Towards a Neural Model of Bonding in Self-Attachment

David Cittern
Imperial College London, UK
Email: david.cittern10@ic.ac.uk

Abbas Edalat
Imperial College London, UK
Email: a.edalat@ic.ac.uk

***Abstract*—We build on a previous neural model of bonding circuitry, in which the orbitofrontal cortex mediates between facilitative and stress reactivity to social stimuli, via the dorsomedial and paraventricular nucleus of the hypothalamus. We integrate recent neuroscientific findings, and consider how the introduction of additional reward could drive a further, counter-conditioning based re-balancing mechanism between activation of these networks, via increasing prefrontal-driven inhibition of the central nucleus of the amygdala. We simulate our model computationally, and hypothesise that such a process may be involved in a particular phase of self-bonding in the newly introduced self-attachment psychotherapy.**

I. Introduction

Early insecure attachment experiences are now believed to have important implications with regards to the development of capacities for self-regulation of emotion and the governing of various aspects of social behaviour. Self-attachment is a new attachment-based psychotherapy that has recently been proposed as a method for re-training an individual's sub-optimal attachment schema [1]. Rooted in neuroscientific theories of attachment, it consists of a number of self administrable protocols which aim to recreate the effects of positive infant-parent Right Hemisphere (RH) to RH interactions using instead internal Left Hemisphere (LH) to RH interactions. With initial success in pre-clinical trials, a key hypothesis with regards to the therapy is that it facilitates the construction of new neural circuitry between the Orbitofrontal Cortex (OFC) and limbic system, in order to increasingly contain pathological and suboptimal neural activity related to early insecure attachment experiences.

We build on previous work and recent attachment-related neuroscientific findings to present a computational model of fear counter-conditioning within neural circuits mediating stress and bonding reactivity to social stimuli. We argue that our model can aid in understanding the dynamics underlying a successful application of a particular phase of the self-attachment therapy, which is concerned with the creation of an abstract, self-directed bond.

The remainder of this paper is organised as follows. In Section II we outline recent findings from neuroscience on early attachment-related brain development, along with the self-attachment psychotherapy. In Section III we detail our new model, and in Section IV we give initial simulation results which we propose correspond to a successful application of a particular phase of the self-attachment protocol. Finally, in Section V, we provide a summary and some suggestions for future work.

II. Background

During the early stages of life, an infant is highly dependent on others for their survival. Attachment theory grew out of the pioneering work of John Bowlby, who proposed that in order to fulfil their basic survival needs each infant had evolved a genetic predisposition to seek out an attachment relationship with a primary caregiver. Moreover, he argued that the nature of these early attachment interactions was significant and had profound implications with regards to the construction of internal working models of self and other.

Whilst sensitive-responsive caregiving within the context of attachment interaction fosters secure-base exploration and optimal cognitive-emotional neural development and integration; neglectful, inconsistent and fear-inducing patterns of early attachment interaction have been linked to the development of avoidant, anxious and disorganised attachment schemas, respectively. According to the secure-base paradigm, insecure attachment hinders early exploratory and social functioning as a result of a diversion of resources towards more immediate and primitive survival needs. Disorganisation is particularly significant, since it has been linked to an increased risk for the later development of serious pathological disturbances such as dissociative [2] and borderline personality disorders [3]. In addition, an intergenerational effect has been observed, whereby these caregivers themselves had disorganised attachment relationships as children [4].

A. Neuroscience of attachment and bonding

The internal working model (attachment schema) has been theorised to be based in unconscious and implicit memories, rooted mainly in RH brain regions centred on the OFC, amygdala and hypothalamus [5] [6, p.139]; areas known to be central to social cognition, emotional processing and fear conditioning. Recent neuroimaging studies which have specifically investigated the neural correlates of attachment, bonding and parenting have confirmed and elaborated on this picture.

The dimension of attachment anxiety has been found to correlate with a relatively over-active amygdala, a region long known to be involved in fear and stress-related reactivity, and more recently posited as playing a general role in salience processing (see Section III). For example, [7] found increased amygdala activation as a function of increasing attachment anxiety when images of facial expressions were used to provide social feedback on performance during a perceptual game. In [8], insecurely attached women showed elevated amygdala activation relative to securely attached women when exposed

978-1-4799-1959-8/15/$31.00

to the crying of a non-related infant. In another study [9], cortisol measurements were collected across an entire day (in a non-laboratory setting), and anxious attachment was found to be correlated with a relatively elevated and flat cortisol profile.

Much evidence from rat studies implicates the Dopamine (DA) reward system in maternal and bonding behaviour ([10]), and a similar picture has started to emerge for humans. For example, [11] used fMRI to look at activity in reward circuitry in response to mothers viewing photographs of their own relative to other people's children. They found increased activation of the Ventral Tegmental Area (VTA) and Substantia Nigra pars Compacta (SNc), which both contain major concentrations of DA-releasing neurons; and the dorsal striatum (Caudate Nucleus (CN) and putamen) and OFC, regions which are known to receive dopaminergic projections. In addition, in [12] relatively low ventral striatum and OFC activity was found in avoidant vs secure mothers in response to seeing own-baby images. Furthermore, reward circuit activation in social feedback appraisal situations appears to be modulated by attachment type. For example, [7] found that increasing attachment avoidance was correlated with decreasing activations in the VTA and ventral striatum in response to facial images conveying social feedback. Some evidence suggests involvement of reward circuitry in response to infants in general (i.e. not necessarily own infants). For example, [13] found activation of the medial OFC (a region implicated in stimulus-reward association) in response to images of unfamiliar infants but not adults.

