An Arousal-Based Neural Model of Infant Attachment

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

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

***Abstract*—We develop an arousal-based neural model of infant attachment using a deep learning architecture. We show how our model can differentiate between attachment classifications during strange situation-like separation and reunion episodes, in terms of both signalling behaviour and patterns of autonomic arousal, according to the sensitivity of previous interaction.**

I. Introduction

Attachment theory, a dominant paradigm in psychology, aims to explain the dynamics of the relationship between an infant and a primary caregiver (often a parent), and the lasting effect that the nature of these early interactions have on the infant's emotional and social development. The theory states that each infant is genetically pre-disposed to seek out an emotionally supportive, dependent relationship with a primary caregiver, to whom they become "attached". Extensive empirical evidence has shown that caregivers who are sensitive and responsive to requests for comfort raise secure children, who are confident in the ability of the caregiver to provide a "secure base" from which they can explore and learn. On the other hand, caregivers who are insensitive to the infant's attachment needs foster various forms of insecure attachment. These organised attachment patterns, which become apparent by the end of the first year, are believed to reflect deeply embedded cognitive-emotional schemas in unconscious and implicit memories, rooted in Right Hemisphere (RH)-biased cortical-subcortical circuits [1].

We build on previous work to present what is, to the best of our knowledge, the first neural model of attachment in an Infant Strange Situation (ISS)-like separation-reunion scenario. We focus in particular on the roles of the amygdala and Orbital Medial Prefrontal Cortex (OMPFC) in encoding previous attachment interactions, and the impact that these encodings have on later approach-avoid attachment behaviours. An understanding of the early development and integration of these circuits, along with the ongoing role that the nature of this early development has on the neural functioning of the individual, is important not only for disciplines such as psychotherapy; but also for making progress in the pursuit of building intelligent machines able to understand the subtle interplay between human emotion and cognition.

A. Attachment Theory And The Strange Situation

The ISS is a controlled laboratory procedure conducted on infants between the ages of 9 and 18 months. It involves a number of stressors (an unfamiliar environment, stranger interaction, and caregiver separation-reunion), aiming to activate attachment systems in the infant. It was originally proposed by Ainsworth, who conducted seminal studies into the correlates between home maternal behaviour and laboratory infant behaviour on reunion following a separation [2]. She found that those infants who went on to be deemed to be secure based on the year's worth of observation explored freely in the presence of the mother during the ISS, and approached, and were almost immediately consoled by, the mother on reunion. Ainsworth found the strongest determinant of infant security to be a measure of caregiver sensitivity in the home (in terms of their ability to perceive the infant's signals accurately, and respond to these signals promptly and appropriately), and a large effect size for sensitivity (according to Ainsworth's definition) on ISS security was later confirmed by a meta-analysis of 4176 infants [3]. Secure infants have subsequently been correlated with caregivers possessing the "free autonomous" Adult Attachment Interview (AAI) profile [4] (an adult attachment classification, determined by a semi-structured interview focusing on childhood experiences aiming to elicit the individual's state of mind with respect to attachment).

On the other hand, caregivers who are consistently dismissive of attachment have been linked to avoidant infants. The infants of these caregivers, who do not approach on reunion, come to learn that their attachment needs will generally not be met by the caregiver, and instead focus on developing alternative, self-coping strategies [5, p.143]. Inconsistent caregivers have been shown to foster ambivalent attachment. These infants, uncertain as to whether or not they can depend on the caregiver for security, spend much of their time in a state of anxiety regarding the proximity of the caregiver: they approach on reunion, but are more difficult to console and are slower to return to exploration.

Caregivers who either convey an inability to provide for the infant's attachment needs ("frightened") or behave in a hostile manner ("frightening") during attachment episodes have been found to foster various forms of disorganised attachment [6], which manifests in bizarre or contradictory infant behaviours (including freezing and stifled screaming) on ISS reunion. Such behaviour has been hypothesised as being the result of a paradoxical situation in which the infant is genetically predisposed to seek out the caregiver as a source of comfort,

but has also by their past experience associated them with being a source of fear ("fear without solution") [7]. This final category is particularly important, since there is evidence for an increased risk of disorganised children developing psychopathologies, such as dissociation [8] and borderline personality disorder [9] later in life. In addition, an inter-generational effect has been observed, whereby these caregivers themselves had disorganised attachment relationships as children [10].

B. Neuroscience of Infant Attachment

The many early, repeated dyadic attachment interactions that the infant and caregiver engage in serve to immediately regulate the infant's emotion, stress and arousal. Moreover, through the mechanism of experience-dependent plasticity, these interactions facilitate the construction of schemas, based in neural circuitry, that form the basis for self-regulation and interactive coping strategies employed later in life in other stressful situations; effecting aspects such as social competence [11] and emotional health and resilience [12] [13]. Emotionally available caregivers who engage in reciprocal interactions foster healthy growth and integration in these key cortical-emotional circuits, whereas those who are unavailable or insensitive can hamper this development. Caregivers who expose the infant to trauma, or excessive fear or unregulated stress during this critical period of development, can "predispose the vulnerable individual to future psychopathology by permanently altering corticolimbic circuits that are implicated in the regulatory failures that underlie the pathophysiology of psychiatric disorders" [1, p.6].

