Department of Computing
Imperial College London

Computational Models of Attachment and Self-Attachment

David Cittern

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London
December 2016

Statement of Originality

The work in this thesis was produced by David Cittern, unless otherwise specified.

Copyright Declaration

Abstract

We explore, using a variety of models grounded in computational neuroscience, the dynamics of attachment formation and change. In the first part of the thesis we consider the formation of the traditional organised forms of attachment (as defined by Mary Ainsworth) within the context of the free energy principle, showing how each type of attachment might arise in infant agents who minimise free energy over interoceptive states while interacting with caregivers with varying responsiveness. We show how exteroceptive cues (in the form of disrupted affective communication from the caregiver) can result in disorganised forms of attachment (as first uncovered by Mary Main) in infants of caregivers who consistently increase stress on approach, but can have an organising (towards ambivalence) effect in infants of inconsistent caregivers. The second part of the thesis concerns Self-Attachment: a new self-administrable attachment-based psychotherapy recently introduced by Abbas Edalat, which aims to induce neural plasticity in order to retrain an individual's suboptimal attachment schema. We begin with a model of the hypothesised neurobiological underpinnings of the Self-Attachment bonding protocols, which are concerned with the formation of an abstract, self-directed bond. Finally, using neuroscientific findings related to empathy and the self-other distinction within the context of pain, we propose a simple spiking neural model for how empathic states might serve to motivate application of the aforementioned bonding protocols.

Acknowledgements

The majority of my thanks must go to my supervisor, Prof. Abbas Edalat, who has been a huge source of guidance and inspiration over the past few years. Much of the content of this thesis originates in his ideas and insights, and it has been a pleasure to work with him. The work in Chapter 3 was completed in collaboration with Dr. Tobias Nolte. I would like to thank him for his contribution, and for the enjoyable meetings that we have shared over the past year and a half. I would also like to express my gratitude to Prof. Dan Levine and Prof. Michael Numan for helpful comments and discussions relating to the work presented in Chapters 4 and 5. Finally, I would like to thank my family for their continued support and encouragement.

Contents

List of Tables

List of Figures

1. Introduction

Attachment theory, a dominant paradigm in psychology, aims to explain the dynamics of the relationship between an infant and a primary caregiver (most often a parent), and the lasting effect that the nature of these early interactions has 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. At the neural level, 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.

In this thesis we will study (in the form of computational models) how various types of attachment might form in infants, along with the proposed neurobiological underpinnings of a new self-administrable attachment-based psychotherapy which has recently been proposed as a method for retraining an individual's attachment schema. The dynamics of attachment are explored using a number of paradigms from computational neuroscience and machine learning: in particular the free energy principle, deep learning and spiking neural networks.

1.1 Chapters

The remainder of this thesis is structured as follows. In Chapter 2 we overview background material on attachment theory (from both psychological and developmental neuroscientific perspectives), and cover existing work in computational and mathematical attachment modelling. In Chapter 3 we present a computational account of organised and disorganised attachment development within the context of the free energy principle. Next, in Chapter 4, we present a model of the hypothesised neurobiological underpinnings of the Self-Attachment bonding protocols, which are believed to provide a method for retraining an individual's

attachment schema. In Chapter 5 we explore the computational neural underpinnings of the Self-Attachment empathy protocols, and how they might motivate application of the bonding protocols. Finally, we offer some concluding remarks in Chapter 6.

1.1.1 Relationship to Published Work

The work in Chapter 4 was presented at IJCNN 2015 and published in the conference proceedings (Cittern and Edalat, 2015c). A paper based on the work presented in Chapter 5 is currently under review by the Journal of Computational Psychiatry. The work in Chapter 3, which was undertaken in collaboration with Dr Tobias Nolte, is as yet unpublished. Other work completed during the course of the PhD includes a neurocognitive model of infant attachment formation (presented at IEEE SSCI 2014 and published in the conference proceedings (Cittern and Edalat, 2014)) and a multi-agent framework for manipulating reward in order to generate new Nash equilibria (presented as a poster at AAMAS 2015 and published as a short paper in the conference proceedings (Cittern and Edalat, 2015a)).

2. Background

During the early stages of life, an infant is highly dependent on others for their survival. Attachment theory posits that each infant is genetically pre-disposed to seek out an emotionally supportive, dependent relationship with a primary caregiver, to whom they turn for comfort and safety during times of stress or perceived threat. The central tenet of the theory is that the nature of early dyadic attachment interactions (and in particular the caregiver's response to the emotionally-charged bids for proximity by the infant) lead to particular, distinguishable attachment styles. These attachment styles, which reflect an internal working model (attachment schema) capturing the extent to which the infant believes they can rely on the caregiver for assistance in emotion and stress regulation, manifest in different patterns of behaviour during times of dysregulation. Early interactions are believed to have a disproportionately large effect on the formation of this schema in the infant, with attachment experiences generalised to other socially challenging and emotionally charged situations that the individual may find themselves in later in life.

Extensive empirical evidence has shown that caregivers who are 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 about the world. On the other hand, a caregiver who is consistently dismissive of attachment need raises an avoidant child, who comes to learn that these needs will generally not be met by the caregiver, and instead focuses on developing alternative, self-coping strategies. Inconsistent caregivers have been shown to raise ambivalent children, who come to spend much of their time in a state of anxiety; uncertain as to whether or not they can depend on the caregiver for security. Caregivers who are either "helpless" (i.e. convey an inability to provide for the infant's attachment needs), or "hostile" (generate additional fear in the infant) have been shown to foster various forms of disorganised attachment. From a clinical perspective, disorganised attachment is particularly significant since (as we will see) it has been linked to a predisposition for the development of various psychological disturbances.

2.1 Roots of Attachment Theory

In the 1940s a psychiatrist and psychoanalyst named John Bowlby began to study early mother-child relationships, and in particular the effects on the child of maternal deprivation and separation. Bowlby strongly opposed the dominant psychoanalytic views of the time, which argued that Freudian drives (such as fulfilment of hunger or thirst) were the key to understanding an infant's tie to their mother (he remarked that "were it true, an infant of a year or two should take readily to whomever feeds him, and this clearly is not the case" (Sandler and Bowlby, 1989, p.104)). Instead, he theorised that it was the nature of the infant's cumulative early experience with its caregiver (usually the mother), and their family life, that was key to understanding this tie along with subsequent emotional development. His initial study into 44 juvenile thieves showed an empirical link between affection-less behaviour and lack of a consistent caregiver in childhood (Bowlby, 1944). Following World War 2 Bowlby collected data on the effects of hospitalisation on children, and his collaborative work with Robertson played a major role in initiating changes in hospital practices so that more parental involvement was encouraged (Bowlby and Robertson, 1953). He was also commissioned to write the World Health Organisation's report on the mental health of homeless children (Bowlby et al., 1951). He concluded that in order for healthy mental development an infant "should experience a warm, intimate, and continuous relationship with his mother (or permanent mother substitute)", which led to widespread changes in the treatment of infants under institutional care.

Bowlby developed attachment theory over the following decades (Bowlby, 1958, 1960b,a, 1969, 1973, 1980, 1982), framing it as a lasting psychological connectedness that had a biological, evolutionary basis. His theory took the view that each child was genetically predisposed to form an attachment with a primary caregiver, to whom they would seek proximity in times of stress, fear or perceived danger. According to Bowlby, the attachment behavioural system (a result of natural selection) interacts with other behavioural systems (fear, exploratory, caregiving) whose collective primary goal is experience of felt security in, and survival of, the infant. He hypothesised that the attachment behavioural system is activated in response to both internal (e.g. pain, hunger) and external (e.g. being threatened or endangered) events, triggering proximity seeking behaviours that in turn activate the caregiving behavioural system (in the caregiver), such that the infant and caregiver are both predisposed to seek and maintain proximity with each other. Central to this attachment behavioural system was the concept of an internal working model: a representation of the attachment relationship and its participants that the infant builds based on its particular experiences. This internal working model is used to generate expectations and make predictions about future attachment-related experiences and caregiving behaviour, based upon which the infant makes decisions about how they should act in order to achieve their goal

of felt security.

The balance between attachment and environmental exploration was first examined by Mary Ainsworth (Ainsworth, 1963), who proposed the idea that infants use their caregivers as a secure base from which to explore the world while monitoring and periodically returning to them as a safe haven. In particular, it was hypothesised that infants who were more assured of their caregiver's availability would be more confident in their exploration. Bowlby later argued that "All of us, from the cradle to the grave, are happiest when life is organised as a series of excursions, long or short, from the secure base provided by our attachment figures" (Bowlby, 1988), where secure forms of attachment entail an internal working model encapsulating confidence in the caregiver's responsiveness and availability which results in effective use of this secure base. The 1960s onwards saw the emergence of a large amount of empirical data supporting Bowlby's theories, and increasing research interest in the differences between, and implications of, the types of attachment that different infant-parent dyads foster.

2.2 An Empirical Classification Scheme for Infant Attachment

The first body of research to directly examine Bowlby's hypotheses and provide empirical evidence for links between maternal caregiving behaviour and infant attachment style came from Mary Ainsworth. In 1967, Ainsworth published a longitudinal study into the patterns of communication and behaviour between parents and their infants in Uganda, by observing the dynamics of 16 families once every two weeks (Ainsworth, 1967). Although a relatively small sample size, the data suggested that the infants gradually came to develop a preference for a particular primary caregiver (in this case the mother), manifesting in a flight to her when they were uncertain or distressed; the use of her as a secure base for exploration; and an active approach on reunion following a separation. In addition, a minority of infants were observed either not being soothed by their mothers on reunion, or not exploring their environment to the same degree. Following interviews, Ainsworth concluded that these differences were the result of disparities in the sensitivity of caregiving (measured in terms of the perception of the subtleties and details of their infant's behaviour). In addition, she found evidence for a positive correlation between infant security and mother's pleasure in breast feeding (Bretherton, 1985). The Uganda study was followed up with similar research in Baltimore, in which 26 mothers (recruited before the birth of their infants) were observed in their home environments over 18 visits (every 3 weeks from 3 to 54 weeks postpartum) lasting 4 hours each. The data collected supported the Uganda observations, and pointed towards a cross-cultural and innate attachment system (Ainsworth et al., 1978).

One difference between the Uganda and Baltimore studies was that Baltimore infants

Episode	Participants	Duration	Description
1	Caregiver, infant, experimenter	$<$ 1 minute	Experimenter brings mother and infant to the room, gives instructions and introduces the room (which contains toys and objects of interest to the infant of this age)
2	Caregiver, infant	3 minutes	Caregiver sits in chair and reads while infant explores. Caregiver responds if approached, but does not initiate
3	Caregiver, infant, stranger	3 minutes	Stranger enters, silent (1 minute), converses with caregiver (1 minute), approaches infant (1 minute)
4	Infant, stranger	3 minutes	First separation: caregiver departs, stranger comforts infant if necessary (or otherwise sits in chair)
5	Caregiver, infant	3 minutes	First reunion: caregiver returns, greets and/or comforts infant, returns to chair and reads
6	Infant	3 minutes	Second separation: caregiver departs, infant is alone
7	Infant, stranger	3 minutes	Second separation: stranger enters, comforts infant if necessary, otherwise sits in chair
8	Caregiver, infant	3 minutes	Second reunion: caregiver returns, greets and/or comforts infant (whilst stranger leaves)

Figure 2.1: Strange situation episodes. Adapted from Ainsworth et al. (1978)

did not display the secure-base behaviour observed in Uganda, with exploration seeming to continue when the mother departed. Ainsworth predicted that the Baltimore infants would display the same pattern of behaviour as the Ugandan infants if they were placed in an unfamiliar environment. As such she devised a controlled laboratory procedure called the Infant Strange Situation (ISS), which has since gone on to become the standard measure of attachment in infants between the ages of 9 and 18 months of age. The ISS consists of eight phases (Fig. 2.1), with 3 stressful components (interaction with a stranger, caregiver separation and an unfamiliar environment) that are designed to activate the infant's attachment system.

Application of the ISS provided results (which have since been replicated many times)

suggesting three distinct attachment categories. These attachment types were termed secure, avoidant and ambivalent, and each was found to be strongly correlated with a different pattern of mother-infant interaction in the home environment (according to observations in the fourth quarter of the first year) [1].

Ainsworth found that those infants who went on to be classified as secure based on the year's worth of observation displayed attachment behaviour in the ISS. These infants exhibited a strong desire to explore their environment in the presence of the caregiver, were distressed on separation, but were almost immediately consoled on reunion with the mother. As a group, secure infants were thus characterised by effective use of the mother as a secure base for exploration. In contrast, two distinct types of insecure attachment emerged. In the first (avoidant) group, infants were observed by Ainsworth to continue to explore despite the mother leaving the room, and avoided the mother completely on reunion. Although they displayed little outward distress on separation, later studies found increased heart rate (Donovan and Leavitt, 1985; Spangler and Grossmann, 1993; Zelenko et al., 2005), decreased heart rate variability (Hill-Soderlund et al., 2008; Smith et al., 2016) and increased cortisol (Spangler and Grossmann, 1993) (although see (Hertsgaard et al., 1995; Spangler, 1998)) for these infants in response to the procedure, suggesting that they do experience separation anxiety but attempt to regulate or repress internal stress and emotion themselves. The second group of insecure infants (ambivalent) explored to a far smaller degree than the secure and avoidant infants, instead remaining preoccupied with the mother's proximity throughout the procedure. Ainsworth identified two types of these ambivalent infants: those who were "angry", who oscillated between actions of rejection and anger towards the mother; and those categorised as "passive", who only expressed faint bids for comfort. Ambivalent infants were observed to be harder to console once they had achieved proximity with the caregiver, and they typically displayed behaviours indicating anger, resistance and/or helplessness. A later meta-analysis of 2104 non-clinical, North American infants that had undertaken the ISS found that 62 percent were classified as secure, 15 percent avoidant, and 9 percent ambivalent van Ijzendoorn et al. (1999a).

