Journal of Consulting and Clinical Psychology
2017, Vol. 85, No. 3, 200–217

0022-006X/17/$12.00 http://dx.doi.org/10.1037/ccp0000165

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

Akihiro Eguchi, Daniel Walters, Nele Peerenboom, Hannah Dury, Elaine Fox, and Simon Stringer
Oxford University

Objective: Cognitive bias modification (CBM) eliminates cognitive biases toward 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. ***Method:*** We investigated 2 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 1-layer neural network, and then we describe simulations in a biologically detailed multilayer neural network model of the ventral visual pathway. ***Results:*** After training with either the continuous transformation learning rule or the trace learning rule, the 1-layer neural network eliminated biases in interpreting neutral stimuli as sad. The multilayer neural network trained with realistic face stimuli was also shown to be able to use continuous transformation learning or trace learning to reduce biases in the interpretation of neutral stimuli. ***Conclusions:*** The simulation results suggest 2 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.

What is the public health significance of this article?
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 modeling 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 optimizing the effectiveness of CBM treatments.

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

Depression is the most common mental health problem, affecting 8%–12% of the adult population (Ustün, Ayuso-Mateos, Chatterji, Mathers, & Murray, 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 & 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 (Greenberg et al., 1999; Hoffman, Dukes, & Wittchen, 2008; Ustün et al., 2004). Similarly, the latest World Health Organization report shows that anxiety and depression lead to a loss of millions of work days (S. 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 toward emotionally negative information, with patients tending to pay attention to negative stimuli, interpret events negatively, and recall negative memories (Mathews & MacLeod, 2005; Roiser, Elliott, & Sahakian, 2012). These biases therefore have been included within cognitive models of depression (Beck, 2008) and anxiety (Mathews & MacLeod, 2005), leading to a growing interest in exploring the causal relationship between these biases, mood states, and clinical symptoms.

This article was published Online First December 19, 2016.

Akihiro Eguchi, Daniel Walters, Nele Peerenboom, Hannah Dury, Elaine Fox, and Simon Stringer, Department of Experimental Psychology, Oxford University.

We thank Bedeho M. W. Mender for invaluable assistance and discussion related to the research.

Correspondence concerning this article should be addressed to Akihiro Eguchi, Department of Experimental Psychology, Oxford University, United Kingdom. E-mail: akihiro.eguchi@psy.ox.ac.uk

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 recognize the clinical potential of these tools, inspiring the development of a family of potential treatments known as cognitive bias modification (CBM; MacLeod, 2012; MacLeod & Mathews, 2012). CBM seeks to eliminate these underlying processing biases through three main varieties of treatment. For example, CBM-Attention (CBM-A), which is also referred to as attentional bias modification, seeks to shift the attention of subjects away from negative stimuli in the environment (Hakamata et al., 2010; MacLeod, Rutherford, Campbell, Ebsworthy, & Holker, 2002), CBM-Interpretation (CBM-I) aims to reduce the tendency for negative interpretation of events (Grey & Mathews, 2000, 2009), and CBM-Memory (CBM-M) seeks to reduce the recall and influence of negative memories (Anderson & Green, 2001; Joormann, Hertel, Brozovich, & Gotlib, 2005). However, CBM as a whole is not without controversy. Most CBM studies so far have focused on CBM-A, with a number of meta-analyses finding the efficacy of CBM-A inconclusive (Cristea, Kok, & Cuijpers, 2015; Hallion & Ruscio, 2011; Mogoaşe, David, & Koster, 2014). CBM-I, on the other hand, has had more promising results (Cristea et al., 2015; Hallion & Ruscio, 2011; Menne-Lothmann et al., 2014).

The negative interpretation bias of facial expression (Bourke, Douglas, & Porter, 2010; Richards, French, Calder, Webb, & Fox, 2002; Surcinelli, Codispoti, Montebarocci, Rossi, & Baldaro, 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 at which participants switched from a categorization of happy to a categorization of angry was therefore determined. 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 & Mathews, 2000; MacLeod et al., 2002). A recent commentary described the problem with current CBM research as a lack of focus on reliably changing the underlying cognitive biases (Fox, Mackintosh, & Holmes, 2014). Fox et al. (2014) argued 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, Notebaert, & MacLeod, 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 have concluded that CBM does not work, despite never successfully changing the bias, in both CBM-I (Micco, Henin, & Hirshfeld-Becker, 2014) and CBM-A (Arditte & Joormann, 2014; Enock, Hofmann, & McNally, 2014). Therefore, it is of crucial importance to investigate the mechanisms behind changing cognitive biases to optimize bias-change procedures, which we do in the current study.

Theory and Modeling 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, 1998; Desimone & Duncan, 1995; Desimone, Wessinger, Thomas, & Schneider, 1990). 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 (Deco & Rolls, 2005; Duncan & Humphreys, 1989). 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).

Although this is one of the biologically reasonable accounts of the mechanism of such biases, West, Anderson, Ferber, and Pratt (2011) recently reported that biased competition may begin as early as the primary visual cortex, and affective prioritization can be solely driven by physical salience of the low-level features in emotional faces themselves. This implies that some degree of prioritized 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 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 was 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 modeling is one useful way to investigate such mechanisms. The current study investigates the potential mechanisms of CBM-I through neural network computer modeling 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, Mathews, Yiend, & Mackin-

tosh, 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 that 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 to the work of Bourke et al. (2010), aiming to change a negative interpretation of facial expressions into a positive interpretation. To achieve such learning without any explicit teaching signal, the new CBM methodologies utilize two previously established biologically plausible synaptic learning mechanisms known as *continuous transformation* (CT) *learning* (Stringer, Perry, Rolls, & Proske, 2006) and *trace learning* (Foldiak, 1991; Wallis & 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 aimed 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 noninvasive 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, Sagi, and Tsodyks (2007) trained participants to classify faces into two categories: friends (F) and nonfriends (NF). Upon reaching good performance, participants were then trained with a sequence of morphed images from F to NF. Participants were tested on how they classified the morphed images. Initially, the first half of the morphed image sequence was classified as F, whereas the second half of the morphed sequence was classified as NF. However, as training progressed, the separation threshold moved toward 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 neighbors, onto the same output neuron. The standard Hebbian learning rule used to modify the feedforward synaptic connections at each time step τ is