Although both hemispheres of the cortex develop rapidly during infancy, it is believed that the RH is dominant in both its development and function during the first 3 years of life, after which there is a shift to the Left Hemisphere (LH) [14] [15]. It is also believed that, in general, the RH is dominant for emotional control and processing (e.g. [16]), and is more directly involved in the regulation of Autonomic Nervous System (ANS) arousal than the LH [1, p.61]. The RH has been found to be associated more with negative emotion and avoidance behaviour, whilst the LH more with positive emotion (plus anger [17]) and approach behaviour [18] [19]. These findings are consistent with differential lateralisation effects observed within the context of attachment classification. For example, in an ISS study of 159 infants aged between 13 and 15 months, electroencephalography measures (taken as a baseline, during play with mother and during play with an experimenter) showed that insecurely attached infants had relatively reduced left frontal activity compared to securely attached infants [20].

The most advanced area of the cortex, in term of both evolution and individual development, is the Prefrontal Cortex (PFC). The PFC is often subdivided into two main regions: the Dorsolateral Prefrontal Cortex (DLPFC), which has been implicated in "cold" executive function such as planning and verbal reasoning; and the OMPFC (encompassing the Orbitofrontal Cortex (OFC) and parts of the Anterior Cingulate Cortex (ACC)), which has been implicated in "hot", emotionally charged executive function such as the regulation of social behaviour [21]. Evidence from a number of imaging studies suggests that both the DLPFC and OMPFC are active in infancy, and that they serve to facilitate this same distinction between types of executive function (see [22] for a review). However, given the relatively long maturation period for the DLPFC (lasting in to the third decade of life [23]), and the dominance of RH-biased cortical areas during the first 18 months, we focus here on the OMPFC and "hot" executive function.

With its extensive, direct and reciprocal connections with the amygdala and hypothalamus and emotionally-biased RH [24], the OMPFC has long been known to play an important integrative role between the limbic system and higher areas of the neocortex. Much evidence points to the OMPFC playing a key role in both emotion regulation and appraisal; functions which are central to early, RH-dominated attachment interactions [25]. The OFC is believed to be involved in the encoding and signalling of expected value [26] and value-driven behaviour [27]. One suggestion, supported by findings that OFC firing is invariant to the available options [28], is that the OFC encodes a form of absolute economic value (rather than a relative preference). Recent theories instead propose that the OFC plays a particularly critical role in goal-directed behaviour, and the encoding of "information about the specific features of expected outcomes" from which a measure of value is derived according to motivational context [29]. This region is believed to enter a critical period of maturation at around the end of the first year of life [1, p.13], corresponding directly to the period at which attachment patterns begin to be reliably observed. For these reasons, it is in the implicit memories of the experience-dependent circuits between the OMPFC and the amygdala, along with circuits connecting to the hypothalamus, that Cozolino tentatively places the attachment schema [5, p.139].

The amygdala, which is highly mature at birth, has long been implicated in Pavlovian fear conditioning; with the lateral nucleus (which receives sensory information from the thalamus and cortex) responding to both conditioned and unconditioned stimuli, and the central nucleus (with its direct links to the hypothalamus and the ANS) controlling the expression of fear in terms of an autonomic fight-or-flight response [30] [31]. The central nucleus can also induce defensive, fear-invoked freezing behaviour via direct links with the dorsal Central (Periaqueductal) Gray (CG) [32]. Whilst the amygdala does also become active in response to positive stimuli, in general it responds more consistently to negative and threatening stimulation [33].

C. Related Work in Computational Attachment Modelling

Petters [34] presents a number of cognitive agent architectures designed to capture empirically observed infant attachment phenomenon of increasing levels of complexity. His work takes Bowlby's view; that the attachment system comprises two key evolutionary adaptations motivating learning and security, and that these manifest in fear and attachment

behavioural systems (to keep the infant safe), and exploratory and socialisation behavioural systems (to foster learning). In each of these architectures, the security goal is activated in part according to an anxiety measure, which is calculated based on a safe-range distance between the infant and caregiver. This safe-range proximity measure, intended to possess an equivalent function to the protection required by an infant in Bowlby's environment of evolutionary adaptedness, is updated by reinforcement following each attachment interaction, based on the caregiver's delay in response to infant signalling.

Hiolle et al [35] design an arousal-based model of an attachment secure-base, in order to control a robot during the exploration of a novel environment. If during the course of its exploration the robot becomes overwhelmed by too many new perceptions, this results in an intolerable level of arousal which in turn triggers calls for attention. The human caregiver is free to interact with the robot, and can choose to either sooth/calm it (to various degrees), or ignore it. The robot will resume exploration once its arousal drops back to within its tolerable threshold. The authors are concerned primarily with the influence of the human's interaction on the robot's learning: they do not consider detailed attachment classifications but instead focus on the fundamental characteristics of the secure-base paradigm. In other recent work, strong attractors in a Hopfield network have been proposed as models of cognitive and attachment schemata [36] [37].

II. The Model

We aim now to develop a neural model capable of eliciting attachment types in an ISS-like scenario. The model presented here is a development in part of both the Reactive Action Learning (RAL) architecture in [34] and the secure-base arousal model in [35], which we attempt to integrate and expand on according to neuroscientific theories of attachment. We design an environment in which both the infant and caregiver are free to move. The infant attempts to learn the objects in their environment, and the degree to which they are successful in doing so (their "surprise") will contribute to their arousal level. In addition, the infant calculates a safe-range distance based on previous responsiveness of the caregiver which, when exceeded by the caregiver, also contributes to an increase in arousal level. The final contributory factor for the arousal level is any fear that the expectation of a caregiver interaction may induce in the infant, again based on a model that the infant has built based on previous interactions.