Another important hormone with regards to attachment is Oxytocin (OXT). Elevated OXT has been linked to a range of pro-social behaviours, including parenting and bonding behaviours, and elevated trust. For example, [14] found that higher levels of OXT predicted greater amounts of maternal behaviours such as gaze towards infant, positive affect and affectionate touch, during the postpartum period. Another study looked at OXT in both new mothers and fathers postpartum [15], finding that both maternal and paternal OXT levels increased across the period. Whilst maternal OXT was related to the amount of affectionate parenting behaviours (as in [14]), paternal OXT correlated with stimulatory parenting behaviours such as "proprioceptive contact, tactile stimulation, and object presentation". Furthermore, OXT levels have been linked to earlier attachment experience and social development. For example, [16] found that adult women exposed to childhood maltreatment had lower OXT levels compared to controls.

Despite strong evidence for the involvement of OXT in encouraging many types of pro-social and attachment-related behaviour, it appears increasingly unlikely that OXT simply acts as a universal bonding hormone. For example, intranasal OXT administration was found to decrease cooperation during trust games, when participants interacted with anonymous strangers compared to familiar persons with whom they had previously been acquainted [17]. Intranasal OXT has also been found to decrease the likelihood of cooperation during a social dilemma game in adults with borderline personality disorder and high levels of attachment anxiety [18]. In [19], securely attached individuals remembered their mother as more caring and close in childhood following intranasal OXT relative to placebo, whereas anxiously-attached individuals remembered their mother as less caring and close (using self-report measures). Such results suggest that, rather than universally promoting bonding behaviour, exogenous OXT administration may instead result in an amplification of the influence of pre-existing interpersonal schemas [20], likely making exogenous OXT administration unsuitable as a treatment for many attachment-related disorders. This provides motivation for the development of therapies in which OXT secretion is naturally enhanced as a side effect of changes to the underlying attachment schema.

B. Self-attachment therapy

Self-attachment [21] is a new attachment-based psychotherapy, rooted in the belief that many affect dysregulation disorders have their primary causes in suboptimal early attachment experience. The therapy aims to naturally stimulate the release of OXT and DA, in order to encourage neural plasticity and increasingly contain suboptimal and pathological neural activity and increase self-agency. It is related to a class of compassion-focused therapies and compassion-focused imagery techniques, which have shown effectiveness with regards to stress and affect regulation [22].

Under the self-attachment paradigm, the self of the individual undergoing the therapy is conceptualised as being split into an inner child and inner adult. The inner child corresponds to the emotional self that becomes dominant under stress, thought to be rooted mainly in the RH. The inner adult corresponds to the more rational self dominant under times of calm and low perceived threat, that is thought to be rooted mainly in the LH. In essence, the therapy aims to recreate the effects of early RH to RH attachment-based interactions between an infant and primary caregiver, using instead LH to RH interactions within the individual's own brain. This is achieved by means of simulating (using imagery techniques) the interactions between a child and secure caregiver, with the inner child and inner adult serving as both the source of, and target for, attachment-based compassion.

In order to complete later stages of the therapy, the individual must first establish an imaginative (but passionate) loving relationship with the inner child, which is subjectively experienced as falling in love with them. There is evidence to suggest that humans are capable of, and indeed driven to, form bonds with both inanimate objects and non-material beings of a more abstract nature. For example, it is common for children to form bonds with inanimate transitional objects that can serve as mother substitutes [23]. In addition, throughout much of history, bonds formed and reinforced through religious practice have been a predominant means of regulating emotion and social behaviour. A key focus for adherents to the Abrahamic religions is on the development of a personal and intimate relationship with God, and it has been argued that such a bond meets many of the criteria of an attachment relationship [24]. An fMRI study of Danish Christians [25] found activation of the CN during formal prayer, suggesting that, as with attachment and bonding to a human, these effects are mediated (at least in part) by the brain's reward system. As discussed previously, the CN receives dopaminergic projections, and has been proposed to play an important role in attachment-related approach behaviours [26].

A key technique in self-attachment is to enhance the bonding process through the use of activities such as song and

dance directed towards the inner child which, it is thought, stimulate the dopaminergic reward system. Music, which has been proposed to play a primary role in synthesis within the mind [27], has long been recognised as a powerful mediator of emotional state, and fMRI studies have shown correlated activation in a wide range of brain regions implicated in emotional processing (e.g. [28]). Furthermore, recent studies have shown evidence for the involvement of the reward system during passive listening to self-reported pleasurable music. For example, in [29], fMRI and PET scans revealed striatal DA release in the CN and Nucleus Accumbens (NAc) in response to anticipation and peak emotional response to the music, respectively. Song, too, has been shown to activate emotion and reward circuitry in both overt and imagined form. [30] compared brain activity while subjects either spoke or sang words to a familiar song. For singing opposed to speaking, they found a relative increase in areas including medial PFC and NAc. In another study [31], professional classical singers were asked to imagine singing an aria (love song), resulting in intense activation in emotional areas (including amygdala, anterior cingulate, and medial prefrontal cortex) along with the CN and putamen.

As discussed above, evidence suggests involvement of reward circuitry during infant interaction, especially for own-infant. In addition, in songbirds at least, it is known that singing directed towards a potential mate, but not undirected singing, results in increased DA concentrations in the VTA [32]. Thus, we hypothesise that song directed towards the conceptualised inner-child is also likely to be a powerful activator of the dopaminergic reward system in humans. In this paper, we begin to explore (in the form of a computational model) some of the neural mechanisms that might underlie the formation of this abstract, self-directed bond. During this phase of the protocol, the volunteer focuses on happy and unhappy images of themselves as a child in order to conceptualise the inner-child, before attempting to create an attachment relationship with this abstract entity. The volunteer is encouraged to sing and/or dance (either overtly or imagined) with the inner child in order to accelerate and enhance this bond making process. The hypothesis presented in this work is that the success of such techniques in driving the self-directed bond formation process may be (at least in part) due to their facilitation of a form of counter-conditioning for fearful associations formed to classes of social stimuli.