In order to assess maternal fourth quarter caregiving behaviour in the home environment, Ainsworth developed four scales: sensitivity-insensitivity, cooperation-interference, accessibility-ignoring and acceptance-rejection (Ainsworth et al., 1978). A mother scoring high on sensitivity is alert to the infant's signals, interprets them accurately, and responds appropriately and promptly. A mother scoring high on cooperation respects the autonomy

[1] Some authors have argued that temperament (in terms of a biological predisposition to a particular behavioural style that is independent of experience) can either partly or wholly account for infant behaviour in the ISS (e.g. Belsky and Rovine (1987); Planalp and Braungart-Rieker (2013)). However, as was noted in the review of a number of meta-analyses, "the empirical evidence is still insufficient to document a causal role of temperament in the development of attachment security" (van IJzendoorn and Bakermans-Kranenburg, 2004, p.234). In this work we take the classical (and most widely held) view of attachment theory: that the biggest factor in infant ISS (and attachment in general) behaviour is the cumulative prior conduct of the caregiver towards the infant, and we do not consider the role of temperament

of the infant and avoids interfering or exerting control over the infant's behaviour whenever possible. An accessible mother was defined as one with high psychological availability and physical accessibility, and low preoccupation with other distracting demands on her attention. Finally, a mother scoring high for acceptance "cheerfully accepts the responsibility of her maternal role" without interacting with the infant as if they were an obstacle to other goals.

Ainsworth found that secure infants tended to have mothers that scored highly on all four of these scales. Although acceptance-rejection was a better discriminant between avoidant and ambivalent ISS classifications, she found that it was the sensitivity scale that best predicted security over insecurity. A later meta-analysis of 4176 infants confirmed the large effect size for sensitive-responsiveness (according to Ainsworth's original definition) on attachment security in the ISS (Wolff and Ijzendoorn, 1997). The study showed that responsiveness alone (without additional consideration of the caregiver's behaviour) was a far weaker predictor than the more encompassing definition of sensitivity, and also found a number of other predictors of attachment security, such as synchrony (the extent to which interactions appeared to be reciprocal and mutually rewarding) and mutuality (including a measure of the extent to which exchanges were positive and involved the infant and caregiver attending to the same thing, and the caregiver's responsiveness and skill in modulating the infant's arousal). A meta-analysis of 13835 children assessed with attachment Q-sort (an alternative procedure to the ISS) also found a strong effect size of caregiver sensitivity on security (van IJzendoorn et al., 2004), and another meta-analysis focusing on caregiver-intervention studies aiming to increase caregiver sensitivity additionally found that successful interventions were strong predictors of enhancement to attachment security (Bakermans-Kranenburg et al., 2003), supporting an effect for sensitivity on attachment security. On the other hand, caregivers of avoidant infants have been found to be dismissing or rejecting of the infant's bids for connection; are emotionally unavailable, distant and seemingly uncomfortable; are more rigid and inflexible in their behaviour; and have been observed to have a propensity to withdraw when the infant was sad. An ambivalent infant attachment style has been found to correlate with an inconsistent style of caregiving that fluctuates between under- and over-involvement (Cozolino, 2014, p.150).

2.3 Disorganisation

The secure and insecure (avoidant, ambivalent) forms of attachment originally uncovered by Ainsworth are typically considered to be organised forms of attachment, in that they manifest in coherent and consistent behavioural strategies in the infant that are thought to result from adaptations to the (attachment related) behaviour of their particular caregiver. A fourth attachment classification was later identified by Mary Main (Main and Solomon,

1986), who noticed that a relatively small subset of infants in the ISS did not fit into one of the three organised classifications in that they seemed to lack a coherent strategy for their attachment behaviour. These infants displayed bizarre or contradictory behaviours in the caregiver's presence and on reunion, including freezing, contradictory approach-avoidance tendencies, backing towards the caregiver, and stifled screaming; behaviours which were displayed without an immediately obvious explanation and often amidst the organised behavioural strategies.

Six categories of behaviour are considered in scoring disorganisation in the ISS (Main and Solomon, 1990; Goldberg, 2000, p.25):

1. Sequential or simultaneous displays of contradictory behaviour (e.g. strong proximity seeking followed by strong avoidance)
2. Undirected, incomplete or interrupted movements and expressions (e.g. seeking proximity to caregiver but turning away before contact is made)
3. Stereotypies (repetitive or ritualistic), asymmetrical and mistimed movements and expressions (e.g. sudden jerky movement)
4. Slow movements and expressions, or freezing (e.g. holding completely still for an extended period)
5. Direct displays of apprehension or fear towards the caregiver (e.g. hands over mouth upon return of caregiver with a fearful facial expression)
6. Direct displays of disorganisation, disorientation or confusion (e.g. aimless wandering)

Disorganised infants (considered insecure since they are unable to effectively use their caregiver as a secure base) are unable to maintain a consistent strategy with regards to attachment behaviour. They are described as either lacking a coherent strategy altogether, or having an inclination towards a particular strategy (secure, avoidant or ambivalent) but being unable to fully realise that preference. Thus, an effort is sometimes made to assign a subcategory to disorganised infants, according to the underlying strategy that best fits their overall behavioural patterns: D-Secure or D-Insecure (D-Avoidant or D-Ambivalent). In cases where there does not appear to be an underlying strategy, infants are typically assigned D-Unclassifiable. In the previously discussed North American meta-analysis (van Ijzendoorn et al., 1999a) disorganised infants were found to comprise approximately 15 percent of the population.

2.3.1 Hypotheses on Causes

There are two main hypotheses with respect to the patterns of caregiving behaviour that give rise to infant disorganisation, which we overview here. These are a) that the caregiver

displays frightened or frightening behaviour towards the infant, and b) that the caregiver displays more general patterns of disrupted affective communication.

2.3.1.1 Frightened/Frightening Behaviour

Main and Hesse proposed that frightened and/or frightening behaviour on the part of the caregiver might be the key to understanding the emergence of disorganised attachment in the infant (Main and Hesse, 1990). They suggested that disorganised infant behaviour results from an unsolvable dilemma, in that the caregiver (i.e. the secure base from whom the infant seeks comfort) has, as a result of past experience, also become associated with being a source of fear (leading to "fear without solution"). In order to test this hypothesis, a coding system including six categories of such infant-directed caregiving behaviours was developed (Main and Hesse, 1992):

1. Anomalous, frightening/threatening behaviour
2. Sexualised behaviour
3. Disorganised/disoriented behaviour
4. Deferential, timid and submissive behaviour
5. Dissociative behaviour
6. Frightened behaviour

A number of studies using this scale have found support for the hypothesis (Jacobvitz et al., 1997; Schuengel et al., 1999; McMahan True et al., 2001; Abrams et al., 2006). Also broadly in support is evidence suggesting a strong association between caregiver maltreatment and infant disorganisation (Carlson et al., 1989; van Ijzendoorn et al., 1999b).

2.3.1.2 Disrupted Affective Communication

Lyons-Ruth et al. built on Main's Frightened/Frightening hypothesis on the origins of disorganised attachment by considering a wider variety of caregiving behaviours and patterns of affective communication mediating the caregiver's overall ability to moderate the infant's distress (Lyons-Ruth et al., 1999). They proposed instead that disorganised attachment arises as a result of a "failure to repair", with the infant unable to organise a consistent strategy for attachment (and for using the caregiver as a source of comfort) when the caregiver comes to be seen as ineffective in modulating their arousal and assisting them to achieve a regulated state. This inability to modulate the infant's arousal was proposed to be caused by the caregiver's own prior attachment experiences: if the caregiver did not themselves experience comfort from their own caregiver during times of distress, then the

infant's distress is proposed to evoke fear in the caregiver. This fear in the caregiver was further proposed to result in competing parental attachment tendencies towards the infant (e.g. communications and behaviours that simultaneously invite and reject the infant), resulting in patterns of disrupted affective communication and misattunement that fail to modulate the infant's arousal. Accordingly, they introduced the Atypical Maternal Behaviour Instrument for Assessment and Classification (AMBIANCE) scale as a method for coding such disrupted (atypical) affective communication in caregivers (Bronfman et al., 1999; Safyer, 2013, Appendix G). Under this scale, which includes many of the frightened, frightening and dissociative behaviours coded for by Main and Hesse (1992), disrupted affective communication has five dimensions: Affective Communication Errors (ACE), Role/Boundary Confusion, Fearful/Disoriented Behaviours, Intrusiveness/Negativity, and Withdrawal.

In their initial study, Lyons-Ruth et al. (1999) investigated disrupted maternal affective communication in 65 mother-infant dyads (22 secure, 13 avoidant and 30 disorganised) during the ISS. Disorganised infants were sub-categorised into D-Secure and D-Insecure subgroups: D-Secure infants tended to continue to approach the caregiver while displaying other disorganised behaviours, whereas D-Insecure infants combined disorganised behaviours with avoidance and resistance. While caregivers of D-Secure infants tended to display higher rates of withdrawal (including compared to mothers of avoidant infants), caregivers of D-Insecure infants were prone to disorientation, negative/intrusive (including frightening) behaviour and role confusion, such that they tended to provide a complex mix of cues that were both involving and rejecting of infant approach. Crucially, across all infants, maternal ACEs were correlated with increased disorganisation, proximity seeking and crying, and ACEs were additionally correlated with increased resistance in infants classified as either organised or D-Insecure. Overall rates of disrupted affective communication on the AMBIANCE scale differentiated mothers of organised and disorganised infants (and also D-Secure from D-Insecure, with more for mothers of the D-Insecure infants), however ACE was the only dimension to individually differentiate mothers of organised and disorganised infants (with no significant difference in rates of ACEs between D-Secure and D-Insecure mothers).

Subsequent studies have found evidence for elevated rates of ACEs in both mothers of ambivalent and disorganised infants, during both free-interaction and while undertaking the ISS. In a sample of 82 strange-situation aged infants (7 avoidant, 27 secure and 48 disorganised, sub-categorised as 12 D-Secure, 21 D-Avoidant, 28 D-Ambivalent), Madigan et al. (2006) investigated disrupted affective communication between infants and their mothers during free interactions both with and without toys. They found heightened prevalence of ACEs and fearful/disoriented behaviours during both interactions in mothers of disorganised compared to organised infants, and elevated rates of role/boundary confusion and negative/intrusive behaviour in the without-toys condition.

Safyer (2013) conducted an ISS with 52 infants (22 categorised as secure, 10 avoidant,

8 ambivalent, 12 disorganised) and used the AMBIANCE scale to measured the prevalence of ACEs in the mother on episodes 2,3,5 and 8 of the procedure. Although no significant difference was found in the amount of ACEs in mothers of disorganised compared to organised (secure, avoidant and ambivalent) infants, significantly higher scores were found for insecure (avoidant, ambivalent, disorganised) versus secure. Furthermore, in pairwise comparisons of the four attachment classifications, it was found that mothers of ambivalent infants displayed significantly more ACEs than mothers of secure, avoidant and disorganised infants. The authors also found higher mean ACE score for mothers of disorganised compared to secure and avoidant infants (although, contrary to the author's prior hypothesis, this difference did not pass the threshold for statistical significance).

Studies have also found evidence for overall heightened levels of disrupted affective communication (including ACE) in mothers of disorganised versus organised, and ambivalent compared to secure or avoidant, infants. Goldberg et al. (2003) studied 197 dyads (152 organised and 45 disorganised, subcategorised as 23 D-Secure and 22 D-Insecure) during episodes 2,5 and 8 of the strange situation. They found that overall disrupted affective communication was lower for mothers of organised compared to disorganised infants, and for mothers of D-Secure compared to D-Insecure infants (although they did not present scores for the ACE dimension alone). The finding of elevated ACEs in mothers of ambivalent infants in Safyer (2013) discussed above is consistent with the results presented in Grienenberger et al. (2005). In that study of 45 mother-infant dyads undertaking the ISS (23 secure, 8 avoidant, 4 ambivalent and 10 disorganised), higher scores for overall disrupted affective communication were found in mothers of ambivalent and disorganised infants compared to mothers of secure and avoidant infants (although they again did not present scores specifically for the ACE dimension).

2.4 Adult Attachment

Although initially focusing on infant development, attachment theory expanded in scope to examine internal working models in adults and their impact on, for example, romantic and peer relationships. The most influential measure of attachment in adults is the Adult Attachment Interview (AAI) (Main et al., 1985; Hesse, 2008), which is a semi-structured interview aiming to elicit an adult's state of mind with respect to attachment. The AAI interview asks questions focusing on early memories of childhood experiences and the relationship the participant had with their parents (such as separation experience and relationship attitudes). The aim is to assess how these past attachment-related experiences are reflected on and evaluated with respect to current functioning, which leads to an attachment classification that is based on both the content and structure (coherence) of the participant's verbal responses to the interview questions.

The AAI defines three adult attachment classifications that mirror the organised (secure, avoidant and ambivalent) infant classifications. Secure-autonomous individuals are those who provide narratives that value attachment relationships and describe them in a balanced way using a coherent and consistent discourse. Those who are classified as Dismissing tend to minimise or devalue the personal impact of attachment relationships, have a defensive discourse, and show memory lapses. Preoccupied individuals provide an incoherent discourse that tends to be over-dependent, and includes angry or ambivalent representations of past attachment experiences. In addition to these classifications an individual might also be classified as either Unresolved or Hostile/Helpless (Lyons-Ruth et al., 2005), which are considered to be disorganised classifications. An unresolved state of mind is one which shows trauma in relation to previous attachment experience (such as loss or abuse), whereas hostile/helpless individuals display contradictory, unintegrated and unexamined emotional evaluations of the caregiver (which may include negative devaluing mental representations of the self and/or caregiver and references to fear).

