from category learning occurred relatively early in the visual information processing stream. Moreover, it may have resulted as a consequence of the dynamic interplay between later brain areas involved in category learning and earlier visual processing areas through feedback mechanisms. This conclusion was based on strong evidence from neuroscience that orientation response ranges become increasingly large as you move further in the visual processing hierarchy hence the conclusion that the perceptual changes may have been caused by changes in the responses of neurons at earlier stages of visual processing with smaller orientation response ranges. Others have argued that this may as well have been a consequence of higher-order brain areas making better use of low-level information subsequent to category learning [12].

The human visual system is a highly complex perceptual system that is also involved in categorization of stimuli. Although a wealth of information has been acquired about its mechanisms, topology and functions, CP in the human visual system is still not well understood. Computational modelling is advantageous because it can be systematically manipulated to uncover operational principles underlying a cognitive and/or neurophysiological phenomenon. Various computational models of CP in the human visual system have been devised demonstrating that many different architectures and computational frameworks can give rise to what has been called *synthetic* CP [13]. Key properties to consider in order to develop a plausible computational model of categorization in the human visual system are 1) feature selection through localized receptive fields and 2) cate-

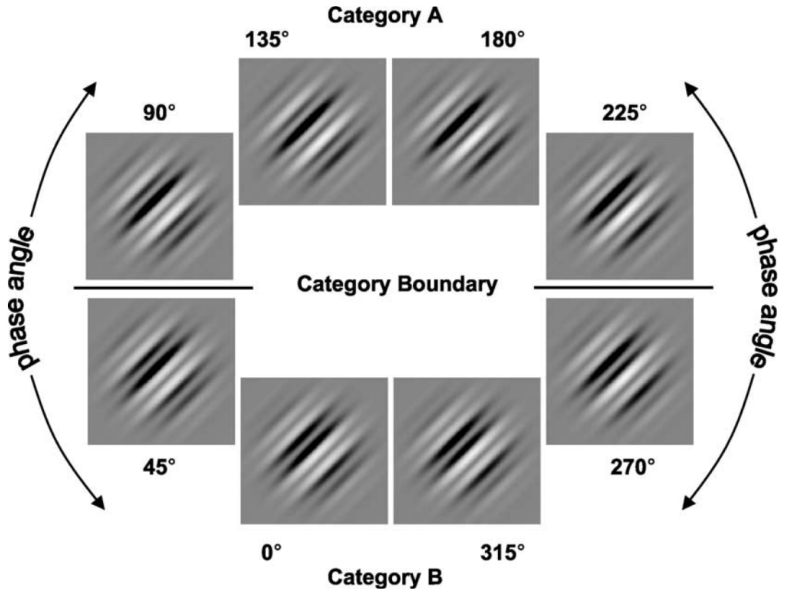

Figure 1: Spatial phase continuum of stimuli for an orientation of 45°. The schema indicates the position of the stimuli in relation to the imposed category boundary and the $3f$ phase is indicated. The images were constructed with a fixed f phase Gaussian grating superimposed on varying $3f$ phases. During discrimination learning and testing, the pairs of stimuli are presented to either visual field where the pair consists of the same stimulus to both visual fields, stimuli from the same category or stimuli from different categories for all possible permutations. Image taken from Notman et al. (2005)

gorization through a global combination of features. Additionally, biological plausibility, hierarchy and modularity may also come into play as motivations for developing or choosing one model over the other.

Casey and Snowden (2012) hypothesized, contrary to the previously mentioned study [11], that neural feedback to the early visual visual cortex or even precortical areas is not necessary for CP and that most of the changes that result in CP occur at the later stages. Finding that other models did not demonstrate the interaction of the different stages of visual processing and task influence, they devised a modular architecture which they argue is 1) biologically plausible and 2) has the appropriate properties to demonstrate task influence (e.g. category learning) [6]. The model was based on previous work by Armony et al. (1995) [14]. The architecture of the computational model (Figure 2) uses uniform layers of identical, interconnected neurons that learn through Hebbian learning and competition supported by lateral inhibition [15], effectively training the abstracted populations of neurons to become sensitive to different overlapping patterns. Crucially, each module can learn to exhibit appropriate local properties (feature selection) whereas the whole model exhibits

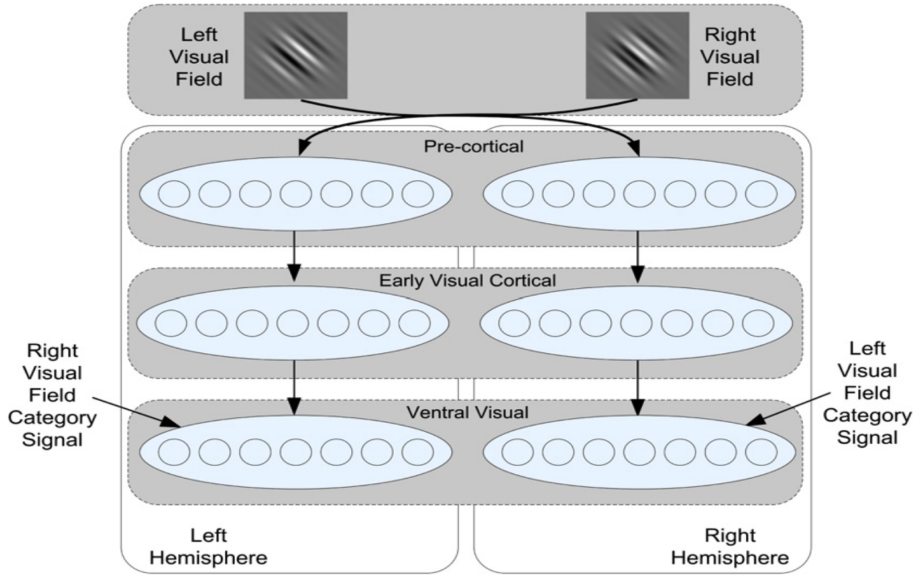

Figure 2: Schematic of the model of learned categorical perception in human vision . The model consists of left and right pathways from the visual fields to the last stage of processing. Information progresses from pre-cortical (PC) modules to early visual cortical (EC) modules and finally to the ventral visual (VV) modules. A binary category signal is provided to VV depending on category membership when the model is trained for category learning. Image taken from Casey and Snowden (2012).

global properties (category learning). The model has been used to make biological processing predictions which have been later confirmed [16] furthermore contributing to enforcing its biological plausibility. A key feature of this model, in light of the hypothesis made by the authors of the study in question, is that its hierarchical and modular nature makes it so that it is possible to explore whether CP emerges as a result of changes to later stages of visual processing only.