C. Related work

A series of recent works have begun to explore the dynamics of attachment using computational models. In [33], a number of cognitive agent architectures are presented, aiming to capture empirically observed behavioural aspects of infant attachment using a winner-take-all competition between fear and attachment behavioural systems (for survival), and exploratory and socialisation systems (for learning). In developmental robotics, the attachment secure-base and dyadic arousal regulation paradigms are now being studied as mechanisms for driving a robot's exploration in a novel environment [34]. Another recent study presented a neural-cognitive architecture in an attempt to explain adaptive infant attachment behaviour and physiology in terms of approach/avoid tendencies mediated by the OFC and fear circuitry in the amygdala [35]. Attachment schemas and prototypes have also been considered within the context of strong patterns in a Hopfield network [36].

III. Model

We now present a computational neural model to attempt to explain how bonding circuitry activation may be strengthened, and stress reactivity dampened, under a counter-conditioning paradigm. This is achieved by introducing and associating additional reward with a class of social stimuli that have previously been conditioned as being fearful/threatening in nature, resulting in a prefrontal-mediated inhibition of stress circuitry and increased release of OXT. The overall architecture is shown in Fig. 2.

Our starting point is Levine's neural model in [37], which identified the OFC as the key region in mediating the relative strength of activity in fight/flight circuitry (focused on the amygdala, Parvocellular part of the Paraventricular Nucleus of the Hypothalamus (PVNp) and Locus Coeruleus (LC)) and bonding circuitry (focused on the reward system). In particular, it is proposed that the OFC sends, via the Dorsomedial Hypothalamus (dmH), different inhibitory strengths to the Magnocellular part of the Paraventricular Nucleus of the Hypothalamus (PVNm) (which controls release of OXT and Vasopressin (VA)) and the PVNp (which controls release of Corticotrophin releasing factor (CRF)). Release of OXT and VA to the reward system enhances the facilitation of the creation of social bonds, whilst CRF release serves to further stimulate activity in the amygdala-PVNp-LC stress loop, inhibiting social approach.

The OFC and Basolateral Amygdala (BLA) are known to have extensive bidirectional connections, which are implemented in our model within a Deep Belief Network (DBN) (using the architecture shown in Fig. 1). The basic building block of a DBN is a Restricted Boltzmann Machine (RBM) [38], which has recently been used for modelling of the psychotherapeutic process [39]. An RBM is a stochastic generative neural network defined by visible units x and hidden units h, and parametrised by $\theta = \{b, c, W\}$, with weights W, hidden biases b and visible biases c. In an RBM h and x are conditionally independent given each other, and each configuration of x and h is assigned a scalar energy $E(x,h) = -\sum_{i,j} W_{ij} h_i x_j - \sum_j c_j x_j - \sum_i b_i h_i$ giving joint-distribution over x and h: $p(x,h) = e^{-E(x,h)} \left(\sum_{x,h} e^{-E(x,h)}\right)^{-1}$. For both x and h binary, units are activated according to a probability given by the logistic function $\sigma(x) = (1 + e^{-x})^{-1}$. A learning procedure called contrastive divergence [40] gives a simple Hebbian-like gradient descent learning rule for parameter updates. To construct a DBN, we first train an RBM and then use its hidden layer activations as the visible layer in another RBM. It has been shown that a variational lower bound on $log\ p(x)$ can always be increased with each additional new layer [41], with successive layers of hidden units coming to learn increasingly higher-order features over the hidden unit activations of the previous layer.

In both humans and animals, the amygdala has long been known to be crucially important in fear conditioning, in which a previously neutral Conditioned Stimulus (CS) is repeatedly paired with a noxious Unconditioned Stimulus (US) that elicits a (fear related) Unconditioned Response (UR). Over time, the

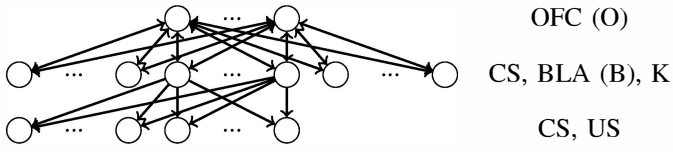

Fig. 1. BLA-OFC DBN architecture

CS comes to elicit the UR independent of the US, as a result of the learnt association. It is believed that the CS and US representations enter and converge on the BLA, which seems to be important in fear acquisition; whilst expressions of the fear response are triggered by the Central Nucleus of the Amygdala (CeA)'s various projections to regions including the hypothalamus and parasympathetic nervous system.

In addition to responding to stimuli with negative emotional valance, it is now known that the amygdala additionally also responds to stimuli with both positive and neutral valance, and furthermore appears to respond preferentially to social vs non-social stimuli. For example, [42] found similar levels of amygdala reactivity to positive and negative faces (social stimuli), but more activation to neutral social relative to neutral non-social stimuli. Such findings support theories for a more general role of the amygdala in the processing of stimuli that are predictive of biologically relevant events (within the context of the individual's current needs).

In our model we therefore consider associations between CS and both appetitive and aversive (biologically salient) US representations. Both CS and US representations enter the BLA at the input level, and the activations of the first layer of hidden units (the BLA layer, B) in the DBN serve as input to the CeA feed-forward network C (Eq. 2). Thus, the BLA layer B (the first hidden layer in the DBN) learns latent representations over associations formed between the CS and US.

