

1 Understanding the Neural Basis of Cognitive Bias Modification as a Clinical Treatment for Depression

2 Akihiro Eguchi, Daniel Walters, Nele Peerenboom,

3 Hannah Dury, Elaine Fox, and Simon Stringer

4 Department of Experimental Psychology, Oxford University, UK

Abstract

Objective: Cognitive Bias Modification (CBM) eliminates cognitive biases towards negative information and is efficacious in reducing depression recurrence, but the mechanisms behind the bias elimination are not fully understood. The present study investigated, through computer simulation of neural network models, the neural dynamics underlying the use of CBM in eliminating the negative biases in the way that depressed patients evaluate facial expressions. **Methods:** We investigated two new CBM methodologies using biologically plausible synaptic learning mechanisms, continuous transformation learning and trace learning, which guide learning by exploiting either the spatial or temporal continuity between visual stimuli presented during training. We first describe simulations with a simplified one-layer neural network, and then we describe simulations in a biologically detailed multi-layer neural network model of the ventral visual pathway. **Results:** After training with either the continuous transformation learning rule or the trace learning rule, the one-layer neural network eliminated biases in interpreting neutral stimuli as sad. The multi-layer neural network trained with realistic face stimuli was also shown to be able to use continuous transformation learning or trace learning in order to reduce biases in the interpretation of neutral stimuli. **Conclusions:** The simulation results suggest two biologically plausible synaptic learning mechanisms, continuous transformation learning and trace learning, that may subserve CBM. The results are highly informative for the development of experimental protocols to produce optimal CBM training methodologies with human participants.

Keywords: depression, cognitive bias modification, visual processing of facial expression, neural network modelling

Public Health Significance Statements: Cognitive Bias Modification (CBM) is a clinical technique aimed at reducing the negative cognitive biases seen in clinical disorders such as anxiety and depression. However, many CBM methodologies fail to adequately alter biases and therefore produce no clinical effect, leading to concern about the treatment's efficacy. This study uses computational modelling to present potential explanations at the neuronal and synaptic level for how a shift in interpretational bias might occur through CBM training. Such an understanding will have a wide impact in helping to guide future research aimed at optimising the effectiveness of CBM treatments.

Understanding the Neural Basis of Cognitive Bias Modification as a Clinical Treatment for Depression

Introduction

Depression is the most common mental health problem, affecting 8 - 12% of the adult population (Ustün et al., 2004). It can lead to a significant reduction in the quality of life for sufferers, and in extreme cases may lead to suicide. It has been related to a number of chronic diseases such as coronary heart disease (Rugulies, 2002; Schneider and Moyer, 2010), and has damaging long term effects on health and well-being. People with anxiety disorder also experience various symptoms similar to those of depression, and both mental health disorders often place a significant burden on psychiatric health services and impact negatively on the economy due to reduced productivity (Ustün et al., 2004; Greenberg et al., 1999; Hoffman et al., 2008). Similarly, the latest WHO report shows that anxiety and depression leads to a loss of millions of work days (Jones, 2016). Consequently, it is of huge importance to discover new more effective treatments for such mental disorders.

One of the common findings in both clinical depression and anxiety is a link to cognitive biases in processing towards emotionally negative information, with patients tending to pay attention to negative stimuli, interpret events negatively, and recall negative memories (Mathews and MacLeod, 2005; Roiser et al., 2012). These biases therefore have been included within cognitive models of depression (Beck, 2008) and anxiety (Mathews and MacLeod, 2005), leading to a growing interest in exploring the causal relationship between these biases, mood states and clinical symptoms.

Cognitive Bias Modification (CBM)

It is thought that the elimination of negative cognitive biases may help to shift the depressed mood state of a patient and reduce anxiety. This led many researchers to recognise the clinical potential of these tools, inspiring the development of a family of potential treatments known as Cognitive Bias Modification (CBM) (MacLeod and Mathews, 2012; MacLeod, 2012). CBM seeks to eliminate these underlying processing biases, with three main varieties of treatment. For example, CBM-Attention (i.e., attentional bias modification - ABM) seeks to shift the attention of subjects away from negative stimuli in the environment (MacLeod et al., 2002; Hakamata et al., 2010), CBM-Interpretation (CBM-I) aims to reduce the tendency for negative interpretation of events (Grey and Mathews, 2009, 2000), and CBM-Memory (CBM-M) seeks to reduce the recall and influence of negative memories (Anderson and Green, 2001; Joormann et al., 2005). However, CBM as a whole is not without controversy. Most CBM studies so far have focused on ABM, with a number of meta-analyses finding the efficacy of ABM inconclusive (Cristea et al., 2015; Hallion and Ruscio, 2011; Mogoşe et al., 2014). CBM-I on the other hand has had more promising results (Cristea et al., 2015; Hallion and Ruscio, 2011; Menne-Lothmann et al., 2014).

The negative interpretation bias of facial expression (Bourke et al., 2010; Richards et al., 2002;

Surcinelli et al., 2006) is one of the examples of clinical disorders where CBM-I intervention can produce a measurable therapeutic outcome (Penton-Voak et al., 2013). In this study, faces were morphed from unambiguously happy to unambiguously angry to give 15 total stimuli. Participants were asked to rate each randomly presented face as either happy or angry, giving a baseline for each participant's emotion recognition along the spectrum of morphs. A balance point was therefore determined, at which participants switched from a categorisation of happy to a categorisation of angry. A CBM training procedure followed in which the previous procedure was repeated, but participants were also given feedback about whether their decision was 'correct' or 'incorrect'. Correct responses were defined as the responses they had previously given in the baseline phase, but with the balance point shifted so that two more faces should now be classified as happy. A final testing phase showed that feedback had shifted participants' balance point in the direction of training.

Nevertheless, it has been less than two decades since the seminal CBM studies, meaning the field is still in its early stages (Grey and Mathews, 2000; MacLeod et al., 2002). A recent commentary describes the problem with current CBM research as a lack of focus on reliably changing the underlying cognitive biases (Fox et al., 2014). They argue that the theoretical assumption behind CBM is the role of negative biases in maintaining clinical symptoms. Indeed, a study working from the same premise found that when a bias change is achieved, so is the change in clinical symptom (Clarke et al., 2014). This implies that there is a necessity to successfully change the bias in the first place to investigate the clinical benefit of CBM. However, a number of studies conclude that CBM does not work, despite never successfully changing the bias, in both CBM-I (Micco et al., 2014) and ABM (Arditte and Joormann, 2014; Enock et al., 2014). Therefore, it is of crucial importance to investigate the mechanisms behind changing cognitive biases in order to optimise bias-change procedures, which we do in the current study.

Theory and Modelling Study

Mathews and Mackintosh (1998) proposed that the negative interpretative biases of emotionally ambiguous expressions in high-trait anxious patients can be explained in the context of the theory of 'biased competition'. The theory of biased competition maintains that any enhancement of attention-related neuronal responses is due to competition among all of the stimuli concurrently displayed in the visual field (Desimone et al., 1990; Desimone and Duncan, 1995; Desimone, 1998). More precisely, the multiple stimuli in the visual field activate cortical neurons that mutually inhibit one another through competitive interactions. At the same time, there are top-down attentional signals from outside the visual cortex. These also influence cortical activity, such that the cells representing the attended stimulus 'win' the competition (Duncan and Humphreys, 1989; Deco and Rolls, 2005). In Mathews and Mackintosh (1998), the 'competition' is between alternate interpretations of emotionally ambiguous stimuli (e.g., sad and happy) with the outcome influenced by a top-down threat-detecting

signal from the amygdala and a cognitive control signal from the rostral anterior cingulate cortex (rACC) and lateral prefrontal cortex (LPFC) (Bishop, 2007).

While this is one of the biologically reasonable accounts of the mechanism of such biases, West et al. (2011) have recently reported that biased competition may begin as early as the primary visual cortex, and affective prioritisation can be solely driven by physical salience of the low-level features in emotional faces themselves. This implies that some degree of prioritised social signals that are already represented in the earlier visual cortex may underlie subsequent discrimination between different emotions. From a theoretical perspective, we believe it is also possible to develop training procedures to achieve CBM-I by modifying the synaptic connections between neurons in order to adjust the flow of electrical signals in the earlier cortical areas that carry information about affective representation. Therefore, the main aim of the current study is to investigate the theoretical ‘front-end’ of the competition account - before top-down signals from the amygdala-prefrontal circuitry in the later biased competition kick in - to provide deeper insight into a more accurate account that guides the development of more effective CBM-I training procedures.

Computational modelling is one useful way to investigate such mechanisms. The current study investigates the potential mechanisms of CBM-I through neural network computer modelling in order to understand how CBM might be achieved from a neurobiological perspective. More precisely, we investigated the underlying plasticity mechanisms and emergent neural dynamics using competitive neural networks, which are unsupervised in that no given activity pattern is imposed on the output neurons during training. In other words, the learning in our model is solely guided by suitable input patterns. A typical CBM-I training procedure involves ‘active training’, where a kind of feedback is provided to the participants to modulate their cognitive bias (Hoppitt et al., 2010). On the other hand, the procedure presented here describes a method of removing the bias without requiring any such feedback. We present here a set of carefully designed sequences of visual images that achieve the synaptic rewiring, which may enhance the effectiveness of the ordinal CBM-I interventions with or without ‘active training’ at the later stage of the processing.

In particular, we present computer simulations to explore two possible CBM-I training methodologies for rewriting previously learned associations. We refer back to the work of Bourke et al. (2010), aiming to change a negative interpretation of facial expressions into a positive interpretation. In order to achieve such learning without any explicit teaching signal, the new CBM methodologies utilise two previously established biologically plausible synaptic learning mechanisms known as *continuous transformation (CT) learning* (Stringer et al., 2006) and *trace learning* (Foldiak, 1991; Wallis and Rolls, 1997). These learning mechanisms are able to guide visual development by exploiting either the spatial continuity or temporal continuity between visual stimuli presented during training. We aim to explore whether both of these learning mechanisms, when combined with carefully designed

sequences of transforming face images presented to the model, will eliminate negative biases in the interpretation of facial expression, which could potentially offer a low-cost and non-invasive treatment, particularly if used in combination with other therapies (e.g., Cognitive Behavioral Therapy (CBT)).

Continuous Transformation Learning

It has been reported that people learn to associate visually similar images together. In an experimental study, Preminger et al. (2007) trained subjects to classify faces into two categories: friends (F) and non-friends (NF). Upon reaching good performance, subjects were then trained with a sequence of morphed images from F to NF. The subjects were tested on how they classified the morphed images. Initially, the first half of the morphed image sequence was classified as F, while the second half of the morphed sequence was classified as NF. However, as training progressed, the separation threshold moved towards NF; that is, an increasing number of frames were classified as F. Eventually, all morphed frames were classified as F.