A meta-analysis of studies (854 infant-caregiver dyads) comparing infant ISS classification (measured at approximately 1 year of age) and caregiver's AAI classification (assessed either prospectively before childbirth, concurrently with the infant classification, or retrospectively) showed a strong predictive power for adult attachment classification on infant attachment type (which was independent of when the AAI classification was undertaken) (van IJzendoorn, 1995). Considering only organised infant/adult attachment types, then on a secure-insecure split adult attachment classification was found to predict infant type with 75% accuracy, whereas on a three-way split amongst these classifications (i.e. pairing secure-autonomous with secure, dismissive with avoidant and preoccupied with ambivalent) predictive accuracy was found to be 70% (or 69% for just those studies in which AAI classification was assessed before childbirth). Considering also disorganised/unresolved infant/adult attachment classifications, then on a secure-insecure split AAI classification predicted infant attachment type with 74% accuracy, while on a four-way split (including unresolved-disorganised pairings) the correspondence was 63% (or 65% for prenatal studies only). In other words, the analysis found that with high probability a caregiver's prenatal adult attachment type will predictively correspond to the attachment type of their infant at one year of age. Unresolved caregivers display a higher prevalence of frightened/frightening infant-directed behaviour (Jacobvitz et al., 2006) and have been directly linked to higher occurrence of infant disorganisation in a number of other studies (Steele et al., 1996; Schuengel et al., 1999), as have hostile-helpless states of mind (Lyons-Ruth et al., 2005). Research has also demonstrated stability in attachment type from infancy to adulthood (see (Cozolino, 2014, p.155) for a summary), with secure to insecure changes typically linked to adverse life experiences or trauma (e.g. Waters et al. (2000)), and insecure to secure transformations also possible (e.g. Hamilton (2000)). Overall, this suggests a tendency towards intergenera-

tional transmission of attachment types, however this transmission is by no means inevitable, especially given intervention.

A number of other measures of adult attachment have been proposed: we briefly describe here those measures used in studies that we refer to later on. The Adult Attachment Projective (George et al., 1999) assesses adult attachment (according to the same attachment groups as in the AAI) by asking participants to provide narratives (tell a story) in response to a standardised set of pictures depicting attachment-relevant events such as solitude, separation or abuse. A number of self-report questionnaires have also been developed. In The Relationship Questionnaire (Bartholomew and Horowitz, 1991), participants are presented with groups of statements regarding attachment and relationships (with each group corresponding to prototypical behaviour for a particular attachment type), with a classification made according to which group of statements the participant reports as best reflecting their stance. The Experience in Close Relationships self-report questionnaire (Brennan et al., 1998) similarly presents attachment-related statements, but asks participants to score how strongly they agree/disagree with each statement. This results in scores for attachment anxiety and avoidance, based on which an overall classification is made (according to a quadrant, with low scores on both scales corresponding to security, high scores to fearful disorganisation, low avoidance and high anxiety to ambivalence, and high avoidance with low anxiety to avoidant attachment).

2.4.1 Summary of Correlations Between Infant and Adult Attachment Classifications

A summary of findings (adapted from Cozolino (2014)) on the correlations between infant and adult attachment types and home caregiving behaviour is given in Fig. 2.2.

2.5 Attachment and Psychological Health

Secure attachment is believed to confer social developmental advantages on the individual: for example, in a longitudinal study infants classified as secure were found to be rated higher in competence by teachers and observed to be less isolated at preschool age, and rated higher in self-confidence and leadership in middle childhood (Sroufe et al., 1999). While secure attachment is thought to be associated with mental resilience and relatively quick recovery from stress, it has been argued that insecure forms of attachment (the prevalence of which correlates with a wide range of issues including depression, clinically significant anxiety, and various personality disorders) can be viewed as a general vulnerability to (although not necessarily a sufficient cause of) mental disorders (with particular symptomatology influenced by other factors including genetic, developmental/experiential, and environmental,

	ISS	AAI	Home Caregiving Behaviour
Secure	Infant seeks proximity, is easily soothed and returns to exploration	Detailed memory, balanced perspective, narrative coherency	Emotionally available, perceptive and effective
Avoidant	Infant does not seek proximity, nor (outwardly) appear stressed	Dismissing/denial, minimising, idealisation, lack of recall	Distant and rejecting
Ambivalent	Infant seeks proximity, is resistant and not easily soothed, slow to return to exploration	Lots of output, intrusions/pressure, preoccupied, idealising or enraged	Inconsistent availability, ACEs
Disorganised	Chaotic, inconsistent, incoherent behaviours	Disoriented, conflictual behaviour, unresolved loss, traumatic history	Disorienting, frightened, frightening, hostile, sexualised behaviours, ACEs

Figure 2.2: Summary of findings from attachment research, adapted from (Cozolino, 2014, p.150).

(Mikulincer and Shaver, 2012)). This general link between attachment insecurity and psychopathology is proposed to be mediated by dysfunctional beliefs about the self and others (e.g. lack of self cohesion, or unstable self esteem), disruptions in the development of capacities for regulation (including self-regulation) of emotion, and problems in interpersonal relationships resulting from this incapacity (Mikulincer and Shaver, 2012).

From both a developmental and clinical perspective, disorganisation is a particularly significant attachment categorisation. Disorganised attachment in infancy has been found to predict teacher ratings of hostile classroom behaviour at five years of age (Lyons-Ruth et al., 1993) along with childhood aggression towards peers (Lyons-Ruth, 1996), and criminality and drug use in young adulthood have also been linked to disorganised models of attachment (Allen et al., 1996). Disorganised attachment has furthermore been linked to Borderline Personality Disorder (BPD), a disorder characterised by affective instability, intense angry emotional outbursts, and difficulties of interpersonal exchange (American Psychiatric Association, 2013; World Health Organization, 1992). A prominent theoretical account of BPD posits a developmental basis for the disorder in early disorganised forms of attachment experience, as a result of a defensive inhibition of mentalization capacities (Fonagy et al., 2000; Fonagy, 2000), which is supported by evidence from a longitudinal study that found disorganised attachment at 18 months (along with maltreatment and maternal hostility) to be predictive of borderline symptoms at 28 years of age (Carlson et al., 2009). As with disorganised attachment, BPD shows a strong intergenerational effect (see discussion in Stepp et al. (2012)), and caregivers with BPD have been linked to the development of disorganised forms of attachment in dependent infants. For example, Hobson et al. (2005) assessed 12-month-old infants of women with and without BPD in three settings including the ISS, and found that a higher proportion of infants of caregivers with BPD (8 out of 10, compared with 6 out of 22 for the control group) were categorised as having disorganised attachment (the remaining 2 infants in the BPD group also showed signs of disorganisation). Women with BPD have additionally been found to display a higher prevalence of hostile-helpless (Lyons-Ruth et al., 2007) and unresolved (Macfie et al., 2014) states of mind in the AAI, which (as we discussed previously) have both otherwise been linked to disorganised infant attachment. Moreover, BPD mothers display an increased tendency for infant-directed behaviour otherwise found to be likely to lead to insecure and disorganised forms of attachment (including frightened and disoriented behaviours and ACEs during the ISS (Hobson et al., 2009)), and a decreased tendency towards positive and affiliative behaviours (including less positive affect and more insensitive behaviour in response to infant distress during ISS reunion (Kiel et al., 2011), and reduced smiling, imitation, touch (White et al., 2011) and sensitivity and responsiveness (Newman et al., 2007) during free-interactions) of the type thought to encourage security. Also supporting the link between infant disorganisation and caregiver BPD, children aged 4-7 of BPD as opposed to non-BPD mothers have been found

to express more negative parent-child relationship expectations (with a perception of the relationship tending towards one that is dangerous, unpredictable and/or involving serious or wilful harm, (Macfie and Swan, 2009)), while a study of 15 year old children found a significant positive correlation between perceptions of maternal hostility reported by the child and increasing borderline characteristics in the mother (Herr et al., 2008).

Disorganised forms of attachment have also been linked to dissociation, which is typically defined as a deficit in the integration of memory, consciousness and identity that manifests in the form of either a lack of attention to the outside environment or sudden breaks in the continuity of discourse, thought or behaviour (of which the individual is unaware) (Liotti, 2004). Dissociative states of mind are associated with both individuals with BPD and those classified as unresolved in the AAI (Liotti, 2004). Previous experiences of unresolved trauma or loss are thought to result in sudden floods of emotion from implicit attachment memories in such individuals, causing rapid shifts in state of mind that might sometimes lead to a trance-like state, and other times to hostile rage (Siegel, 2012, p.140). As we discussed in Section 2.3, disorganised infants display many behaviours within the context of attachment that have been noted as similar to those indicative of dissociation in adults, including sudden immobility, freezing and unresponsiveness towards the caregiver; contradictory patterns of movement (displayed either simultaneously or in quick succession); sudden dazed or trance-like expressions; disorientation; and aggressive gestures performed suddenly in the middle of affectionate behaviours displayed towards the caregiver (Liotti, 2004; Main and Morgan, 1996). It has been proposed that these behaviours might be the first instance of dissociative reactions during life (Liotti, 2004), and that early disorganised attachment experiences increase the vulnerability for dissociative reactions to other traumas later in life (Liotti, 1992, 2004).

2.6 Neuroscience of Attachment

In this section we briefly outline the key neural and regulatory circuits, systems and abstractions currently believed to underlie various infant attachment phenomena. Much of what follows is based on the integrative work of Schore (2002, 2003a,b), Siegel (2012) and Cozolino (2010, 2014) who, over the past few decades, have worked to develop a neuroscientific framework for attachment. We focus here particularly on key functionality of the networks thought to be centrally involved in attachment from the developmental perspective of the infant. More detail on these networks (including recent studies that have correlated patterns of activity with independently assessed attachment classifications), along with the neuroscience of parenting and caregiving, is introduced in Chapters 4 and 5.

The newborn baby has only a partially developed brain, with the brainstem and a significant part of the limbic system (including the amygdala) highly mature at birth (Cozolino,

2010; Humphrey, 1968; Nikolić and Kostović, 1986; Ulfig et al., 2003), but cortical networks that regulate these areas developing later, over critical periods of varying length and onset (Cozolino, 2010). As Cozolino explains:

> "It is an unfortunate twist of evolutionary fate that the amygdala is mature before birth while the systems that inhibit it take years to develop. This leaves us vulnerable to overwhelming fear with little or no ability to protect ourselves. On the other hand, evolution has also provided us with caretakers who allow us to link into their developed cortex until our own is ready" (Cozolino, 2010, p.253).

At this early stage of development these authors view the brain primarily as a social organ, facilitating communication and emotional appraisal in order to achieve homeostatic regulation. The many early, repeated attachment interactions that an infant and caregiver engage in are thought to be processed primarily in the emotionally-biased RH (which, according to Schore, is both dominant and undergoing a period of critical development during the first year of life (Schore, 2003a, p.74)), and the internal working model (attachment schema) has been theorised to be based in unconscious and implicit memories, rooted mainly in RH-biased networks centred on the Orbitofrontal Cortex (OFC), insular and cingulate cortices, amygdala and hypothalamus (Schore, 2003a; Cozolino, 2014, p.53,143); areas known to be central to emotional processing, social cognition, fear conditioning and homeostatic regulation. The Anterior Insular (AI) and Anterior Cingulate Cortex (ACC) are involved in a wide range of social information and emotion processing tasks (Cozolino, 2014, p.104,238), and in Chapter 5 we will explore the role of these regions in facilitating empathic states. Within the context of early attachment experience, the OFC and amygdala are thought to be crucial regions in the assignment of emotions to caregiver representations and the prediction of interaction outcomes, along with the regulation of homeostatic state via connectivity with the hypothalamus. We explore the functions of these regions in more detail in the following sections, in order to lay the ground work for our model of self-directed bonding in Chapter 4.

Two key systems which dyadic attachment interactions serve to regulate are the arousal and stress systems which (as we will see in Chapter 4) are strongly interlinked and mutually stimulatory. Arousal (controlled by the Autonomic Nervous System (ANS)) is a physiological and psychological state of being alert or reactive to environmental stimuli. Of particular relevance is the Sympathetic Nervous System (SNS): a subsystem of the ANS responsible for mobilising the body for "fight or flight" in response to perceived threat or danger. During the fight or flight response, Norepinephrine (NR) (which increases alertness, attention, heart rate and anxiety) is released from the Locus Coeruleus (LC) in response to stimulation by the amygdala and hypothalamus (both of which receive inputs from the OFC) [2].

[2]Another subsystem of the ANS, the Parasympathetic Nervous System (PSNS), serves an opposite,

The body's reaction to stress (loosely defined as a reaction to perceived challenge) is controlled by the hypothalamic–pituitary–adrenal axis. When stimulated by the amygdala, the hypothalamus releases Corticotropin-releasing Hormone (CRH), which in turn results in the release of adrenocorticotropic hormone from the pituitary gland, which causes the release of cortisol from the adrenal cortex. Cortisol serves to increase blood sugar, suppress the immune system, and aid in metabolism, but high sustained (chronic) levels in the body can lead to high blood pressure and muscle damage. Evidence from animal and human studies furthermore suggests that chronic stress might lead to a variety of effects on the brain, including cell destruction, changes in proportions of cell types, and decreased plasticity (Leuner and Shors, 2013) in regions including the hippocampus (Frodl et al., 2010; Woon et al., 2010; Krugers et al., 2006), Medial Prefrontal Cortex (mPFC) (Radley et al., 2005, 2006) and OFC (Varga et al., 2017).

Crucially, according to these authors, dyadic attachment interactions serve not only to immediately regulate the infant's emotion, stress and arousal; but moreover, through a mechanism of experience-dependent plasticity, facilitate the construction of networks that form the basis for self-regulation and interactive coping strategies employed later in life in other stressful situations (Wallin, 2007, p.24). The appropriate caregiver empathically attunes to the infant's distressed state before guiding them towards positive emotional states in which stress and arousal are returned back to within tolerable levels. This repair in the infant's internal state is achieved with affectively charged interactions and communications involving, for example, the physical expression of positive emotion, touch and singing. As we will see in Chapter 4, such communications activate areas in the infant's brain involved in positive emotion and reward (including the OFC). Emotionally available caregivers who engage in these reciprocal interactions aimed at repairing the infant's internally distressed state are believed to foster healthy growth and integration in these key cortical-emotional circuits, and these dyadic transitions from positive to negative and back to positive affect are thought to form the basis for the emergence of emotional resilience and imprint in the infant expectations that ruptures in internal state will be repaired (Schore, 2001). On the other hand, caregivers who are unavailable and insensitive to infant attachment need can hamper this development, and in extreme cases where the caregiver exposes the infant

dampening function to the SNS by preparing the body for states of "rest and digest" by lowering blood pressure and heart rate. The hypothalamus, which receives strong projections from the OFC and amygdala, is involved in the regulation of both the SNS and PSNS systems (Saper and Lowell, 2014). According to Polyvagal Theory, the PSNS consists of a phylogentically older system (the unmyelinated branch of the vagus nerve) involved in defensive immobilisation, and a phylogenetically newer system (the myelinated branch) which rapidly inhibits SNS and hypothalamic–pituitary–adrenal axis activity to allow for social engagement (Porges, 2011), and it has been proposed that insecure attachment experience could disrupt OFC integration with and regulation of this vagal system (Schore, 2003a, p.34). Some attachment theorists have proposed secure attachment to involve a good balance between SNS and PSNS activity, avoidant attachment a bias towards PSNS activity, ambivalence a bias towards SNS activity, and disorganisation to involve simultaneously high levels of both SNS and PSNS activation (Cozolino, 2014, p.152), (Siegel, 2012, p.318).

to trauma, excessive fear or unregulated stress during this critical period of development, representations of a self interacting with a highly dysregulating other are stored in the infant's brain (Schore, 2003a, p.35) predisposing them to future psychopathology (Schore, 2003a, p.6).