When the infant's arousal level rises above their tolerance threshold, they probabilistically decide whether or not to seek an attachment interaction - a decision which is based on their model of the caregiver's effectiveness in helping them to lower their arousal level. Accordingly, the infant will choose to either approach the caregiver for assistance, or self-regulate and wait for their arousal to drop. The caregiver response is defined along two dimensions - responsiveness (how quickly they respond, as in the reactive architectures in [34]), and appropriateness (their effectiveness in lowering the infant's arousal). We assume here that the infant approaches the caregiver with the sole intention of seeking assistance in arousal regulation. Thus, whilst it is clearly possible to imagine cases in which a caregiver could raise the arousal of the infant in a positive way, we consider all attachment interactions resulting in arousal reduction to be "positive" in emotional valance, and all interactions resulting in an increase in arousal to be "negative" in valance (and inducing a proportional degree of fear in the infant). We investigate the particular forms of attachment that emerge following exposure to different types of caregiver, as measured by the safe-range distance and signalling behaviour of the infants, along with the relative patterns of arousal and the extent to which the objects in the environment are learnt. In addition to providing a neuroscientific basis, we extend previous works by looking at how previously unconsidered disorganised forms of attachment may emerge within the context of frightening, dysregulating caregiver behaviour.

A. Caregiver Types

We define caregiver types according to a joint density over two independent random variables: appropriateness $\phi \in \mathbb{R}$ and responsiveness $\eta \geq 0$, intended to cover the sensitivity-insensitivity scale originally defined by Ainsworth. Appropriateness refers to the nature of the interaction (i.e whether or not the caregiver responds in a way that comforts the infant and lowers their arousal level, or distresses them and raises their arousal), and is normally distributed. Responsiveness refers to the promptness with which they respond to the infant's signalling and requests for comfort, and is distributed according to an Inverse Gaussian ($\mathcal{IG}$) distribution. We assume that responsiveness also captures the notion of the caregiver's ability to recognise the infant's signalling requests. We define four caregiver profiles, labelled according to the typical AAI classifications (Fig. 1). Free-autonomous caregivers provide consistently appropriate, low-latency responses to the infant's signalling and requests for attachment. Dismissing caregivers are consistently unresponsive, with low (but mostly positive) appropriateness and high delay. The ambivalent profile defines a caregiver that generally provides responses of an appropriate nature, but in an inconsistent manner. Finally, the disorganised caregiver is inconsistent in both appropriateness and delay. They have a positive mean appropriateness of response (larger than the dismissing caregiver), and thus there will be some positively appropriate responses. However, because of the significantly larger standard deviation, a large proportion of responses will be of a highly inappropriate nature, resulting in elevated fear levels in the infant.

B. Infant Architecture

We attempt to model the infants unsupervised acquisition of beliefs about the capability of the caregiver in moderating their high-arousal states. A widely used artificial neural network in cognitive-emotional modelling is the Adaptive Resonance Theory classifier [38], in which a notion of vigilance determines whether top-down patterns match bottom-up inputs. Here we use a stochastic generative neural network called

	Appropriateness	Responsiveness
Free-autonomous	$\phi \sim \mathcal{N}(1, 0.005)$	$\eta \sim \mathcal{IG}(0.1, 10)$ (std = 0.01)
Dismissing	$\phi \sim \mathcal{N}(0.025, 0.005)$	$\eta \sim \mathcal{IG}(2, 80000)$ (std = 0.01)
Ambivalent	$\phi \sim \mathcal{N}(0.75, 0.75)$	$\eta \sim \mathcal{IG}(0.1, \frac{2}{1125})$ (std = 0.75)
Disorganised	$\phi \sim \mathcal{N}(0.25, 0.75)$	$\eta \sim \mathcal{IG}(0.1, \frac{2}{1125})$ (std = 0.75)

Fig. 1. Caregiver profiles.

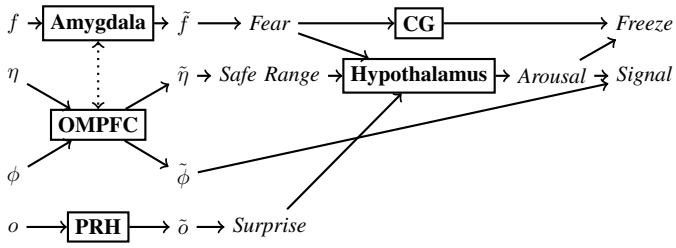

Fig. 2. Infant architecture.

the Restricted Boltzmann Machine (RBM), which has recently been used in cognitive-emotional modelling within the context of psychotherapy [39]. For visible $x \in \{0, 1\}$ and hidden units $h \in \{0, 1\}$ a RBM is parametrised by weights between each visible and hidden unit W, hidden unit biases b and visible biases c. Each configuration of x and h is assigned a scalar energy $E(x, h) = -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j$ so that the joint-distribution over x and h is given by $p(x, h) = e^{-E(x,h)} \left(\sum_{x,h} e^{-E(x,h)}\right)^{-1}$.

Training a RBM corresponds to adjusting W, b and c so that low energy (and thus high probability) is assigned to those configurations that the network has seen. In this way, the RBM provides an account of Hebbian association, and allows us to model a form of unsupervised learning on the part of the infant. A small change to the energy function and activations can be made so that the RBM can learn real valued input [40]. Furthermore, RBMs can be stacked vertically to form a powerful high-dimensional model called a Deep Belief Network (DBN) that is able to generate samples in a computationally efficient manner. It is now popular to regard the cortex as having some form of hierarchical generative model that is able to synthesise sensory inputs [41]: evidence for this comes from, for example, Charles Bonnet syndrome [42] (the experience of complex visual hallucinations in people with acquired blindness), which has been modelled with a Deep Boltzmann Machine [43] (a close relative of the DBN). We use the conventional 1-step contrastive divergence algorithm [44] with online parameter updates for training all networks. The infant's full neural architecture is shown in Fig. 2, and below we detail each individual component.