Continuous transformation (CT) learning is an invariance learning mechanism that may provide an insight into the mechanism of such memory reconstruction via ordinary Hebbian learning at the neuronal level (Stringer et al., 2006). It associatively remaps the feedforward connections between successive neural layers while keeping the same initial set of output neurons activated as the input patterns are gradually changed. Consider a set of stimuli that can be arranged into a continuum, in which each successive stimulus in the continuum has a degree of overlap – a number of features in common – with the previous stimulus in the continuum. CT learning can exploit this feature overlap between successive stimuli to form a single percept of all, or at least a large subset, of the stimuli in the stimulus set.

Specifically, when an output neuron responds to one of the input patterns, the feedforward connections from the active input neurons to the active output neuron are strengthened by associative (Hebbian) learning. Then, when the next similar (overlapping) input pattern is presented, the same output neuron is again activated due to the previously strengthened connections. Now the second input pattern is associated with the same output neuron through further associative learning. This process can continue to map a sequence of many gradually transforming input patterns, where each input pattern has a degree of spatial overlap with its neighbours, onto the same output neuron. The standard Hebbian learning rule used to modify the feedforward synaptic connections at each timestep τ is

$$\delta w_{ij}^\tau = k r_i^\tau r_j^\tau \quad (1)$$

where r_j^τ is the firing rate of input neuron j at time τ , r_i^τ is the firing rate of output neuron i at time τ , δw_{ij}^τ is the change in the synaptic weight w_{ij}^τ from input neuron j to output neuron i at time τ , and k is a constant called the learning rate that governs the amount of synaptic weight change.

To prevent the same few neurons always winning the competition, the synaptic weight vector of each output neuron i is renormalized to unit length after each learning update for each training pattern by setting

$$\sqrt{\sum_j w_{ij}^2} = 1. \quad (2)$$

Neurophysiological evidence for synaptic weight normalization has been described by Royer and Paré (2003).

We hypothesised that this CT learning will eliminate negative biases in the interpretation of facial expression when combined with carefully designed sequences of transforming face images presented to the model. In particular, we will exploit the remapping capabilities of CT learning by morphing very happy faces, which are associated with a positive output representation, into neutral faces during training. This may cause the strong efferent connections from the neutral faces to be remapped to the positive output representation by associative learning operating in the feedforward connections. This should result in positive output neurons firing to both positive (happy) and neutral faces, and negative output neurons only firing to negative (sad) faces.

Trace Learning

Other psychological studies have shown that sequential presentation of the different views of an object, which produces temporal continuity, can facilitate view invariant object learning, where the different views of an object occurring close together in time are bound onto the same output representation (Perry et al., 2006). In contrast, systematically switching the identity of a visual object during such sequential presentation impairs position-invariant representations (Cox et al., 2005). Li and DiCarlo (2008) have reported a neuronal evidence of similar temporal association of visual objects that are presented close together in time. In their study, monkeys were first trained to track an object that was shifted around on a screen. In the experimental condition, the target object was swapped to a different object when the object was at a particular retinal location for the monkeys. As a result, individual neurons in IT that were originally selective to the target object started to respond also to the different object at the specific retinal location. These results show that the temporal statistics of object presentations should play a key role in the development of transform-invariant object representations in the visual brain.

Trace learning is a biologically plausible mechanism to achieve such temporal association by incorporating a memory trace of recent neuronal activity into the learning rule used to modify the feedforward synaptic connections (Foldiak, 1991; Wallis and Rolls, 1997). This encourages output neurons to learn to respond to input patterns that occur close together in time. Stimuli that are experienced close together in time are likely to be strongly related; for instance, successive stimuli could be different views of the same object. If a mechanism exists to associate together stimuli that

tend to occur close together in time, then a network will learn that those stimuli form a single percept. Trace learning provides one such mechanism by incorporating a temporal memory trace of postsynaptic cell activity \bar{r}_i into a standard Hebbian learning rule. In this paper, the form of trace learning rule implemented at each timestep τ is

$$\delta w_{ij}^\tau = k \bar{r}_i^{\tau-1} r_j^\tau \quad (3)$$

where r_j^τ is the firing rate of presynaptic neuron j at time τ , $\bar{r}_i^{\tau-1}$ is the trace of postsynaptic neuron i at time $\tau - 1$, δw_{ij}^τ is the change in the synaptic weight w_{ij}^τ from presynaptic neuron j to postsynaptic neuron i at time τ , and k is the learning rate. The trace term is updated at each time step according to

$$\bar{r}_i^\tau = (1 - \eta) \bar{r}_i^{\tau-1} + \eta r_i^\tau. \quad (4)$$

where η is a parameter anywhere in the interval $[0, 1]$ that controls the relative balance in the trace term \bar{r}_i^τ of the current postsynaptic cell firing rate, r_i^τ , and the previous trace of postsynaptic cell firing, $\bar{r}_i^{\tau-1}$. For the simulations described below, η was set to 0.8. The synaptic weight vector of each output neuron i is renormalized to unit length according to Equation (2) after each learning update for each training pattern.

We propose that such innate trace learning mechanisms may also be exploited to eliminate negative biases in the interpretation of facial expression when combined with carefully designed sequences of transforming face images presented to the model. In particular, if during training with a trace learning rule, a neutral face is presented in temporal proximity with many other very happy faces that are associated with a positive output representation, then this should encourage these positive output neurons to learn to respond to the neutral face as well. When the neutral face is subsequently presented, the positive output representation should suppress the negative output representation by competition mediated by inhibitory interneurons. By implementing a trace learning rule and presenting the network with occasional neutral faces amongst many happy faces, we expect to see positive output neurons learning to respond to both positive and neutral faces.

Overview of Simulation Studies Carried Out in this Paper

We first describe simulations with a simplified one-layer neural network architecture in order to test the two hypothesised CBM learning mechanisms in a highly controlled manner in the section describing Experiment 1. This is an important step to take to clearly illustrate the exact underlying mechanisms of CBM in as simple a model as possible. Then, we present simulation results in which realistic face stimuli are used to train a more biologically detailed multi-layer neural network computer model, VisNet, of the ventral visual pathway in the primate brain (Wallis and Rolls, 1997), which has

recently been used to show how the visual system may learn to represent facial expressions (Tromans et al., 2011; Eguchi et al., 2016) in the section describing Experiment 2.

In both sections, we extend these previous modelling studies involving synaptic plasticity and learning to the problem of understanding the neurobiological basis of CBM training by both CT learning (Experiment 1a and 2a) and trace learning (Experiment 1b and 2b). Specifically, we show that both of these learning mechanisms can be used to eliminate negative biases in the interpretation of facial expression. That is, a subpopulation of ‘sad’ output neurons that initially responds to both sad and neutral faces before learning will only respond to the sad faces after CBM training. On the other hand, a subpopulation of ‘happy’ output neurons that initially responds to just happy faces before learning will respond to both happy and neutral faces after training.

Experiment 1: One-Layer Network

In this section, we aim to demonstrate how CT learning and trace learning may each be used to carry out CBM within a one-layer competitive neural network. These simulations used a highly idealised network architecture and input stimulus representations in order to provide a very controlled way of investigating and testing the underlying computational hypotheses described in the sections in the Introduction above.

In particular, we show how the responses of a one-layer competitive neural network may be remapped, through CBM training, from a negatively biased state to an unbiased state. We first demonstrate the remapping using CT learning in Experiment 1a, then we demonstrate the remapping using trace learning in Experiment 1b.

One-Layer Model Description

The network architecture and activation equations are common to the models described in the sections below about Experiment 1a and 1b. The network, depicted in Figure 1a, comprises a single layer of input cells which drive activity in a layer of two output cells through feedforward synapses. The output neurons compete with each other so that only one such neuron can remain active at a time when an input pattern is presented to the network. In the brain, such competition between neurons within a layer is implemented by inhibitory interneurons.

We describe this architecture as a one-layer network because there is only a single layer of synapses in the model. The 1-dimensional layer of input cells provide a highly idealised representation of facial expressions ranging continuously from happy to sad. In the simulations, the input neurons have binarised (0/1) firing rates. Each input neuron responds selectively to a small localised region of the unidimensional space of facial expressions, with the entire space of expressions from happy to sad covered by the input layer. Consequently, the input layer represents each facial expression of a

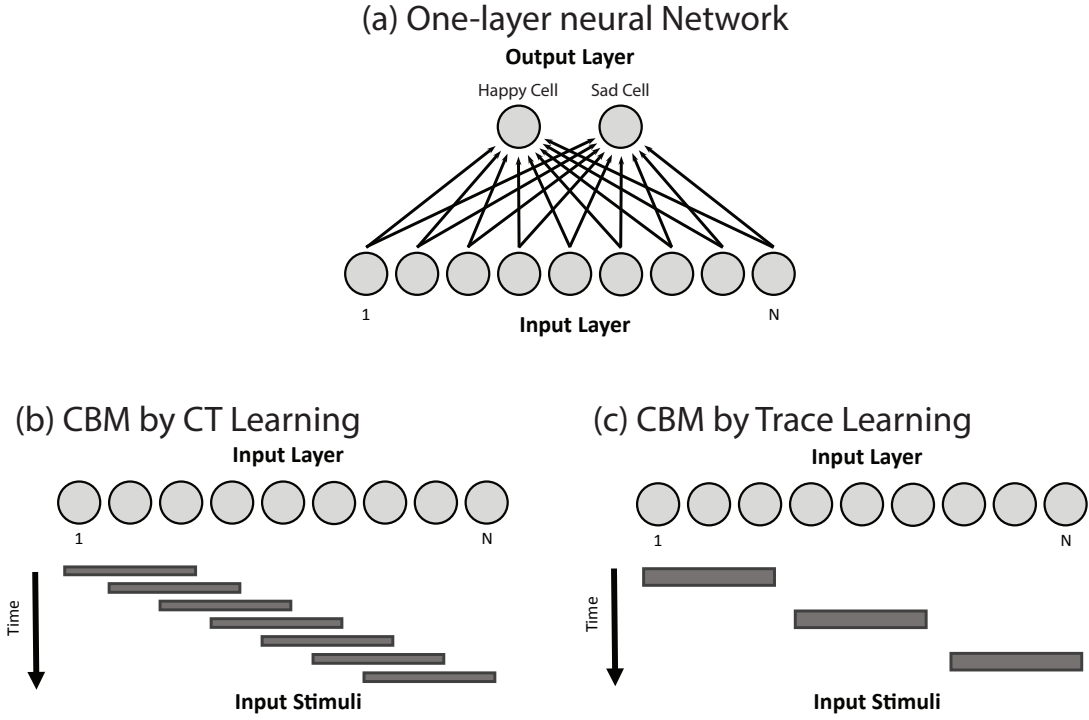


Figure 1. (a) The one-layer neural network architecture used in the models described in the sections about Experiment 1. A single layer of N input cells drives activity in the two output cells through the feedforward synapses (black arrows). The input layer cells respond to stimuli that range from happy to sad, with different simulations requiring different numbers of input cells, as detailed in the sections about Experiment 1a and 1b. There are always two output cells in the network, with the left output cell responding to happy stimuli, and the right output cell responding to sad stimuli. (b) The training protocol for the one-layer network trained by CT learning. The input layer contains a total of $N = 600$ neurons. During each training step, the current input stimulus is represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal grey lines. We refer to the length of the stimulus as its *stride*, which was set to be 100 input neurons. The firing rates of all other input cells are set to 0. At each successive training step, the input stimulus is advanced by 1 input cell in order to ensure that successive stimuli are varied in a continuous manner, i.e. successive stimuli overlap with each other, which is a requirement of CT learning. During each training epoch, the input stimuli are shifted once through the whole continuum from happy to sad. (c) The training protocol for the one-layer network trained by trace learning. The input layer contains a total of $N = 900$ neurons. During each training step, the current input stimulus is represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal grey lines. The length of each stimulus, its stride, was set to be 100 input neurons. The firing rates of all other input cells are set to 0. In order to prevent CT-like learning effects from occurring, the input stimuli do not overlap with each other. During training, the most happy input stimuli are closely interleaved with more neutral input stimuli from the middle of the stimulus range, while the most sad stimuli are shown without temporal interleaving with the neutral stimuli. This stimulus presentation order enables trace learning to associate together the happy and neutral stimuli onto the same happy output cell.

particular emotional valence by the co-activation of a localised cluster of input neurons at the corresponding position within the layer.