In addition to feeding forward into the CeA, BLA activations at B also serve as input into to the OFC (the top hidden layer of the DBN, O), along with the CS representation and $n = 10$ softmax units representing associated rewards, giving a modified DBN architecture. The activation of the OFC's hidden layer O in turn serves as input to the Intercalated Cells of the Amygdala (ITC) feed-forward network I (Eq. 3), whilst the activation state of the OFC's reward nodes K serves as input to the dmH-PVNp feed-forward network D (Eq. 4). This flow of information from the BLA to the OFC, with the BLA hidden unit activations B in turn influencing the OFC hidden unit activations O, is consistent with a recent study that found a dominant directional influence from BLA to OFC at the point of decision making [43].

A. Stress/Arousal System

The stress/arousal system is intended to encapsulate the basic dynamics of the CeA, PVNp and LC loop described in [37]. It is excited by input from the BLA (capturing BLA-CeA excitation), inhibited by the dmH (capturing dmH-PVNp inhibition), and in turn excites the BLA (capturing Norepinephrine (NE)-based stimulation by the LC). This creates a positive feedback loop which serves to further enhance stress levels once this circuit is activated, until stimuli inputs significantly change. The stress level at time t is given by:

$$S(t) = \max(0, C(t) - I(t) - D(t)) \tag{1}$$

for $C(t)$ the activation strength along the BLA-CeA pathway at time t, and $I(t)$ and $D(t)$ the inhibitory strengths of the ITC and dmH at time t, respectively. The activation strength of the BLA-CeA pathway at time t is:

$$C(t) = \sum_i \tanh\left(\sum_j W_{ij} B_j(t)\right) + \gamma S(t-1) \tag{2}$$

where $B_i \in \{0, 1\}$ is the first hidden-layer activation in the BLA-OFC DBN (i.e. the BLA layer), and $S(t-1)$ is the stress/arousal level for the previous timestep. Thus, the BLA activates the CeA at a strength dependent on feedback from the stress/arousal system in the previous timestep (with $\gamma = 0.1$). Inhibitory strength of the ITC at time t is given by:

$$I(t) = \sum_i \tanh\left(\sum_j W_{ij} O_j(t)\right) \tag{3}$$

where $O_i \in \{0, 1\}$ is the top hidden-layer activation in the BLA-OFC DBN (i.e. the OFC layer). Similarly, inhibitory strength of the dmH on the PVNp pathway at time t is given by $D(t)$, where:

$$D(t) = \sum_i \tanh\left(\sum_j W_{ij} K_j(t)\right) \tag{4}$$

for $K_i \in \{0, 1\}$ the state of the OFC's associative reward node i at time t (determined by running a Gibbs chain). OXT levels $\phi(t)$ at time t are calculated based on the strength of the inhibitory input to the PVNm:

$$\phi(t) = \max\left(1, q - g\left(\frac{q-1}{M_{max}}\right)\right) \tag{5}$$

where $q = 2$ is a parameter controlling maximum OXT level, $g \sim \mathcal{N}(M(t), 0.05)$ (to introduce a very small amount of random noise in OXT levels), and $M(t)$ is the strength of the inhibition on the PVNm at time t:

$$M(t) = \sum_i \tanh\left(\sum_j W_{ij} K_j(t)\right) \tag{6}$$

and $M_{max} = \sum_i \tanh(W_{i1})$ is the maximum inhibition on the PVNm (i.e. inhibition on the PVNm for minimum reward input from the OFC).

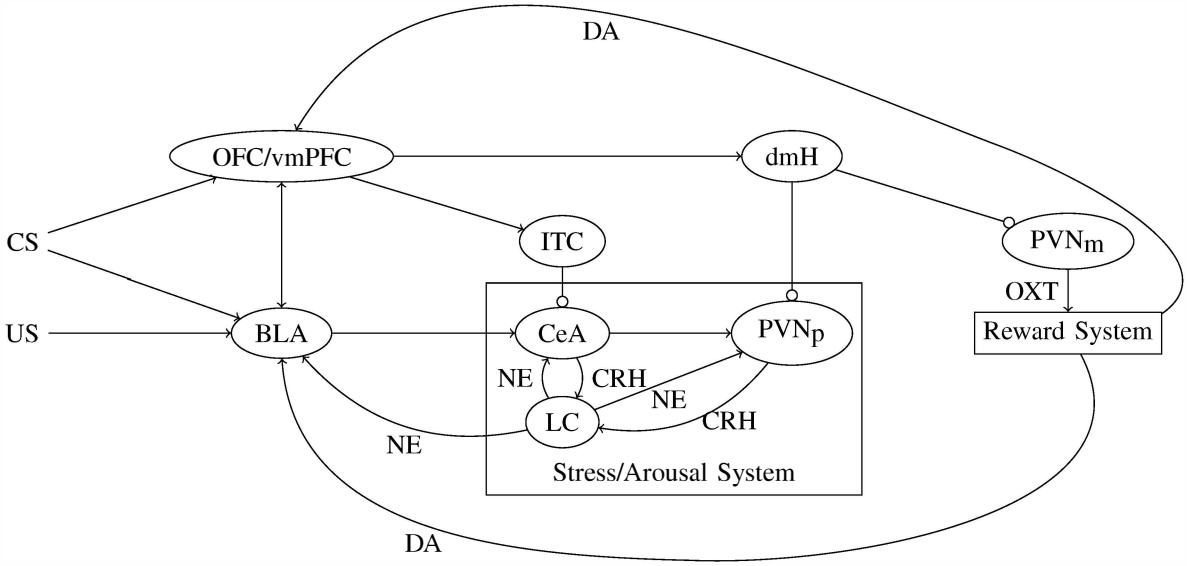

Fig. 2. The overall neural architecture, based on both [37] and [44]. Conditioned (CS) and unconditioned (US) stimuli representations enter the basolateral amygdala (BLA), with CS representations also entering the orbitofrontal/ventromedial prefrontal cortex (OFC/vmPFC). Stress reactivity in the central nucleus of the amygdala (CeA) - parvocellular part of the paraventricular nucleus of the hypothalamus (PVNp) - locus coeruleus (LC) loop is stimulated by the BLA, and inhibited by OFC/vmPFC - intercalated cells (ITC) and dorsomedial hypothalamus (dmH) - PVNp pathways. Oxytocin (OXT) release is controlled by the OFC/vmPFC - dmH - magnocellular part of the paraventricular nucleus of the hypothalamus (PVNm). Dopamine (DA) release drives inhibitory learning in vmPFC-ITC pathways.