1) Care level: When the infant is signalling to the caregiver and in their proximity, they may receive some form of care

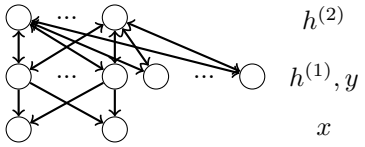

Fig. 3. Infant's OMPFC (DBN).

$\phi_t \in \mathbb{R}$ from them, where $\phi_t \sim \mathcal{N}$ according to the caregiver's appropriateness distribution (Fig. 1). We calculate the infant's care level C_t at timestep t as:

$$C_t = \begin{cases} \beta \cdot C_{t-1}, & \text{if } \phi_t = 0 \\ \phi_t, & \text{otherwise} \end{cases} \quad (1)$$

where $0 < \beta = 0.5 < 1$ is a care decay rate parameter.

2) Perirhinal Cortex: The infant explores their environment and attempts to learn the objects within it, generating a "surprise" value which is indicative of the extent to which they have mastered some particular object. We place this functionality in the Perirhinal Cortex (PRH): a part of the medial temporal lobe that is thought to be particularly important for the recognition and identification of environmental stimuli, and item memory and familiarity [45] (although other regions such as the Parahippocampal Cortex may also be involved [46]).

The PRH is modelled as a RBM which takes as input an object id o (a randomly generated binary string of length 50), and has 500 binary hidden units. Both hidden and visible units are stochastically activated according to the standard logistic activation function, with a learning rate of $l_{prh} = 0.001$. A surprise level S_t at time t is calculated based on the RBM's generative model, as a measure of item familiarity. For some particular object o that the infant is currently learning in the environment, the PRH produces a generative sample $\tilde{o}_t$, with the Markov chain starting from o at the visible layer. Then the surprise at time t is given by:

$$S_t = \sum_{i=1}^{length(o)} |o_i - \tilde{o}_{t_i}| \quad (2)$$

3) Orbital Medial Prefrontal Cortex: The OMPFC is modelled as a DBN which associates details of previous attachment encounters that the infant has had with the caregiver. The role of the OMPFC is thus to learn a model of the caregiver's effectiveness in regulating the infant's arousal levels, from which the value of attachment behaviours can be derived. The DBN has three layers of neurons (Fig. 3), with length$(x) = 2$, length$(h^{(1)}) = 50$ and length$(h^{(2)}) = 50$. The length of the binary caregiver id (which symbolises a nondescript sensory representation of the caregiver) is length$(y) = 10$.

The network is trained on input $x \in \mathbb{R}^2$ following each attachment encounter; the input at x_1 is the caregiver delay to the attachment request, and the input to x_2 is the quality (appropriateness) of care that was received. Thus the input to x is distributed according to a joint Gaussian, Inverse-Gaussian

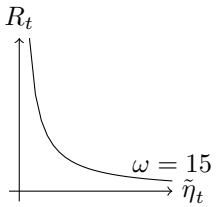

Fig. 4. Safe-range for generative responsiveness samples.

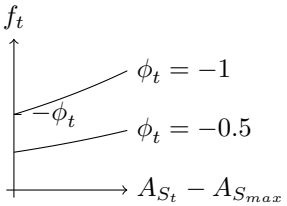

Fig. 5. Fear signals for appropriateness of response -0.5 and -1.

distribution. Input data is standardised to zero mean and unit variance before training, based on all previously seen data. Then we use Gaussian activations at x for the down-pass [47], and noisy rectified linear units for the up-pass in to $h^{(1)}$ [48]. At the top-level associative memory we use stochastic logistic activation for the down-passes in to $h^{(1)}$ and y, and noisy rectified linear units for the up-pass in to $h^{(2)}$. When a generative sample of delay and care is required, the state of the caregiver neurons y are held constant and equal to the caregiver id during Gibbs sampling at the top level associative memory. There are 50 hidden units at both $h^{(1)}$ and $h^{(2)}$, and we use a learning rate of $l_{ompfc} = 0.0001$.

The infant calculates a safe-range R_t at each timestep t. In [34] the safe-range was updated according to reinforcement only following each interaction. We instead recalculate R_t at every timestep based on a generative responsiveness sample $\tilde{\eta}_t$ from the OMPFC (with the caregiver's id y held fixed at the top level associative memory during Gibbs sampling), according to:

$$R_t = max\left(\left({env_w}^2 + {env_h}^2\right)^{\frac{1}{2}}, \frac{\omega}{max(\epsilon, \tilde{\eta}_t)}\right) \tag{3}$$

for $0 < \omega = 15$ a safe-range parameter, ϵ a small positive number, and $env_w = env_h = 50$ the dimensions of the environment.

4) Amygdala: Here we use a RBM as a model for the amygdala's role in fear conditioning, taking as input a concatenation of a fear signal and the caregiver id. Whilst it is well known that the amygdala and OMPFC can have inhibiting effects on each other, for simplicity we do not consider such effects in this initial model. Gaussian activations are used for the top-down pass in to the visible units, and rectified linear units for the bottom-up pass in to the hidden units, with a learning rate of $l_{am} = 0.001$.

After every attachment encounter in which a negative amount of care is received, we calculate the fear signal according to:

$$f_t = -\phi_t \delta^{\gamma(A_{S_t} - A_{S_{max}})} \tag{4}$$

with $1 < \delta = 1.5$ and $0 < \gamma = 0.75$ fear signal parameters, A_{S_t} the infant's current sustained arousal level (Eq. 7) and $A_{S_{max}} = 0.8$ the infant's upper arousal threshold. For example, for care received $\phi_t = -1$, we would have a fear signal as shown in Fig. 5. This fear signal f_t is then appended with the binary caregiver id y, and given as an input pattern to the amygdala (via the lateral nucleus). The caregiver id can be seen as a conditioned stimulus, and the fear signal an unconditioned stimulus (e.g. as a result of the caregiver causing physical pain). In line with theories suggesting that emotionally-driven implicit memories are encoded more strongly in hyper and hypo states of arousal [49, p.88], we have that f_t is a function of the amount by which the infant's current arousal level A_{S_t} exceeds their upper threshold $A_{S_{max}}$.