Recall that infants with insecure (particularly ambivalent and disorganised) attachment types have been found to exhibit elevated cortisol levels compared to secure infants in response to the ISS (Section 2.2), and that during this procedure insecure infants have shown relatively high increases in heart rate on separation and slower return to baseline on reunion. In Chapter 4 we will additionally review evidence for heightened baseline cortisol activity in adults classified as preoccupied (ambivalent), suggesting a lasting impact for failures in attachment-based regulation of arousal and stress systems during insecure infant attachment experiences.

2.6.1 Hemispheric Lateralisation of Emotional Processing and Control

One proposal in Schore's neuroscientific theory of attachment (outlined here) is a dominant role for RH regions during early development and attachment experiences, which are thought to be primarily emotional (rather than cognitive) in nature. Although both hemispheres develop rapidly during an infant's early years, Schore highlights studies of both electrical activity (using Electroencephalography (EEG), (Giudice et al., 1987)) and blood flow (using Dynamic Single Photon Emission Computed Tomography (Chiron et al., 1997)) in the brain suggesting that the right cerebral hemisphere is dominant in both its development and processing during the first 3 years of life (corresponding to the ages of infants in ISS data), after which there is a shift to the Left Hemisphere (LH). Across all age groups, particular distinctions across the lateral dimension have been noted in emotional control and processing, for which it is believed that (in general) the RH is dominant. This is supported anatomically in the greater relative connection that the RH has to subcortical regions compared to the LH (Schore, 2003a, p.61), and also by (for example) studies of facial affect recognition ability in children with hemispheric lesions (Voeller et al., 1988). It also appears that the RH is more directly involved with the regulation of cortisol (Wittling and Pflüger, 1990), and the ANS and arousal (Schore, 2003b) than the LH.

A large number of studies (reviewed by Schore) have found evidence supporting the idea of lateralised valence, where RH activation is associated with more negative emotion whilst LH activation is associated with more positive emotion. For example, EEG recordings have shown that negative mood and depression are associated with greater frontal cortex activity in the RH (Henriques and Davidson, 1991; Flor-Henry et al., 2004), and EEG recordings from the left and right frontal and parietal regions of 10 month old infants showed that RH activation level correlated with the visible level of distress shown by the infant during

maternal separation (Davidson and Fox, 1989). Lesions to the LH have been found to be accompanied by depression, and RH lesions with elevated mood (Carran et al., 2003; Braun et al., 2008). A Functional Magnetic Resonance Imaging (fMRI) study found enhanced RH activation when participants were shown negative emotional images, and higher LH activation when they were shown positive images (Lee et al., 2004). Another fMRI study found a strong bias towards the RH in the role of processing pain (Symonds et al., 2006). Some neuroscientists are nonetheless cautious as to the extent to which the generalisation of lateralised valance holds. In particular, studies have found a greater relative frontal LH activation during states of anger, which would seem to contradict the model (Harmon-Jones et al., 2010). There is, however, largely a consensus that the RH generally mediates emotion resulting in avoidance behaviour, whereas the LH is involved in emotion resulting in approach (see for example (Hane et al., 2008), which studied EEG measures in infants at 4 and 9 months). This is consistent with both the findings of the anger studies and those used to support the idea of lateralised valance.

A few studies have investigated differential lateralisation effects with respect to attachment classifications. In an ISS study of 159 infants aged between 13 and 15 months, EEG measures (taken as a baseline, during play with mother and during play with an experimenter) showed that insecurely attached infants were found to exhibit relatively reduced left frontal brain activity compared to securely attached infants (Dawson et al., 2001), broadly consistent with theory of emotional valance lateralisation. In an EEG study on adults, avoidant individuals made more errors when judging positive attachment-related words presented to the RH using divided visual field methods [3] (Cohen and Shaver, 2004). In another EEG study, adults were shown attachment-related video clips aimed at inducing happiness, sadness and fear (Rognoni et al., 2008). Participants rated as more avoidant or ambivalent were found to have a higher resting asymmetry towards the RH relative to secure adults. In response to positive stimuli, ambivalent individuals showed a greater relative LH activation, whilst avoidant showed greater relative RH activation. In contrast, in response to negative stimuli, ambivalent individuals showed higher relative RH activation (especially to fear-inducing stimuli), and avoidant a higher relative LH activation.

2.6.2 Amygdala

The amygdala is thought to be a key subcortical region within the context of attachment. It plays an important role in the initial appraisal of external stimuli (Rolls, 1990), and the learning and storage of these appraisals and emotional events in the form of unconscious, implicit memories. The emotional significance that the amygdala assigns to external stimuli

[3] In Divided Visual Field experiments, visual stimuli are presented to either the left or right visual field. If the visual stimulus appears on the left visual field, the visual information is initially projected to the right hemisphere, whereas if the visual stimulus is presented to the right visual field the visual information is initially received by the left hemisphere.

is projected to other regions involved in cognitive functioning, such as attention, perception and explicit (conscious) memory. This can result in various modulating effects, for example enhancing the memory of an emotionally salient event in the hippocampus, or re-directing cortical attention (LeDoux and Phelps, 1993). From the perspective of social interaction, the primary role of the amygdala can be summarised as the modulation of vigilance and attention, remembering emotionally significant events and people, and preparing the body for a fight or flight response (Cozolino, 2014, p.175-178).

It is widely acknowledged that one of the amygdala's primary roles is to remember a threat, and generalise this threat to other possible future situations (Cozolino, 2010, p.254). The amygdala has long been known to be involved in triggering a state of fear (Feinstein et al., 2011): in particular, it has been strongly implicated in Pavlovian fear conditioning, in which an emotionally neutral Conditioned Stimulus (CS) is repeatedly paired with an Unconditioned Stimulus (US) that triggers a state of fear, until the CS itself comes to elicit a fear response in absence of the US (LeDoux, 2003). The lateral nucleus (a part of the Basolateral Amygdala Complex, consisting of the Lateral, Basolateral (Basal) and Basomedial (Accessory Basal) nuclei (BLA)) is a major input region to the amygdala, receiving sensory information from the thalamus and cortex, and responds to both CS and US input. The BLA then projects to the central nucleus, which has direct projections to the hypothalamus and ANS (which allow for the rapid translation of appraisals into stress and autonomic fight or flight responses (Davis, 1992)) along with the dorsal central (periaqueductal) gray, which can induce defensive, fear-invoked freezing behaviour (LeDoux et al., 1988). This pathway is often called the "fast-path" for threat response, and allows the body to respond to a potential threat even before conscious awareness, and while the "slow-path" (mediated from the thalamus through the neocortex and hippocampus) analyses the extent of the treat posed by the external stimulus within context (LeDoux, 1997).

As we will explore in more detail in Chapter 4, insecure (particularly ambivalent) forms of adult attachment have been found to correlate with relatively high levels of activity in the amygdala in a number of social and attachment-related contexts.

2.6.3 Hypothalamus

The hypothalamus is an integrative region which brings together a range of inputs relating to the internal environment, comparing them to setpoints (ideal ranges) and activating autonomic, endocrine and behavioural responses in order to maintain the internal state within these ranges (Saper and Lowell, 2014). Receiving strong projections from the OFC and amygdala, the hypothalamus is summarised by Cozolino as being a key region within the context of attachment in terms of its function in translating social information into hormonal secretions and bodily processes (Cozolino, 2014, p.48). As we will discuss in more detail in Chapter 4, the hypothalamus plays an important role in both stress and arousal

(fight/flight) responses to perceived threat: briefly, when threat is detected the Central Nucleus of the Amygdala (CeA) stimulates the Parvocellular part of the Paraventricular Nucleus of the Hypothalamus (PVNp) to release CRH (the precursor to cortisol, the stress hormone), which causes the LC to release NR and stimulate the SNS. The hypothalamus also plays important roles in attachment from the perspective of the caregiver: as we will see in Chapters 4 and 5, the Magnocellular part of the Paraventricular Nucleus of the Hypothalamus (PVNm) (when indirectly stimulated by the OFC) releases Oxytocin (OXT) (a hormone and neurotransmitter that is involved in bonding, and can enhance prosocial motivation), while the Medial Preoptic Area (mPOA) (a part of the anterior hypothalamus which receives projections from the mPFC) facilitates caregiving behaviour via onward projections to the mesolimbic Dopamine (DA) pathway.

2.6.4 Orbitofrontal Cortex

In general, the OFC has been highlighted as playing an important role in emotional appraisal (i.e. the evaluation of the significance of internal and external stimuli, so that stimuli appraised as being motivationally significant will evoke an emotional reaction), in associating sensory input with reward, and in the evaluation of interpersonal interactions in terms of internal states (Siegel, 2012, p.312-314), although the precise nature of its role in these functions is an open research question. The OFC, which is particularly large in the RH (Schore, 2003a, p.50), has direct and reciprocal connections with the amygdala and hypothalamus (Barbas, 2007), and receives polysensory interoceptive and exteroceptive information about the internal and external environment (Kringelbach and Rolls, 2004). This region is thought by attachment researchers to have an important role in early social communication, attachment behaviour and affect regulation (Schore, 2002, 2003a,b; Siegel, 2012; Cozolino, 2010, 2014): in what follows we briefly overview some of the evidence motivating that claim.

In contrast to more dorsal and lateral areas of the prefrontal cortex which primarily mediate LH-biased "cold" executive functions (Chan et al., 2008; Zelazo and Müller, 2002) (e.g. working memory (Barbey et al., 2013) and planning (Mushiake et al., 2006)), orbital and medial regions (including the OFC), in concert with regions such as the ACC and AI to which they are strongly connected, are more involved in the mediation of RH-rooted "hot" executive functions (such as the regulation of emotion and social behaviour, and decision making involving emotional interpretation). Evidence from a number of imaging studies suggests that both lateral and orbital-medial regions are active in infancy, and that they serve to facilitate this same distinction between types of executive function (see Grossmann (2013) for a review). However, while "cold" areas such as the dorsolateral prefrontal cortex have a long maturation period lasting up until the third decade of life (Goldman-Rakic, 1987; Luna et al., 2001), according to Schore the OFC (which is argued to be directly involved in

a range of attachment functions) enters a critical period of maturation [4] at 10 to 12 months of age, which corresponds to the period at which attachment patterns begin to be reliably observed (Schore, 2003a, p.13,47).

A large body of evidence implicates (particularly lateral parts of) the OFC in being involved in the inhibition and suppression of emotion in a broad range of contexts, including physical sensation (e.g. pain), selective attention, emotion regulation, decision making and social relationships (see Hooker and Knight (2006) for a review). For example, the lateral OFC is thought to be involved in the reappraisal and suppression of negative emotion (Golkar et al., 2012), and individuals assessed as having heightened emotion dysregulation have been found to have reduced grey matter volume in this region (Petrovic et al., 2015). Reappraisal of negative scenes in order to decrease emotional response results in increased lateral OFC and adjacent ventrolateral prefrontal cortex and decreased amygdala activity, with activity in OFC and amygdala inversely correlated (Ochsner et al., 2002, 2004; Phan et al., 2005). One particularly relevant inhibitory pathway (that we will discuss in Chapter 4) extends from the Medial Orbitofrontal Cortex (mOFC) and overlapping/adjacent Ventromedial Prefrontal Cortex (vmPFC) through to the Intercalated Cells of the Amygdala (ITC) and CeA, and is involved in inhibition of the conditioned fear response.

A number of proposals have been made regarding the specific function of the OFC with respect to learning and decision making, and how representations stored by the OFC might relate to emotional states. For example, the influential somatic marker hypothesis proposes that the mOFC is involved in learning and updating associations between particular situations (stimuli) and changes in physiological state (e.g. heart rate). The hypothesis proposes that these learned associations give rise to emotions in similar future situations that occur or are considered, and that these emotions guide decision making (by focusing attention on stimuli predicting positive states, and rejecting stimuli predicting negative outcomes) (Damasio et al., 1996).

A number of proposals on OFC function focus on a role for this region in reinforcement learning. Reinforcement learning is concerned with how an agent should select actions in their environment in order to maximise long-run cumulative reward (where a reward is defined as anything an agent will work to acquire). Delivery of unexpected reward serves to increase the probability of future repetition of the action within the current environment state upon which its receipt was contingent, while unexpected punishment decreases this probability, and a prominent account of reinforcement learning in the brain proposes that the reward-prediction error (i.e. difference between expected and actual reward) is represented in the phasic firing of DA neurons (Niv, 2009). A recent proposal of OFC function suggests

[4] A critical period of maturation is one in which the region is especially sensitive to environmental stimuli. Schore argues that the OFC is in such a period at 10-12 months of age based on evidence of neuron maturation along with synaptic excess (with experience during this period determining which synapses will be preserved or pruned), and also based on evidence for the OFC serving as a central region for attachment and homeostatic regulatory processes. See (Schore, 2003a, p.13) for detailed summary and references.

that this region is involved in decision making processes in terms of representing hidden states of the environment within the context of reinforcement learning, particularly in cases where hidden states are partially observable (i.e. not determinable based on sensory input alone) (Wilson et al., 2014; Schuck et al., 2016).