B. Reward System

There is strong evidence to suggest that phasic DA firing signals an appetitive reward prediction error (i.e. unexpected rewards) [45]. Midbrain DA neurons in the VTA project to the OFC and Ventromedial Prefrontal Cortex (vmPFC) via the mesocortical pathway, and evidence from rodent studies suggests a critical role for DA in the prefrontal cortex for the consolidation and retrieval of fear extinction (see [46] for an overview). This may be at least in part due its phasic signalling of appetitive reward prediction errors, strengthening vmPFC-ITC connections known to result in inhibition of the CeA fear response [47] [44].

Although mechanisms underlying the signalling of aversive outcomes and unexpected punishments are less clear, DA projections from the VTA do also reach the amygdala (both directly, and via the NAc and mesolimbic pathway), and there is some evidence to suggest a role for DA in fear acquisition [46]. A recent theory proposes that whilst DA neurons in general respond to unexpected cues, there are in fact two distinct types of DA neuron population [48]. Under this view, one population of DA neurons, found in the ventromedial SNc and throughout the VTA, is involved in value learning, with increases in phasic firing for unexpected reward and slight decreases for unexpected punishment. These signals are projected to areas involved in value learning, such as the Nucleus Accumbens Shell (NAcs), dorsal striatum and vmPFC. A second population of DA neurons, found in the dorsolateral SNc and medial VTA, is involved in signalling salience, with increased phasic firing for highly salient events (regardless of valence). These signals are believed to be sent to areas involved in orienting, cognitive processing and general motivation, such as the Nucleus Accumbens Core (NAcc), dorsal striatum and dorsolateral prefrontal cortex.

In our model, we assume that reward predictions are computed by the OFC and signalled to the reward system. We consider value-signalling DA neurons, which encapsulate appetitive and aversive reward prediction errors in phasic firing patterns that are projected to the OFC/vmPFC and amygdala. In the model, reward predictions K are associative activations, resulting from a Gibbs chain initiated at a representation composed of the CS and BLA hidden unit activation, with only the reward (K) nodes un-clamped during the chain. The OFC's reward nodes K use a 1-of-n encoding (i.e. they encode n reward categories). They are activated according to a softmax function, and we assign the index of the activated node to a particular reward prediction $P(t) \in \mathbb{Z}$. More explicitly, we use $n = 20$ reward units and distribute these equally amongst positive and negative rewards. For indexed-reward values increasingly uniformly from $-n/2$ to $n/2$, exclusive of zero, then the predicted reward at time t is given by:

$$P(t) = \begin{cases} j - \frac{n}{2} - 1 & \text{if } j \leq \frac{n}{2} \\ j - \frac{n}{2} & \text{otherwise} \end{cases} \tag{7}$$

for $j \in \mathbb{N}^+$ the index of the activated unit in the OFC's softmax reward units at time t (i.e. $K_j(t) = 1$ and $K_{i \neq j}(t) = 0$). The reward-prediction temporal difference error $F(t)$ at time t is then given by:

$$F(t) = \phi(t)(R(t) + \delta P(t) - P(t-1)) \quad (8)$$

for $\delta = 0.1$ the temporal-difference discount factor, $R(t)$ the reward at time t, $P(t)$ the predicted reward at time t and $P(t-1)$ the predicted reward at time $t-1$.

Here we assume a modulating effect of OXT on phasic DA firing (the reward prediction error). In [49], it is suggested that OXT could influence both salience and valance attribution based on such an effect on the two distinct DA neuron populations discussed previously. This is based on evidence suggesting a role for OXT in increasing the salience of social cues, on the location of OXT receptors throughout the mesocorticolimbic system, on recent evidence suggesting that activation of OXT neurons that target the VTA stimulate DA neurons, and on behavioural evidence suggesting a role for OXT in facilitating shifts in valance attribution (see [49] for a discussion). In our model, OXT can modulate valance attribution by increasing phasic DA firing $F(t)$ for positive unmodulated reward prediction error $R(t)+\delta P(t)-P(t-1) > 0$, and decreasing $F(t)$ for unmodulated negative error.

C. Counter-conditioning

As discussed previously, in individuals with high attachment anxiety, stress circuitry appears to be over-active and could be triggered for a large and general class of social stimuli. A key part of self-attachment therapy is pairing classes of previously fearful or stress-inducing social stimuli with alternative representations (e.g. music) that naturally induce reward. Thus, we hypothesise that self-attachment therapy serves, at least in part, to counter-condition negative and fearful associations learnt for classes of social stimuli as a result of previous relational trauma. Based on findings from fear extinction, we consider here a mechanism which (in addition to the OFC-dmH-PVNp inhibitory pathway originally detailed in [37]) may inhibit stress circuitry.

In fear extinction, the CS is presented without the aversive US until it no longer elicits the UR. Rather than CS:US fear memories being "forgotten", it seems that fear extinction predominantly involves the creation of new CS:no-US memories, formed in descending connections from the vmPFC to the ITC, that in turn inhibit the CeA [47]. In contrast to fear extinction, in a counter-conditioning paradigm the CS is repeatedly paired with a qualitatively different (i.e. appetitive for fear) US, until the CS no longer elicits the UR. Although few studies have explicitly looked at the neural mechanisms underpinning counter-conditioning, behavioural experiments suggest that counter-conditioning to a positive (or even neutral) US is more effective in inhibiting a fear response compared to extinction alone (e.g. [50]).