In this paper, I explore the model presented by Casey and Snowden (2012) because 1) I find their claim to be controversial and 2) I believe that although a category signal to late processing stages may induce CP without the need for feedback mechanisms to earlier stages, their model does not convincingly show that this is the case in real brains. Nonetheless, the authors have made an important and novel contribution to CP research by 1) evaluating whether an abstract, system-level model of visual processing can sufficiently model early visual analysis such that it exhibits both local (feature detection) and global behaviour (categorization), 2) whether CP can be induced through task influence to the later stages of processing and 3) by comparing their results systematically to behavioural experiments. The authors chose to construct their experiments with the model using the same experimental framework as Notman et al. (2005) whose hypothesis they sought to refute.

I first discuss the mathematical framework of the model, implementation details and methodology (section 2). Later, I proceed to describe the results I obtained with my simulation of the model (section 3) and show that CP is observed at earlier stages with a category signal. Lastly, I discuss the results in relation to those obtained by Casey and Snowden (2012) and other findings in the literature (section 4).

2. The Model

The chosen model architecture has the advantage of being biologically plausible at the population level yet simple enough to implement and analyze. It affords a systems level interpretation of visual processing in the human brain through its hierarchical and modular structure. The model was chosen to explore the influence of category signals at later stages of visual processing and the ability/sufficiency of such signals to induce CP. The idea of modelling a category signal was taken by Casey and Snowden (2012) from Armony et al.'s (1995) computational model of fear conditioning [14]. The response of these modules was modulated through a conditioning signal. In the case of CP, we can reformulate this signal as a category signal given to latest processing areas. This feedback mechanism is akin to what has previously been observed between the prefrontal cortex and inferotemporal cortex of monkeys trained to categorize images of cats and dogs [8]. Next, the input representation used to train and test the model is discussed.

2.1 Input representation

In the original human study by Notman et al. (2005), human participants were shown eight image pairs of Gaussian windowed gratings with combined spatial frequencies ($f + 3f$) to form compound Gabors varying in the relative spatial phase of the two components (see Figure 1). Two sets of stimuli with orientations of 45°and $-45°$ respectively were generated. The spatial phase f was set to 0°for all stimuli whereas the $3f$ spatial phase varied systematically from 0°to 315°. Within each set, four Gabors belonged to category A with spatial phases $3f_A \in \{90°, 135°, 180°, 225°\}$ and four Gabors belonged to category B with spatial phases $3f_B \in \{90°, 135°, 180°, 225°\}$. Participants were first tested on discrimination ability by being presented image pairs restricted to either orientation set where they had to make within-category comparisons with six pairs with $3f$ spatial phases {90°/135°, 135°/ 180°, 180°/225°, 45°/0°, 0°/135°, 31°/270°} and between-category comparisons with two pairs of $3f$ spatial phases {90°/45°, 225/270°}. Following the first phase of the experiment,

participants were trained to categorize through trial and error with feedback on both sets of images and were subsequently tested again on their ability to discriminate between-category and within-category signals.

It would be highly desirable to present the actual images to the model but this would require extensive preprocessing of the images and is not necessarily pertinent to the questions asked in the present paper. Consequently, the stimuli were generated similarly to the way in which Casey and Snowden (2012) originally constructed them. In accord with Notman et al's (2005) experiment, the generated stimuli only varied the $3f$ phase with $p \in P = \{0°, 45°, ..., 315°\}$. The stimuli were presented as patterns of phase activity for $3f$ phases and for completeness, the f phase was included in the stimuli with a constant value of 0 resulting in 9-dimensional vector representations. Additionally, the stimuli orientations were restricted to a constant value of $S_o = 45°$.

Consequently, inputs to the model such that,

$$x_{po} = e^{-\Lambda_p(p-S_p)^2-\Lambda_o(o-S_o)^2} \tag{1}$$

$$\Lambda_p = \frac{-\ln 1/2}{(\lambda_p/2)^2} \tag{2}$$

$$\Lambda_o = \frac{-\ln 1/2}{(\lambda_o/2)^2} \tag{3}$$

where x_{po} is an input x with spatial phase $p \in P$ and $o \in O = \{0°, 15°, 30°, 40°, 43°, 45°, 47°, 50°, 60°, 75°, 90°\}$. For each stimulus of a given spatial phase, the pattern of activity is represented as a Gaussian centered at the appropriate stimulus phase S_p and with a bandwidth $\lambda_p = 106°$. This value was chosen in accord with Casey and Snowden (2012) who argue that this approximately matches the orientation selectivity of early cortical neurons [6]. The pattern of activity decreases in strength as the difference in phase increases (e.g. moving away from the mean of the Gaussian) wrapping around such that a phase of 360° is equal to a phase of 0° as can be seen in Figure 3.