In the framework outlined by Rolls (2013), which we broadly follow for our models of Self-Attachment in Chapters 4 and 5, the OFC is involved in emotion, (reinforcement) learning and decision making as a result of two primary functions. The first of these functions is its representation of the reward (anything that an animal will work to acquire) and punishment (anything that they will work to avoid) value of primarily reinforcing stimuli (stimuli such as food or pain that are positively or negatively reinforcing, i.e. will increase or decrease the probability of repeating behaviours paired with them, innately and without learning). Evidence reviewed by Rolls from both human and primate studies suggests that the OFC represents reward and punishment associated with many different types of primary reinforcer (including taste, pleasant and painful touch, and visual and auditory stimuli), and in particular that its activity encodes reward *value* within the current motivational context (e.g. neuron firing in response to food stimuli is positively correlated with hunger) (Rolls, 2013, p.71). The second function is the OFC's involvement in the rapid learning and reversal of associations between previously neutral stimuli and primary reinforcers, such that these previously neutral stimuli become (as a result of their association with primary reinforcers) secondary reinforcers (i.e. rewarding or punishing themselves). Rolls argues that emotions are states elicited by reinforcers, defined in terms of the withholding or administration of rewards and punishments (for example, non-delivery of an expected reward might lead to anger, while delivery of a punishment might elicit a state of fear) (Rolls, 2013, p.18).

The OFC (along with the amygdala) outputs information on the reinforcing (emotional) value of stimuli to four main targets in order to drive behaviour: the hypothalamus (which, as we have discussed above, controls endocrine and autonomic reactions such as increased heart rate), the basal ganglia (which is involved with habitual stimulus-response learning and behaviour), the ACC (which is involved in learning response-outcome (where the outcome is a motivationally relevant reward or punishment) associations used in goal-directed behaviour), along with other areas involved in conscious declarative multi-step planning and decision making (Rolls, 2013, p.209). In terms of the OFC's interaction with the basal ganglia and the habitual learning and response system, Rolls proposes in particular that the OFC is responsible for the computation of reward prediction errors, and that the midbrain DA neurons (whose phasic firing encapsulates such errors (Schultz et al., 1997; Niv, 2009)) receive these signals from the OFC (possibly via the ventral striatum) (Rolls, 2015, p.129). This proposal is made on the basis of the OFC containing both neurons that signal the reward expected and reward obtained (unlike midbrain DA neurons), and furthermore neurons that respond to mismatches between these signals.

As we will explore in Chapter 4, the OFC (along with other areas associated with reward processing, including the ventral striatum and Ventral Tegmental Area (VTA)) has been found to be relatively underactive in individuals classified avoidant for attachment in various social interaction-related contexts.

2.7 Existing Work in Computational Modelling of Attachment

Although remaining relatively fertile ground, a series of recent works have begun to explore the dynamics of attachment using mathematical and computational models.

2.7.1 Dynamical Systems Models

Stevens and Zhang (2009) present a dynamical systems model of physiological infant attachment in terms of opioid and arousal variables. The role of the caregiver is considered to be a regulator of the infant's internal physiological state, in particular their opioid (natural pain killers, such as endorphins and enkephalins) and arousal (e.g. norepinephrine, cortisol) systems. High levels of opioid activity are associated with stimulating seeking and exploratory behaviour (for low arousal levels), or states of euphoria (for high arousal levels). On the other hand, when the infant's opioid levels drop below a desirable threshold this is subjectively experienced as anxiety or fear, and is usually associated with increases in the arousal system, triggering attachment behaviour. This attachment behaviour is an attempt on the part of the infant to induce comforting behaviour from the caregiver, which stimulates opioid activity in the infant that decreases their arousal level and helps to return them to a state of equilibrium. The infant's internal state is then given only by joint levels of activity in these opioid and arousal systems. Avoidant infants (who show low levels of separation distress and continued exploration in the ISS) are proposed to be represented by higher levels of sensitivity to opioids than arousal. On the other hand, ambivalent infants (who show high levels of separation distress, and a relatively long time to be soothed on reunion) are proposed to be represented by higher levels of sensitivity to arousal than opioids. Secure infants (who are distressed on separation, but quickly soothed on reunion) are represented by equal sensitivity to opioids and arousal.

Another dynamical systems model of attachment is presented in Buono et al. (2006), which identifies the infant's anxiety (increases in which imply decreases in comfort) along with their emotional distance from the caregiver as quantities of interest. The authors begin by considering the infant's anxiety as a variable driven by parameters representing the insensitivity of the caregiver to the infant's needs and the distance between the infant and caregiver, along with infant-specific parameters governing how they return to baseline

following a stressful episode and their emotional stability. For some parameter values peak anxiety (following separation) can be sustained with slow decay, whereas for other values it returns quickly to baseline, which is proposed to be in correspondence with patterns observed in the organised attachment types during the strange situation along a secure-insecure split. The model is then extended to consider short-term variation in the ability of the caregiver to down-regulate infant stress, by changing the parameter representing the emotional distance of the infant from the caregiver (described above) into a variable. This emotional distance variable itself changes according parameters governing the caregiver's inconsistency and insensitivity, along with a parameter representing the infant's intrinsic curiosity levels and the variable (described above) representing the infant's anxiety level. The equation describes an infant who will seek to increase distance (when it is low) from an insensitive caregiver (who either does not respond to, or makes worse, infant high-anxiety states), who has a natural tendency to reduce distance (proportional to their anxiety), and who has a natural drive to explore (create distance) during low anxiety states. These two equations (infant anxiety and emotional distance) are then combined to give a complete model capable of capturing the three organised forms of infant attachment (secure, avoidant and ambivalent) in terms of anxiety recovery in response to high anxiety and low distance states, and the distance infants put between themselves and caregivers with differing consistency and sensitivity profiles.

2.7.2 Control Systems Model

Buono et al. (2006) propose a simple model of attachment based on a feedback system grounded in control theory. The system represents the infant, whereas the controller (regulating system activity) models the relationship between the caregiver and infant. The inputs comprise a disturbance (representing some external stressful event) and a reference signal (corresponding to a baseline stress level), and the output is the infant's overall stress level. In the absence of the caregiver the infant will amplify externally induced stress, whereas the caregiver is proposed to have a regulating effect on infant stress in reducing the discrepancy between actual stress and baseline levels. The regulator comprises three gains (positive or negative) on infant stress: a proportional gain (assumed to be random) representing the current ability of the caregiver to regulate infant stress, an integral gain representing the healthiness of the past relationship between infant and caregiver (so that caregivers will be more likely to reduce stress if they have previously done so), and a differential gain representing caregiver consistency.

2.7.3 Attachment Types as Strong Patterns

Attachment schemas and prototypes have also been considered within the context of strong (i.e. multiply learned) patterns in a Hopfield network (Edalat and Mancinelli, 2013; Edalat,

2013a). It is shown both mathematically and in simulations that strong patterns give rise to strongly stable attractors, with a large basin of attraction compared to simple (i.e. singly learned) patterns. For a single strong pattern it is proven mathematically that a square law property holds, such that the storage capacity for retrieving a strong pattern exceeds that for retrieving a simple pattern by a multiplicative factor that equals the square of the number of times that the strong pattern has been stored. This strong stability of strong patterns, along with their large basins of attraction for random patterns, is proposed to provide a conceptual model of how an infant might come to conform to a particular pattern of attachment. Psychotherapeutic-driven changes in attachment type are proposed to be modelled by the creation of new strong attractors, which result from the learning of a new strong pattern that increasingly weakens an existing strong pattern as it competes for a larger basin of attraction.

2.7.4 Goal-Based Agent Architectures

Petters (2006a,b) presents a number of goal-based cognitive agent architectures designed to capture empirically observed attachment phenomena comprising 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 innate drives manifest in fear and attachment behavioural systems (to keep the infant safe), and exploratory and socialisation behavioural systems (to foster learning).

The first architecture (Fig. 2.3), called the goal-switching architecture, is intended to capture a regular and repeated interchange of behaviour that alternates between exploration and proximity seeking. This architecture consists of 3 subsystems: goal activation, goal selection and action. The goal activation subsystem contains 4 goal activator modules, each of which represents a different implicit goal: Exploration (which senses the environment, and passes increasing activation when a target of exploration is near), Socialisation (concerned with learning about other agents, with a internal architecture to the exploration activator), Security (which includes caregiver proximity anxiety, object wariness and unknown agent wariness sub-activators), and Physical need (intended to represent the amalgamation of all types of physical need, e.g. feeding and warmth needs). In simulations based on this architecture, the caregiver remains static whilst the infant is free to sense and explore the environment. The infant decides how to behave according to a winner-take-all policy based on the goal with highest activation level: for Exploration the highest activator they move towards a toy, for Physical Need, Security or Socialisation (with caregiver target) activators highest they move towards the caregiver and signal to them, whereas for Socialisation (with stranger target) activator highest they move towards, and signal to, the stranger.

Next, building on this goal-switching architecture, an attempt is made to describe information processing structures that an infant might possess that would enable them to assess

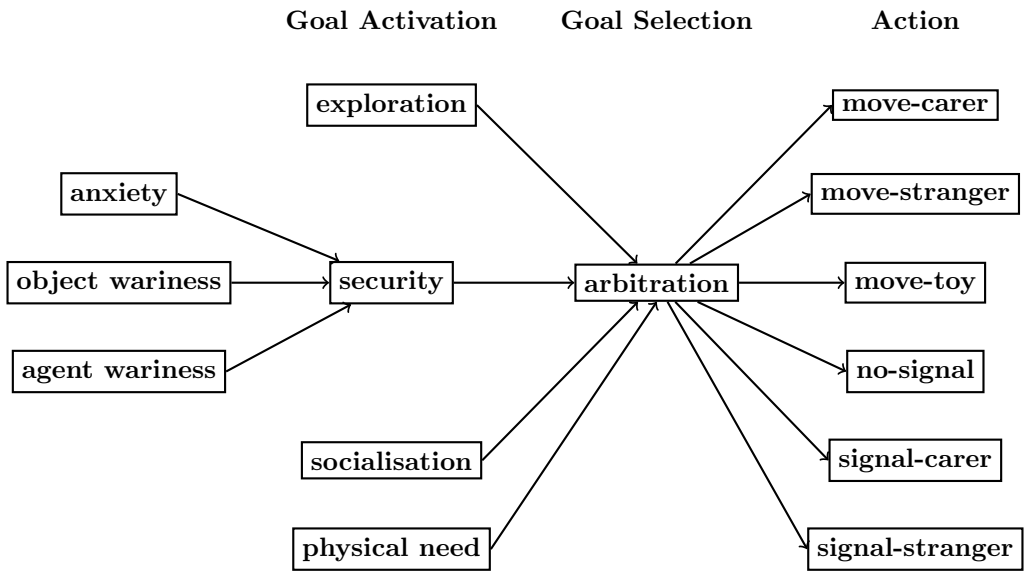

Figure 2.3: Petter's goal-switching agent architecture. See text for details.

carer sensitivity, and form an appropriate (secure or insecure) pattern of behaviour based on this assessment. The caregiver is now assumed to have two goals to manage: exploring their environment and responding to the infant's signalling requests. In addition, the infant can now adjust their proximity safe-range based on the responsiveness (delay) of the caregiver following infant signalling. The infant's goal activation and action selection are as before, but the caregiver agent now only responds to the infant agent when the infant agent's signals are above some particular threshold. In simulatory experiments, there was found to be a critical point for this threshold, determining whether the dyad developed into a secure or insecure (non-specific between avoidant and ambivalent) style.

Finally, two architectures are presented to reproduce ISS behaviour seen in secure, avoidant and ambivalent types of attachment. The first of these is similar in structure to the proceeding architecture, although there is an additional avoid-pain goal activator, along with a mutual inhibition arbitration scheme. Secure attachment occurs in infants who have had a history of close physical contact predisposing the avoid-pain subsystem to low activation, whereas avoidancy occurs in infants who have experienced rejection (resulting in high levels of activation of the avoid-pain subsystem, which inhibits the security goal activator leaving exploration as the dominant goal in a process akin to ethological displacement), and ambivalence in infants with lower avoid-pain goal activation than avoidant infants (such that only the security goal is activated). The second of these architectures additionally contains a deliberative subsystem (inspired in part by Joseph LeDoux's slow-

path/fast-path theory of emotional processing (LeDoux, 1997)), and also includes a comfort sensor input (which senses the comfort that physical proximity to the caregiver provides, and stores this in memory). In this second architecture, avoidant attachment is explained in terms of an inhibitory effect that the deliberative system has based on recall of memories of a caregiver who previously provided low levels of comfort. In contrast to our work (to follow in the next chapter), the architectures did not attempt to capture disorganised forms of infant attachment.

2.7.5 Arousal Secure-Base Model

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. Hiolle et al. (2012, 2014) design an arousal-based model of an attachment secure base, in order to study the influence of a human's interaction on the robot's learning. The robot has a single goal, which is to learn the best model of its environment that it can. At the same time it aims to keep in balance an internal measure of arousal, which is defined by the degree to which the robot is memorising and recalling the current percepts, and which dictates the robot's behaviour. When the arousal level is low it continues to explore and learn, but when it is too high it seeks comfort from a human attachment figure (who 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. This drive to moderate the arousal level is motivated by the inverted U-shape theory of arousal, which states that higher-mammals attempt to maintain their average arousal level at a middle point that gives optimal physiological conditions for learning (Anderson, 1990; Baldi and Bucherelli, 2005).

The robot architecture has 3 main subsystems: for learning, arousal and action selection. The learning system comprises two neural networks: a Hopfield network used for the measure of memory recall, and a Kohonen self-organising map used for the measure of classification. The input to both networks is a binary matrix containing discretised sensor values for the robot's current percepts, and the robot's arousal level is calculated using metrics from the Hopfield network and Kohonen map, reflecting their performance. The first metric is the discrepancy between the current stimuli and the output of the Hopfield network, called surprise (since it decreases as a function of the familiarity). The second metric is the categorisation adjustment, which is the sum of the variations of the weights of the Kohonen map, corresponding to the difficulty of the categorisation of the input. Arousal is also dependent on caregiver interaction (with any received comfort reducing the arousal level). The robot then chooses actions based on this arousal level and an interval threshold: if arousal is below the threshold the robot is not stimulated enough and searches for a new stimulus, whereas if arousal is within the threshold they will continue to focus on and attempt to learn the current stimulus. If arousal exceeds the threshold interval then the

robot seeks out the caregiver for comfort.