At the beginning of the simulation, the feedforward synaptic connection weights are initialised such that the left output cell (happy output cell) responds to happy stimuli, and the right output cell (sad output cell) responds to sad stimuli. A negative cognitive bias can be introduced in the network by initialising the synaptic connections such that the more neutral input stimuli are initially responded to by the sad output neuron rather than the happy output neuron. Then, by modifying the strengths of the feedforward synaptic weights from the input cells to the output cells through CBM training, it is possible to alter the response characteristics of the output neurons in the network. In particular, we show that CBM training by either CT learning or trace learning can shift the network away from a negative bias to a situation in which the happy output cell responds to the majority of the input stimuli including both happy and more neutral stimuli.

At each timestep during simulation of the network, an input stimulus of a particular emotional valence is selected to be presented to the network. During CBM training, the input stimuli are presented in accordance with the spatio-temporal statistics required by either CT learning or trace learning, as described below in the sections about Experiment 1a and 1b respectively. Then the input cell firing rates, r_j , are set to be either 0 or 1 according to the training and testing protocols described in the sections about Method for Experiment 1a and 1b. The output cell firing rates, r_i , are calculated by setting the activation level, h_i , of each output cell i to

$$h_i = \sum_j w_{ij} r_j \quad (5)$$

where w_{ij} is the synapse from presynaptic input cell j to postsynaptic output cell i , and the sum is taken over all presynaptic input cells j . The output cell firing rates are then set by applying winner-take-all inhibition so that the output cell with the highest activation level is given a firing rate of 1 and the other output cell is given a firing rate of 0.

During CBM training, after the firing rates of the output cells have been computed, the synaptic weights are then updated by either the Hebbian learning rule (1) in Experiment 1a or the trace learning rule (3) in Experiment 1b.

Initial Setup of the Network

Before the network undergoes CBM training, the feedforward synaptic weights to the sad and happy output cells are set manually to control whether or not there is a pre-existing cognitive bias.

In order to establish the synaptic connectivity without an initial bias, the synaptic weights to the sad output cell, $w_{\text{SAD}j}$, are set so that

$$w_{\text{SAD}j} = \frac{1}{1 + \exp(-2\beta(\epsilon_j - \alpha))} \quad (6)$$

The parameter $\epsilon_j \in [-3, +3]$ represents the preferred stimulus location of input cell j within the sad to happy continuum, with most sad $= -3$ and most happy $= +3$. The input neurons are distributed evenly throughout the sad to happy stimulus continuum. The slope β is set to an appropriate value (described in Table 1a), and the threshold α is set to 0. The synaptic weights to the happy output cell, $w_{\text{HAPPY}j}$, are set to be

$$w_{\text{HAPPY}j} = 1 - w_{\text{SAD}j}. \quad (7)$$

The effect of setting the weights in this manner is that all input cells send feedforward synaptic weights to both of the output cells, but the sad output cell receives stronger synaptic weights from the input cells representing the sad end of the input continuum, and the happy output cell receives stronger

synaptic weights from the input cells representing the happy end of the input continuum. In particular, with $\alpha = 0$, the feedforward synaptic connections are unbiased in that the happy output cell and sad output cell receive mirror-symmetric distributions of afferent synaptic connections covering the entire stimulus space. This can be seen in the left plot of Figure 2a for the first simulation with CT learning (Experiment 1a) and Figure 2d for the second simulation with Trace learning (Experiment 1b).

In order to introduce a negative bias in the synaptic weights such that the sad output cell will also respond to most of the middle, more neutral, portion of the input continuum, the synaptic weights from the input cells to the sad output cell are set according to Equation (6) with the threshold α set to a negative value (described in Table 1a for Experiment 1a and Table 1b for Experiment 1b). The synaptic weights from the input cells to the happy output cell are then set according to Equation (7). As can be seen in the left plot of Figure 2b for the first simulation (Experiment 1a) and Figure 2e for the second simulation (Experiment 1b), this results in the sad output cell receiving stronger synaptic weights from a greater proportion of the input cells than the happy output cell does.

Experiment 1a: CBM by CT Learning

In this section, we simulate CBM in the one-layer network by the continuous transformation (CT) learning mechanism described in the Introduction. It associatively remaps the feedforward connections between successive neural layers while keeping the same initial set of output neurons activated as the input patterns are gradually changed. We will exploit this mechanism by morphing happy input stimuli, which are strongly associated with the positive output representation, i.e. the happy output neuron, into more neutral stimuli during training. This causes the efferent connections from the neutral stimuli to be remapped to the positive output representation by associative learning operating in the feedforward connections. When the neutral stimuli are presented again after training, the positive output representation should respond and also suppress the negative output representation by competition mediated by inhibitory interneurons.

Method. Figure 1b shows the setup for training the one-layer network with CT learning. The input layer contains a total of $N = 600$ neurons. The layer of input cells represent a continuum of facial expressions from happy (left) to sad (right). The input stimulus presented to the network at any given training step is represented by the firing rates of a contiguous subblock of input cells being set to 1, as illustrated by the horizontal grey lines in Figure 1b. We refer to the length of the stimulus as its *stride*, which was set to be 100 input neurons. The firing rates of all other input cells are set to 0. In this simulation with CT learning, the Hebb learning rule (1) is used. Since the Hebb learning rule does not contain a memory trace of previous neuronal activity, this ensures that any observed bias modification is the result of CT learning and not the result of trace learning.

During training of the network, illustrated in Figure 1b, the input stimulus is moved

Table 1
Parameters used in the simulations

| Parameter | Value | | | |
|---------------------------------------|---------------------------------|------------------|------------------|------------------|
| (a) One-Layer Network (CT) | | | | |
| No. of Input Cells N | 600 | | | |
| Stride | 100 | | | |
| Sigmoid Slope (β) | 0.5 | | | |
| Biased Sigmoid Threshold (α) | -1 | | | |
| Learning Rate (k) | 0.001 | | | |
| Training Epochs | 100 | | | |
| (b) One-Layer Network (Trace) | | | | |
| No. of Input Cells N | 9 | | | |
| Stride | 100 | | | |
| Sigmoid Slope (β) | 0.5 | | | |
| Biased Sigmoid Threshold (α) | -1 | | | |
| Learning Rate | 0.01 | | | |
| Eta (η) | 0.8 | | | |
| Training Epochs | 100 | | | |
| (c) VisNet | | | | |
| Gabor: Phase shift (ψ) | 0, π | | | |
| Gabor: Wavelength(λ) | 2 | | | |
| Gabor: Orientation(θ) | 0, $\pi/4$, $\pi/2$, $3\pi/4$ | | | |
| Gabor: Spatial bandwidth (b) | 1.5 octaves | | | |
| Gabor: Aspect ratio (γ) | 0.5 | | | |
| No. of Layers | 4 | | | |
| Retina | $256 \times 256 \times 16$ | | | |
| | 1st layer | 2nd layer | 3rd layer | 4th layer |
| Dimension | 128×128 | 128×128 | 128×128 | 128×128 |
| Num. of fan-in connections | 201 | 100 | 100 | 100 |
| Fan-in radius | 24 | 24 | 36 | 48 |
| Sparseness of activations | 1 % | 44 % | 32 % | 25 % |
| Sigmoid slope (β) | 15 | 99 | 146 | 207 |
| Learning rate (k) | 1.0 | 1.0 | 1.0 | 1.0 |
| Training Epochs | 20 | 20 | 20 | 20 |
| Excitatory Radius (σ_E) | 1.4 | 1.1 | 0.8 | 1.2 |
| Excitatory Contrast (δ_E) | 5.35 | 33.15 | 117.57 | 120.12 |
| Inhibitory Radius (σ_I) | 4.94 | 13.88 | 9.72 | 14.80 |
| Inhibitory Contrast (δ_I) | 1.5 | 1.5 | 1.6 | 1.4 |

continuously through the layer of input cells, advancing one input cell per learning update of the network. At each stimulus presentation, the activations of the output neurons are first updated according to Equation (5), the firing rates of the output neurons are then computed using winner-take-all competition, and then the feedforward synaptic weights are modified according to Equations (1) and (2). One epoch of training is completed after the input stimulus has been shifted through the whole continuum from happy to sad. Upon reaching the specified number of training epochs, the training phase is finished and the testing phase begins, which follows the same protocol as the training phase with the exception that the weight update and normalization equations, Equations (1) and (2), are not simulated. The simulation is then complete. A one-layer neural network model was simulated with the parameters given in Table 1a.

Results. First, the network was simulated with the synaptic weights initially hardwired to unbiased values according to Equations (6) and (7) with the threshold α set to 0. Next, the network was simulated with a negative cognitive bias introduced by hardwiring the synaptic weights according

to Equations (6) and (7) with the threshold α set to -1 . This ensured that the sad output cell responded not only to very sad stimuli but also to the majority of the more neutral stimuli. In the final simulation, the negative bias in the previous biased network was eliminated by CBM training using CT learning. This had the effect of remapping the feedforward synaptic weights so that the happy output cell took over responding to the majority of the neutral stimuli.

Untrained Network Performance (Before and After Biases are Added). The network was simulated with the synaptic weights initially hardwired to unbiased values. The left plot of Figure 2a shows the unbiased weights from the input cells to the output cells. The sad output cell receives the strongest synaptic weights from the input cells representing the sad end of the stimulus continuum, and the happy output cell receives the strongest synaptic weights from the input cells representing the happy end of the stimulus continuum. The two output cells receive equal, albeit mirror symmetric, distributions of synaptic weights from the input cells representing the middle, more neutral, portion of the stimulus continuum. The right plot of Figure 2a shows the firing rates of the two output cells in response to presentation of the input stimuli. The happy output cell responds strongly to very happy input stimuli, the sad output cell responds strongly to very sad input stimuli, and most importantly both output cells respond to equal sized regions of the more neutral intermediate input stimuli. These responses are to be expected given the unbiased feedforward synaptic weight profiles between the input cells and the output cells.