Here we consider a similar vmPFC-ITC extinction inhibitory process to occur as a result of the counter conditioning-like aspects of the self-attachment protocol. Our computational model of this process is based on the recent fear extinction model in [44]. In that model, the learning of conditioned fear responses in the CeA is driven by prediction errors encapsulating unexpected punishment, that serve to strengthen activation based on signals coming from the BLA. On the other hand, extinction learning is driven by prediction errors encapsulating less punishment than expected, which strengthen activation of the ITC (which in turn inhibits the CeA) based on signals from the vmPFC. Also considered is the role of hippocampal inputs to the BLA and vmPFC in terms of context modulation, although we exclude those inputs here for simplicity.

Similarly to [44], we assume that positive reward prediction errors (i.e. signalling unexpected rewards) strengthen inhibitory activation along the vmPFC-ITC pathway, whilst negative reward prediction errors (i.e. unexpected punishments) strengthen activation of the BLA-CeA pathway. As in that model, we update vmPFC-ITC weights in a Hebbian manner according to:

$$\triangle W_{ij}(t) = \begin{cases} \alpha_{itc} F(t) O_i(t) I_j(t) & \text{if } F(t) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

and similarly, BLA-CeA weights according to

$$\triangle W_{ij}(t) = \begin{cases} -\alpha_{cea} F(t) B_i(t) C_j(t) & \text{if } F(t) < 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Thus, as compared to an extinction, our model will afford more rapid strengthening of ITC inhibition, driven by the relatively larger prediction errors (and thus weight updates) between the vmPFC and ITC. In our model, CS-representations are hypothesised to correspond to a general class of social stimuli that have previously been paired with negatively-valenced US-. US+ representations are related to the reward-inducing activities (such as song), and positively-valenced infant perceptions, that are utilised during the protocol. This gives a reasonable first approximation to a counter-conditioning model, although the finding of enhanced fear inhibition in counter-conditioning to a neutral US as compared to extinction may suggest a more intricate process in the brain [50].

IV. Simulation results

We first generated 40 random binary stimuli, each of length 200 bits (where each bit was activated independently with probability 0.1). These stimuli were split evenly amongst the four stimulus categories (CS+/- and US+/-), and each US+/- stimulus was assigned a reward/punishment with a surjective mapping to $\{x|x \in \mathbb{Z}, 1 \leq x \leq 10\}$ and $\{x|x \in \mathbb{Z}, -1 \geq x \geq -10\}$, respectively. We then created equal quantities of US+ paired with CS+ and random stimuli; and US- paired with CS- and random stimuli, to create an input dataset of size 200. The DBN was then trained on this input data (with 500 hidden units in the second and third layers, a learning rate of 0.05, a sparsity target of 0.01, an L2 weight decay penalty of 0.02, and batch size 50, for 2000 epochs), so that it learnt sparse latent representations for the CS and US associations.

The BLA-CeA feed-forward network was trained to have a high activation for CS-:US- pairs, using as input the first hidden layer (BLA) activations in the DBN, and the weight update rule described above with temporal difference errors proportional to the corresponding punishment for the US-.

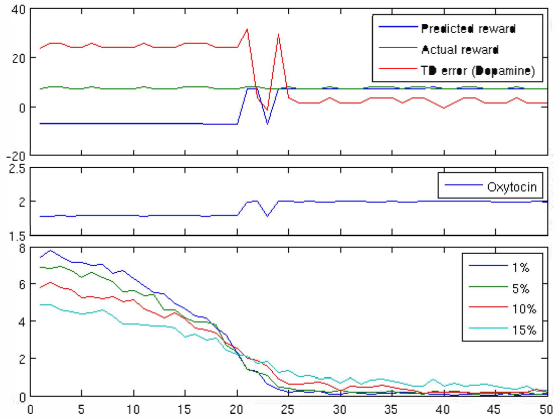

Fig. 3. Predicted/actual reward, Dopamine (top) and Oxytocin (middle) levels at each time-step (x-axis) of a typical run through the bonding counter-conditioning protocol. The bottom chart shows the reduction in average stress reactivity to the CS- permuted with 1, 5, 10 and 15% random noise, taken as an average over 100 activations at each time-step of the protocol.

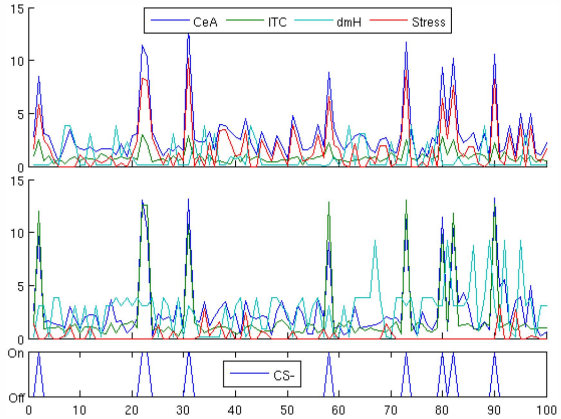

Fig. 4. Stress circuit reactivity for the counter-conditioned CS- before and after bonding protocols. Before application of the protocol, presentation of the the CS- (bottom) at random discrete timesteps is associated with large CeA and stress spikes (top, red). After counter-conditioning, ITC and dmH inhibition reduces associated stress reactivity (middle, red).