The values Λ_p and Λ_o are chosen so that the associated bandwidth is achieved with the Gaussian at half the height of the curve [11]. Moreover, $o \in O$ are used to extend the experiment and test for generalization of CP to different orientations from the constant one of 45°selected for training. This is to mirror the second experiment carried out by Notman et al (2005) on the subjects that had previously learnt to categorize the Gabors. In the case of the present computational model, the Gaussian patterns of activity are extended in the second phase of the simulation during testing. The patterns are such that for increasing differences between $o \in O$ and $S_o = 45°$, maximal activity

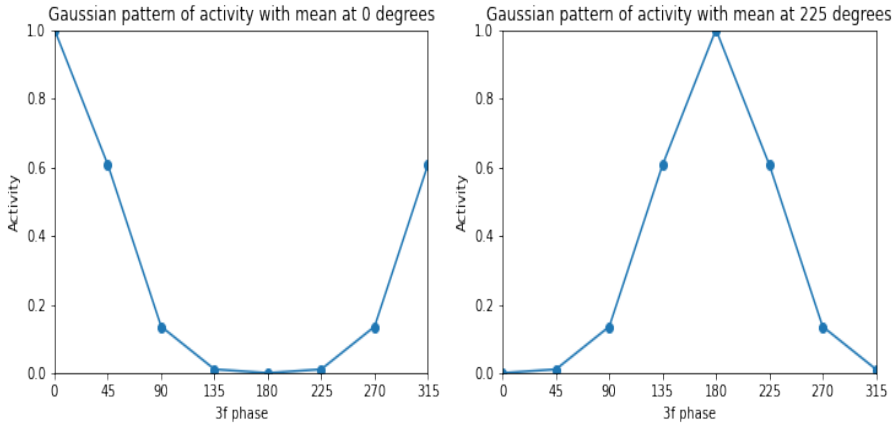

Figure 3: An example of two 9-dimensional stimuli as they were presented to the model. The left image corresponds to a stimulus with $3f$ phase at 0°whereas the one on the right corresponds to a stimulus with $3f$ phase at 180°. These stimuli are abstractions of the Gabor patches described earlier and represent Gaussian patterns of activity with mean centered at the appropriate stimulus phase with bandwidth $\lambda_p = 106°$ to match human data on phase selectivity. The patterns of activity wrap around such that a stimulus phase of 0°is equivalent to a stimulus phase of 360°.

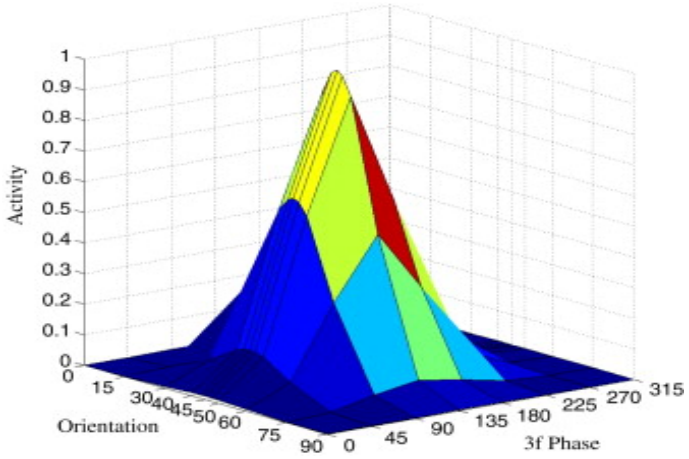

Figure 4: Surface plot showing an example input with a $3f$ phase of 135°and stimulus orientation of 45°, spanning orientations 0°to 90°. The orientation values are meant to introduce orientation differences with respect to Gaussian patterns of activity centered at 45°. These stimuli were used in Casey and Snowden (2012) to test for CP generalization ability of the model as was originally done in the human experiments. Image taken from Casey and Snowden (2012).

at the center of the Gaussian decreases as well as mean activity, as is illustrated in Figure 4. Lastly, an orientation bandwidth of $\lambda_o = 30°$ was chosen for the Gaussian patterns of activity to match human data [6].

The model was therefore presented with four inputs at any given time. As seen in Figure 2, the model consists of two processing streams, one for each visual field. As a result, the model is

presented with a Gaussian pattern of activity representing the Gabors to each visual field as well as a category input to the latest stage of processing. The category signal is represented as a binary value such that an input for category A is 0 and one for B is 1 and this signal is only presented during the category training phase.

2.2 Model architecture and computational aspects

The computational model of visual processing used in this paper was constructed such that it has three distinct layers or modules (see Figure 2) each representing a neuronal population of the hierarchy of visual processing in the following order: precortical (PC) processing such as occurs in the retina and lateral geniculate nucleus (LGN), early visual cortical (EC) processing such as in V1 and V2 and finally ventral visual (VV) processing such as in the posterior and anterior inferotemporal cortex. In order to keep the model simple, it is restricted to contralateral unit responses. The category signal is provided to VV because it it thought that the areas it represents contain neurons that become tuned along relevant category dimensions [10].

As mentioned previously, the model mirrors the competitive learning architecture previously developed by Armony et al (1995) where each of the modules represents a single layer of rate coded neurons fully connected to their respective inputs. Each unit j in each layer of the model integrates over respective d-dimensional inputs x such that its activation and output are as follows:

$$u_j = \sum_{i=1}^{d} x_i w_{ij}(t) \tag{4}$$

$$y_j = \begin{cases} f(u_j) & if\ j = argmax_i f(u_i) \\ f(u_j - \mu_k y_{win}) & otherwise \end{cases} \tag{5}$$

$$f(u_j) = \begin{cases} 1 & u_j \geq 1 \\ u_j & 0 < u_j < 1 \\ 0 & u_j \leq 0 \end{cases} \tag{6}$$

where y_j represents the output of the neurons, u_j the integration of inputs in a neuron and $f(u_j)$ the activation of a given neuron. The integration of inputs to the neurons is a weighted sum of inputs x_i using weights $w_{ij}(t)$ from input i to neuron j. Weights were randomly initialized for all iterations of the model using a normal distribution centered at 0 with a standard deviation of 0.5. The winning neuron $y_{win} = \max_i f(u_i)$ is selected according to its activation and inhibits all other

neurons with inhibition rate μ_k for each module $k \in \{0, 1, 2\}$ where PC is 0, EC is 1 and VV is 2. The time step t is such that $1 \leq t \leq N$ where N is the number of epochs, a multiple of the number of inputs presented to the model.