A first experiment aimed to uncover whether the behaviour of the human caregiver changed the learning experience of the robot or not. The experimenter behaved in 2 prototypically different ways towards the robot: responding to every call for attention in a dedicated way under the first prototype, and providing only initial comfort under the second. The result was a larger mean and standard deviation of categorisation adjustment for the non-cared for robot than the cared-for robot, and a larger surprise for the non-cared for robot. The authors interpreted this as the cared-for robot having learnt its environment better than the non-cared for robot, and was explained as being the result of the robot tending to lose focus whilst calling for attention often, thus not allowing for its networks to converge and stabilise. A second experiment, conducted with public participation at the London Science Museum, aimed to assess whether the robot would elicit appropriate caregiving behaviour from humans. Two robot profiles were tested: a "needy" robot (which called for attention and looked for a human face when its arousal was high) and an "independent" robot (which required assistance less often). The authors found that participants preferred to interact with the needy robot. Subsequently, the model was updated in Hiolle et al. (2014) to allow the robot to adapt itself along the needy/independent axis, according to responsiveness of the caregiver.

2.7.6 Arousal-Based Neural Model of Infant Attachment

We attempted a synthesis of the work in Petters (2006a) and Hiolle et al. (2012) with an arousal-based neurocognitive model (Cittern and Edalat, 2014). This model aimed to explain basic infant attachment behaviour and physiology in a simple (ISS-like) separation and reunion scenario, following episodes of learning based on secure-base exploration and attachment approach. In the model, arousal levels are driven by a measure of novelty (proposed to involve the perirhinal cortex) during exploration representing the degree to which the infant is overwhelmed by environmental perceptions, along with a safe-range distance (which, as in Petters (2006a), adapts based on experience of the ability of the caregiver to relieve attachment distress, in this case excessively high arousal levels), plus fear circuitry activation (centred on the amygdala) on retrieval of memories of previously hostile caregiving. The model was able to capture very basic elements of secure, avoidant, ambivalent and disorganised infant attachment behaviour and physiology, in infants who had previously experienced attachment episodes with caregivers with varying degrees of sensitivity and responsiveness with respect to infant arousal reduction.

2.7.7 Decision and Game Theoretic Models

Buono et al. (2006) presents game theoretic models of attachment to show how secure, avoidant and ambivalent forms of attachment might emerge as equilibrium decision choices. The authors begin by considering a single player game (i.e. a decision theoretic model) of an infant's choice as to whether to approach or avoid a caregiver who might either attend to or ignore them. The infant is assumed to have stress above some tolerable level, and the payoffs correspond to changes in stress (i.e. increases or decreases). When the model includes a guarded form of approach (corresponding to resistant and guarded forms of attachment seen in ambivalent infants) which dampens the effect of attention/rejection on decreases/increases in infant stress, the authors show mathematically how approach, guarded approach or avoidance behaviour (corresponding to the three organised forms of attachment) emerge as optimal responses to caregivers with responsiveness profiles that fall into three distinct regions (see Chapter 3 for more detail). This single player game is then extended to consider also payoffs for the caregiver given these joint-action outcomes. The authors showed how pure Nash equilibria corresponding to secure and avoidant attachment relationships emerge according to payoff configurations, and additionally determined a payoff configuration for which a mixed strategy equilibrium reflecting ambivalent attachment emerges, with the probability that the infant seeks comfort proportional to the cost of caregiving (i.e. the higher the caregiving cost the clingier the infant will be), and the probability of caregiving attention proportional to the stress the infant receives when rejected.

2.7.8 Reinforcement Learning as a Model of Attachment-Based Psychotherapy

Edalat and Lin (2014) present a computational neural model of mentalization-based psychotherapy from an attachment-theory perspective. The work is based on an existing neural model for pathways of cognitive-emotional decision making, which involves key attachment-related brain areas. Under the model, interaction between the OFC and amygdala mediates heuristic (emotional) decisions, whereas an OFC-dorsolateral prefrontal cortex loop controls deliberative (cognitive) decision making. Both of these networks are implemented with Restricted Boltzmann Machines (RBMs), with the type of decision made based on output from the two networks (and mediated by the ACC, which has a documented role in error detection). Input to these networks comes from a Hopfield network representing the hypothalamus (capturing the individual's current biologically-based needs), which categorises six basic emotions (representing needs for cognitive closure) along with a mentalization pattern (representing a need for cognition). Reinforcement learning is applied along with the notion of strong (i.e. multiply learned) patterns in order to gradually increase the need for cognition in the needs network, which in turn gradually results in a shift away from heuristic

and towards deliberative decision making.

In Cittern and Edalat (2015a,b) we presented a mathematical framework for the generation of a new Nash equilibrium within a multi-agent reinforcement learning setting. Inspired by evidence suggesting the possibility that both cognitive (Yang et al., 2013) and emotional (Yang et al., 2014) reappraisal of reward-prediction errors might be possible in the brain, we showed how convergence to a new equilibrium outcome (that has previously been determined as desirable in some wider context) could be achieved as a result of reward manipulation (here in the form of application of a simple multiplicative factor). We applied the framework to one of the games in Buono et al. (2006) describing an avoidant attachment relationship, to show how the framework could result in a more desirable secure form of attachment. In Edalat (2017a) this process was adapted in order to describe the application of Self-Attachment therapy (which is introduced in Chapters 4 and 5), with both agents partaking in the game fully internalised within the self.

3. Attachment as Free Energy Minimisation

In this chapter we explore infant attachment formation within the context of the free energy principle, showing how each of the classical organised attachment types might arise in infant agents who minimise free energy over interoceptive states while interacting with caregivers with varying responsiveness. We then consider how affective communication errors (in the form of misleading or ambiguous exteroceptive cues from the caregiver) might have an organising effect in infant agents who interact with caregivers who are inconsistent with regards to fulfilling infant attachment needs, but a disorganising effect in infants interacting with caregivers who consistently increase the infant's stress levels on approach.

3.1 The Free Energy Principle

The free energy principle is a theory of brain function proposing that the brain resists a tendency to disorder by aiming to restrict itself to a small number of physiological and sensory states that it a priori prefers to occupy (Friston, 2010). The theory argues that the only tractable way the brain can restrict itself to this small number of preferred states is by minimising a quantity called free energy, which provides an upper bound on a measure of surprise (that increases as a function of the undesirability of states that are occupied). According to the theory, action, perception and learning are all fundamentally driven by this minimisation of free energy, with the resulting process (describing loops of interaction between an agent and its environment) referred to as Active Inference.

We follow the mathematical models outlined in Friston et al. (2015) and FitzGerald et al. (2015) based on a finite discrete partially observable Markov decision process (see Appendix A.2 for further details). We thus have a finite set of W observations (or observable outcomes) $\widetilde{o} \in O = \{1, ..., W\}$, a finite set of J discrete hidden states $\widetilde{s} \in S = \{1, ..., J\}$ and a finite set of L discrete actions $\widetilde{a} \in \Omega = \{1, ..., L\}$, where $\sim$ denotes a sequence of variables over time. The generative process R defining the environment dynamics up to current time t is then:

$$R(\widetilde{o}, \widetilde{s}, \widetilde{a}) = Pr(\{o_0, ..., o_t\} = \widetilde{o}, \{s_0, ..., s_t\} = \widetilde{s}, \{a_0, ..., a_t\} = \widetilde{a}) \tag{3.1}$$

The agent is assumed to have an internal model of this generative process, called their "generative model" (i.e. an internal model of how hidden causes produce sensory data). The agent's generative model over observations $\widetilde{o} \in O$, hidden states $\widetilde{s} \in S$ and control states $\widetilde{u} \in U$ is:

$$P(\widetilde{o}, \widetilde{s}, \widetilde{u}) = Pr(\{o_0, ..., o_T\} = \widetilde{o}, \{s_0, ..., s_T\} = \widetilde{s}, \{u_0, ..., u_T\} = \widetilde{u}) \tag{3.2}$$

which (unlike the generative process) includes beliefs about future states (up to time $T > t$). Under the model, actions (a, a variable that acts on the generative process) are distinguished from control states (u, the corresponding hidden cause in the generative model). A policy $\pi \in \{1, ..., K\}$ indexes a sequence of future control states $(\widetilde{u} \mid \pi) = (u_t, ..., u_T | \pi)$, and it is assumed that the agent has an approximate posterior distribution Q over hidden and control states:

$$Q(\widetilde{s}, \widetilde{u}) = Pr(\{s_0, ..., s_T\} = \widetilde{s}, \{u_0, ..., u_T\} = \widetilde{u}) \tag{3.3}$$

which is parametrised by $(\widehat{s}, \widehat{\pi})$, where $\widehat{s} \in [0, 1]^J$ is a $J \times 1$ vector of hidden state expectations, and $\widehat{\pi} \in [0, 1]^K$ is a $K \times 1$ vector of policy expectations.

The variational free energy F is:

$$\begin{aligned} F(\widetilde{o}, \widehat{s}, \widehat{\pi}) &= \mathbb{E}_Q[-\ln P(\widetilde{o}, \widetilde{s}, \widetilde{u}) - H[Q(\widetilde{s}, \widetilde{u})] \\ &= -\ln P(\widetilde{o}) + KL[Q(\widetilde{s}, \widetilde{u}) || P(\widetilde{s}, \widetilde{u} \mid \widetilde{o})] \end{aligned} \tag{3.4}$$

where:

$$H[P(x)] = \mathbb{E}_{P(x)}[-\ln P(x)] \tag{3.5}$$

is the Shannon entropy, and:

$$KL[Q(x)||P(x)] = \mathbb{E}_{Q(x)}[\ln Q(x) - \ln P(x)] \tag{3.6}$$

is the Kullback-Leibler divergence. The agent is assumed to have a prior distribution specifying the utility (preference) of each outcome at time $\tau > t$:

$$P(o_\tau) = C_\tau \tag{3.7}$$

and the free energy principle argues that agents fundamentally aim to minimise a quantity called surprise $-\ln P(\widetilde{o})$. According to the theory, the only tractable way to do this is by

minimising free energy which gives an upper-bound on this surprise (see second rearrangement in Eq. 3.4, which upper-bounds surprise since the KL divergence term cannot be less than zero).

The following factorisation is assumed for the generative model:

$$P(\widetilde{o}, \widetilde{x} \mid \widetilde{a}) = P(\widetilde{o} \mid \widetilde{s}, A)P(\widetilde{s} \mid \widetilde{a}, B, D)P(\widetilde{u} \mid \gamma)P(\gamma|\alpha, \beta)P(A|\theta)P(B|\phi)P(D|\xi) \tag{3.8}$$

with $\widetilde{x}=\widetilde{s}, \widetilde{u}, \gamma, A, B, D$. The first factor $P(\widetilde{o} \mid \widetilde{s}, A) = P(o_0|s_0, A)P(o_1|s_1, A)...P(o_t|s_t, A)$, defining observations given hidden states, is encoded in matrix form (such that i.e. column j of A, i.e. $A_{\bullet j}$, encodes the likelihood of observations given hidden state j):

$$P(o_t = i|s_t = j, A) = A_{ij} \tag{3.9}$$

The second factor $P(\widetilde{s} \mid \widetilde{a}) = P(s_t|s_{t-1}, a_t, B)...P(s_1|s_0, a_1, B)P(s_0|D)$ defines hidden state transitions (and the initial hidden state), and is encoded in matrix form as:

$$P(s_{t+1} = i|s_t = j, u_t, B) = B(u_t)_{ij} \tag{3.10}$$

$$P(s_0 = i|D) = D_i \tag{3.11}$$

The third factor $P(\widetilde{u} \mid \gamma) = \sigma(\gamma \cdot \mathbf{Q})$ expresses beliefs about sequences of control states (i.e. policies), with σ a softmax function. Here, $\mathbf{Q}$ is a $K \times 1$ vector containing the expected negative free energy of each policy at the current time t, so that row $\mathbf{Q}(\pi)$ gives the negative free energy expected under some particular policy π:

$$\mathbf{Q}(\pi) = \sum_{\tau=t+1}^{T} \mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau)] + H[Q(s_\tau|\pi)] \tag{3.12}$$

where $Q(o_\tau, s_\tau|\pi) = P(o_\tau|s_\tau)Q(s_\tau|\pi) = \mathbb{E}_{Q(s_t)}[P(o_\tau, s_\tau|s_t, \pi)]$ is a posterior predictive distribution over future states and outcomes.
The fourth factor $P(\gamma|\alpha, \beta)$ expresses a prior over precision γ (encoding confidence in prior beliefs), which is assumed to have a gamma distribution with shape and rate parameters α and β:

$$P(\gamma|\alpha, \beta) = Gamma(\alpha, \beta) \tag{3.13}$$

The fifth factor $P(A|\theta)$ is a Dirichlet prior (with concentration parameters θ) over the multinomial distributions $A_{\bullet j}$ (encoding the likelihood of observations given hidden state

j):

$$P(A_{\bullet j} \mid \theta) = Dirichlet(\theta_{\bullet j}) \tag{3.14}$$

Similarly, the sixth factor $P(B|\phi)$ is a Dirichlet prior (with concentration parameters ϕ) over the multinomial distributions $B(u)_{\bullet j}$ encoding the likelihood of hidden states at $t+1$ given that the hidden state at time t is j:

$$P(B(u)_{\bullet j} \mid \phi(u)) = Dirichlet(\phi(u)_{\bullet j}) \tag{3.15}$$

The final factor $P(D|\xi)$ is a Dirichlet prior (with concentration parameters ξ) over the multinomial distribution encoding the initial hidden state:

$$P(D \mid \xi) = Dirichlet(\xi) \tag{3.16}$$

For the approximate posterior Q the following factorisation is assumed:

$$Q(\widetilde{x} \mid \widehat{x}) = Q(s_0|\widehat{s}_0)...Q(s_T|\widehat{s}_T)Q(u_t,...,u_T|\ \widehat{\pi})Q(\gamma|\widehat{\gamma})Q(A|\widehat{\theta})Q(B|\widehat{\phi})Q(D|\widehat{\xi}) \tag{3.17}$$