When the infant signals to the caregiver, they produce a generative sample $\tilde{f}_t$ from the amygdala, giving fear level F_t at timestep t:

$$F_t = \begin{cases} \varphi \cdot F_{t-1}, & \text{if signalling and } \tilde{f}_t \leq 0 \\ \tilde{f}_t, & \text{if signalling and } \tilde{f}_t > 0 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

which decays in a similar way to arousal, according to the fear decay rate parameter $0 < \varphi = 0.5 < 1$. Thus, the fear level can be seen as a conditioned fear response to the caregiver. This fear level is projected (via the central nucleus) to both the CG and hypothalamus.

5) Hypothalamus: The hypothalamus is responsible for generating an arousal level A_t at each timestep t. For euclidean distance between infant and caregiver at t given by D_t, the arousal level is given by:

$$A_t = F_t + S_t + max(0, D_t - R_t) \tag{6}$$

We calculate instantaneous and sustained measures of this arousal (as in [35]). Instantaneous arousal A_{I_t} is given by $A_{I_t} = \frac{A_{I_T} \cdot A_{I_{t-1}} + A_t}{1 + A_{I_T}}$, where $A_{I_T} = 100$ is a parameter signifying the time window on which the instantaneous arousal is calculated. Sustained arousal is calculated according to:

$$A_{S_t} = \begin{cases} \frac{A_{S_T} \cdot A_{S_{t-1}} + A_{I_t}}{1 + A_{S_T}}, & \text{if } C_t < 0.01 \text{ and } A_{I_t} > 0.01 \\ A_{S_{t-1}} - \alpha C_t, & \text{otherwise} \end{cases} \tag{7}$$

where $A_{S_T} = 50$ is the sustained arousal time window and $\alpha = 0.45$ is the sustained arousal decay rate when there is care. Increasing α increases the rate at which sustained arousal drops when care is being provided. For fixed α, higher levels of instantaneous arousal can result in sustained arousal beginning to drop more quickly in response to no care than in response to consistently low positive care. The sustained arousal level drives the action choice that the infant makes at each timestep:

- If $A_{S_{min}} = 0.2 > A_{S_t}$ (and they has been learning the target object for some minimum number of timesteps) then the infant is deemed to be under-stimulated and will choose and approach a new target object in the environment to learn.
- If $A_{S_{min}} \leq A_{S_t} \leq A_{S_{max}}$ then the infant will continue to learn the current object
- If $A_{S_t} > A_{S_{max}}$ and $F_t < \kappa$ (with $\kappa = 2$ a freezing parameter), then the infant will either signal to the caregiver with probability p, where:

$$p = \left(1 + m \cdot e^{n \cdot \tilde{\phi}_t}\right)^{-1} \quad (8)$$

or stop learning the current object (with probability $1-p$) until their sustained arousal level drops. $\tilde{\phi}_t$ is a generative appropriateness sample from the OMPFC at time t, and $0 < m = 3$ and $n = -10$ are signalling parameters.

- If $A_{S_t} > A_{S_{max}}$ and $F_t \geq \kappa$ then the infant freezes (a fear induced defence triggered via the CG)

The logistic signalling probability function (Eq. 8) with these particular parameters results in a small positive probability of signalling for low $\tilde{\phi}$. The shape parameter n can be seen as the extent to which the generative sample plays an influence in the decision as to whether or not to signal, and the scale parameter m can be seen as controlling the amount of care expected in order for the infant to want to signal (high m gives a stronger aversion to signalling when anticipating negative care).

III. Simulation Results

We ran 100 simulations for each of the four caregiving profiles, with each simulation consisting of two phases: a pre-training phase corresponding to a home environment, and a separation-reunion phase corresponding to episodes 6 (second separation) and 8 (second reunion) of the ISS.

The pre-training phase consisted of 10000 timestep iterations in which the infant explored and learnt their environment, and decided whether or not to seek an attachment interaction when their arousal level was above their upper threshold (Eq. 8). At each timestep in which the infant was not signalling the caregiver moved randomly with uniform probability 0.2. When the infant was signalling, the caregiver approached the infant with probability defined by their particular caregiving profile (Fig. 1). During this phase, infants of free-autonomous caregivers had (on average) the largest safe-range distance, with mean 63.5 and standard deviation 17.1. Infants of dismissing caregivers had the lowest safe-range, with mean 15.9 and standard deviation 18.4, whereas ambivalent and disorganised infants acquired similar safe-ranges, with means of 44.3 and 42.9, and standard deviations of 27.6 and 28.4 respectively. Fig. 7 plots, for each infant, the mean length of timewindows in which $A_{S_t} > A_{S_{max}}$, and the mean percentage of these high-arousal timewindows that the infant spent signalling. Infants of free-autonomous caregivers are densely clustered around points with high signalling probability and low mean high-arousal timewindow length, whereas infants of dismissive caregivers are densely clustered according to signalling, but sparsely clustered by length of timewindow: in other words, infants of dismissive caregivers varied more than other infants in the mean amount of time required to recover from high-arousal states. Interestingly, ambivalent and disorganised infants cluster into both relatively high and low signalling groups. Infants appearing in the low signalling cluster (predominantly disorganised) typically had very inappropriate (negative) early attachment experiences, resulting in very low subsequent signalling behaviour. In terms

	Mean surprise		
Caregiver	Object 1	Object 2	Object 3
Free-autonomous	13.50	14.24	16.31
Dismissing	21.10	18.92	17.35
Ambivalent	18.67	18.70	19.06
Disorganised	19.23	18.81	19.41

Fig. 6. Mean object surprise for infants of each caregiver profile at the end of the pre-training phase (Eq. 2).