The network was simulated with a negative cognitive bias introduced by hardwiring the synaptic weights. The left plot of Figure 2b shows the synaptic weights after a bias has been applied. The sad output cell receives stronger synaptic weights from the sad end of the input range and most of the more neutral input cells, and the happy output cell now receives stronger synaptic weights from only the input cells representing the happy end of the input continuum. The effect of this bias is that the sad output cell now responds not only to very sad stimuli but also to the majority of the more neutral stimuli, whereas the happy output cell does not. This can be seen in the right plot of Figure 2b.

Learned (Remapped) Network Performance. The negative bias in the previous biased network was eliminated by CBM training using CT learning. After CT learning, the synaptic weights should remap such that the happy output cell now receives stronger synaptic weights from the input cells representing a larger portion of the intermediate, more neutral, stimuli than the sad output cell. The effect of this learned remapping is that the happy output cell responds to a greater proportion of the input stimulus space than the sad output cell does. That is, the happy output cell now responds to the majority of the intermediate neutral stimuli. This can be seen in the right plot of Figure 2c (c.f. the right plot of Figure 2b). This represents CBM, where the bias in the network has been shifted from negative to positive by CT learning.

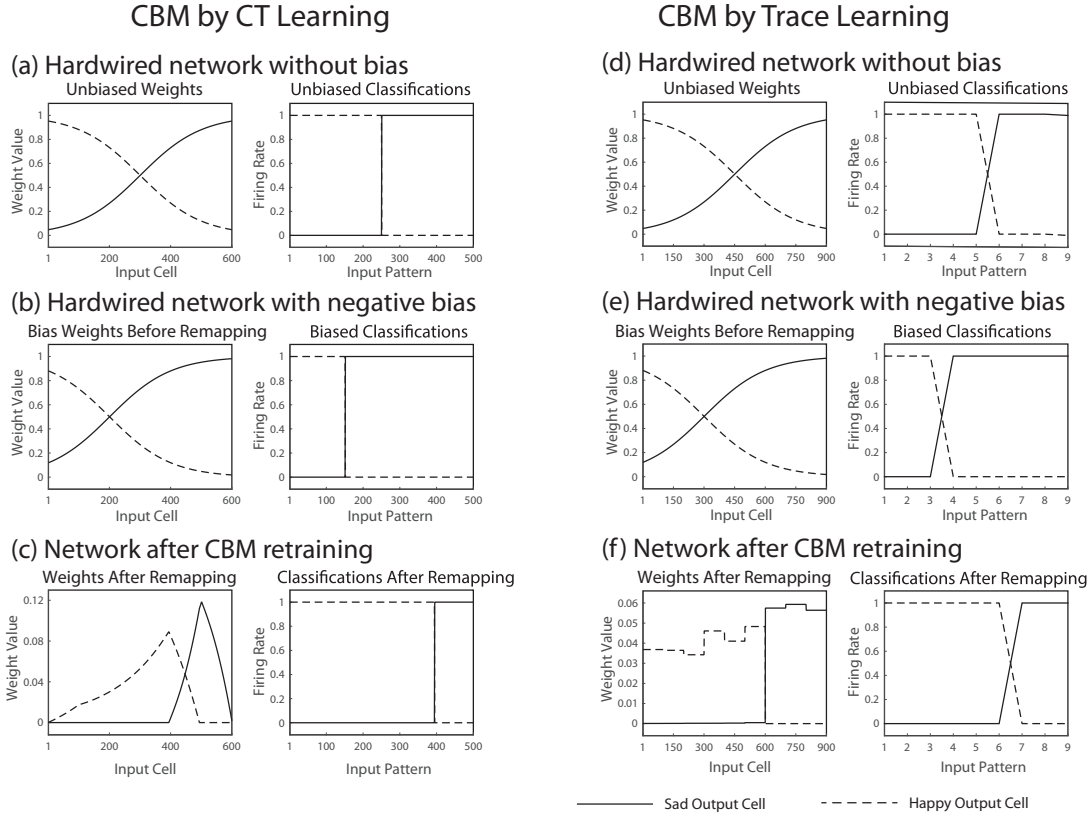


Figure 2. Demonstration of CBM in a one-layer network using CT learning (a-c) and trace learning (d-f) to remap the synaptic weights. The figure shows the feedforward synaptic weights (left column) and firing rates of the output cells (right column) at various stages of the simulation. **(a,d)** Results of testing the initial, unbiased hardwired network. The lack of bias in the synaptic weights results in the happy and sad output cells responding to equal numbers of the input patterns. **(b,e)** Results of testing the biased hardwired network. After the negative bias is introduced to the synaptic weights, the sad output cell now responds to the majority of the input patterns. **(c,f)** Results of testing the network after remapping the synaptic weights through CBM training with CT learning (c) and with trace learning (f). The learning has effected a remap in the synaptic weights such that the happy output cell now has stronger synaptic weights from the majority of the input cells. The effect of this remapping is that the happy output cell now responds not only to the most happy stimuli but also to the majority of the more neutral input patterns.

Experiment 1b: CBM by Trace Learning

Having shown how CBM may be accomplished through CT learning, we now show how it may also be accomplished using a different learning paradigm: trace learning. In this section, we simulate CBM in the one-layer network by the trace learning mechanism described in the Introduction. Trace learning is an invariance learning mechanism which utilises a trace learning rule, Equation (3) with weight vector normalisation Equation (2) to modify the feedforward synaptic connections. Trace learning incorporates a memory trace $\bar{r}_i^{\tau-1}$ of recent neuronal activity into the learning rule used to modify the feedforward synaptic connections. This encourages output neurons to learn to respond to input patterns that occur close together in time. If, during training, a neutral stimulus is presented in temporal proximity with many other very happy stimuli that are associated with the positive output representation, i.e. the happy output neuron, then this should encourage the positive output representation to respond to the neutral stimulus as well. When the neutral stimulus is subsequently presented, the positive output representation should suppress the negative output representation by competition, which in the brain is mediated by inhibitory interneurons.

Method. The setup for training the one-layer network with trace learning is shown in Figure 1c. The input layer contains $N = 900$ neurons. The input layer represents a range of facial expressions from happy (left) to sad (right). Each input stimulus shown to the network is represented by the firing rates of a contiguous subblock of input cells being set to 1, as illustrated by the horizontal grey lines in Figure 1c. The length of each stimulus presented to the network was set to be 100 input neurons, while the firing rates of all other input cells were set to 0.

In contrast to the training protocol used for the above simulations with CT learning described in the section describing Method for Experiment 1a, the input stimuli used for trace learning in this section do not overlap as they advance through the input space. This prevents any CT-like learning effects from occurring, and so ensures that any bias modification that occurs is the result of trace learning and not the result of CT learning. The training protocol with trace learning is shown in Figure 1c.

During training of the network, illustrated in Figure 1c, the input stimuli are divided into two separate groups: one group containing stimuli from the most happy and more neutral (middle) parts of the input stimulus range; and one group containing only stimuli from the sad end of the input stimulus range. During an epoch of training, one of the two stimulus groups is selected at random. If the stimulus group contains only the sad stimuli, these stimuli are shown to the network in a random order. If the stimulus group contains both the happy and more neutral stimuli, then the happy stimuli are interleaved with the neutral stimuli such that a happy stimulus is shown followed by a neutral stimulus, but with these stimuli paired in a random order. After presentation of the first group of stimuli (happy/neutral, or sad), the second group of stimuli is shown to the network. During the presentation of each stimulus, the activations of the output neurons are updated by Equation (5), the firing rates of the output neurons are then computed according to winner-take-all competition, and the synaptic weights are then updated according to the trace learning rule Equation (3) with weight vector normalisation Equation (2). After all stimuli have been presented, an epoch of training is complete and the next epoch of training begins. The order of the stimulus groups, and the order of stimulus presentation within the group, are randomly selected for each training epoch. Upon reaching the specified number of epochs, the training phase is finished and the testing phase begins, during which the input stimuli are presented one at a time to the network, ranging from happy to sad. The weight update and normalization equations, Equations (3) and (2), are not simulated during the testing phase. After the testing phase, the simulation is complete. A one-layer neural network model was simulated with the parameters given in Table 1b.

Results. The network was first simulated with the synaptic weights manually set to unbiased values according to Equations (6) and (7) with $\alpha = 0$. Next, the network was simulated with a negative bias introduced by hardwiring the synaptic weights according to Equations (6) and (7) with $\alpha = -1$.

This caused the sad output neuron to respond to most of the more neutral stimuli in addition to the sad stimuli. Lastly, the negative bias in the previous network was eliminated by CBM training using trace learning. This resulted in the happy output neuron now responding to most of the neutral stimuli as well as the happy stimuli.

Untrained Network Performance (Before and After Biases are Added). The network was simulated with unbiased hardwired synaptic weights. Figure 2d (left) shows the unbiased synaptic weights. The sad output cell receives the strongest synaptic weights from the sad end of the stimulus range, while the happy output cell receives the strongest synaptic weights from the happy end of the stimulus range. The two output cells receive equal, albeit mirror symmetric, distributions of synaptic weights from the intermediate neutral portion of the stimulus continuum. Figure 2d (right) shows the firing rate responses of the two output cells to the full range of input stimuli. The happy output cell responds to happy stimuli, the sad output cell responds to sad stimuli, while both output cells respond to equal numbers of the more neutral intermediate stimuli.

The network was then simulated with a negative cognitive bias introduced by hardwiring the synaptic weights. Figure 2e (left) shows the synaptic weights. The sad output cell receives stronger synaptic weights from the sad end of the input range and most of the more neutral input cells, while the happy output cell receives stronger synaptic weights from only the happy end of the input range. Figure 2e (right) shows the firing rate responses of the two output neurons to the full range of input stimuli. Due to the biased synaptic weights, the sad output cell responds to the majority of the more neutral stimuli in addition to the sad stimuli, whereas the happy output cell only responds to the more happy stimuli.

Learned (Remapped) Network Performance. The negative bias in the previous biased network was eliminated by CBM training using trace learning. After trace learning, the feedforward synaptic weights remap so that the happy output neuron receives stronger synaptic weights from input neurons representing the happy stimuli and the majority of the more neutral stimuli, while the sad output cell receives strong synaptic weights only from the sad end of the input stimulus range. This can be seen in the left plot of Figure 2f. The effect of this remapping is that the happy output cell now responds to stimuli from the happy to middle neutral region of the input stimulus range, while the sad output cell responds only to stimuli from the sad end of the input stimulus range, which can be seen in the right plot of Figure 2f. Thus, trace learning has produced CBM, where the bias in the network has been shifted from negative to positive.

Experiment 2: VisNet Simulation

In this section, we test computational hypotheses described in the Introduction using realistic face stimuli presented to an established, biologically detailed, hierarchical neural network model,

VisNet, of the primate ventral visual pathway (Wallis and Rolls, 1997; Stringer et al., 2006). The simulations with VisNet were carried out in two stages of training as follows.