As demonstrated by Equations 4, 5 and 6, the model implemented here displays competition such that feature detectors are formed. It is a variant of the competitive learning algorithm originally developed by Rumelhart & Zipser (1986) [15]. At each module, once integration of inputs and activation is calculated according to Equations 4 and 6, a winner neuron is selected and it suppresses the outputs of other neurons. This output is then fed to the following layer (for PC and EC).

2.3 Learning in the model

A simple Hebbian learning rule is implemented to achieve competitive learning in each module of the model as follows:

$$w'_{ij}(t) = \begin{cases} w_{ij}(t) + \eta x_i y_j & if\ x_i > \rho \bar{x} \\ w_{ij}(t) & otherwise \end{cases} \tag{7}$$

$$\bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i \tag{8}$$

$$w_{ij}(t+1) = \frac{w'_{ij}(t)}{\sum_{k=1}^{d} w'_{kj}(t)} \tag{9}$$

where $w'_{ij}(t)$ represents an intermediate step in which weights are increased by a factor of η, the learning rate, if they have above average inputs. The threshold for increasing weights is determined according to $\bar{x}$ and the factor ρ which controls the threshold above which to increase weights. Finally, each neuron j's weights $w'_{ij}(t)$ are normalized by weights $w'_{kj}(t)$ generating the updated weights $w_{ij}(t+1)$ at time $t+1$ (Equation 9). The above learning rule defines a way for units in the modules to learn the features of the stimuli the model processes. A final component of this computational learning framework is the influence of the category being learnt during the category learning phase of the simulation. The category signal in this model is akin to feedback given to human participants during category learning through trial and error. In order to incorporate this aspect into the learning process, a category signal is provided to VV (see Figure 2) only during category learning. The category input weight W_c is fixed to a constant value and is not subject to learning but nonetheless influences learning through normalization (Equation 9). The category

signal is 0 at all times except during category learning when the stimulus presented is from category B and during testing after category learning.

Learning in the model is divided into two phases: pre-training and category training. During the pre-training phase, the model is not given any category signal while during the category training phase, the model is given the category signal and weight normalization takes the additional weight W_c into account.

2.4 CP measure in the model versus in the human study

As was discussed, the present model is meant to mirror experiments conducted by Notman et al. (2005) with human subjects. In order to measure the CP effect of category learning, discrimination performance was measured before and after category training by counting the number of Hit and False Alarm responses to the same-different image task. Hits are measured separately for within and between category image pairs and counted as $H(W)$ when a participant correctly identified the images as belonging to the same category and as $H(B)$ when a participant correctly identified images as from different categories. A False Alarm is counted as F if the subject identified the images as being different when they were identical. As in Casey & Snowden (2012), discrimination performance was calculated as an A' score [17]. This score is a non-parametric measure of the area under the single-point Receiver Operating Characteristic (ROC) curve calculated for within and between category discriminations separately:

$$A' = \begin{cases} \frac{1}{2} & if\ H < F \\ \frac{1}{2} + \frac{(H-F)(1+H-F)}{4H(1-F)} & otherwise \end{cases} \tag{10}$$

where H is the probability of a Hit ($H(W)$ or $H(B)$) and F is the probability of a False Alarm. In order to obtain $A'(W)$ and $A'(B)$ in the model for between- and within-category discrimination respectively, discrimination testing was assessed after the pre-training phase and after the category training phase. To obtain Hits and False Alarms, the outputs of each pair of modules (one module from one hemisphere compared with the corresponding module in the other hemisphere) were compared after testing on all possible within- and between-category pairings. Outputs of each of the left and right modules are summed separately to give Y' and min-max normalized for each pair of images presented during the testing to yield Y. The values obtained for the left and right modules are then compared according to some threshold value δ to determine whether or not they

are different according to the following:

$$Y' = \sum_{i=1}^{m} y_i \tag{11}$$

$$Y = \frac{Y' - \min_i(y_i)}{\max_i(y_i) - \min_i(y_i)} \tag{12}$$

$$|Y_{left} - Y_{right}| > \delta \tag{13}$$

The model was simulated and implemented using Python in a Jupyter Notebook. The main librairies used were the following: `Scipy`, `Numpy`, `Sklearn` and `Matplotlib`. The code for the simulation is available on GitHub

3. Simulation and Results

In this section, experimental procedures for the computational model are discussed. The aim here was both to model the experiments in human studies as well as attempt to replicate the findings of Casey and Snowden (2005). As mentioned in Section 2.3, the model's testing is divided into two phases for a single orientation. In the first phase (pre-training), stimuli are presented to the model without a category signal to VV. Discrimination performance is subsequently tested (as described in Section 2.4). In the second phase (category training), stimuli are presented to the model with the category signal and discrimination performance is tested again subsequently.

3.1 Simulation phases and model parameters

In the pre-training phase, the model is presented with the stimuli in order to achieve a stable model before category training. The training set chosen for this phase is the set of all possible stimuli. This choice is made in order to train each module in the model to be capable of discrimination of stimuli with different phases. At each epoch, all eight phase inputs are presented in random order to either visual field which is also chosen at random. The visual field not chosen for a given epoch receives the zero vector as input to ensure that the model can discriminate between an input and no input. An example of responses from the left PC of a model before and after pre-training can be seen in Figure 5.

The category training phase of the experiment consisted of blocks of double training and single training. During double training, stimuli pairs from the set of permutations of within-category stimuli (twelve in total for each category) and between-category stimuli (thirty-two in total) were

presented in random order to the left and right visual fields. During single training, the model was presented with each of the eight phase inputs in random order to a randomly chosen visual field while the other visual field received the zero-vector input representing no stimulus. One category training epoch consisted of fifty-six double and three times eight single inputs chosen in random order for a total of eighty pairs of images per epoch. The model was trained for thirty pre-training epochs and then sixteen category training epochs. Details of the parameters used are outlined in Table 1. Both simulation phases were restricted to training with a single stimulus orientation $S_o = 45°$.