Similarly, we trained the vmPFC-ITC network to have high activation for CS+:US+ pairs, using as input the second hidden layer (OFC/vmPFC) activations in the DBN. Both the BLA-CeA and vmPFC-ITC feed-forward networks were configured with 200 hidden units, a learning rate of 75×10^{-6} and weights initialised randomly in the interval $[0, 10^{-4}]$. Finally, the weights in the OFC-dmH-PVNp feed-forward network were initialised to have high activation (inhibition) for increasingly large reward representations in the OFC (i.e. generally higher activation in the OFC-dmH-PVNp network corresponded to activation of units with decreasing index in the OFC's softmax reward units). Similarly, the OFC-dmH-PVNm network was initialised to have increasingly high activation (inhibition) for increasingly large punishment representations (corresponding to lower indices in the OFC's reward units). Both the OFC-dmH-PVNp and OFC-dmH-PVNm feed-forward networks had the same number of hidden units as the BLA-CeA and vmPFC-ITC networks described above, however weights were fixed after this initialisation and not subject to any learning. This gave a network that responded with relatively high stress to CS- (and slightly permuted CS-) inputs, but relatively low stress for CS+ (and slightly permuted CS+) inputs, representing an individual with highly aversive stress reactivity to a particular set of inputs (which are here taken to correspond to a class of attachment-related social stimuli).

We then attempted to simulate the counter-conditioning elements of the bonding phase of the self-attachment protocol, for the same CS- shown in Fig. 4 (top). On every iteration, we paired this CS- with one of either two US+ stimuli (with corresponding rewards of 7 or 8), and propagated these into the OFC to compute the predicted reward (Eq: 7). Based on this predicted reward, and the actual reward associated with the US+, we computed the temporal difference error (corresponding to a phasic DA signal) as in Eq. 8, and used this to update the weights in either the BLA-CeA or OFC-ITC feed forward networks (according to Eq. 9 and Eq. 10). The CS-:US+ was then also stored in the DBN along with the categorical representation for the received reward, using online learning (i.e. a single epoch).

The onset of the protocol results in large temporal-difference errors (phasic DA release), which spikes when the OFC's reward-prediction shifts before settling to a lower baseline level (Fig. 3, top). This spike in DA coincides with a spike in OXT (Fig. 3, middle), and a large drop in average stress levels associated with the CS- randomly permuted at 1,5,10 and 15% levels (Fig. 3, bottom). The effect can be seen in Fig. 4, which shows 100 time-steps in which either the counter-conditioned CS- stimulus, or a random stimulus, is presented to the network along with a random pattern for the US input, before (top) and after (bottom) application of the protocols. Before counter-conditioning, stages in which the CS- is presented correspond with large spikes in CeA input, low dmH inhibition on the PVNp, and a correspondingly high stress level. Inhibition from the dmH and ITC are increased following completion of counter-conditioning, with corresponding reduced stress reactivity.

V. Conclusion

We presented a computational neural model of OFC-mediated stress and bonding reactivity to social stimuli. In particular, we considered how this balance could be driven by both OFC-dmH inhibition on PVNp, and vmPFC-ITC inhibition on CeA. We showed how a counter-conditioning procedure, which we likened to a particular phase of the new self-attachment psychotherapy, could be used to drive a re-balance of activity between these circuits by increasing both inhibition in these pathways, and OXT release. Based on recent evidence, we considered here a role for OXT in modulating DA, although we note that DA might also have a modulating effect on OXT [49]. Recent studies have also suggested a role for OXT in direct modulation of stress reactivity to fearful stimuli, potentially mediated via receptors in the amygdala, although such effects are not explored in this initial model. For simplicity, we did not consider the role of the hippocampus in either fear context modulation [44] or short term memory effects of bonding [37]. Neither did we consider VA, which is believed to play a role in selective attention and memory modulation [37]. Future work should consider such effects.

REFERENCES

[1] A. Edalat. (2013) Self-attachment: A new and integrative psychotherapy (presented at institute of psychiatry, kings college london). [Online]. Available: http://www.doc.ic.ac.uk/~ae/papers/iop-talk.pdf

[2] G. Liotti, *Disorganized/disoriented attachment in the psychotherapy of the dissociative disorders*. Analytic Press, Inc, 1995.

[3] P. Fonagy *et al.*, "Attachment and borderline personality disorder: A theory and some evidence," *Psychiatric Clinics of North America*, vol. 23, no. 1, 2000.

[4] J. Holmes, "Disorganized attachment and borderline personality disorder: A clinical perspective," *Attachment & human development*, vol. 6, no. 2, 2004.

[5] A. N. Schore, *Affect Dysregulation and Disorders of the Self (Norton Series on Interpersonal Neurobiology)*. WW Norton & Company, 2003, vol. 1.

[6] L. Cozolino, *The neuroscience of human relationships: Attachment and the developing social brain*. WW Norton & Co, 2006.

[7] P. Vrtička et al., "Individual attachment style modulates human amygdala and striatum activation during social appraisal," *PLoS One*, vol. 3, no. 8, 2008.

[8] M. M. Riem *et al.*, "Attachment in the brain: adult attachment representations predict amygdala and behavioral responses to infant crying," *Attachment & human development*, vol. 14, no. 6, 2012.

[9] T. Kidd *et al.*, "Adult attachment style and cortisol responses across the day in older adults," *Psychophysiology*, vol. 50, no. 9, 2013.

[10] M. Numan and D. S. Stolzenberg, "Medial preoptic area interactions with dopamine neural systems in the control of the onset and maintenance of maternal behavior in rats," *Frontiers in neuroendocrinology*, vol. 30, no. 1, 2009.

[11] A. Bartels and S. Zeki, "The neural correlates of maternal and romantic love," *Neuroimage*, vol. 21, no. 3, 2004.

[12] L. Strathearn *et al.*, "Adult attachment predicts maternal brain and oxytocin response to infant cues," *Neuropsychopharmacology*, vol. 34, no. 13, 2009.

[13] M. L. Kringelbach *et al.*, "A specific and rapid neural signature for parental instinct," *PLoS One*, vol. 3, no. 2, 2008.