In the first training stage, VisNet was trained on a set of randomised computer generated face images, where the identity and expression of each face was chosen randomly. Eguchi et al. (2016) have reported that this led to the development of separate sub-populations of output neurons that responded selectively to either facial identity or expression. Such neurons have been experimentally observed in single unit recording neurophysiology studies on the primate brain (Hasselmo et al., 1989).

The second stage of training involved CBM by either CT learning or trace learning, similar to that described above for the one-layer network. Specifically, we tested whether the initial negative bias in the synaptic connectivity developed in the pretraining could be shifted from sad to happy after CBM retraining on new, specially designed sequences of face images. In these second stage simulations, the sequences of face images used for CBM retraining were constructed in accordance with the spatio-temporal stimulus statistics required by either the CT learning (Experiment 2a) or trace learning hypotheses (Experiment 2b).

VisNet Model Description

The simulation studies presented below are conducted with a hierarchical neural network model of the primate ventral visual pathway, VisNet, which was originally developed by Wallis and Rolls (1997). The standard network architecture is shown in Figure 3a. It is based on the following: (i) A series of hierarchical competitive networks with local graded lateral inhibition. (ii) Convergent connections to each neuron from a topologically corresponding region of the preceding layer, leading to an increase in the receptive field size of neurons through the visual processing areas. (iii) Synaptic plasticity based on a local associative learning rule such as the Hebb rule or trace rule.

In past work, the hierarchical series of 4 neuronal layers of VisNet have been related to the following successive stages of processing in the ventral visual pathway: V2, V4, the posterior inferior temporal cortex, and the anterior inferior temporal cortex. However, this correspondence has always been quite loose because the ventral pathway may be further subdivided into a more fine grained network of distinct sub-regions.

Each layer consists of 128×128 cells, and the forward connections to individual cells are derived from a topologically corresponding region of the preceding layer, using a Gaussian distribution of connection probabilities. These distributions are defined by a radius which will contain approximately 67% of the connections from the preceding layer. The values used in the current studies are given in Table 1c. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman and Simoncelli, 2011; Pasupathy, 2006; Pettet and Gilbert, 1992).

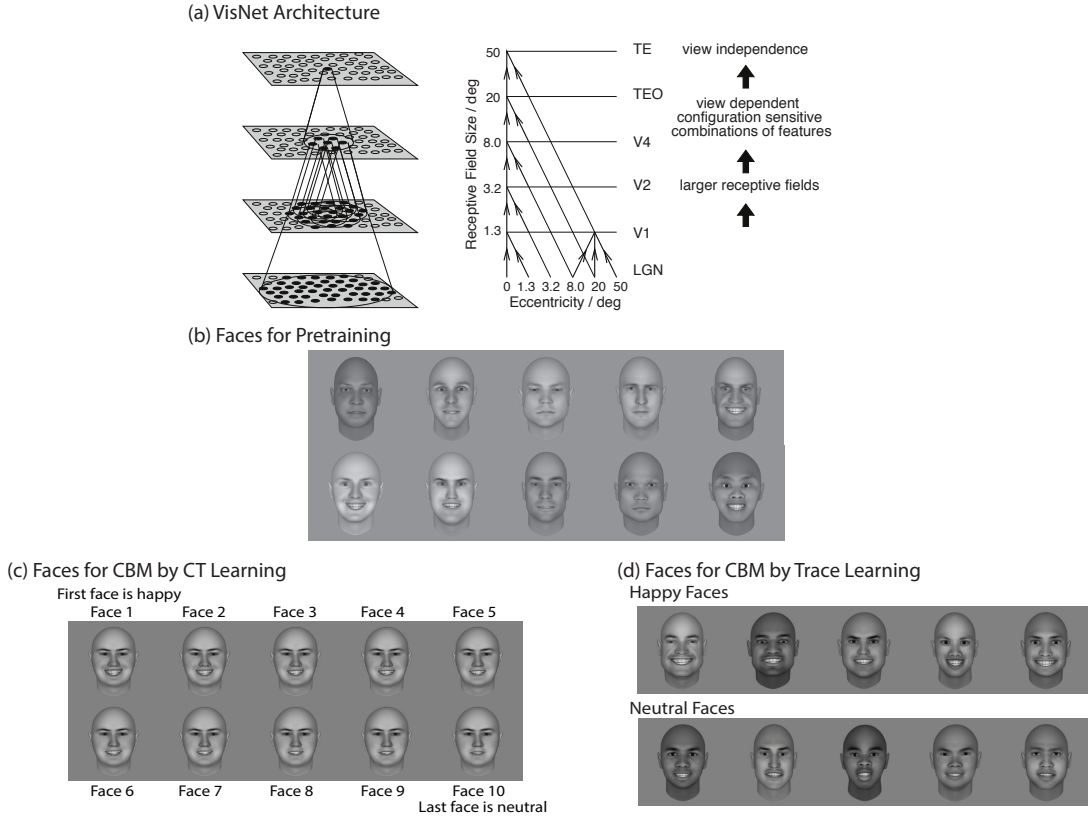


Figure 3. (a) Left: Stylised image of the four layer VisNet architecture. Convergence through the network is designed to provide fourth layer neurons with information from across the entire input retina. Right: Convergence in the visual system V1: visual cortex area V1; TEO posterior inferior temporal cortex, TE inferior temporal cortex (IT) (b) Examples of the face stimuli used to pretrain VisNet. 100 realistic human faces were randomly generated with different identities, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. (c) Examples of the face stimuli used to perform CBM retraining on VisNet through CT learning. The image set is constructed from 5 different facial identities. For each of these facial identities, ten face images were constructed by sampling ten evenly-spaced expressions between happy and neutral. (d) Examples of the face stimuli used to perform CBM retraining on VisNet through trace learning. The image set consisted of 25 faces with a happy expression and 25 faces with a neutral expression. Each of these 50 faces had a different randomly generated identity. The figure presents some examples of these images.

During training with visual objects, the strengths of the feedforward synaptic connections between successive neuronal layers are modified by biologically plausible local learning rules, where the change in the strength of a synapse depends on the current or recent activities of the pre- and post-synaptic neurons. A variety of such learning rules, in this case both Hebbian learning (Equation (1)) and trace learning (Equation (3)), may be implemented with different learning properties.

Pre-processing of the visual input by Gabor filters. Before the visual images are presented to VisNet's input layer 1, they are pre-processed by a set of input filters that accord with the general tuning profiles of simple cells in V1. The filters provide a unique pattern of filter outputs for each image, which is passed through to the first layer of VisNet. In this paper, the input filters used are Gabor filters. These filters are known to provide a good fit to the firing properties of V1 simple cells, which respond to local oriented bars and edges within the visual field (Jones and Palmer, 1987; Cumming and Parker, 1999). The input filters used are computed by the following equations:

$$g(x, y, \lambda, \theta, \psi, b, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \quad (8)$$

with the following definitions:

$$\begin{aligned} x' &= x \cos \theta + y \sin \theta \\ y' &= -x \sin \theta + y \cos \theta \\ \sigma &= \frac{\lambda(2^b+1)}{\pi(2^b-1)} \sqrt{\frac{\ln 2}{2}} \end{aligned} \tag{9}$$

where x and y specify the position of a light impulse in the visual field (Petkov and Kruizinga, 1997). The parameter λ is the wavelength ($1/\lambda$ is the spatial frequency), σ controls number of such periods inside the Gaussian window based on λ and spatial bandwidth b , θ defines the orientation of the feature, ψ defines the phase, and γ sets the aspect ratio that determines the shape of the receptive field. In the experiments in this paper, an array of Gabor filters is generated at each of 256×256 retinal locations with the parameters given in Table 1c.

The outputs of the Gabor filters are passed to the neurons in layer 1 of VisNet according to the synaptic connectivity given in Table 1c. That is, each layer 1 neuron receives connections from 201 randomly chosen Gabor filters localised within a topologically corresponding region of the retina.

Calculation of cell activations within the network. Within each of the neural layers 1 to 4 of the network, the activation h_i of each neuron i is set equal to a linear sum of the inputs r_j from afferent neurons j in the preceding layer weighted by the synaptic weights w_{ij} according to Equation (5).

Self-organising map. In this paper, we have run simulations with a self-organising map (SOM) (Von der Malsburg, 1973; Kohonen, 1982) implemented within each layer. In the SOM architecture, short-range excitation and long-range inhibition are combined to form a Mexican-hat spatial profile and is constructed as a difference of two Gaussians as follows:

$$I_{a,b} = -\delta_I \exp\left(-\frac{a^2 + b^2}{\sigma_I^2}\right) + \delta_E \exp\left(-\frac{a^2 + b^2}{\sigma_E^2}\right) \tag{10}$$

Here, to implement the SOM, the activations h_i of neurons within a layer are convolved with a spatial filter, I_{ab} , where δ_I controls the inhibitory contrast and δ_E controls the excitatory contrast. The width of the inhibitory radius is controlled by σ_I while the width of the excitatory radius is controlled by σ_E . The parameters a and b index the distance away from the centre of the filter. The lateral inhibition and excitation parameters used in the SOM architecture are given in Table 1c.

Contrast enhancement of neuronal firing rates within each layer. Next, the contrast between the activities of neurons with each layer is enhanced by passing the activations of the neurons through a sigmoid transfer function as follows:

$$r = f^{sigmoid}(h') = \frac{1}{1 + \exp(-2\beta(h' - \alpha))} \tag{11}$$

where h' is the activation after applying the SOM filter, r is the firing rate after contrast enhancement, and α and β are the sigmoid threshold and slope respectively. The parameters α and β are constant within each layer although α is adjusted within each layer of neurons to control the sparseness of the firing rates. For example, to set the sparseness to 4%, the threshold is set to the value of the 96th percentile point of the activations within the layer. The parameters for the sigmoid activation function are shown in Table 1c. These are general robust values found to operate well. They are similar to the standard VisNet sigmoid parameter values that were previously optimised to provide reliable performance (Stringer et al., 2006, 2007; Stringer and Rolls, 2008).

Information analysis. A single cell information measure was applied to the trained network of Eguchi et al. (2016) in order to identify the different subpopulations of output (4th layer) neurons that responded selectively to either happy faces or sad faces regardless of facial identity. Full details on the application of this measure to VisNet are given by Rolls and Milward (2000). In particular, the magnitude of the information measure reflects the extent to which a neuron responds selectively to a particular stimulus category such as a happy or sad expression, but also responds invariantly to different examples from that category such as different face identities.

The single cell information measure is applied to individual cells in layer 4, and measures how much information is available from the response of a single cell about which stimulus category, i.e. a happy expression or a sad expression, was shown. For each cell, the single cell information measure used was the maximum amount of information a cell conveyed about any one stimulus category. This is computed using the following formula with details given by Rolls et al. (1997) and Rolls and Milward (2000). The stimulus-specific information $I(s, R)$ is the amount of information the set of responses R has about a specific stimulus category s , and is given by

$$I(s, R) = \sum_{r \in R} P(r|s) \log_2 \frac{P(r|s)}{P(r)} \quad (12)$$

where r is an individual response from the set of responses R .