$$\delta w_{ij}^{\tau} = k r_i^{\tau} r_j^{\tau}, \tag{1}$$

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

$$\sqrt{\sum_j w_{ij}^2} = 1. \tag{2}$$

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

We hypothesized 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 exploited 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 firing to only 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 (e.g., Perry, Rolls, & Stringer, 2006). In contrast, systematically switching the identity of a visual object during such sequential presentation impairs position-invariant representations (Cox, Meier, Oertelt, & DiCarlo, 2005). Li and DiCarlo (2008) 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 had 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 Inferotemporal (IT) cortex 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 & 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 article, the form of trace learning rule implemented at each time step τ is

$$\delta w_{ij}^{\tau} = k\bar{r}_i^{\tau-1} r_j^{\tau}, \tag{3}$$

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

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

where η is a parameter anywhere in the interval [0, 1] that controls the relative balance in the trace term $\bar{r}_i{}^{\tau}$ 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 in the next section, η was set to .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 among many happy faces, we expected to see positive output neurons learning to respond to both positive and neutral faces.

Overview of Simulation Studies Carried Out in This Article

We first describe simulations with a simplified one-layer neural network architecture to test the two hypothesized 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 multilayer neural network computer model, VisNet, of the ventral visual pathway in the primate brain (Wallis & Rolls, 1997), which has recently been used to show how the visual system may learn to represent facial expressions (Eguchi, Humphreys, & Stringer, 2016; Tromans, Harris, & Stringer, 2011), in the section describing Experiment 2.

In both sections, we extend these previous modeling studies involving synaptic plasticity and learning to the problem of understanding the neurobiological basis of CBM training by both CT learning (Experiments 1a and 2a) and trace learning (Experiments 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 respond to the sad faces only 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 idealized network architecture and input stimulus representations to provide a controlled way of investigating and testing the underlying computational hypotheses described in the sections in the introduction.

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 about Experiments 1a and 1b. The network, depicted in Figure 1a, comprises a single layer of input cells that 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 one-dimensional layer of input cells provides a highly idealized representation of facial expressions ranging continuously from happy to sad. In the simulations, the input neurons have binarized (0/1) firing rates. Each input neuron responds selectively to a small localized 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 particular emotional valence by the coactivation of a localized cluster of input neurons at the corresponding position within the layer.

At the beginning of the simulation, the feedforward synaptic connection weights are initialized 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 initializing the synaptic connections such that the more neutral input stimuli are initially responded to by

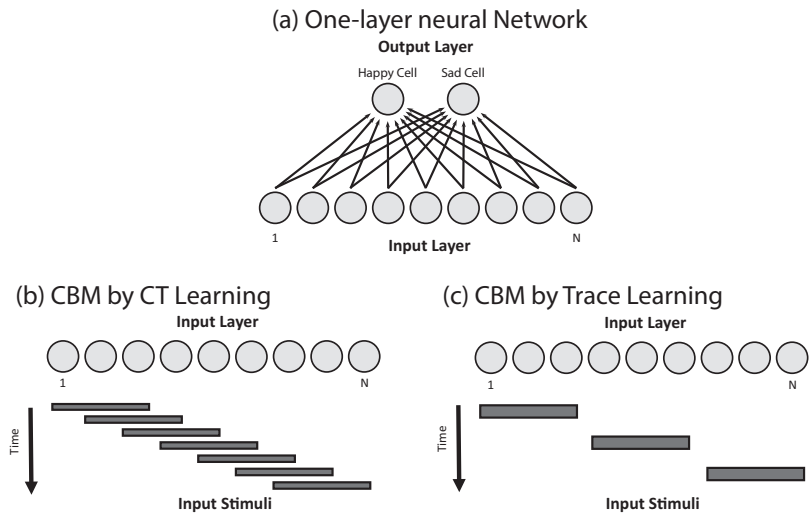

Figure 1. Panel a: The one-layer neural network architecture used in the models described in the sections about Experiment 1. A single layer of N input cells drove activity in the two output cells through the feedforward synapses (black arrows). The input layer cells responded to stimuli that ranged from happy to sad, with different simulations requiring different numbers of input cells, as detailed in the sections about Experiments 1a and 1b. There were 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. Panel 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 was represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal gray 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 were set to 0. At each successive training step, the input stimulus was advanced by one input cell to ensure that successive stimuli were varied in a continuous manner, that is, successive stimuli overlapped with each other, which is a requirement of CT learning. During each training epoch, the input stimuli were shifted once through the whole continuum from happy to sad. Panel c: The training protocol for the one-layer network trained by trace learning. The input layer contained a total of $N = 900$ neurons. During each training step, the current input stimulus was represented by the firing rates of a contiguous subblock of input cells being set to 1 as illustrated by the horizontal gray lines. The length of each stimulus, its stride, was set to be 100 input neurons. The firing rates of all other input cells were set to 0. To prevent CT-like learning effects from occurring, the input stimuli did not overlap with each other. During training, the *most happy* input stimuli were closely interleaved with more neutral input stimuli from the middle of the stimulus range, whereas the *most sad* stimuli are shown without temporal interleaving with the neutral stimuli. This stimulus presentation order enabled trace learning to associate together the *happy* and *neutral* stimuli onto the same *happy* output cell. CBM = cognitive bias modification; CT = continuous transformation.

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

$$h_i = \sum_j w_{ij} r_j, \tag{5}$$

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 were then set by applying winner-take-all inhibition so that the output cell with the highest activation level was given a firing rate of 1 and the other output cell was given a firing rate of 0.

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

Initial Setup of the Network

Before the network underwent CBM training, the feedforward synaptic weights to the *sad* and *happy* output cells were set manually to control whether there was a preexisting cognitive bias.

To establish the synaptic connectivity without an initial bias, the synaptic weights to the *sad* output cell, w_{SADj}, were set so that

$$w_{SADj} = \frac{1}{1 + \exp[-2\beta(\epsilon_j - \alpha)]}. \quad (6)$$

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 were distributed evenly throughout the *sad* to *happy* stimulus continuum. The slope β was set to an appropriate value (described in the top section of Table 1), and the threshold α was set to 0. The synaptic weights to the *happy* output cell, w_{HAPPYj}, were set to be

$$w_{HAPPYj} = 1 - w_{SADj}. \quad (7)$$

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 α = 0, the feedforward synaptic connections were unbiased in that the *happy* output cell and *sad* output cell received 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).