[14] R. Feldman *et al.*, "Evidence for a neuroendocrinological foundation of human affiliation plasma oxytocin levels across pregnancy and the postpartum period predict mother-infant bonding," *Psychological Science*, vol. 18, no. 11, 2007.

[15] I. Gordon *et al.*, "Oxytocin and the development of parenting in humans," *Biological psychiatry*, vol. 68, no. 4, 2010.

[16] C. Heim *et al.*, "Lower csf oxytocin concentrations in women with a history of childhood abuse," *Molecular psychiatry*, vol. 14, no. 10, 2008.

[17] C. H. Declerck *et al.*, "Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information," *Hormones and Behavior*, vol. 57, no. 3, 2010.

[18] J. Bartz *et al.*, "Oxytocin can hinder trust and cooperation in borderline personality disorder," *Social cognitive and affective neuroscience*, 2010.

[19] J. A. Bartz *et al.*, "Effects of oxytocin on recollections of maternal care and closeness," *Proceedings of the National Academy of Sciences*, vol. 107, no. 50, 2010.

[20] M. Olff *et al.*, "The role of oxytocin in social bonding, stress regulation and mental health: An update on the moderating effects of context and interindividual differences," *Psychoneuroendocrinology*, vol. 38, no. 9, 2013.

[21] A. Edalat, "Introduction to self-attachment and its neural basis," in *Neural Networks (IJCNN), 2015 International Joint Conference on*, 2015.

[22] H. Rockliff *et al.*, "A pilot exploration of heart rate variability and salivary cortisol responses to compassion-focused imagery," *Journal of Clinical Neuropsychiatry*, vol. 5, 2008.

[23] C. J. Litt, "Theories of transitional object attachment: An overview," *International Journal of Behavioral Development*, vol. 9, no. 3, 1986.

[24] L. A. Kirkpatrick, *Attachment, evolution, and the psychology of religion*. Guilford Press, 2005.

[25] U. Schjødt *et al.*, "Rewarding prayers," *Neuroscience letters*, vol. 443, no. 3, 2008.

[26] J. R. Villablanca, "Why do we have a caudate nucleus," *Acta Neurobiol Exp (Wars)*, vol. 70, no. 1, 2010.

[27] L. Perlovsky, "Musical emotions: Functions, origins, evolution," *Physics of life reviews*, vol. 7, no. 1, 2010.

[28] S. Koelsch *et al.*, "Investigating emotion with music: an fmri study," *Human brain mapping*, vol. 27, no. 3, 2006.

[29] V. N. Salimpoor *et al.*, "Anatomically distinct dopamine release during anticipation and experience of peak emotion to music," *Nature neuroscience*, vol. 14, no. 2, 2011.

[30] K. Jeffries *et al.*, "Words in melody: an h215o pet study of brain activation during singing and speaking," *Neuroreport*, vol. 14, no. 5, 2003.

[31] B. Kleber *et al.*, "Overt and imagined singing of an italian aria," *Neuroimage*, vol. 36, no. 3, 2007.

[32] Y.-C. Huang and N. A. Hessler, "Social modulation during songbird courtship potentiates midbrain dopaminergic neurons," *PloS one*, vol. 3, no. 10, 2008.

[33] D. Petters, "Designing agents to understand infants," Ph.D. dissertation, School of Computer Science, The University of Birmingham, 2006.

[34] A. Hiolle *et al.*, "Arousal regulation and affective adaptation to human responsiveness by a robot that explores and learns a novel environment," *Frontiers in neurorobotics*, vol. 8, 2014.

[35] D. Cittern and A. Edalat, "An arousal-based neural model of infant attachment," in *Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on*, Dec 2014.

[36] A. Edalat and F. Mancinelli, "Strong attractors of hopfield neural networks to model attachment types and behavioural patterns," in *Neural Networks (IJCNN), The 2013 International Joint Conference on*. IEEE, 2013.

[37] D. S. Levine, "Neural networks of human nature and nurture," *Avances en psicología latinoamericana*, vol. 26, no. 1, 2008.

[38] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," 1986.

[39] A. Edalat and Z. Lin, "A neural model of mentalization/mindfulness based psychotherapy," in *Neural Networks (IJCNN), 2014 International Joint Conference on*, July 2014.

[40] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, 2002.

[41] G. E. Hinton *et al.*, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, 2006.

[42] P. Vrtička *et al.*, "Lateralized interactive social content and valence processing within the human amygdala," *Frontiers in human neuroscience*, vol. 6, 2012.

[43] R. L. Jenison, "Directional influence between the human amygdala and orbitofrontal cortex at the time of decision-making," *PloS one*, vol. 9, no. 10, 2014.

[44] A. A. Moustafa *et al.*, "A model of amygdala–hippocampal–prefrontal interaction in fear conditioning and extinction in animals," *Brain and cognition*, vol. 81, no. 1, 2013.

[45] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, 2009.

[46] A. D. Abraham *et al.*, "Dopamine and extinction: A convergence of theory with fear and reward circuitry," *Neurobiology of learning and memory*, vol. 108, 2014.

[47] M. R. Milad and G. J. Quirk, "Fear extinction as a model for translational neuroscience: ten years of progress," *Annual review of psychology*, vol. 63, 2012.

[48] E. S. Bromberg-Martin *et al.*, "Dopamine in motivational control: rewarding, aversive, and alerting," *Neuron*, vol. 68, no. 5, 2010.

[49] T. M. Love, "Oxytocin, motivation and the role of dopamine," *Pharmacology Biochemistry and Behavior*, vol. 119, 2014.

[50] A. K. Raes *et al.*, "The effect of counterconditioning on evaluative responses and harm expectancy in a fear conditioning paradigm," *Behavior therapy*, vol. 43, no. 4, 2012.