The maximum amount of information that can be attained is $\log_2(N)$ bits, where N is the number of stimulus categories. For the case of two stimulus categories, i.e. happy and sad expressions, the maximum amount of information is 1 bit.

Pretraining VisNet

In the first stage of the simulations, VisNet was pretrained on a set of 100 randomised computer generated face images, which were created using the software package FaceGen. FaceGen allows for controlled production of realistic face stimuli, developed from a series of photographs of real people. The faces were randomly generated with different identities, and the expressions of individual faces were also randomly set along a continuous dimension between happy and sad. Examples of these face

images are shown in Figure 3b.

The pretraining stage was carried out using the Hebbian learning rule (1) with weight vector normalisation (2). The presentation of the 100 randomised faces constituted one epoch of training, and the network was trained for a total of 20 training epochs during this stage.

The network was then tested by presenting 100 happy faces all with different facial identities, and then presenting 100 sad faces with different facial identities. For each presentation of a face, the firing rates of all of the output neurons were recorded. Information analysis was then used to identify whether any output neurons carried high levels of information about facial expression. That is, whether these neurons had learned to respond to either happy expressions regardless of identity, or sad expressions regardless of identity.

Figure 4b shows the single cell information carried by all output (4th layer) neurons before and after pretraining on the randomised face images. The plot shows the information carried by the 4th layer neurons about either happy or sad expressions, where the neurons are plotted in rank order along the abscissa. The maximum amount of information possible for the simulation is $\log_2(N)$ bits where N is the number of categories (Happy or Sad), that is 1 bit. The dashed line represents the untrained network while the solid line represents the trained network. The result shows that pretraining VisNet on many randomly generated faces has significantly increased the amount of single cell information carried by 4th layer neurons about the facial expression as originally reported in Eguchi et al. (2016).

These computed information values enabled us to identify two different subpopulations of output neurons that had learned to respond to either happy or sad expressions regardless of facial identity.

Figure 4c shows the response profiles of five Happy output neurons and five Sad output neurons recorded in response to the matrix of test faces shown in Figure 4a directly after the initial stage of pretraining (solid line). The plots show the average firing rate of the cells in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. These neurons have approximately monotonic response profiles, with Happy neurons (top row) responding maximally to the most happy faces and Sad neurons (bottom row) responding maximally to Sad faces, as previously reported in the simulation study of Eguchi et al. (2016). Interestingly, these authors showed that these neurons were actually encoding particular spatial relationships between the facial features that correlated with facial expression. For a more detailed analysis of the neuronal firing properties that developed during the pretraining stage, please refer to this previous publication.

In the next sections, we show how to remap the feedforward synaptic connections to these two subpopulations of output neurons by either CT learning or trace learning in order to shift the cognitive bias from negative to positive.

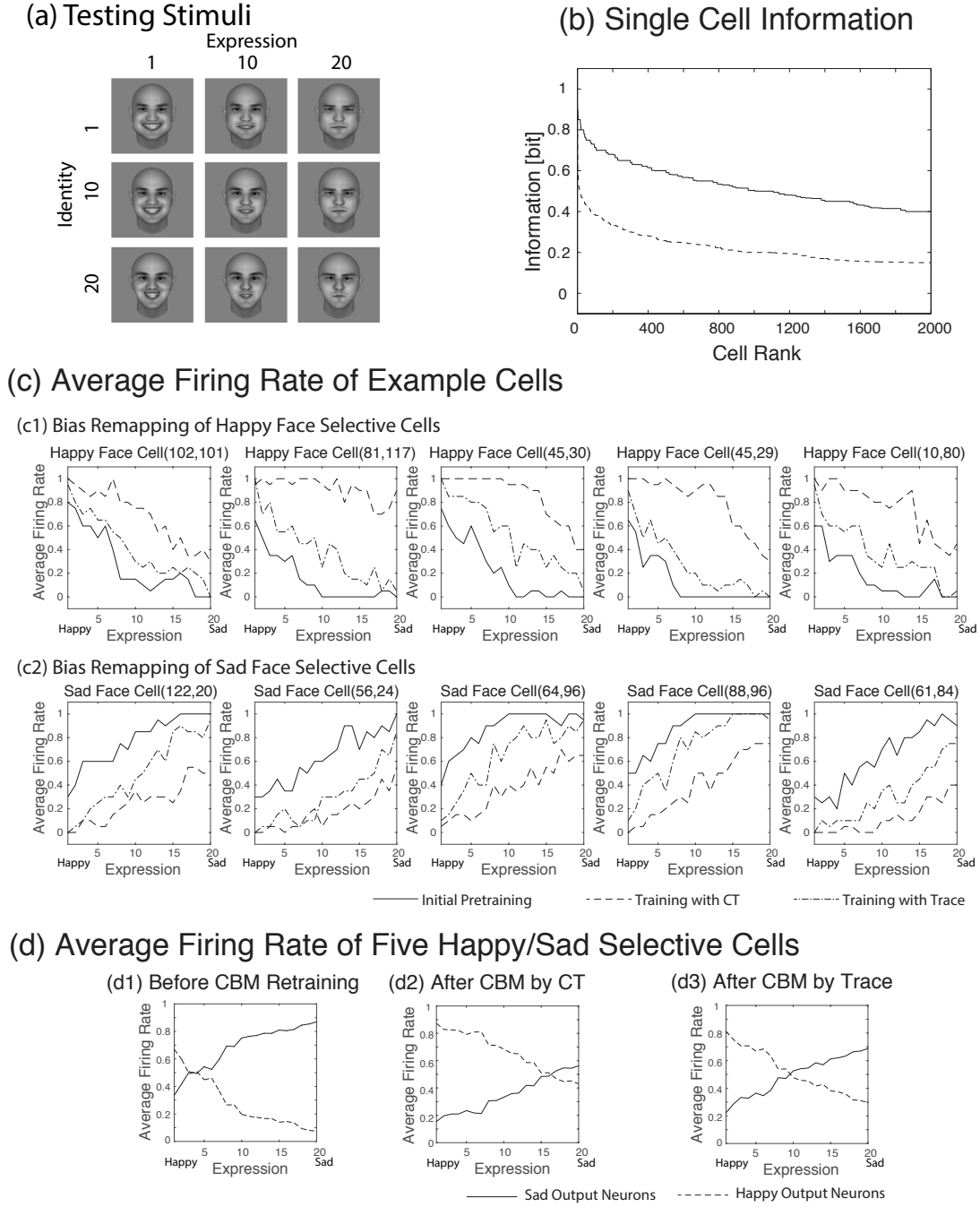


Figure 4. (a) The face stimuli used to test VisNet. A 1-dimensional space of 20 different facial identities, which varied gradually from Identity A to Identity B, were constructed. Then each of these identities was varied over a 1-dimensional space of 20 different expressions which varied gradually from sad to happy. (b) The amount of information carried by output (4th layer) neurons after pretraining VisNet. The plot shows the information carried by all of the 4th layer neurons about either happy or sad expressions, where the neurons are plotted in rank order along the abscissa. (c) Demonstration of CBM by CT learning (c1) and trace learning (c2) in VisNet. The firing rates of five Happy output neurons and five Sad output neurons are recorded in response to the matrix of test faces shown in (a) directly before and after CBM retraining. The plots show the average firing rate of the cells in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. (d) The plots show the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. The subplot (d1) shows the output of the network directly before CBM retraining, and the subplot (d2) and (d3) shows the output of the network after CBM retraining with CT learning and with trace learning, respectively.

Experiment 2a: CBM by CT Learning

Method. After pretraining VisNet on 100 randomised faces as described above in the section below, VisNet then underwent a stage of CBM retraining by CT learning. During this, the network was retrained on continuously transforming face images with the Hebbian learning rule (1) with weight vector renormalisation (2). Figure 3c shows examples of the face stimuli used to perform CBM retraining by CT learning. The image set is constructed from 5 different facial identities. For each of these facial identities, ten face images were constructed by sampling ten evenly-spaced expressions between happy and neutral. Figure 3c shows a subset of these images corresponding to one particular facial identity morphed through 10 equispaced expressions from happy (top left) to neutral (bottom right). During CBM retraining, the first facial identity was presented transforming continuously through the 10 expressions from happy to neutral. Then the second facial identity was similarly presented transforming continuously through the 10 expressions from happy to neutral. This was repeated for all 5 facial identities in turn. This constituted one epoch of training. The network underwent a total of 50 training epochs. In this situation, CT learning (Stringer et al., 2006) will begin to remap the feedforward synaptic connections through successive neuronal layers within the network according to the computational hypothesis described in the Introduction. That is, when the happy face is presented, this stimulates the happy output (4th layer) neurons to respond. Then, as the face is gradually morphed from happy to neutral, the happy output cells continue to respond due to the CT learning mechanism operating in the feedforward synaptic connections between successive layers. At the same time, the later more neutral faces are remapped onto the happy output neurons through the Hebbian learning rule (1) with weight vector renormalisation (2). This retraining is carried out for each of the 5 different facial identities over 100 training epochs. In this way, the low-level features representing more neutral faces in the lower layers of the network become remapped onto the more happy output representations. Thus, CBM occurs.

We wanted to assess how well CBM retraining remapped the more neutral faces away from the sad output neurons and onto the happy output neurons. In order to do this, we began by reanalysing the amount of information that individual output neurons carried about either happy or sad expressions directly before the CBM retraining stage. Specifically, we identified the subset of 1,000 neurons that carried the most information about the presence of a happy expression, and another subset of 1,000 neurons that carried the most information about the presence of a sad expression. In this way, we identified two separate subsets of output neurons: i.e. Happy vs Sad subpopulations. The performance of the CBM retraining was assessed by recording and analysing the firing rates of the Happy and Sad subpopulations of output neurons in response to the set of test faces shown in Figure 4a directly before and after CBM retraining. This is the same set of face images as used in the simulation study conducted by Eguchi et al. (2016). In particular, a 1-dimensional space of 20 different facial identities,

which varied gradually from one Identity A to another Identity B, was constructed. Each of these facial identities was then varied over a 1-dimensional space of 20 different expressions which varied gradually from sad to happy. This produced a matrix of 400 face stimuli constructed from 20 identities \times 20 expressions. By recording the responses of the Happy and Sad subsets of output neurons to these test faces directly before and after CBM retraining we were able to assess how well the CBM retraining had remapped the more neutral faces away from the Sad neurons and onto the Happy neurons.

Results. After pretraining the network on the set of 100 randomly generated faces (Figure 3b), we identified the subset of five output neurons that carried the most information about a happy expression, and another subset of five output neurons that carried the most information about a sad expression. Figure 4c shows the average firing rates of the five Happy output neurons (top row) and five Sad output neurons (bottom row) recorded in response to the matrix of test faces shown in Figure 4a directly before and after CBM retraining. The plots show the average firing rate of the cells after the initial pretraining (solid line), after the remapping with CT learning (dashed line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. It can be seen that the Happy output neurons respond with a greater average firing rate across the space of expressions after CBM training by CT learning. In particular, CBM retraining has remapped the more neutral faces away from the Sad output neurons and onto the Happy output neurons.