To introduce a negative bias in the synaptic weights such that the *sad* output cell would 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 were set according to Equation 6, with the threshold α set to a negative value (described in the top section of Table 1 for Experiment 1a and middle section of Table 1 for Experiment 1b). The synaptic weights from the input cells to the *happy* output cell were 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 resulted in the *sad* output cell's receiving stronger synaptic weights from a greater proportion of the input cells than the *happy* output cell did.

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 exploited this mechanism by morphing *happy* input stimuli, which are strongly associated with the positive output representation,

Table 1
Parameters of the Three Different Simulation Studies (Two One-layer network studies and one VisNet Study)

Parameter	Value	1st layer	2nd layer	3rd layer	4th layer
One-layer network (CT)					
No. of input cells	600				
Stride	100				
Sigmoid slope (β)	.5				
Biased sigmoid threshold (α)	−1				
Learning rate (*k*)	.001				
Training epochs	100				
One-layer network (trace learning)					
No. of input cells	9				
Stride	100				
Sigmoid slope (β)	.5				
Biased sigmoid threshold (α)	−1				
Learning rate (*k*)	.01				
Eta (η)	.8				
Training epochs	100				
VisNet					
Gabor: Phase shift (Ψ)	0, π				
Gabor: Wavelength (λ)	2				
Gabor: Orientation (θ)	0, π/4, π/2, 3π/4				
Gabor: Spatial bandwidth (*b*)	1.5 octaves				
Gabor: Aspect ratio (γ)	.5				
No. of layers	4				
Retina	256 × 256 × 16				
Dimension		128 × 128	128 × 128	128 × 128	128 × 128
No. 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	.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

Note. CT = continuous transformation.

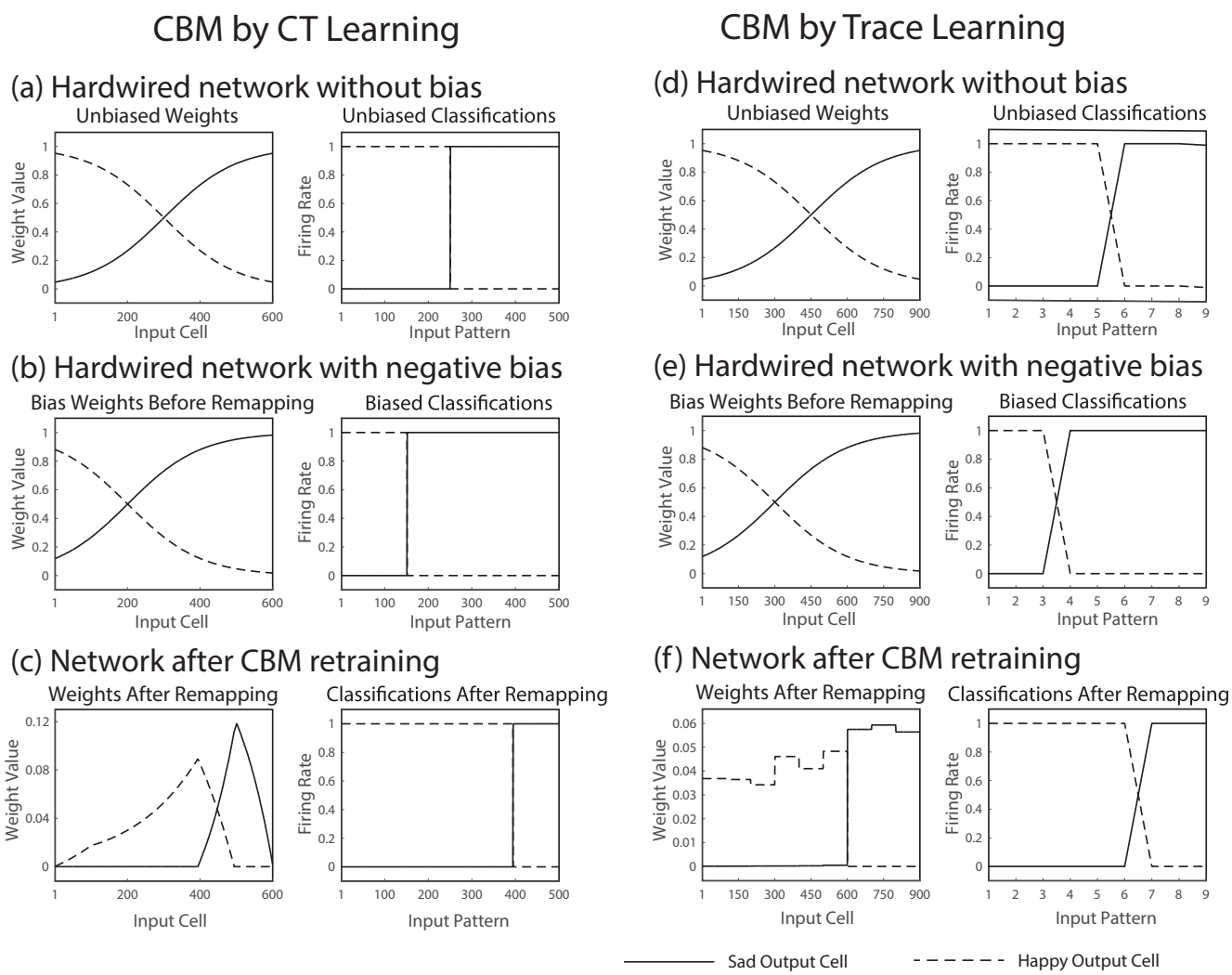

Figure 2. Demonstration of CBM in a one-layer network using CT learning (Panels a–c) and trace learning (Panels d–f) to remap the synaptic weights. The figure shows the feedforward synaptic weights (left column) and firing rates of the output cells (right column) at various stages of the simulation.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

that is, 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 represents 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 gray 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 (see Equation 1) is used. Because 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 was moved 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 were first updated according to Equation 5, the firing rates of the output neurons were then computed using winner-take-all competition, and then the feedforward synaptic weights were modified according to Equations 1 and 2. One epoch of training was completed after the input stimulus had been shifted through the whole continuum from happy to sad. Upon reaching the specified number of training epochs, the training phase was finished and the testing phase began, which followed the same protocol as the training phase with the exception that the weight update and normalization equations, Equations 1 and 2, were not simulated. The simulation was then complete. A one-layer neural network model was simulated with the parameters given in the top section of Table 1.