with parameters $\widehat{x}=\widehat{s}, \widehat{\pi}, \widehat{\gamma}, \widehat{\theta}, \widehat{\phi}, \widehat{\xi}$ and:

$$Q(\gamma|\widehat{\gamma}) = Gamma(\alpha, \widehat{\beta}= \alpha/\ \widehat{\gamma}) \tag{3.18}$$

$$Q(A|\widehat{\theta}) = Dirichlet(\widehat{\theta}) \tag{3.19}$$

$$Q(B|\widehat{\phi}) = Dirichlet(\widehat{\phi}) \tag{3.20}$$

$$Q(D|\widehat{\xi}) = Dirichlet(\widehat{\xi}) \tag{3.21}$$

Given these factorisations, it can be shown (see Appendix A.2) that the variational updates to $\widehat{x}$ that minimise free energy are given by:

$$\widehat{s}_t = \begin{cases} \sigma(\widehat{A} \cdot o_t + \widehat{D}) & \text{if } t = 1 \\ \sigma(\widehat{A} \cdot o_t + \widehat{B}\ (a_{t-1})\widehat{s}_{t-1}) & \text{otherwise} \end{cases} \tag{3.22}$$

$$\widehat{\pi} = \sigma(\widehat{\gamma} \cdot \mathbf{Q}) \tag{3.23}$$

$$\widehat{\gamma} = \alpha/(\beta - \mathbf{Q} \cdot \widehat{\pi}) \tag{3.24}$$

$$\widehat{\theta}_{ij} = \theta_{ij} + \sum_{t=1}^{T} o_{ti} \widehat{s}_{tj} \tag{3.25}$$

$$\widehat{\phi}\ (u)_{ij} = \phi(u)_{ij} + \sum_{t=2}^{T} [u = a_{t-1}] \cdot \widehat{s}_{ti} \widehat{s}_{t-1j} \tag{3.26}$$

$$\widehat{\xi} = \xi + \widehat{s}_1 \tag{3.27}$$

for $\widehat{A}_{ij} = \mathbb{E}_Q[\ln A_{ij}] = \psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})$, $\widehat{B}_{ij} = \mathbb{E}_Q[\ln B_{ij}] = \psi(\widehat{\phi}_{ij}) - \psi(\sum_i \widehat{\phi}_{ij})$ and $\widehat{D}_i = \mathbb{E}_Q[\ln D_i] = \psi(\widehat{\xi}_i) - \psi(\sum_i \widehat{\xi}_i)$, with ψ the digamma function, and the Iverson brackets $[\cdot]$ returning one if the expression is true and zero otherwise.

The first three of these updates (Eqs. 3.22 - 3.24) are inference updates, and are iterated until convergence (or, and as here, N times) before each step of an episode (see Algorithm A1 for full details). Briefly, following an observation the agent iterates these inference updates before selecting an action that they expect to minimise free energy (sampled from $\widehat{\pi}$), then on performance of this action the environment will transition to a new hidden state and provide the agent with a new observation. These two steps repeat until the end of the episode. The variational updates involved in perception (inference of the hidden state, Eq. 3.22) have been associated with the prefrontal cortex, while the updates involved in action (Eq. 3.23) have been linked with activity in the striatum, and the expected precision (Eq. 3.24) has been associated with dopaminergic projections from the VTA and Substantia Nigra pars Compacta (SNc) (Friston et al., 2015, Fig.1). The final three updates (Eqs. 3.25 - 3.27) are Hebbian-like learning updates with implicit learning rates determined by the amount of prior experience, and are typically performed following each length-T trial (episode) (FitzGerald et al., 2015).

3.2 Decision Theoretic Model of Attachment

Buono et al. (2006) proposes both one and two player game theoretic models of attachment (the one-player game is a decision theoretic model, which we examine here). The decision-theoretic problem with a half-go action (corresponding to the guarded request for comfort seen in ambivalent infants) is given in Fig. 3.1.

The probability that the caregiver attends is $0 \leq q \leq 1$. When the infant seeks comfort

		Caregiver's action	
		Attend	Ignore
		q	$1-q$
Infant's action	Go	1	$-s$
	Don't Go	0	0
	Half Go	h	$-t$

Figure 3.1: Decision theoretic model of infant attachment

(i.e. chooses action Go) but the caregiver ignores them, the payoff is $-s$. If the infant is stressed by this rejection then $s > 0$, whereas if they are comforted by being close to the caregiver (even though the caregiver ignores them) then $s < 0$ (in this second case it is assumed that $s > -1$, i.e. if the infant is ignored then they receive less comfort than if the caregiver attends). If the infant does not go to the caregiver for comfort (i.e. they choose Don't Go) then they receive no comfort, regardless of what the caregiver chooses to do. Finally, if the infant approaches the caregiver in a guarded fashion (Half Go) then outcomes are parametrised by h and t. It is assumed that $0 < h < 1$, i.e. comfort received from the Half-Go action is less than for the Go action, but more than for the Don't Go action. As for s, there are two cases for the sign of t: if $t < 0$ then the infant receives comfort from being around the caregiver even if the caregiver ignores them (in this case it is assumed that $-1 < s < t < 0$). If $t > 0$ then the infant is stressed by the caregiver ignoring them (in this case it is assumed that $0 < t < s$).

We consider here the cases of q for $0 < t < s$, $0 < h < 1$ and either $h > t/s$ or $h < t/s$, which allows for the emergence of the three actions (and thus three attachment types) as optimal responses to the caregiver with a known q. The expected payoffs for the infant for each action are $P_{\text{Go}} = q - (1-q)s$, $P_{\text{Don't Go}} = 0$, $P_{\text{Half Go}} = hq - (1-q)t$. For $t \leq 0$ then Go is a dominant strategy, and the infant should never choose Don't Go or Half Go, whereas for $0 < t < s$ there is no dominant strategy. In particular, when $q = 0$ then $P_{\text{Go}} = -s$, whereas when $q = 1$ then $P_{\text{Go}} = 1$, and $P_{\text{Go}} = 0$ when $q = s/(1+s)$. For the Half Go action we have that when $q = 0$ then $P_{\text{Half Go}} = -t$ whereas when $q = 1$ then $P_{\text{Half Go}} = h$, and $P_{\text{Half Go}} = 0$ when $q = t/(h+t)$. For the Don't Go action, when $q \in \{0, 1\}$ then $P_{\text{Don't Go}} = 0$. We also have that $P_{\text{Go}} = P_{\text{Half Go}}$ for $q - s(1-q) = hq - t(1-q)$ i.e. $q = (s-t)/(1+s-h-t)$. Thus, for $0 < t < s$, $0 < h < 1$ and $t/(h+t) > s/(1+s)$ (i.e. $h < t/s$), then $P_{\text{Don't Go}} > \max(P_{\text{Go}}, P_{\text{Half Go}})$ for $q \in [0, s/(1+s))$ and $P_{\text{Go}} > \max(P_{\text{Don't Go}}, P_{\text{Half Go}})$ for $q \in (s/(1+s), 1]$. On the other hand, for $0 < t < s$, $0 < h < 1$ and $t/(h+t) < s/(1+s)$ (i.e. $h > t/s$) then $P_{\text{Don't Go}} > \max(P_{\text{Go}}, P_{\text{Half Go}})$ for $q \in [0, t/(h+t))$, $P_{\text{Half Go}} > \max(P_{\text{Go}}, P_{\text{Don't Go}})$ for $q \in (t/(h+t), (s-t)/(1+s-h-t))$, and $P_{\text{Go}} > \max(P_{\text{Don't Go}}, P_{\text{Half Go}})$ for $q \in ((s-t)/(1+s-h-t), 1]$. In other words, if $h < t/s$ then for low q Don't Go is optimal, and for high q Go is optimal; whereas if $h > t/s$ then for low q Don't Go is optimal, for mid q Half Go is optimal, and for high q Go is

optimal.

We extend the analysis of Buono et al. (2006) to the more general case in which the payoff for the infant who chooses Go when the caregiver Attends is parametrised by g. Now $P_{\text{Go}} = gq - (1-q)s$, with $P_{\text{Half Go}} = hq - (1-q)t$ and $P_{\text{Don't Go}} = 0$ as before. When $q = 0$ then $P_{\text{Go}} = -s$, whereas when $q = 1$ then $P_{\text{Go}} = g$, and $P_{\text{Go}} = 0$ when $q = s/(g+s)$. As before, for the Half Go action we have that when $q = 0$ then $P_{\text{Half Go}} = -t$ whereas when $q = 1$ then $P_{\text{Half Go}} = h$, and $P_{\text{Half Go}} = 0$ when $q = t/(h+t)$. For the Don't Go action, when $q \in \{0, 1\}$ then $P_{\text{Don't Go}} = 0$. We also have that $P_{\text{Go}} = P_{\text{Half Go}}$ for $gq - s(1-q) = hq - t(1-q)$ i.e. $q = (s-t)/(g+s-h-t)$. Thus, for $0 < t < s$, $0 < h < g$ and $t/(h+t) > s/(g+s)$ (i.e. $h < gt/s$), then $P_{\text{Don't Go}} > \max(P_{\text{Go}}, P_{\text{Half Go}})$ for $q \in [0, s/(g+s))$ and $P_{\text{Go}} > \max(P_{\text{Don't Go}}, P_{\text{Half Go}})$ for $q \in (s/(g+s), 1]$. On the other hand, for $0 < t < s$, $0 < h < g$ and $t/(h+t) < s/(g+s)$ (i.e. $h > gt/s$) then $P_{\text{Don't Go}} > \max(P_{\text{Go}}, P_{\text{Half Go}})$ for $q \in [0, t/(h+t))$, $P_{\text{Half Go}} > \max(P_{\text{Go}}, P_{\text{Don't Go}})$ for $q \in (t/(h+t), (s-t)/(g+s-h-t))$, and $P_{\text{Go}} > \max(P_{\text{Don't Go}}, P_{\text{Half Go}})$ for $q \in ((s-t)/(g+s-h-t), 1]$.

3.3 Free Energy Model of Attachment

The argument of the free energy principle (that agents act, perceive and learn in order to restrict themselves to some limited number of a priori preferred states) aligns with the idea of the infant's early developing brain directing action, performing emotional appraisal and learning about the characteristics of their attachment caregiver in order to achieve homeostatic regulation. Thus, using the decision theoretic model outlined above as a starting point, we now attempt to formulate a basic model of attachment in terms of free energy minimisation, with an infant agent who has prior preferences for interoceptive observations related to low stress states. Evidence suggesting that the physiological stress response is related to subjective estimates of uncertainty (de Berker et al., 2016) also fits with our use of the free energy principle and active inference, in the sense that this framework inherently involves a drive towards the resolution of uncertainty (Friston et al., 2015).

To begin with (in order to explore the parameter space) we assume that the infant agent has a perfect generative model of the environment (i.e. they know the caregiver's value q), before later considering how the generative model might be learned as experience accumulates. As in the decision theoretic model outlined in the previous section, the value q here encapsulates the probability that (on any particular timestep) the caregiver will respond in such a way (i.e. attentively) that effectively lowers the infant's internal stress levels should they approach. We refer to this as caregiving "responsiveness" (with responsiveness increasing as a function of q).

3.3.1 Environment

We consider the environment for the infant agent in terms of the control states, observations (to begin with, corresponding only to interoceptive states relating to relative stress levels), contexts, and hidden states.

3.3.1.1 Control States

The control states $u \in U = \{U_1, U_2, U_3\}$ correspond to the actions that the infant can choose. These are Seek (U_1, corresponding to Go in the decision theoretic model), Guarded Seek (U_2, corresponding to Half-Go), and Avoid (U_3, corresponding to Don't Go). We assume that the infant must, at each timestep, choose one of these actions. Control states and allowable sequences of control states (i.e. allowable policies) are shown in Fig 3.2.

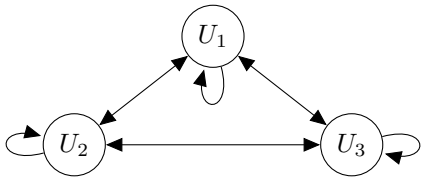

Figure 3.2: Control states and allowable sequences of control states: U_1 = Seek, U_2 = Guarded Seek, U_3 = Avoid.

3.3.1.2 Observations

To begin with, the observations for the infant agent are interoceptive observations related to its internal state (Fig. 3.3). Later on, in addition to these interoceptive observations, we will also consider exteroceptive observations emanating from the caregiver. The preference distribution over interoceptive observations corresponds to the payoffs in the decision theoretic model, which represent the amount of comfort or stress reduction (relative to the previous timestep) received by the infant when the caregiver either attends to or ignores them. In particular, solid green represents the payoff g, checkered green is h, solid red is $-s$, checkered red is $-t$, and solid blue is 0.

Figure 3.3: The set of (interoceptive) observations corresponding to payoffs in the decision theoretic model.

3.3.1.3 Caregiving Behaviours

There are two possible caregiving behaviours $X = \{Attend, Ignore\}$, encoded in the environment dynamics, which correspond to the caregiver's behaviour towards the infant. Each behaviour assigns interoceptive observations to control states. For both Attend and Ignore, control state U_3 (Avoid) maps to the blue observation, which is the internal state in which there is no change in stress reduction relative to the previous timestep. For the Attend behaviour (Fig. 3.4), control state U_1 (Seek) maps to the solid green observation (reduction in stress of g), and control state U_2 (Guarded Seek) maps to the checkered green observation (reduction in stress of h with $g > h > 0$).

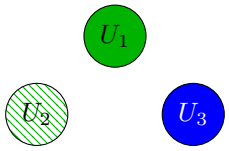

Figure 3.4: Mappings from control states to interoceptive observations for Attend.

On the other hand, when the caregiver chooses to Ignore (Fig. 3.5), control state U_1 (Seek) maps to the solid red observation (stress increase of s relative to the previous timestep), and control state U_2 (Guarded Seek) maps to the checkered red observation (stress increase of t with $0 < t < s$).

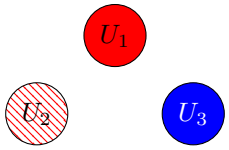

Figure 3.5: Mappings from control states to interoceptive observations for Ignore.