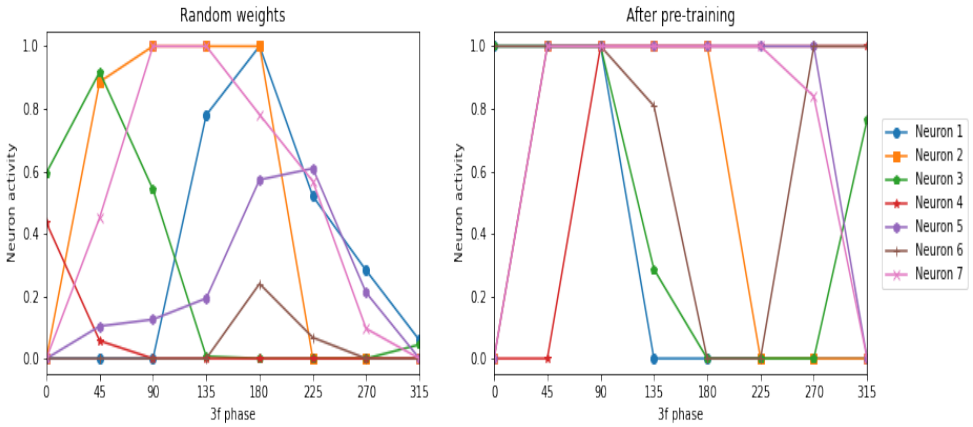

Figure 5: Neuron activity in the left PC module of example model with 7 units for random weights and after 20 epochs of pre-training. The response of each neuron is shown for all $3f$ phases tested. As can be seen, structure (e.g. feature detectors) emerges within the PC module early during training due to the competitive learning algorithm that the model implements.

It should be noted that the the original parameters used by Casey and Snowden (2005) in their model were not all set to identical values in the model presented here for reasons that will become obvious in Section 3.2.

3.2 Acquired CP in the model

One hundred models were trained and tested for CP using the parameters outlined in Table 1. All the models were randomly assigned different weights at time $t = 1$ using the method describe in Section 2.2. The model first underwent the pre-training phase which was then followed by the category training phase. Examples of neuron activity in VV before and after pre-training as well as after category training is illustrated in Figure 7. Strong categorical activity was observed in the VV module following category learning. Discrimination testing matched that of the original model and in fact outperformed it due to some minor adjustments. The learning rate and number

Parameter	Value	
Neurons per module	M_{PC}	7
	M_{EC}	7
	M_{VV}	7
Inhibition rate	μ_{PC}	0.6
	μ_{EC}	0.4
	μ_{VV}	0.2
Category input fixed weight	W_c	0.4
Learning rate	η	0.01
Pre-training epochs	N_p	30
Category training epochs	N_c	16
Weight change threshold	ρ	1
Difference threshold	δ	0.2

Table 1: The parameters used for the experiments carried out with every iteration of the model.

of epochs for pre-training in the original model were found to be unnecessarily large. The model clearly converged to feature detectors as early as the first few epochs (see Figure 5). This may have been caused by a minor difference in the stimulus presentation scheme. Here, the whole set of possible permutations of stimuli for within and between categories were presented whereas in the original experiment, within and between comparisons were chosen so as to lie on what was termed a "circular" continuum [6]. The training sets in the original studies were considerably smaller as a result (see Section 2.1).

Discrimination performance was assessed using the methods and equations described earlier (Section 2.4). As predicted by the authors of the original paper, a CP effect was observed in the VV following category training. $A'(W)_{before} = 0.67$ and $A'(B)_{before} = 0.68$ values were assessed before category training and $A'(W)_{after} = 0.51$ and $A'(B)_{after} = 1.0$ were assessed after category training. These results demonstrate that the CP effect is indeed observed in later processing stages through a category signal. The CP signature is demonstrated here by three key changes in mean A' score values. Firstly, the low mean A' score before category training for both within- and between-category discrimination indicates that strong category representations had not been learnt yet since discrimination ability was not very high and very similar for both within- and between-category discriminations. Secondly, the drastic increase in the mean A' score for between-category

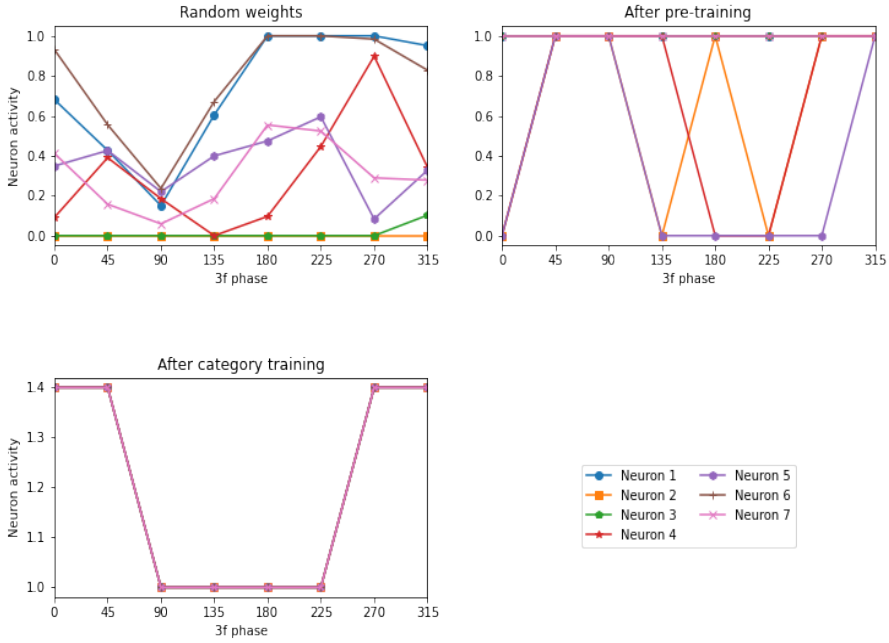

Figure 6: Neuron responses in the VV module for an example model. Responses are shown before any training (e.g. random weights), after pre-training and after category training. The bottom left image clearly demonstrates the CP effect and the learnt categorical representations in VV induced as a result of the feedback category signal.