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 were 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 received the strongest synaptic weights from the input cells representing the *sad* end of the stimulus continuum, and the *happy* output cell received the strongest synaptic weights from the input cells representing the *happy* end of the stimulus continuum. The two output cells received 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 responded strongly to *very happy* input stimuli, the *sad* output cell responded strongly to *very sad* input stimuli, and most important, both output cells responded 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 was applied. The *sad* output cell received 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 received 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 responded not only to *very sad* stimuli but also to the majority of the more neutral stimuli, whereas the *happy* output cell did 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 have remapped such that the *happy* output cell now received stronger synaptic weights from the input cells representing a larger portion of the intermediate, more neutral, stimuli than did the *sad* output cell. The effect of this learned remapping is that the *happy* output cell responded to a greater proportion of the input stimulus space than the *sad* output cell did. That is, the *happy* output cell now responded to the majority of the intermediate neutral stimuli. This can be seen in the right plot of Figure 2c (cf. 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.

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 that utilizes a trace learning rule, Equation 3 with weight vector normalization 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, that is, 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 gray lines in Figure 1c. The length of each stimulus presented to the network was set to be 100 input neurons, whereas the firing rates of all other input cells were set to 0.

In contrast to the training protocol used for the previous simulations with CT learning described in the Method section 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 were divided into two separate groups: one group contain-

ing stimuli from the *most happy* and more neutral (middle) parts of the input stimulus range and one group containing stimuli from only the *sad* end of the input stimulus range. During an epoch of training, one of the two stimulus groups was selected at random. If the stimulus group contained only the *sad* stimuli, these stimuli were shown to the network in a random order. If the stimulus group contained both the *happy* and more neutral stimuli, then the *happy* stimuli were interleaved with the neutral stimuli such that a *happy* stimulus was 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 was shown to the network. During the presentation of each stimulus, the activations of the output neurons were updated by Equation 5, the firing rates of the output neurons were then computed according to winner-take-all competition, and the synaptic weights were then updated according to the trace learning rule Equation 3 with weight vector normalization Equation 2. After all stimuli had been presented, an epoch of training was complete and the next epoch of training began. The order of the stimulus groups and the order of stimulus presentation within the group were randomly selected for each training epoch. Upon reaching the specified number of epochs, the training phase was finished and the testing phase began, during which the input stimuli were presented one at a time to the network, ranging from happy to sad. The weight update and normalization equations, Equations 3 and 2, were not simulated during the testing phase. After the testing phase, the simulation was 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. Last, the negative bias in the previous network was eliminated by CBM training using trace learning. This resulted in the *happy* output neuron's now responding to most of the neutral stimuli as well as the *happy* stimuli.

Untrained network performance (before and after biases were added). The network was simulated with unbiased hardwired synaptic weights. Figure 2d (left side) shows the unbiased synaptic weights. The *sad* output cell received the strongest synaptic weights from the *sad* end of the stimulus range, whereas the *happy* output cell received the strongest synaptic weights from the *happy* end of the stimulus range. The two output cells received equal, albeit mirror-symmetric, distributions of synaptic weights from the intermediate neutral portion of the stimulus continuum. Figure 2d (right side) shows the firing rate responses of the two output cells to the full range of input stimuli. The *happy* output cell responded to *happy* stimuli and the *sad* output cell responded to *sad* stimuli, whereas both output cells responded 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 side) shows the synaptic weights. The *sad* output cell received stronger synaptic weights from the *sad* end of the input range and most of the more neutral input cells, whereas the *happy* output cell received stronger synaptic weights from only the *happy* end of the input range. Figure 2e (right side) 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 responded to the majority of the more neutral stimuli in addition to the *sad* stimuli, whereas the *happy* output cell responded to only 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 remapped so that the *happy* output neuron received stronger synaptic weights from input neurons representing the *happy* stimuli and the majority of the more neutral stimuli, whereas the *sad* output cell received strong synaptic weights from only 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 responded to stimuli from the *happy* to middle neutral region of the input stimulus range, whereas the *sad* output cell responded to stimuli from only the *sad* end of the input stimulus range, which can be seen in the right plot of Figure 2f. Thus, trace learning produced CBM, where the bias in the network was 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 (Stringer et al., 2006; Wallis & Rolls, 1997). The simulations with VisNet were carried out in two stages of training as explained in the following paragraphs.

In the first training stage, VisNet was trained on a set of randomized computer-generated face images where the identity and expression of each face was chosen randomly. Eguchi et al. (2016) reported that this led to the development of separate subpopulations 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, Rolls, & Baylis, 1989).

The second stage of training involved CBM by either CT learning or trace learning, similar to that described previously 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 spatiotemporal stimulus statistics required by either the CT learning (Experiment 2a) or trace learning (Experiment 2b) hypotheses.

VisNet Model Description

The simulation studies presented next were 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 (shown in Figure 3a) is based on the following: (a) a series of hierarchical competitive networks with local graded lateral inhibition; (b) 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