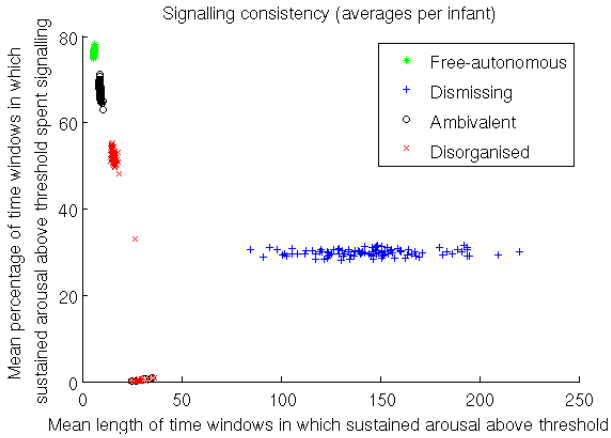

Fig. 7. Mean length of timewindows in which sustained arousal was above the upper threshold, and mean percentage of these high-arousal timewindows that the infant spent signalling, during the pre-training phase. Each point is an individual infant.

of exploration, infants of free-autonomous caregivers were significantly more successful in learning the objects in their environment, as measured by the surprise value for each object based on generative PRH samples at the end of the phase (Fig. 6). Since infants of free-autonomous caregivers had the lowest proportion of high-arousal timesteps over pre-training, this result would be predicted by the inverted-U theory of arousal and learning.

In the separation-reunion phase, we first separated the infant and caregiver for 50 timesteps by moving the infant to position (1,1) in the environment, and the caregiver to position $(env_w + 1,\ env_h + 1)$, which caused the infant's sustained arousal to rise since the caregiver was now outside of their safe-range. The caregiver was then brought back into the center of the environment for a further 50 timesteps. In accordance with the controlled nature of the ISS protocol, all caregivers responded in the same way to attachment bids from their respective infants ($\forall t$: $\eta_t = 0$ and $\phi_t = 1$).

Sustained arousal levels rose least rapidly on average for infants of free-autonomous caregivers (who had the largest mean safe-range) during separation, and they also recovered most quickly to within their comfortable threshold following reunion (Fig. 9). This faster mean recovery during reunion can be explained by the more frequent signalling exhibited on average by these infants (Fig. 8). The slower mean recovery for infants of ambivalent caregivers is in accordance with an extensive body of observational data suggesting that

Caregiver	Pretrain	Reunion
Free-autonomous	76.38 (15.11)	97.64 (1.45)
Dismissing	29.91 (5.34)	27.72 (6.27)
Ambivalent	65.71 (17.73)	75.61 (15.38)
Disorganised	45.27 (21.77)	48.49 (24.88)

Fig. 8. Mean (standard deviation) of percentage of timewindows (in which $A_{S_t} > A_{S_{max}}$) that the infant spent signalling.

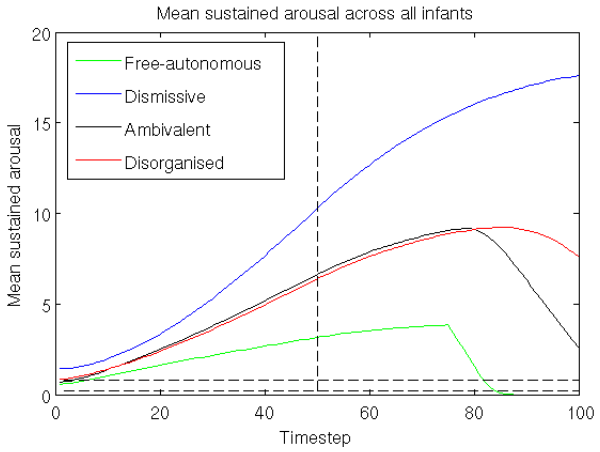

Fig. 9. Mean sustained arousal for infants pretrained on each of the four caregiver types, during the separation-reunion phase. The dashed vertical line distinguishes the separation and reunion episodes. The dashed horizontal lines mark the infant's sustained arousal thresholds.

ambivalent infants are slower to be soothed and to return to exploration on reunion. Sustained arousal for infants of dismissing caregivers rose more rapidly than for infants of free-autonomous caregivers during separation and was slower to recover during reunion. This more rapid rise is explained by the larger safe-range of these infants, and the slower recovery by the fact that they signalled less frequently on average during the reunion episode.

These patterns are also supported by a number of physiological studies of the ISS. Infant heart rate is known to increase on separation regardless of attachment classification, but when [50] controlled for behaviour they found significant relative increase for both avoidant and disorganised infants over secure infants. Increased cortisol levels following the final reunion episode have been found in avoidant [50], ambivalent [51] and disorganised [50] [52] infants relative to secure. In a sample of secure and avoidant infants, respiratory measures found significantly higher vagal withdrawal (and thus Sympathetic Nervous System (SNS) activity) for avoidant infants during separation, and greater (although non-significant) withdrawal during reunion [53]. In addition, α-amylase measures found significantly greater SNS activity (including baseline) for avoidant infants across the entire procedure. Although these findings of secure relative to insecure autonomic activation are supported by our model, the precise nature of the relationship between the arousal of avoidant, ambivalent and disorganised infants is predictive and open to further empirical investigation.