Furthermore, we identified the subset of 1,000 neurons that carried the most information about a happy expression, and another subset of 1,000 neurons that carried the most information about a sad expression. The firing rates of the subpopulation of Happy output neurons and subpopulation of Sad output neurons were then recorded in response to the matrix of test faces shown in Figure 4a directly before and after CBM retraining. Figure 4d shows the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). The left plot shows the output of the network directly before CBM retraining, and the right plot shows the output of the network after CBM retraining with CT learning. It can be seen that directly before CBM retraining, the subpopulation of Sad output neurons respond more strongly on average than the Happy output neurons to all facial expressions greater than 4 on the happiness scale (1-20) represented along the abscissa. However, after CBM retraining, the Sad output neurons respond more strongly than the Happy output neurons only to facial expressions greater than 16 on the happiness scale. Thus, CBM retraining has remapped the more neutral faces away from the sad output neurons and onto the happy output neurons. In particular, CBM retraining is able to shift the bias in the network from negative to positive using a biologically plausible Hebbian learning rule (1) with weight vector renormalisation (2) when the faces are presented transforming continuously from happy to sad as shown in Figure 3c.

Experiment 2b: CBM by Trace Learning

Method. In this section, VisNet underwent a stage of CBM retraining by trace learning after the initial stage of pretraining VisNet on 100 randomised faces as described above. During this, the network was retrained on faces with either happy or neutral expressions, with the synapses modified using the trace learning rule (3) with weight vector renormalisation (2). Figure 3d shows examples of the face stimuli used to perform CBM retraining by trace learning. The image set consisted of 25 faces with a happy expression and 25 faces with a neutral expression. Each of these 50 faces had a different randomly generated identity. Figure 3d shows some examples of these images. The top row shows a selection of 5 happy faces, while the bottom row shows 5 neutral faces. During CBM retraining, faces with happy or neutral expressions were shown alternately in an interleaved fashion. That is, the presentation order was happy face 1, neutral face 1, happy face 2, neutral face 2, and so on until eventually happy face 25, neutral face 25. The ordered presentation of all 50 faces constituted one epoch of training. The network underwent a total of 50 training epochs. In this situation, trace learning (Foldiak, 1991; Wallis and Rolls, 1997) will encourage the happy output neurons to learn to respond to both the happy faces and more neutral faces that are presented in temporal proximity. That is, the neurons that are originally selective to only happy faces may start to respond also to the more neutral faces based on temporal associations. In this way, the low-level features representing more neutral faces in the lower layers of the network become remapped onto the more happy output representations. Hence, CBM takes place.

Results. The network performance was assessed in a similar manner to that described above for CT learning in Experiment 2a. After pretraining the network on the set of 100 randomly generated faces (Figure 3b), we identified the subset of five neurons that carried the most information about a happy expression, and another subset of five neurons that carried the most information about a sad expression. Figure 4c shows the average firing rates of the five Happy output neurons (top row) and five Sad output neurons (bottom row) recorded in response to the matrix of test faces shown in Figure 4a directly before and after CBM retraining. The plots show the average firing rates of the cells after the initial training (solid line), and after the remapping with trace learning (dash-dot line), in response to 20 different facial expressions ranging from very happy (1) to very sad (20). For each facial expression, the firing rates are averaged over the 20 different facial identities. It can be seen that the Happy output neurons respond with a greater average firing rate across the space of expressions after CBM training by trace learning. In particular, CBM retraining has remapped the more neutral faces away from the Sad output neurons and onto the Happy output neurons.

Also, we identified the subset of 1,000 neurons that carried the most information about a happy expression, and another subset of 1,000 neurons that carried the most information about a sad expression. These were exactly the same subsets of Happy and Sad output cells that were identified for

the CT learning simulation described in the section describing Experiment 2a. The firing rates of the subpopulation of Happy output neurons and subpopulation of Sad output neurons were then recorded in response to the matrix of test faces shown in Figure 4a directly before and after CBM retraining. Figure 4 shows the average firing rate of all the Happy output cells (dashed line) and all the Sad output cells (solid line) in response to 20 different facial expressions ranging from very happy (1) to very sad (20). The subplot (d1) shows the output of the network directly before CBM retraining, and the subplot (d3) shows the output of the network after CBM retraining with trace learning. It can be seen that directly before CBM retraining, the subpopulation of Sad output neurons respond more strongly on average than the Happy output neurons to all facial expressions greater than 3 on the happiness scale (1-20) represented along the abscissa. However, after CBM retraining, the Sad output neurons respond more strongly than the Happy output neurons only to facial expressions greater than 18 on the happiness scale. Hence, the more neutral faces have been remapped away from the sad output neurons and onto the happy output neurons by the CBM retraining. In particular, CBM retraining has shifted the bias in the network from negative to positive using a biologically plausible trace learning rule (3) with weight vector renormalisation (2) when the faces are presented with the happy and neutral expressions shown in Figure 3d interleaved.

Discussion

In this paper we have described and modelled two alternative CBM training mechanisms; continuous transformation (CT) learning (Stringer et al., 2006) and trace learning (Foldiak, 1991; Wallis and Rolls, 1997). These learning mechanisms were previously used to model how the primate ventral visual pathway learns to perform transform invariant visual object recognition. CT learning binds together input stimuli onto the same categorical output representation using spatial continuity, while trace learning binds together stimuli using temporal continuity. Experimental support for these two learning mechanisms has been provided by previous psychophysical studies, which have confirmed that human subjects bind together different images onto a single categorical representation using a mixture of both spatial continuity (CT learning) and temporal continuity (trace learning) (Perry et al., 2006). Our current simulations have shown that these same learning mechanisms may be implemented in neural network computer models to rewire the synaptic connectivity in order to eliminate the kind of negative cognitive biases associated with clinical depression.

To the authors' knowledge, this is the first study that has modelled the application of the CT learning and trace learning mechanisms to CBM-Interpretation. Previous experimental studies have found that CBM-Interpretation can reduce negative cognitive biases in human participants (Grey and Mathews, 2000; Mathews and Mackintosh, 2000), which in turn can reduce the risk for depression recurrence (Holmes et al., 2009). This paper provides potential explanations at the neuronal and

synaptic level for how such a shift in interpretational bias might occur through CBM training.

Understanding the way in which biases can be shifted is crucial at present, given the mixed results seen in CBM research so far (Fox et al., 2014). In this paper, we have successfully demonstrated how computational models can be used to explore and exploit existing psychological phenomena in order to optimise a CBM procedure.

Implications and Future Work

The results of these simulations are highly informative for the development of experimental protocols to develop optimal CBM training methodologies with human participants. We aim to develop two separate experiments using the stimuli from these simulations; presenting them to participants in the order in which they have been shown to induce CT and also trace learning. A pilot investigation will explore whether a bias change will occur under the passive viewing methodology described above, or whether participants will be required to actively engage in the task in order to ensure that their attention on the task is maintained. If so, the task will resemble a modified version of Penton-Voak et al. (2013), where participants are asked to rate facial expressions in order to determine their baseline emotional bias. However, the learning will still remain unsupervised in that no feedback will be given. Using our stimuli and the required presentation order, we will investigate whether or not the predicted bias change will occur, and also whether or not a concurrent reduction in clinical symptoms arises. Thus, there are important clinical implications of the current modelling work in helping clinical investigators design and implement novel and more optimal CBM interventions.

We also believe that the development of well specified computational models help to guide future research aimed at optimising the effectiveness of CBM interventions. For example, the simulations presented in this paper utilized either CT learning or trace learning, but not both together, to effect a shift in the cognitive bias from negative to positive. On the other hand, psychophysical studies have shown that human subjects bind together different images onto a single categorical representation using a mixture of both spatial continuity and temporal continuity (Perry et al., 2006). Wallis and Bühlhoff (2001) have also shown that both spatial and temporal continuity seem to play a key role for modifying recognition memory. In addition, a recent modelling study has predicted that invariance learning in the primate ventral visual pathway may be most effective when CT learning and trace learning are combined together simultaneously (Spoerer et al., 2016). Therefore, in future work we will investigate CBM training methodologies that combine together both CT learning and trace learning simultaneously for maximum therapeutic effect. Furthermore, the future work could look at other types of learning, such as reinforcement learning, to optimise CBM procedures using feedback.

We will also explore various architectural extensions to the model, in order to more accurately reflect the known neuroanatomy of relevant brain areas. One such extension could be the addition of a

biased competition account. Based on this theory, Mathews and Mackintosh (1998) proposed a model to explain the negative interpretative biases of emotionally ambiguous expressions in high-trait anxious patients. In their scenario, the competition is between alternate interpretations of emotionally ambiguous stimuli (e.g., sad and happy), similar to the basis of the mechanism proposed in our current study. However, they also include a top-down threat-detecting signal from the amygdala, and a cognitive control signal from the rostral anterior cingulate cortex (rACC) and lateral prefrontal cortex (LPFC) (Bishop, 2007) to implement the biased competition. As a result, the negative interpretation is more likely to win the competition when such biased signals are present (Mathews and Mackintosh, 1998).

Their model does not necessarily exclude any other mechanism that may influence the relevant representations developed in the earlier stages of visual processing. The current study has investigated the potential mechanism to modify such neural representations of affective visual inputs developed at the earlier stages. Therefore, the model of amygdala-prefrontal circuitry with biased competition (Mathews and Mackintosh, 1998) is not mutually exclusive with the model proposed in the current study, but instead is compatible and rather complementary. Our model provides the theoretical ‘front-end’ of the competition account, before such top-down signals are explored.

Although it does not simulate the rostral regions further than IT, Deco and Rolls (2005) have previously presented a single unified model of hierarchical processing with attentional modulation mechanisms via backprojection in VisNet. In terms of physiology, there exist bidirectional connections between TE and the further rostral areas such as amygdala and orbitofrontal cortex (OFC). Grabenhorst and Rolls (2010, 2011) proposed that these connections may form autoassociative networks, which are suitable for implementing the biased competitions. With such extensions of the model, it is possible to further investigate how prioritised emotional signals from earlier stages of visual processing may influence the nature of competition in the latter stages, where signals from IT and areas such as amygdala and ACC meet. This would provide deeper insight into a more accurate account that guides the development of more effective CBM-Interpretation training procedures.

The purpose of CBM interventions, after all, is to ‘re-train’ a response to stimuli. One interesting question to ask is whether the process of acquiring and removing biases shares similar mechanisms. In the current study, we presented two potential mechanisms to enhance the CBM-I intervention without ‘active training’ but simply by presenting carefully designed sequences of the artificial visual inputs to the network. As the original negative biases of patients also occurs without requiring ‘active training,’ it might be that the proposed mechanisms also bear some relation to the causative process of acquiring negative biases.