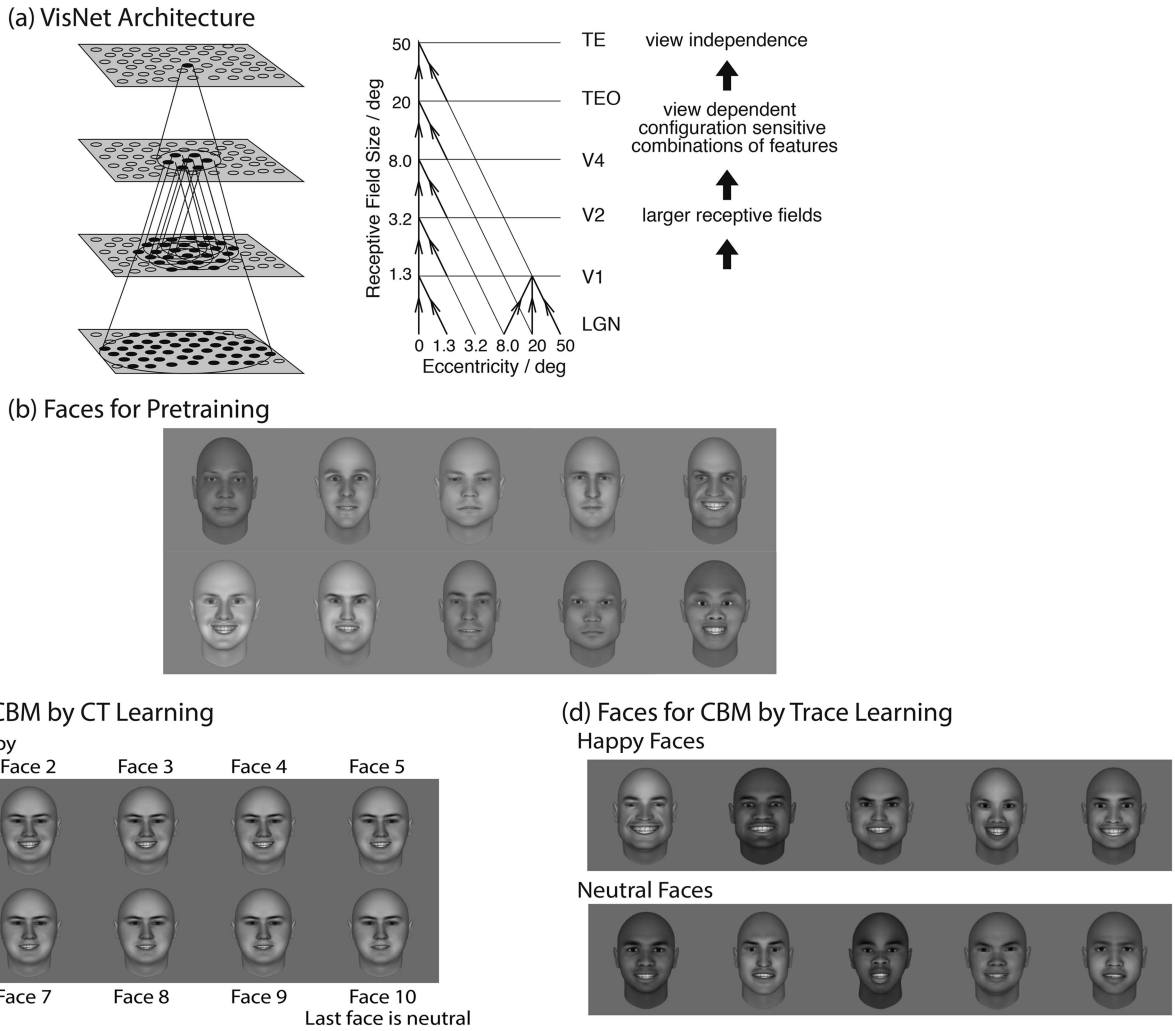

Figure 3. Panel a (left side): Stylized image of the four-layer VisNet architecture. Convergence through the network was designed to provide fourth-layer neurons with information from across the entire input retina. Panel a (right side): Convergence in the visual system V1: visual cortex area V1; TEO posterior inferior temporal cortex; and TE inferior temporal cortex. Panel b: Examples of the face stimuli used to pretrain VisNet. Here 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. Panel c: Examples of the face stimuli used to perform CBM retraining on VisNet through CT learning. The image set was constructed from five different facial identities. For each of these facial identities, 10 face images were constructed by sampling 10 evenly spaced expressions between happy and neutral. Panel 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. CBM = cognitive bias modification; CT = continuous transformation; LGN = lateral geniculate nucleus; deg = degree.

through the visual processing areas; and (c) synaptic plasticity based on a local associative learning rule such as the Hebb rule or trace rule.

In past work, the hierarchical series of four 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 (Wallis and Rolls, 1997). However, this correspondence has always been quite loose because the ventral pathway may be further subdivided into a finer grained network of distinct subregions.

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 that contains approximately 67% of the connections from the preceding layer. The values used in the current studies are

given in the bottom section of Table 1. The gradual increase in the receptive field of cells in successive layers reflects the known physiology of the primate ventral visual pathway (Freeman & Simoncelli, 2011; Pasupathy, 2006; Pettet & Gilbert, 1992).

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 postsynaptic neurons. A variety of such learning rules, in this case both Hebbian learning (see Equation 1) and trace learning (see Equation 3), may be implemented with different learning properties.

Preprocessing of the visual input by Gabor filters. Before the visual images are presented to VisNet's Input Layer 1, they are preprocessed 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 article, the input filters used were 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 (Cumming & Parker, 1999; J. P. Jones & Palmer, 1987). The input filters used were computed by the following equations:

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

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 & 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 article, an array of Gabor filters was generated at each of 256 × 256 retinal locations with the parameters given in the bottom section of Table 1.

The outputs of the Gabor filters were passed to the neurons in Layer 1 of VisNet according to the synaptic connectivity given in the bottom section of Table 1. That is, each Layer 1 neuron received connections from 201 randomly chosen Gabor filters localized 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 was 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-organizing map. In this article, we ran simulations with a self-organizing map (SOM; Kohonen, 1982; Von der Malsburg, 1973) implemented within each layer. In the SOM architecture, short-range excitation and long-range inhibition were combined to form a Mexican-hat spatial profile and were 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 were convolved with a spatial filter, I_{ab}, where δ_I controlled the inhibitory contrast and δ_E controlled the excitatory contrast. The width of the inhibitory radius was controlled by σ_I, whereas the width of the excitatory radius was controlled by σ_E. The parameters a and b index the distance away from the center of the filter. The lateral inhibition and excitation parameters used in the SOM architecture are given in the bottom section of Table 1.

Contrast enhancement of neuronal firing rates within each layer. Next, the contrast between the activities of neurons with each layer was 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 were constant within each layer, although α was adjusted within each layer of neurons to control the sparseness of the firing rates. For example, to set the sparseness to 4%, the threshold was 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 the bottom section of Table 1. These are general robust values found to operate well. They are similar to the standard VisNet sigmoid parameter values that were previously optimized to provide reliable performance (Stringer et al., 2006; Stringer & Rolls, 2008; Stringer, Rolls, & Tromans, 2007).