discrimination indicates that following category learning, the models learnt strong categorical representations resulting in robust differences in activity for members of category A and B. This is akin to results in humans where members of different categories become more perceptually dissimilar after category learning (e.g. *acquired distinctiveness*). Lastly, the decrease of the mean A' score for within-category discrimination to barely above chance level indicates that similar activity patterns were learnt for members of the same category resulting in strong categorical representations as opposed to individual representations for each member (e.g. *acquired equivalence*). These results mirror the usual results in CP research both in humans and computational models. Independent samples t-tests showed that the differences in mean A' scores before and after for all modules were significantly different ($p < 0.0000001$). Although significant differences were obtained for the EC and PC modules as well, these results do not support the claim that CP was induced in those modules since they lack the key properties mentioned above for VV. Figure 7 illustrates the changes induced by category training to the VV module of the models trained.

The analysis described here demonstrates that it is possible to induce CP computationally through a feedback category signal provided to the VV module, the latest processing stage in the model. The VV module exhibits both a decrease in discrimination performance for Gabor patterns of activity of the same category and an increase in discrimination performance for the stimuli of different categories. This effect is only seen in VV whereas in other modules there is only an increase in discrimination performance for both within and between category comparisons. As claimed by Casey and Snowden (2012), CP feedback mechanisms need not be extended further to earlier stages of processing in order to obtain a CP effect. Nonetheless, the model lacks any feedback mechanisms from later stages to earlier stages and so may not be fit to demonstrate that such mechanisms are not present in real brains and that they do not indeed induce CP to earlier stages as well. In order to test what the influence of the PC and EC is on the results discussed here, Casey and Snowden (2012) interchanged the PC and EC weights following pre-training and found that this did not change the strength of the observed CP effect in the VV module [6] which was confirmed in my simulation as well. Furthermore, they note that no CP effect is observed in VV prior to category learning which they argue may indicate that the VV module learns to use the encoded phase selective outputs from the PC and EC modules that arise as a result of pre-training.

Although these results indicate a potential for the existence of the computational architecture of this model in real brains, whether a category input to earlier modules would induce CP was tested and results confirmed that this was the case. Fifty iterations of the model were pre-trained without a category signal and then tested with the category signal to EC. When provided with the category signal to EC, similar patterns of activation were found as is seen in VV (see Figure 7). This indicates that the category signal biases the response of a module without any prior learning of categorical information and that the potential for a feedback mechanism to earlier stages is a possibility for the CP effect. These results support the hypotheses of other researchers who tested models that had both feed-forward and feedback mechanisms and observed CP at each level of processing [18]. Furthermore, this may provide evidence to refute the claim of Casey and Snowden (2012) that the VV module is using the category signal and encoded phase selective inputs from prior processing stages to effectively learn categorical representations removing the need for such representations at earlier stages of processing.

Fifty iterations of the computational model with random weight initialization were also tested for orientation generalization following the methods described in Section 2.1. The models were

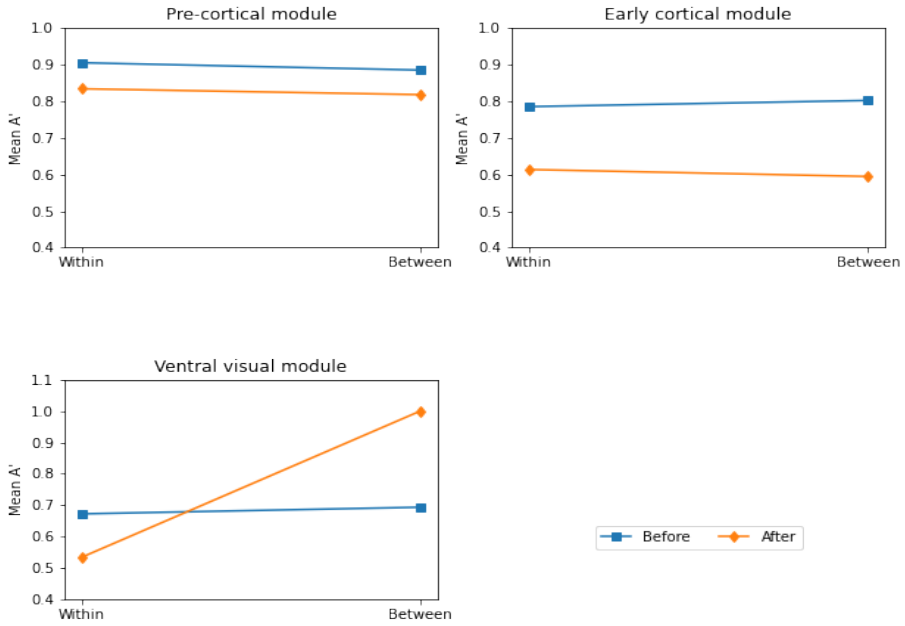

Figure 7: Mean $A'(W)_{before}$, $A'(W)_{after}$, $A'(B)_{before}$ and $A'(B)_{after}$ scores for each module. The CP effect is clearly seen in the bottom left figure representing discrimination performance in the VV module before and after category learning for within and between category discrimination.

trained using the same parameters as outlined in 1 while varying orientations. Mean $A'(W)_{diff} = A'(W)_{after} - A'(W)_{_{before}}$ and mean $A'(B)_{diff} = A'(B)_{after} - A'(B)_{_{before}}$ were calculated for the VV module to investigate whether the CP effects generalized to these other orientations $o \in O = \{0°, 15°, 30°, 40°, 43°, 45°, 47°, 50°, 60°, 75°, 90°\}$. It was found that the computational model generalized relatively well to most orientations (e.g. from 15°to 75°) but that a sharp decline in the mean difference between before and after category training discrimination performance occurred for other more extreme values (Figure 8). Additionally, the CP effect is slightly higher for immediate neighbouring orientation values surrounding the single orientation value for which the model was trained. This can be solved by decreasing the difference threshold to $\delta = 0.01$ as was done by Casey and Snowden (2012). The results obtained here do not predict human results as it was found that humans could not generalize well to these other orientation values after having learnt to categorize stimuli restricted to an orientation of 45°, supporting Notman et al's (2005) conclusion that CP effects occur in the earlier stages of visual processing [11]. In fact, the results of Notman et al (2005) showed that the CP effect was specific to orientations in a range of 6.5°surrounding the orientation

of 45°used for training. Casey and Snowden (2012) obtained similar results to the ones discussed here in their analysis [6].