IV. Conclusion

In line with theories in developmental neuroscience that highlight the early dominance of the emotionally-biased RH in ISS-aged infants, we have focused in particular on the role of the OMPFC and amygdala in encoding previous attachment interactions in order to direct approach-avoidance reunion behaviour. However, the encoding presented here is likely to be a simplification. As discussed previously, it is now believed that the OFC plays a particularly important role in goal-directed decision making, which also involves the DLPFC. Indeed, it appears as though a clear distinction can be made between the contributions of these circuits, with the DLPFC encoding task-specific mappings between stimuli and responses, and the OFC encoding representations of states relevant to reward.

In [54] a neural model is presented for cognitive-emotional decision making, with heuristic rules encoded predominantly between the amygdala and OFC, and deliberative rules between the DLPFC and OFC. The particular decision rule that is employed is influenced by current need: when lower needs (such as security) are fulfilled then higher needs (such as cognition) are prioritised, and the network will bias towards deliberative rules. In contrast, when low-level needs are not fulfilled, heuristic rules will tend to be employed. In [55] a Bayesian network formulation for goal-directed decision making is presented, including action (corresponding to motor cortex representations), state (corresponding to action-outcome representations in the parietal and medial temporal cortices) and value (OFC) variables. Optimal deterministic policies (representing DLPFC encoding) can then be inferred given a current-state observation. Extensions to this model could attempt to integrate such frameworks, however care should be taken since there is currently no evidence to suggest that infants younger than 18 months are capable of goal-directed decision making [56]. Other possible avenues for future work include a closer correspondence with the ISS protocol, and the consideration of other factors of caregiver behaviour (such as synchrony and mutuality) that have also been found to be effective determinants of ISS behaviour [3].

References

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

[2] M. D. S. Ainsworth, M. C. Blehar, E. Waters, and S. Wall, *Patterns of attachment: A psychological study of the strange situation*. Psychology Press, 1978.

[3] M. S. Wolff and M. H. Ijzendoorn, "Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment," *Child development*, vol. 68, no. 4, 1997.

[4] M. Van IJzendoorn, "Adult attachment representations, parental responsiveness, and infant attachment: a meta-analysis on the predictive validity of the adult attachment interview." *Psychological bulletin*, vol. 117, no. 3, 1995.

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

[6] K. Lyons-Ruth, E. Bronfman, and G. Atwood, "A relational diathesis model of hostile-helpless states of mind," *Attachment disorganization*, 1999.

[7] M. Main and E. Hesse, "Disorganized/disoriented infant behavior in the strange situation, lapses in the monitoring of reasoning and discourse during the parents adult attachment interview, and dissociative states," *Attachment and psychoanalysis*, 1992.

[8] G. Liotti, "Disorganized/disoriented attachment in the psychotherapy of the dissociative disorders." 1995.

[9] P. Fonagy, M. Target, and G. Gergely, "Attachment and borderline personality disorder: A theory and some evidence," *Psychiatric Clinics of North America*, vol. 23, no. 1, 2000.

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

[11] N. S. Weinfield, L. A. Sroufe, B. Egeland, and E. A. Carlson, "The nature of individual differences in infant–caregiver attachment." 1999.

[12] A. N. Schore, "Advances in neuropsychoanalysis, attachment theory, and trauma research: Implications for self psychology," *Psychoanalytic Inquiry*, vol. 22, no. 3, 2002.

[13] P. Fonagy, G. Gergely, and E. L. Jurist, *Affect regulation, mentalization and the development of the self*. Karnac Books, 2003.

[14] S. Giudice, R. Thatcher, and R. Walker, "Human cerebral hemispheres develop at different rates and ages," *Science*, vol. 236, 1987.

[15] C. Chiron, I. Jambaque, R. Nabbout, R. Lounes, A. Syrota, and O. Dulac, "The right brain hemisphere is dominant in human infants." *Brain*, vol. 120, no. 6, 1997.

[16] K. K. Voeller, J. A. Hanson, and R. N. Wendt, "Facial affect recognition in children a comparison of the performance of children with right and left hemisphere lesions," *Neurology*, vol. 38, no. 11, 1988.

[17] E. Harmon-Jones, C. K. Peterson, and C. Harmon-Jones, "Anger, motivation, and asymmetrical frontal cortical activations," in *International handbook of anger*. Springer, 2010.

[18] C. M. Braun, R. Daigneault, S. Gaudelet, and A. Guimond, "Diagnostic and statistical manual of mental disorders, symptoms of mania: which one (s) result (s) more often from right than left hemisphere lesions?" *Comprehensive psychiatry*, vol. 49, no. 5, 2008.

[19] A. A. Hane, N. A. Fox, H. A. Henderson, and P. J. Marshall, "Behavioral reactivity and approach-withdrawal bias in infancy." *Developmental Psychology*, vol. 44, no. 5, 2008.

[20] G. Dawson, S. B. Ashman, D. Hessl, S. Spieker, K. Frey, H. Panagiotides, and L. Embry, "Autonomic and brain electrical activity in securely-and insecurely-attached infants of depressed mothers," *Infant Behavior and Development*, vol. 24, no. 2, 2001.

[21] R. C. Chan, D. Shum, T. Toulopoulou, and E. Y. Chen, "Assessment of executive functions: Review of instruments and identification of critical issues," *Archives of Clinical Neuropsychology*, vol. 23, no. 2, 2008.

[22] T. Grossmann, "Mapping prefrontal cortex functions in human infancy," *Infancy*, vol. 18, no. 3, 2013.