Information analysis. A single-cell information measure was applied to the trained network of Eguchi et al. (2016) to identify the different subpopulations of output (fourth 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 was applied to individual cells in Layer 4 and measured how much information was available from the response of a single cell about which stimulus category, that is, 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 was computed using the following formula, with details given by Rolls, Treves, Tovee, and Panzeri (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

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

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

The maximum amount of information that can be attained was $\log_2(N)$ bits, where N is the number of stimulus categories. For the

case of two stimulus categories, that is, 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 randomized computer generated face images, which were created using the software package FaceGen (2013). 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 (see Equation 1) with weight vector normalization (see Equation 2). The presentation of the 100 randomized 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 (fourth layer) neurons before and after pretraining on the randomized face images. The plot shows the information carried by the fourth-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 are 1 bit. The dashed line represents the untrained network, whereas the solid line represents the trained network. The result shows that pretraining VisNet on many randomly generated faces significantly increased the amount of single-cell information carried by fourth-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 were 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). It is interesting that 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 to shift the cognitive bias from negative to positive.

Experiment 2a: CBM by CT Learning

Method. In this section we describe how, after pretraining VisNet on 100 randomized faces as described earlier, 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 (see Equation 1) with weight vector renormalization (see Equation 2). Figure 3c shows examples of the face stimuli used to perform CBM retraining by CT learning. The image set was constructed from five different facial identities. For each of these facial identities, 10 face images were constructed by sampling 10 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 and then transformed 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 five facial identities in turn, which constituted one epoch of training. The network underwent a total of 50 training epochs. In this situation, CT learning (Stringer et al., 2006) began 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 was presented, it stimulated the *happy* output (fourth layer) neurons to respond. Then, as the face gradually morphed from happy to neutral, the *happy* output cells continued 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 were remapped onto the happy output neurons through the Hebbian learning rule (see Equation 1) with weight vector renormalization (see Equation 2). This retraining was carried out for each of the five 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 became remapped onto the *more happy* output representations. Thus, CBM occurred.

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. To do this, we began by reanalyzing 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: that is, *happy* versus *sad* subpopulations. The performance of the CBM retraining was assessed by recording and analyzing the firing rates of the *happy* and *sad* subpopulations of output neurons in response to the

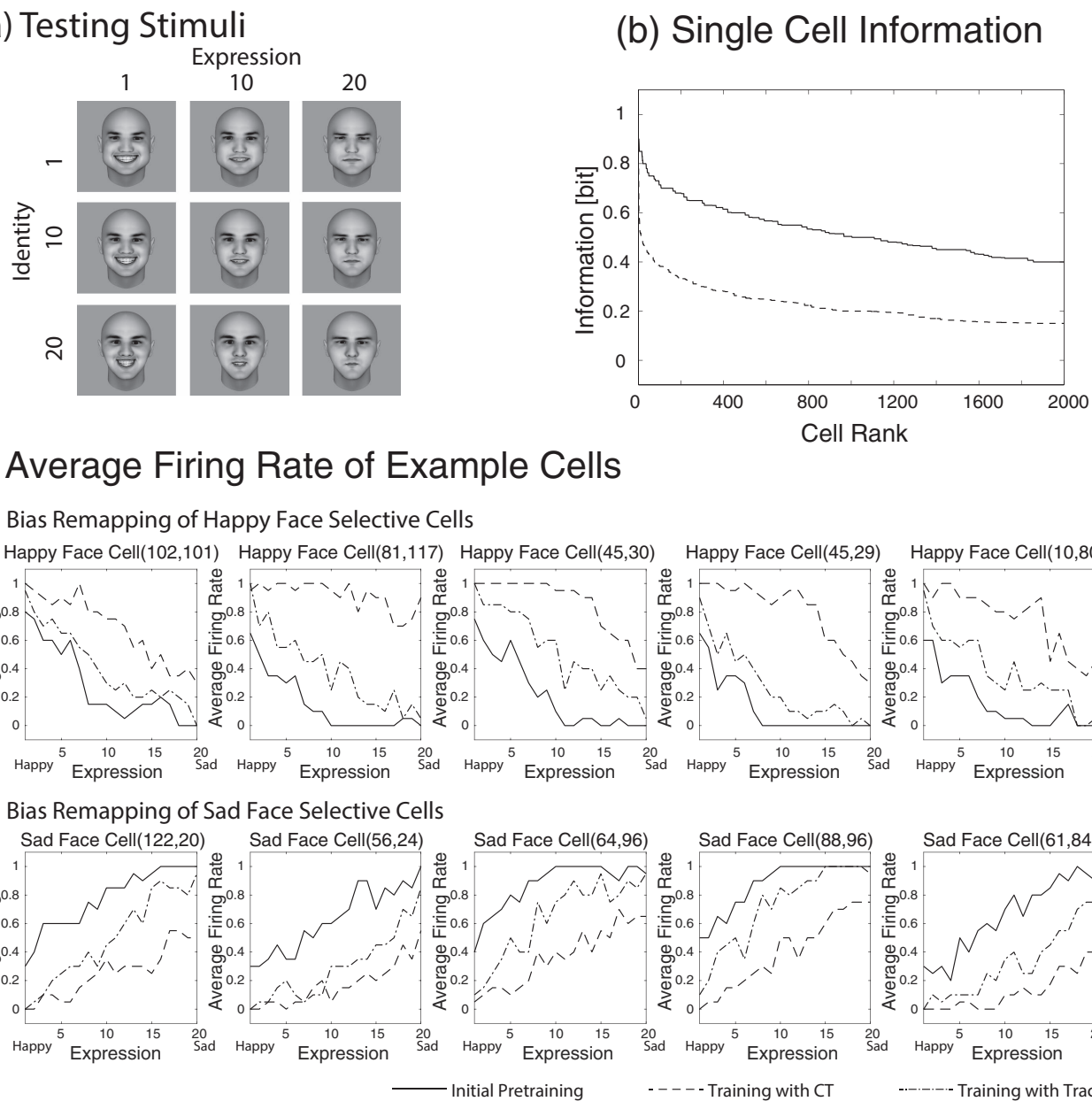

Figure 4 (opposite)