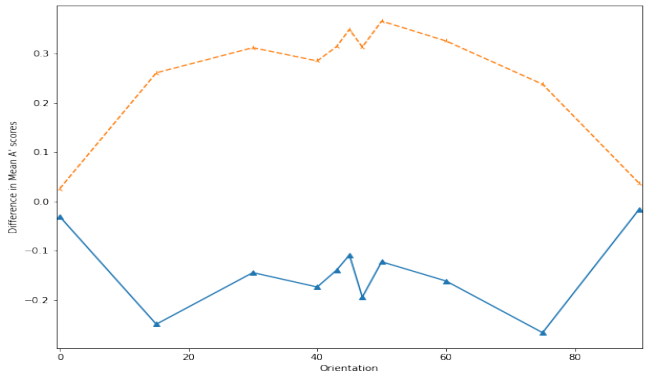

Figure 8: Mean $A'(W)_{before}$, $A'(W)_{after}$, $A'(B)_{before}$ and $A'(B)_{after}$ scores for each module. The CP effect is clearly seen in the bottom left figure representing discrimination performance in the VV module before and after category learning for within and between category discrimination.

4. Discussion

Various models of categorical perception have been devised ranging from many connectionist neural network models [13] [18] [20] [21] [22] to biophysical models [3] and even a dynamical systems model [19]. The simple connectionist feedforward model presented here has the advantage of being simple to analyze while maintaining the abstract properties of biological brains through its competitive learning architecture. Moreover, it demonstrates hierarchical processing and modularity, both key properties of biological brains. The model exhibits local properties in each module or layer (feature detection) and global properties as a system (category learning). In Casey and Snowden's (2012) view, this model was a first step in demonstrating that CP was possible without the need for elaborate feedback mechanisms to earlier stages of sensory processing. Indeed, through a category signal to the VV module, it was shown that the CP effect can occur within that module only and that it is capable of using the encoded sensory information from earlier stages to display categorical activity. At the neuronal level, units respond with similar activity when the model is presented with stimuli that belong to the same category whereas there is a sharp difference in response when it is presented with stimuli from different categories. This phenomenon is grounded in biological evidence as has been supported by recent neurophysiological studies. As an important

step towards elucidating the neural mechanisms of CP effect, the neural basis of categorization has been increasingly researched. Several studies have shown that category learning modifies the individual neuronal properties of the inferotemporal cortex [10] [23] and other studies of visual categorization in non-human primates suggest that changes in neural representations due to category learning occur in higher-order visual areas and higher association cortex but not earlier sensory cortex [9]. However, other studies in humans and rodents point towards a plausible alternative to this view: that category biasing during categorization tasks can occur in signals in earlier visual and auditory cortices [24] [25]. As has been shown computationally in the present simulation, a category signal, potentially being fed back from later processing stages, may also induce the CP effect on earlier populations of neurons. The advantage therefore of the present model is that it provides computational evidence for the plausibility of multiple downstream effects of category learning through feedback. More recent work in computational modelling from Freedman et al. (2020) has shown through a biophysical model of top-down signalling grounded in neurophysiological evidence that a CP effect may emerge as a result of integration of feedforward and feedback signals in an association area[3]. This association area is thought to be implicated in improving perceptual stability in noisy environments thereby demonstrating a novel functional role for CP. In their model, sensory encoding change would eventually manifest itself to all stages in the downstream information processing [3].

Although the model discussed in this paper broadly provides computational evidence for potential mechanisms of CP in real brains, it is limited in several ways. Firstly, the model is quite simplistic since it is an abstraction of extremely complex neural processes. The hierarchy introduced is arbitrarily defined to represent vaguely the different levels of processing that occur in the human visual processing stream. Although it was shown that the computational simulation discussed here broadly maps to human data, it is also sensitive to initial parameters such as the number of training epochs, learning rate and a variety of other parameters not discussed extensively here. Nonetheless, the model may potentially make predictions about complex neurophysiological processes, namely that CP might occur, at least sometimes, without the need for feedback mechanisms to early perceptual processing stages. Compared to biophysical models, the present model, although biologically plausible at a systems level, lacks key neurophysiological properties. Other models such as Freedman et al's (2020) neural circuit of top-down signalling much more closely match biological reality but have the disadvantage of being more complex to analyze and simulate. Secondly, the model does

not mirror the orientation selectivity that occurs in humans when they are trained to do the task explored in this paper. This may indicate the need for at least some feedback mechanisms that, although not inducing CP, could enable this specificity of category learning in humans [26]. Lastly, an important weakness of the model is that, as discussed in Section 3.2, it seems that CP-like effects (e.g. activity-wise) can be induced whenever any given module is given the category signal, with or without category training. It seems therefore that the results may have been biased by the way in which the category signal was defined. Further investigation into the learning aspect of the category signal and perhaps a higher-order learning module representing areas like the prefrontal cortex would be necessary.

5. Conclusion

In this paper, a computational model of categorical perception in human vision is explored. The simulations carried out support prior explorations of the model by replicating findings that category learning can be restricted to an influence on later stages of processing and still generate a CP effect. These findings are based on a broadly biologically inspired model that exhibits local properties, hierarchy and global properties. Furthermore, alternative simulations to the ones carried out by Casey and Snowden (2012) were carried out in this paper. These experiments provided support for a category signal potentially inducing CP at any stage of processing. Although this claim is not aiming to specifically argue for an alternative biological process, it demonstrates that this model may be limited in its ability to support the inferences it was used to make. Lastly, the model fails to map to human data on the specificity of the CP effect. Contrary to those results, the model generalizes well to other orientations.