[23] B. Luna, K. R. Thulborn, D. P. Munoz, E. P. Merriam, K. E. Garver, N. J. Minshew, M. S. Keshavan, C. R. Genovese, W. F. Eddy, and J. A. Sweeney, "Maturation of widely distributed brain function subserves cognitive development," *Neuroimage*, vol. 13, no. 5, 2001.

[24] H. Barbas, "Flow of information for emotions through temporal and orbitofrontal pathways," *Journal of Anatomy*, vol. 211, no. 2, 2007.

[25] A. Etkin, T. Egner, and R. Kalisch, "Emotional processing in anterior cingulate and medial prefrontal cortex," *Trends in cognitive sciences*, vol. 15, no. 2, 2011.

[26] S. E. Morrison and C. D. Salzman, "The convergence of information about rewarding and aversive stimuli in single neurons." *The Journal of Neuroscience*, 2009.

[27] N. H. Kalin, S. E. Shelton, and R. J. Davidson, "Role of the primate orbitofrontal cortex in mediating anxious temperament," *Biological Psychiatry*, vol. 62, no. 10, 2007.

[28] C. Padoa-Schioppa and J. A. Assad, "The representation of economic value in the orbitofrontal cortex is invariant for changes of menu," *Nature neuroscience*, vol. 11, no. 1, 2008.

[29] G. Schoenbaum, Y. Takahashi, T.-L. Liu, and M. A. McDannald, "Does the orbitofrontal cortex signal value?" *Annals of the New York Academy of Sciences*, vol. 1239, no. 1, 2011.

[30] M. Davis, "The role of the amygdala in fear and anxiety," *Annual review of neuroscience*, vol. 15, no. 1, 1992.

[31] J. LeDoux, "The emotional brain, fear, and the amygdala," *Cellular and molecular neurobiology*, vol. 23, no. 4-5, 2003.

[32] J. E. LeDoux, J. Iwata, P. Cicchetti, and D. Reis, "Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear," *The Journal of Neuroscience*, vol. 8, no. 7, 1988.

[33] M. Davis, P. J. Whalen *et al.*, "The amygdala: vigilance and emotion," *Molecular psychiatry*, vol. 6, no. 1, 2001.

[34] D. Petters, "Implementing a theory of attachment: A simulation of the strange situation with autonomous agents," in *Proceedings of the Seventh International Conference on Cognitive Modelling*, vol. 7, 2006.

[35] A. Hiolle, L. Cañamero, M. Davila-Ross, and K. A. Bard, "Eliciting caregiving behavior in dyadic human-robot attachment-like interactions," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 2, no. 1, 2012.

[36] A. Edalat, "Capacity of strong attractor patterns to model behavioural and cognitive prototypes," in *Advances in Neural Information Processing Systems*, 2013, pp. 2661–2669.

[37] 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, pp. 1–10.

[38] G. A. Carpenter and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer vision, graphics, and image processing*, vol. 37, no. 1, 1987.

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

[40] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, 2010.

[41] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[42] D. P. Reichert, P. Seriès, and A. J. Storkey, "Charles bonnet syndrome: evidence for a generative model in the cortex?" *PLoS computational biology*, vol. 9, no. 7, p. e1003134, 2013.

[43] P. Series, D. P. Reichert, and A. J. Storkey, "Hallucinations in charles bonnet syndrome induced by homeostasis: a deep boltzmann machine model," in *Advances in Neural Information Processing Systems*, 2010, pp. 2020–2028.

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

[45] M. Brown, G. Barker, J. Aggleton, and E. Warburton, "What pharmacological interventions indicate concerning the role of the perirhinal cortex in recognition memory," *Neuropsychologia*, vol. 50, no. 13, 2012.

[46] C. B. Martin, D. A. McLean, E. B. O'Neil, and S. Köhler, "Distinct familiarity-based response patterns for faces and buildings in perirhinal and parahippocampal cortex," *The Journal of Neuroscience*, vol. 33, no. 26, 2013.

[47] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, 2006.

[48] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.

[49] A. N. Schore, *The Science of the Art of Psychotherapy (Norton Series on Interpersonal Neurobiology)*. WW Norton & Company, 2012.

[50] G. Spangler and K. E. Grossmann, "Biobehavioral organization in securely and insecurely attached infants," *Child development*, vol. 64, no. 5, 1993.

[51] G. Spangler, "Emotional and adrenocortical responses of infants to the strange situation: The differential function of emotional expression," *International Journal of Behavioral Development*, vol. 22, no. 4, 1998.

[52] L. Hertsgaard, M. Gunnar, M. F. Erickson, and M. Nachmias, "Adrenocortical responses to the strange situation in infants with disorganized/disoriented attachment relationships," *Child development*, vol. 66, no. 4, 1995.

[53] A. L. Hill-Soderlund, W. R. Mills-Koonce, C. Propper, S. D. Calkins, D. A. Granger, G. A. Moore, J.-L. Gariepy, and M. J. Cox, "Parasympathetic and sympathetic responses to the strange situation in infants and mothers from avoidant and securely attached dyads," *Developmental Psychobiology*, vol. 50, no. 4, 2008.

[54] D. S. Levine, "Brain pathways for cognitive-emotional decision making in the human animal," *Neural Networks*, vol. 22, no. 3, 2009.

[55] M. Botvinick and J. An, "Goal-directed decision making in prefrontal cortex: a computational framework," in *Advances in Neural Information Processing Systems*, 2009, pp. 169–176.

[56] U. M. Klossek and A. Dickinson, "Rational action selection in 1 1/2- to 3-year-olds following an extended training experience," *Journal of experimental child psychology*, vol. 111, no. 2, pp. 197–211, 2012.