Acknowledgements

The authors wish to thank B.M.W. Mender for invaluable assistance and discussion related to the research.

References

- Anderson, M. C. and Green, C. (2001). Suppressing unwanted memories by executive control. *Nature*, 410:366–369.
- Arditte, K. A. and Joormann, J. (2014). Rumination moderates the effects of cognitive bias modification of attention. *Cognitive Therapy and Research*, 38.
- Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, 165:969–977.
- Bishop, S. J. (2007). Neurocognitive mechanisms of anxiety: an integrative account. *Trends in Cognitive Sciences*, 11(7):307–316.
- Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52:383–394.
- Bourke, C., Douglas, K., and Porter, R. (2010). Processing of facial emotion expression in major depression: A review. *Australian and New Zealand Journal of Psychiatry*, 44:681–696.
- Clarke, P. J. F., Notebaert, L., and MacLeod, C. (2014). Absence of evidence or evidence of absence: Reflecting on therapeutic implementations of attentional bias modification. *BMC Psychiatry*, 14.
- Cox, D. D., Meier, P., Oertelt, N., and DiCarlo, J. J. (2005). ‘Breaking’ position-invariant object recognition. *Nature neuroscience*, 8(9):1145–1147.
- Cristea, I. A., Kok, R. N., and Cuijpers, P. (2015). Efficacy of cognitive bias modification interventions in anxiety and depression: Meta-analysis. *British Journal of Psychiatry*, 206:7–16.
- Cumming, B. G. and Parker, A. J. (1999). Binocular neurons in v1 of awake monkeys are selective for absolute, not relative, disparity. *The Journal of Neuroscience*, 19:5602–5618.
- Deco, G. and Rolls, E. T. (2005). Neurodynamics of Biased Competition and Cooperation for Attention: A Model With Spiking Neurons. *Journal of Neurophysiology*, 94(1):295–313.
- Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373):1245–1255.
- Desimone, R. and Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, 18(1):193–222.

- Desimone, R., Wessinger, M., Thomas, L., and Schneider, W. (1990). Attentional Control of Visual Perception: Cortical and Subcortical Mechanisms. *Cold Spring Harbor Symposia on Quantitative Biology*, 55:963–971.
- Duncan, J. and Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458.
- Eguchi, A., Humphreys, G. W., and Stringer, S. M. (2016). The visually-guided development of facial representations in the primate ventral visual pathway: a computer modelling study. *Psychological Review*, 123(6).
- Enock, P. M., Hofmann, S. G., and McNally, R. J. (2014). Attention bias modification training via smartphone to reduce social anxiety: A randomized, controlled multi-session experiment. *Cognitive Therapy and Research*, 38.
- Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, 3:194–200.
- Fox, E., Mackintosh, B., and Holmes, E. (2014). Travellers’ tales in cognitive bias modification research: A commentary on the special issue. *Cognitive Therapy Research*, 38:239–247.
- Freeman, J. and Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience*, 14:1195–1201.
- Grabenhorst, F. and Rolls, E. T. (2010). Attentional Modulation of Affective Versus Sensory Processing: Functional Connectivity and a Top-Down Biased Activation Theory of Selective Attention. *Journal of Neurophysiology*, 104(3):1649–1660.
- Grabenhorst, F. and Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):56–67.
- Greenberg, P. E., Sisitsky, T., Kessler, R. C., Finkelstein, S. N., Berndt, E. R., Davidson, J. R. T., Ballenger, J. C., and Fyer, A. J. (1999). The Economic Burden of Anxiety Disorders in the 1990s. *The Journal of Clinical Psychiatry*, 60(7):427–435.
- Grey, S. and Mathews, A. (2000). Effects of training on interpretation of emotional ambiguity. *The Quarterly Journal of Experimental Psychology Section A*, 53:1143–1162.
- Grey, S. and Mathews, A. (2009). Cognitive bias modification – Priming with an ambiguous homograph is necessary to detect an interpretation training effect. *Journal of Behavior Therapy and Experimental Psychiatry*, 40(2):338–343.

- Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E., Ernst, M., and Pine, D. S. (2010). Attention Bias Modification Treatment: A Meta-Analysis Toward the Establishment of Novel Treatment for Anxiety. *Biological Psychiatry*, 68:982–990.
- Hallion, L. S. and Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological Bulletin*, 137:940–958.
- Hasselmo, M. E., Rolls, E. T., and Baylis, G. C. (1989). The role of expression and identity in the face-selective responses of neurons in the temporal visual cortex of the monkey. *Behavioural Brain Research*, 32:203–218.
- Hoffman, D. L., Dukes, E. M., and Wittchen, H.-U. (2008). Human and economic burden of generalized anxiety disorder. *Depression and Anxiety*, 25(1):72–90.
- Holmes, E. A., Lang, T. J., and Sham, D. M. (2009). Developing interpretation bias modification as a “Cognitive Vaccine” for depressed mood: Imagining positive events makes you feel better than thinking about them verbally. *Journal of Abnormal Psychology*, 118:76–88.
- Hoppitt, L., Mathews, A., Yiend, J., and Mackintosh, B. (2010). Cognitive Bias Modification: The Critical Role of Active Training in Modifying Emotional Responses. *Behavior Therapy*, 41(1):73–81.
- Jones, J. P. and Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58:1187–1211.
- Jones, S. (2016). 50 million years of work could be lost to anxiety and depression. *The Guardian*.
- Joormann, J., Hertel, P. T., Brozovich, F., and Gotlib, I. H. (2005). Remembering the good, forgetting the bad: Intentional forgetting of emotional material in depression. *Journal of Abnormal Psychology*, 114(4):640–648.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.
- Li, N. and DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, 231:1502–1507.
- MacLeod, C. (2012). Cognitive bias modification procedures in the management of mental disorders: *Current Opinion in Psychiatry*, 25:114–120.
- MacLeod, C. and Mathews, A. (2012). Cognitive bias modification approaches to anxiety. *Annual Review of Clinical Psychology*, 8:189–217.

- MacLeod, C., Rutherford, E., Campbell, L., Ebsworthy, G., and Holker, L. (2002). Selective attention and emotional vulnerability: Assessing the causal basis of their association through the experimental manipulation of attentional bias. *Journal of Abnormal Psychology*, 111:107–123.
- Mathews, A. and Mackintosh, B. (1998). A Cognitive Model of Selective Processing in Anxiety. *Cognitive Therapy and Research*, 22(6):539–560.
- Mathews, A. and Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of Abnormal Psychology*, 109:602–615.
- Mathews, A. and MacLeod, C. (2005). Cognitive Vulnerability to Emotional Disorders. *Annual Review of Clinical Psychology*, 1:167–195.
- Menne-Lothmann, C., Viechtbauer, W., Höhn, P., Kasanova, Z., Haller, S. P., Drukker, M., van Os, J., Wichers, M., and Lau, J. Y. F. (2014). How to boost positive interpretations? A meta-analysis of the effectiveness of cognitive bias modification for interpretation. *PLoS One*, 9:e100925.
- Micco, J. A., Henin, A., and Hirshfeld-Becker, D. R. (2014). Efficacy of interpretation bias modification in depressed adolescents and young adults. *Cognitive Therapy and Research*, 38.
- Mogoşe, C., David, D., and Koster, E. H. W. (2014). Clinical efficacy of attentional bias modification procedures: An updated meta-analysis. *Journal of Clinical Psychology*, 70:1133–1157.
- Pasupathy, A. (2006). Neural basis of shape representation in the primate brain. *Progress in brain research*, 154:293–313.
- Penton-Voak, I. S., Thomas, J., Gage, S. H., McMurran, M., McDonald, S., and Munafo, M. M. (2013). Increasing recognition of happiness in ambiguous facial expressions reduces anger and aggressive behavior. *Psychological Science*, 24:688–697.
- Perry, G., Rolls, E. T., and Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research*, 46:3994–4006.
- Petkov, N. and Kruizinga, P. (1997). Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological Cybernetics*, 76:83–96.
- Pettet, M. W. and Gilbert, C. D. (1992). Dynamic changes in receptive-field size in cat primary visual cortex. *Proceedings of the National Academy of Sciences*, 89:8366–8370.
- Preminger, S., Sagi, D., and Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects. *Vision Research*, 47:965–973.

- Richards, A., French, C. C., Calder, A. J., Webb, B., and Fox, R. (2002). Anxiety-related bias in the classification of emotionally ambiguous facial expressions. *Emotion*, 2:273–287.
- Roiser, J. P., Elliott, R., and Sahakian, B. J. (2012). Cognitive mechanisms of treatment in depression. *Neuropsychopharmacology*, 37:117–136.
- Rolls, E. T. and Milward, T. (2000). A model of invariant object recognition in the visual system: Learning rules, activation functions, lateral inhibition, and information-based performance measures. *Neural Computation*, 12:2547–2572.
- Rolls, E. T., Treves, A., Tovee, M., and Panzeri, S. (1997). Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *Journal of Computational Neuroscience*, 4:309–333.
- Royer, S. and Paré, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature*, 422:518–522.
- Rugulies, R. (2002). Depression as a predictor for coronary heart disease: a review and meta-analysis. *American Journal of Preventive Medicine*, 23:51–61.
- Schneider, S. and Moyer, A. (2010). Depression as a predictor of disease progression and mortality in cancer patients. *Cancer*, 116:3304–3304.
- Spoerer, C. J., Eguchi, A., and Stringer, S. M. (2016). A computational exploration of complementary learning mechanisms in the primate ventral visual pathway. *Vision Research*, 119:16–28.
- Stringer, S. M., Perry, G., Rolls, E. T., and Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics*, 94:128–142.
- Stringer, S. M. and Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural networks: the official journal of the International Neural Network Society*, 21:888–903.
- Stringer, S. M., Rolls, E. T., and Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network*, 18:161–187.
- Surcinelli, P., Codispoti, M., Montebanocci, O., Rossi, N., and Baldaro, B. (2006). Facial emotion recognition in trait anxiety. *Anxiety Disorders*, 20:110–117.
- Tromans, J. M., Harris, M., and Stringer, S. M. (2011). A computational model of the development of separate representations of facial identity and expression in the primate visual system. *PLoS ONE*, 6:e25616.

- 959 Ustün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., and Murray, C. J. L. (2004). Global
960 burden of depressive disorders in the year 2000. *The British Journal of Psychiatry: The Journal of*
961 *Mental Science*, 184:386–392.
- 962 Von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex.
963 *Kybernetik*, 14:85–100.
- 964 Wallis, G. and Bülthoff, H. H. (2001). Effects of temporal association on recognition memory.
965 *Proceedings of the National Academy of Sciences*, 98:4800–4804.
- 966 Wallis, G. and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress*
967 *in Neurobiology*, 51:167–194.
- 968 West, G. L., Anderson, A. A. K., Ferber, S., and Pratt, J. (2011). Electrophysiological Evidence for
969 Biased Competition in V1 for Fear Expressions. *Journal of Cognitive Neuroscience*,
970 23(11):3410–3418.