References

[1] Harnad, S. (2017). To Cognize is to Categorize. *In Handbook of Categorization in Cognitive Science* (pp. 21–54). Elsevier. https://doi.org/10.1016/B978-0-08-101107-2.00002-6

[2] Neitz, J., Neitz, M. (2017). Evolution of the circuitry for conscious color vision in primates.*Eye*, 31(2), 286–300. https://doi.org/10.1038/eye.2016.257

[3] Min, B., Bliss, D. P., Sarma, A., Freedman, D. J., Wang, X.-J. (2020). A neural circuit mechanism of categorical perception: Top-down signaling in the primate cortex [Preprint]. *Neuroscience*. https://doi.org/10.1101/2020.06.15.151506

[4] Harnad, S. (1987). *Categorical perception: the groundwork of cognition*. Cambridge, UK: Cambridge University Press.

[5] Goldstone, R. L., Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(1), 69–78. https://doi.org/10.1002/wcs.26

[6] Casey, M. C., Sowden, P. T. (2012). Modeling learned categorical perception in human vision. *Neural Networks,* 33, 114–126. https://doi.org/10.1016/j.neunet.2012.05.001

[7] Pérez-Gay, F., Thériault, C., Gregory, M., Sabri, H., Rivas, D., Harnad, S. (n.d.). How and Why Does Category Learning Cause Categorical Perception? *Scholarly Publishing*, 32.

[8] Freedman, D. J., Riesenhuber, M., Poggio, T., Miller, E. K. (2003). A Comparison of Primate Prefrontal and Inferior Temporal Cortices during Visual Categorization. *The Journal of Neuroscience*, 23(12), 5235. https://doi.org/10.1523/JNEUROSCI.23-12-05235.2003

[9] Freedman, D. J., Riesenhuber, M., Poggio, T., Miller, E. K. (2006). Experience-Dependent Sharpening of Visual Shape Selectivity in Inferior Temporal Cortex. *Cerebral Cortex*, 16(11), 1631–1644. https://doi.org/10.1093/cercor/bhj100

[10] Sigala, N., Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, 415(6869), 318–320. https://doi.org/10.1038/415318a

[11] Notman, L. A., Sowden, P. T., Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition*, 95(2), B1–B14. https://doi.org/10.1016/j.cognition.2004.07.002

[12] Mollon, J. D., Danilova, M. V. (1996). Three remarks on perceptual learning. *Spatial Vision*, 10(1), 51–58. https://doi.org/10.1163/156856896X00051

[13] Damper, R. I., Harnad, S. R. (2000). Neural network models of categorical perception. *Perception Psychophysics*, 62(4), 843–867. https://doi.org/10.3758/BF03206927

[14] Armony, J. L., Servan-Schreiber, D., Cohen, J. D., LeDoux, J. E. (1995). An anatomically constrained neural network model of fear conditioning. *Behavioral Neuroscience*, 109(2), 246–257. https://doi.org/10.1037/0735-7044.109.2.246

[15] Rumelhart, D. E., Zipser, D. (1985). Feature Discovery by Competitive Learning*. *Cognitive Science*, 9(1), 75–112. https://doi.org/10.1207/s15516709cog0901_5

[16] Armony, J. (1997). Stimulus generalization of fear responses: Effects of auditory cortex lesions in a computational model and in rats. *Cerebral Cortex*, 7(2), 157–165. https://doi.org/10.1093/cercor/7.2.157

[17] Pollack, I., Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, 1(1–12), 125–126. https://doi.org/10.3758/BF03342823

[18] Spratling, M. W., Johnson, M. H. (2006). A feedback model of perceptual learning and categorization. *Visual Cognition*, 13(2), 129–165. https://doi.org/10.1080/13506280500168562

[19] Beer, R. D. (2003). The Dynamics of Active Categorical Perception in an Evolved Model Agent. *Adaptive Behavior*, 11(4), 209–243. https://doi.org/10.1177/1059712303114001

[20] Salminen, N. H., Tiitinen, H., May, P. J. C. (2009). Modeling the categorical perception of speech sounds: A step toward biological plausibility. *Cognitive, Affective, Behavioral Neuroscience*, 9(3), 304–313. https://doi.org/10.3758/CABN.9.3.304

[21] Thériault, C., Pérez-Gay, F., Rivas, D., Harnad, S. (2018). Learning-induced categorical perception in a neural network model. *ArXiv*:1805.04567 [Cs, Stat]. http://arxiv.org/abs/1805.04567

[22] Schouten, M. E. H., van Hessen, A. J. (1992). Modeling phoneme perception. I: Categorical perception. *The Journal of the Acoustical Society of America*, 92(4), 1841–1855. https://doi.org/10.1121/1.403841

[23] Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P. A. (2008). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6), 1126–1141. https://doi.org/10.1016/j.neuron.2008.10.043

[24] Ester, E. F., Sprague, T. C., Serences, J. T. (2020). Categorical Biases in Human Occipitoparietal Cortex. *The Journal of Neuroscience*, 40(4), 917. https://doi.org/10.1523/JNEUROSCI.2700-19.2019

[25] Xin, Y., Zhong, L., Zhang, Y., Zhou, T., Pan, J., Xu, N. (2019). Sensory-to-Category Transformation via Dynamic Reorganization of Ensemble Structures in Mouse Auditory Cortex. *Neuron*, 103(5), 909–921.e6. https://doi.org/10.1016/j.neuron.2019.06.004

[26] Gilbert, C. D., Sigman, M., Crist, R. E. (2001). The Neural Basis of Perceptual Learning. Neuron, 31(5), 681–697. https://doi.org/10.1016/S0896-6273(01)00424-X