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 one-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 one-dimensional space of 20 different expressions that varied gradually from sad to happy. This produced a matrix of 400 face stimuli constructed from 20 identities × 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 (see 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 were averaged over the 20 different facial identities. It can be seen that the *happy* output neurons responded with a greater average firing rate across the space of expressions after CBM training by CT learning. In particular, CBM retraining 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 responded more strongly on average than did 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 responded more strongly than did the *happy* output neurons to only facial expressions greater than 16 on the happiness scale. Thus, CBM retraining remapped the more neutral faces away from the *sad* output neurons and onto the *happy* output neurons. In particular, CBM retraining was able to shift the bias in the network from negative to positive using a biologically plausible Hebbian learning rule (see Equation 1) with weight vector renormalization (see Equation 2) when the faces were 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 randomized faces as described earlier. During this, the network was retrained on faces with either happy or neutral expressions, with the synapses modified using the trace learning rule (see Equation 3) with weight vector renormalization (see Equation 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 five happy faces, whereas the bottom row shows five 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 reaching Happy Face 25 and 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 & Rolls, 1997) encourages 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

Figure 4 (*opposite*). Panel a: The face stimuli used to test VisNet. A one-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 one-dimensional space of 20 different expressions that varied gradually from sad to happy. Panel b: The amount of information carried by output (fourth layer) neurons after pretraining VisNet. The plot shows the information carried by all of the fourth-layer neurons about either happy or sad expressions, where the neurons are plotted in rank order along the abscissa. Panel c: Demonstration of CBM by CT learning (c1 row) and trace learning (c2 row) in VisNet. The firing rates of five *happy* output neurons and *five sad* output neurons were recorded in response to the matrix of test faces shown in Panel 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 were averaged over the 20 different facial identities. Panel 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 were averaged over the 20 different facial identities. The subplot (d1) shows the output of the network directly before CBM retraining, and the subplots d2 and d3 show the output of the network after CBM retraining with CT learning and with trace learning, respectively. CBM = cognitive bias modification; CT = continuous transformation.

network become remapped onto the *more happy* output representations. Hence, CBM takes place.

Results. The network performance was assessed in a manner similar to that described earlier for CT learning in Experiment 2a. After pretraining the network on the set of 100 randomly generated faces (see 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 were averaged over the 20 different facial identities. It can be seen that the *happy* output neurons responded with a greater average firing rate across the space of expressions after CBM training by trace learning. In particular, CBM retraining 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 (see Figure 4d1) shows the output of the network directly before CBM retraining, and the subplot (see Figure 4d3) 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 responded more strongly on average than did 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 responded more strongly than did the *happy* output neurons to only facial expressions greater than 18 on the happiness scale. Hence, the more neutral faces had been remapped away from the *sad* output neurons and onto the *happy* output neurons by the CBM retraining. In particular, CBM retraining had shifted the bias in the network from negative to positive using a biologically plausible trace learning rule (see Equation 3) with weight vector renormalization (see Equation 2) when the faces were presented with the happy and neutral expressions shown in Figure 3d interleaved.

Discussion

In this article we described and modeled two alternative CBM training mechanisms: continuous transformation (CT) learning (Stringer et al., 2006) and trace learning (Foldiak, 1991; Wallis & 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, whereas 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 to eliminate the kind of negative cognitive biases associated with clinical depression.

To our knowledge, this is the first study to model 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 & Mathews, 2000; Mathews & Mackintosh, 2000), which in turn can reduce the risk for depression recurrence (Holmes, Lang, & Sham, 2009). This article 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 article, we successfully demonstrated how computational models can be used to explore and exploit existing psychological phenomena to optimize 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 earlier or whether participants will be required to actively engage in the task 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 were asked to rate facial expressions 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 the predicted bias change will occur and also whether a concurrent reduction in clinical symptoms arises. Thus, there are important clinical implications of the current modeling 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 helps to guide future research aimed at optimizing the effectiveness of CBM interventions. For example, the simulations presented in this article 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 Bülthoff (2001) have also shown that both spatial and temporal continuity seem to play a key role for modifying recognition memory. In addition, a recent modeling 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, Eguchi, & Stringer, 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 optimize CBM procedures using feedback.

We will also explore various architectural extensions to the model, 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 was 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 included 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 was more likely to win the competition when such biased signals were present (Mathews & 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 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 & 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) 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. 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 prioritized 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 retrain 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. Because the original negative biases of patients also occur without requiring active training, it might be that the proposed mechanisms also bear some relation to the causative process of acquiring negative biases.

References

Anderson, M. C., & Green, C. (2001, March 15). Suppressing unwanted memories by executive control. *Nature, 410,* 366–369. http://dx.doi.org/10.1038/35066572

Arditte, K. A., & Joormann, J. (2014). Rumination moderates the effects of cognitive bias modification of attention. *Cognitive Therapy and Research, 38,* 189–199. http://dx.doi.org/10.1007/s10608-013-9581-9

Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry, 165,* 969–977. http://dx.doi.org/10.1176/appi.ajp.2008.08050721

Bishop, S. J. (2007). Neurocognitive mechanisms of anxiety: An integrative account. *Trends in Cognitive Sciences, 11,* 307–316. http://dx.doi.org/10.1016/j.tics.2007.05.008

Bourke, C., Douglas, K., & 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., & MacLeod, C. (2014). Absence of evidence or evidence of absence: Reflecting on therapeutic implementations of attentional bias modification. *BMC Psychiatry, 14,* 8. http://dx.doi.org/10.1186/1471-244X-14-8

Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). "Breaking" position-invariant object recognition. *Nature neuroscience, 8,* 1145–1147.

Cristea, I. A., Kok, R. N., & 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., & Parker, A. J. (1999). Binocular neurons in V1 of awake monkeys are selective for absolute, not relative, disparity. *Journal of Neuroscience, 19,* 5602–5618.

Deco, G., & Rolls, E. T. (2005). Neurodynamics of biased competition and cooperation for attention: A model with spiking neurons. *Journal of Neurophysiology, 94,* 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,* 1245–1255.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18,* 193–222.

Desimone, R., Wessinger, M., Thomas, L., & Schneider, W. (1990). Attentional control of visual perception: Cortical and subcortical mechanisms. *Cold Spring Harbor Symposia on Quantitative Biology, 55,* 963–971.

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458.

Eguchi, A., Humphreys, G. W., & Stringer, S. M. (2016). The visually-guided development of facial representations in the primate ventral visual pathway: A computer modeling study. *Psychological Review, 123,* 696–739. http://dx.doi.org/10.1037/rev0000042

Enock, P. M., Hofmann, S. G., & 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,* 200–216. http://dx.doi.org/10.1007/s10608-014-9606-

FaceGen. (2013). (Version 3.C.1) [Computer software]. Toronto, ON, Canada: Singular Inversions Inc.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3,* 194–200.

Fox, E., Mackintosh, B., & Holmes, E. (2014). Travellers' tales in cognitive bias modification research: A commentary on the special issue. *Cognitive Therapy Research, 38,* 239–247. http://dx.doi.org/10.1007/s10608-014-9604-1

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14,* 1195–1201.

Grabenhorst, F., & 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,* 1649–1660.

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences, 15,* 56–67.

Greenberg, P. E., Sisitsky, T., Kessler, R. C., Finkelstein, S. N., Berndt, E. R., Davidson, J. R. T., & Fyer, A. J. (1999). The economic burden of anxiety disorders in the 1990s. *Journal of Clinical Psychiatry, 60,* 427–435.

Grey, S., & Mathews, A. (2000). Effects of training on interpretation of emotional ambiguity. *Quarterly Journal of Experimental Psychology Section A, 53,* 1143–1162.

Grey, S., & 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,* 338–343.

Hakamata, Y., Lissek, S., Bar-Haim, Y., Britton, J. C., Fox, N. A., Leibenluft, E., . . . 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., & 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., & 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 Brian Research, 32,* 203–218.

Hoffman, D. L., Dukes, E. M., & Wittchen, H.-U. (2008). Human and economic burden of generalized anxiety disorder. *Depression and Anxiety, 25,* 72–90.

Holmes, E. A., Lang, T. J., & 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., & Mackintosh, B. (2010). Cognitive bias modification: The critical role of active training in modifying emotional responses. *Behavior Therapy, 41,* 73–81.

Jones, J. P., & 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, April 12). 50 million years of work could be lost to anxiety and depression. *Guardian*. Retrieved from https://www.theguardian.com/global-development/2016/apr/12/50-million-years-work-lost-anxiety-depression-world-health-organisation-who

Joormann, J., Hertel, P. T., Brozovich, F., & Gotlib, I. H. (2005). Remembering the good, forgetting the bad: Intentional forgetting of emotional material in depression. *Journal of Abnormal Psychology, 114,* 640–648.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics, 43,* 59–69.

Li, N., & DiCarlo, J. J. (2008, September 12). 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., & 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., & 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., & Mackintosh, B. (1998). A cognitive model of selective processing in anxiety. *Cognitive Therapy and Research, 22,* 539–560.

Mathews, A., & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of Abnormal Psychology, 109,* 602–615.

Mathews, A., & 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., . . . 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., & Hirshfeld-Becker, D. R. (2014). Efficacy of interpretation bias modification in depressed adolescents and young adults. *Cognitive Therapy and Research, 38,* 89–102. http://dx.doi.org/10.1007/s10608-013-9578-4

Mogoaşe, C., David, D., & 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., & 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., & Stringer, S. M. (2006). Spatial vs temporal continuity in view invariant visual object recognition learning. *Vision Research, 46,* 3994–4006.

Petkov, N., & 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., & Gilbert, C. D. (1992). Dynamic changes in receptive-field size in cat primary visual cortex. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 89,* 8366–8370.

Preminger, S., Sagi, D., & 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., & Fox, R. (2002). Anxiety-related bias in the classification of emotionally ambiguous facial expressions. *Emotion, 2,* 273–287.

Roiser, J. P., Elliott, R., & Sahakian, B. J. (2012). Cognitive mechanisms of treatment in depression. *Neuropsychopharmacology, 37,* 117–136.

Rolls, E. T., & 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., & 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., & Paré, D. (2003, April 3). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature, 422,* 518–522. http://dx.doi.org/10.1038/nature01530

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., & Moyer, A. (2010). Depression as a predictor of disease progression and mortality in cancer patients: A meta-analysis. *Cancer, 116,* 3304. http://dx.doi.org/10.1002/cncr.25318

Spoerer, C. J., Eguchi, A., & 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., & Proske, J. H. (2006). Learning invariant object recognition in the visual system with continuous transformations. *Biological Cybernetics, 94,* 128–142.

Stringer, S. M., & Rolls, E. T. (2008). Learning transform invariant object recognition in the visual system with multiple stimuli present during training. *Neural Networks, 21,* 888–903.

Stringer, S. M., Rolls, E. T., & Tromans, J. M. (2007). Invariant object recognition with trace learning and multiple stimuli present during training. *Network, 18,* 161–187.

Surcinelli, P., Codispoti, M., Montebarocci, O., Rossi, N., & Baldaro, B. (2006). Facial emotion recognition in trait anxiety. *Anxiety Disorders, 20,* 110–117.

Tromans, J. M., Harris, M., & 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.

Ustün, T. B., Ayuso-Mateos, J. L., Chatterji, S., Mathers, C., & Murray, C. J. L. (2004). Global burden of depressive disorders in the year 2000. *British Journal of Psychiatry, 184,* 386–392.

Von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striate cortex. *Kybernetik, 14,* 85–100.

Wallis, G., & Bülthoff, H. H. (2001). Effects of temporal association on recognition memory. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 98,* 4800–4804.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

West, G. L., Anderson, A. A. K., Ferber, S., & Pratt, J. (2011). Electrophysiological evidence for biased competition in V1 for fear expressions. *Journal of Cognitive Neuroscience, 23,* 3410–3418.

Received April 29, 2016
Revision received October 18, 2016
Accepted October 19, 2016 ■

Correction to Eguchi et al. (2016)

In the article "Understanding the Neural Basis of Cognitive Bias Modification as a Clinical Treatment for Depression" by Akihiro Eguchi, Daniel Walters, Nele Peerenboom, Hannah Dury, Elaine Fox, and Simon Stringer (*Journal of Consulting and Clinical Psychology*, Advance online publication. December 19, 2016. http://dx.doi.org/10.1037/ccp0000165), there was an error in the **Discussion** section's first paragraph for **Implications and Future Work**. The in-text reference citation for Penton-Voak et al. (2013) was incorrectly listed as "Blumenfeld, Preminger, Sagi, and Tsodyks (2006)". All versions of this article have been corrected.

http://dx.doi.org/10.1037/ccp0000193