3.3.1.4 Hidden States

The hidden states correspond to the four control states, along with the two possible caregiving behaviours (i.e. whether the caregiver will Attend or Ignore). Although we call the set of all of these states "hidden states", only states involving blue interoceptive observations cannot be determined with certainty based on the observation (the remaining states are fully observable).

$$U_1 \quad U_2 \quad U_3 \quad \otimes \quad \text{Attend} \quad \text{Ignore}$$

Figure 3.6: The set of hidden states is given by the tensor product of control states and caregiving behaviours.

such that the six hidden states are 1: (Seek,Attend), 2: (Seek,Ignore), 3: (Guarded Seek,Attend), 4: (Guarded Seek, Ignore), 5: (Avoid, Attend), 6: (Avoid, Ignore). The hidden state transition probabilities (generative process) are given by:

$$R(s_{t+1}|s_t, u_t) = G(u_t) \tag{3.28}$$

where:

$$G(u_t = U_i \in U) = M(i, 0, L, L) \otimes \begin{bmatrix} q & q \\ (1-q) & (1-q) \end{bmatrix} \tag{3.29}$$

with $M(i, j, m, n) \in \{j, 1\}^{m \times n}$ is the $m \times n$ matrix with all elements in row i equal to 1, and all other elements equal to j; and $\otimes$ the Kronecker tensor product; so that (for example):

$$\begin{aligned} G(u_t = U_2) &= \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \otimes \begin{bmatrix} q & q \\ (1-q) & (1-q) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ q & q & q & q & q & q \\ (1-q) & (1-q) & (1-q) & (1-q) & (1-q) & (1-q) \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned} \tag{3.30}$$

meaning that all hidden states can be entered into or exited from within the time horizon.

For reasons of simplicity, we consider all hidden states to represent states for which the infant's stress levels are (arbitrarily) above some tolerable threshold (i.e. states in which their attachment system is active). We don't explicitly consider the return to a baseline state (in which the infant's attachment system is deactivated), since the addition of such a state would require us to define additional transition probabilities from the current hidden states. Empirical studies measuring cortisol (Spangler and Grossmann, 1993; Hertsgaard et al., 1995; Spangler, 1998) and heart rate (Donovan and Leavitt, 1985; Spangler and Grossmann, 1993; Zelenko et al., 2005; Hill-Soderlund et al., 2008; Smith et al., 2016) during the strange situation experiment (i.e. with controlled high caregiving responsiveness) consistently show a quicker return to baseline for secure infants on final reunion in the strange situation, however data on relative times to return to baseline for avoidant, ambivalent and disorganised infants is (on the whole) currently inconclusive, and moreover there is currently

no such data for interactions in which caregiver responsiveness is uncontrolled.

3.3.1.5 Initial Hidden State Distribution

We assume that the infant starts in the hidden state involving the Avoid control state (U_3) representing physical and/or emotional distance from the caregiver, and that initial caregiving behaviour is determined by the probability q governing their overall responsiveness. The actual initial hidden state distribution (generative process) is thus given by:

$$R(s_0) = \sigma([0 \;\; 0 \;\; 1] \otimes [q \;\; (1-q)])^\top \tag{3.31}$$

3.3.1.6 Outcomes

The set of all possible outcomes (total W) for the infant is given by the tensor product of control states and interoceptive observations:

Figure 3.7: The set of outcomes is given by the tensor product of control states and interoceptive observations.

The (generative process) distribution of outcomes (rows) given hidden states (columns) is given by:

$$R(o_t|s_t) = \begin{bmatrix} O_1 \\ O_2 \\ O_3 \end{bmatrix} \tag{3.32}$$

with:

$$O_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \; O_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \; O_3 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 1 & 1 \end{bmatrix} \tag{3.33}$$

and the remaining elements of $R(o_t|s_t)$ are zero. The infant's outcome preferences are given by:

$$P(o_\tau) = C = \sigma(\{1\}^{1\times L} \otimes [g \quad h \quad -s \quad -t \quad 0])^\top \tag{3.34}$$

i.e. the infant is assumed to be indifferent with respect to control states, but not with respect to interoceptive observations. We can argue for these priors from an evolutionary perspective, under which (given the negative effects on both brain and body) preference for prolonged and chronic states of stress would be deselected for.

3.4 Simulations

We begin by performing free energy minimisation in an infant agent that interacts with environments defined by various values of q: in each case, the infant is assumed to have a perfect generative model of the environment (i.e. they know the actual value q), and no generative model learning occurs. Then we consider agents that start with a generative model reflecting no knowledge of caregiver responsiveness and show that, by additionally performing learning on their generative model, distinct behavioural policies corresponding to secure, avoidant and ambivalent attachment will emerge from this common starting point. In all simulations we set prior gamma parameters $\alpha = 250$ and $\beta = 1$, and number of variational iterations $N = 4$, with presented results averaged over 100 independent repetitions. Our implementation is built using the SPM12 toolkit (SPM12, 2014), which contains a number of routines for the underlying discrete free energy minimisation scheme described here.

3.4.1 Parameter Space Exploration for Perfect Generative Models

We begin by exploring the parameter space for relative stress changes resulting from states involving infant seeking and caregiver attention (with preference parameter g), infant seeking and caregiving ignoring (s), infant guarded seeking and caregiver attention (h), infant guarded seeking and caregiver ignoring (t), and infant avoidance irrespective of caregiving behaviour (preference 0). Our aim is to find configurations for these interoceptive observation preference parameters that result in distinct regions of q values corresponding to the three organised forms of attachment (secure, ambivalent and avoidant), with the degree of attachment organisation measured by the proportion that the corresponding action (Seek, Guarded Seek, Avoid) is chosen by the infant in each case.

In order to conduct this initial preference parameter analysis we assume that the infant has a perfect generative model B of the generative process G governing hidden state transitions. The generative model is:

$$\widehat{B}\,(u_t) \approx G(u_t) = R(s_{t+1}|s_t, u_t) \tag{3.35}$$

which is achieved with the following Dirichlet concentration parameters:

$$\phi(u_t) = G(u_t) \circ \{1000\}^{J\times J} + \{\epsilon\}^{J\times J} \tag{3.36}$$

with $A \circ B$ the Hadamard (element-wise) matrix product (i.e. $(A \circ B)_{ij} = A_{ij}B_{ij}$) and $\epsilon = 10^{-10}$ a small positive number (added since Dirichlet parameters must be greater than zero). Similarly, we begin by assuming that the infant's model D of the true initial state distribution is also accurate:

$$\widehat{D} \approx R(s_0) \tag{3.37}$$

according to parameters:

$$\xi = R(s_0) \circ \{1000\}^{J\times 1} + \{\epsilon\}^{J\times 1} \tag{3.38}$$

We also assume that the infant has a perfect generative model A of the generative process for interoceptive observations given hidden states:

$$\widehat{A} \approx R(o_t|s_t) \tag{3.39}$$

parametrised by:

$$\theta = R(o_t|s_t) \circ \{1000\}^{W\times J} + \{\epsilon\}^{W\times J} \tag{3.40}$$

As we discussed previously, in the single-step decision theoretic model of attachment the parameter configuration $h > gt/s$ results in an interval of q values $(t/(h+t), (s-t)/(g+s-h-t))$ such that the Half-Go (Guarded Seek) action is optimal, whereas for $h < gt/s$ there is no such interval. Recall also that organised (i.e. Secure, Ambivalent and Avoidant) forms of attachment are characterised by coordinated behaviours aimed at achieving either proximity or distance from the caregiver in response to attachment need, compared to disorganised forms of attachment which is characterised by (amongst other

things) simultaneous or sequential conflicting/contradictory behaviours. Our aim here is thus to examine how the infant chooses to behave when they minimise free energy, and whether similar bounds (over internal infant stress response and caregiving behaviour) apply in terms of resulting in sequentially consistent infant behaviour corresponding to the three organised forms of attachment.

Accordingly, we ran simulations for nine equally spaced configurations of responsiveness parameter $q \in \{0.1, 0.2, ..., 0.9\}$, for $h \in \{0.05, 0.5, 1, 1.5, 1.95\}$, $g = 2$, and varying value-pairs of $(s, t) \in \{(0.1, 0.05),\ (0.5, 0.075),\ (0.9, 0.225),\ (1.3, 0.455),\ (1.7, 0.765),\ (2.1, 1.155),\ (2.5, 1.625),\ (2.9, 2.175),\ (3.3, 2.805)\}$. Values of t were chosen such that, for s increasing in equal increments of 0.4 (from 0.1 to 3.3), (s, t) pairs have values of gt/s that increase in increments of 0.2 (from 0.1 to 1.7), which allows us to examine infant behaviour in response to increasing $h - (gt/s)$ (for some fixed value of h). Simulation results are for a single iteration of depth $T = 4$, averaged over 100 repetitions.

Fig. 3.8 shows the mean proportion of selection for Seek, Guarded Seek and Avoid during iterations (averaged over the repetitions), for $g = 2$ and varying these values of q, h, s and t. For fixed h, increasing q increases the proportion of selections of Seek (for all values of gt/s), whereas decreasing q increases the proportion of selections of Avoid (for high gt/s). Guarded Seek selection increases with increasing h (and is highest for lower-mid values of q). Note from the top row of plots (for which $\forall(s, t) : h = 0.05 < gt/s$) that Guarded Seek is chosen (in proportions up to 0.97 for $q = 0.2$, $s = 0.9$ and $t = 0.225$) relatively frequently even when $h < gt/s$, although (for fixed q, s and t) the highest selection proportions for this action are focused in regions for which $h > gt/s$ and increase as h increases. The key point to note is that there are many parameter configurations for which highly consistent sequential selection of the three actions (corresponding to the three organised attachment types) emerges as q (caregiver responsiveness) is varied.

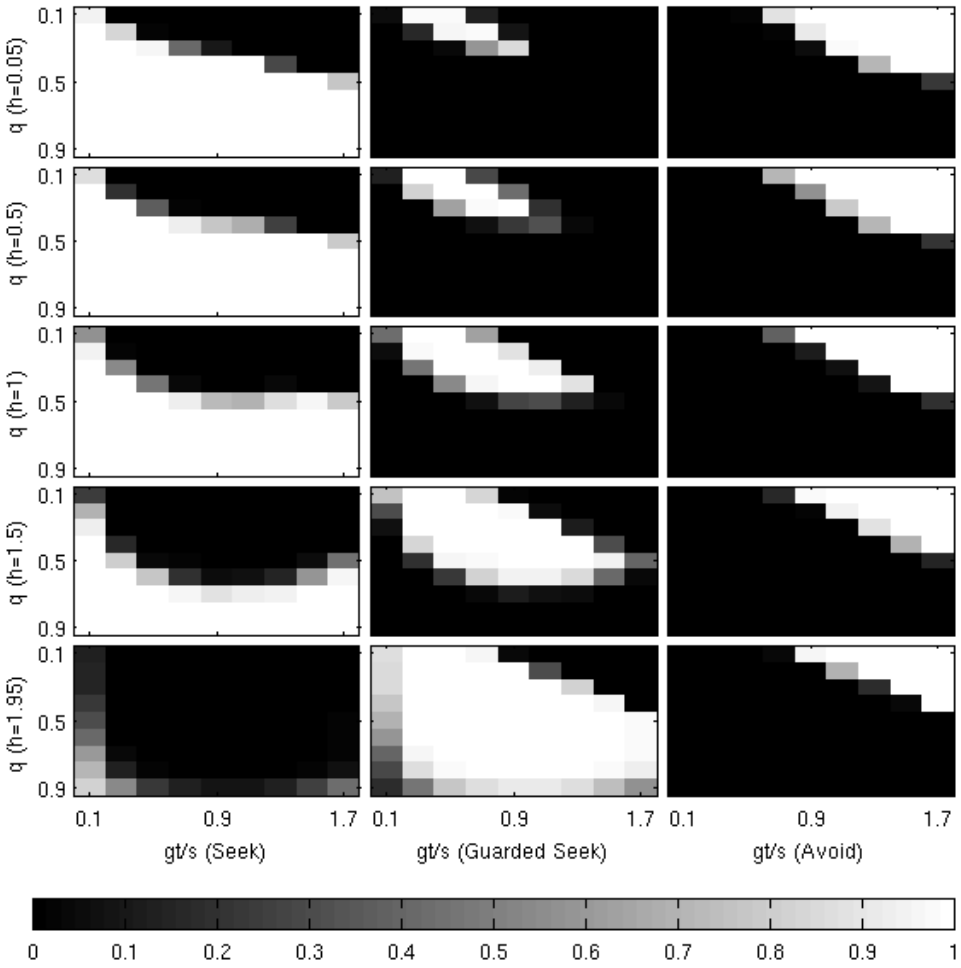

Figure 3.8: Heatmap of action selection proportions (black=0, white=1) per 4-step iteration of free energy minimisation (averaged over repetitions) for Seek (**Left column**), Guarded Seek (**Middle column**) and Avoid (**Right column**), for varying values of gt/s (**x-axis**) and q (**y-axis**). Proportions are shown (from top to bottom) for $h = 0.05$, $h = 0.5$, $h = 1$, $h = 1.5$ and $h = 1.95$.

Based on the above, in the simulations that follow (involving generative model learning) we choose parameters $g = 2$, $h = 0.75$, $t = 0.9$ and $s = 2$ for the agent's interoceptive observation preference distribution. In our free energy model, for high q conditions (approaching $q = 0.9$), the Seek action (which has the highest mean expected negative free energy) is preferentially and increasingly selected as free energy minimisation progresses (Fig. 3.9). On the other hand, for mid q conditions ($0.3 \leq q \leq 0.4$) Guarded Seek has the highest mean expected negative free energy on the final step and is preferentially and increasingly selected

over steps, whereas for low q conditions (approaching $q = 0.1$) Avoid has the highest mean expected negative free energy on the final step and is preferentially and increasingly selected as free energy minimisation progresses. Thus, these parameters result in three intervals of q for which infants consistently sequentially choose behaviours corresponding to the three organised forms of attachment. Precision parameter α controls the gradient of the curves and thus the extent to which attachment is organised, such that increasing α (i.e. increasing prior expected precision) increases the extent to which Seek is chosen for $q \geq 0.5$, Guarded Seek is chosen for $0.3 \leq q \leq 0.4$, and Avoid is chosen for $q \leq 0.2$.

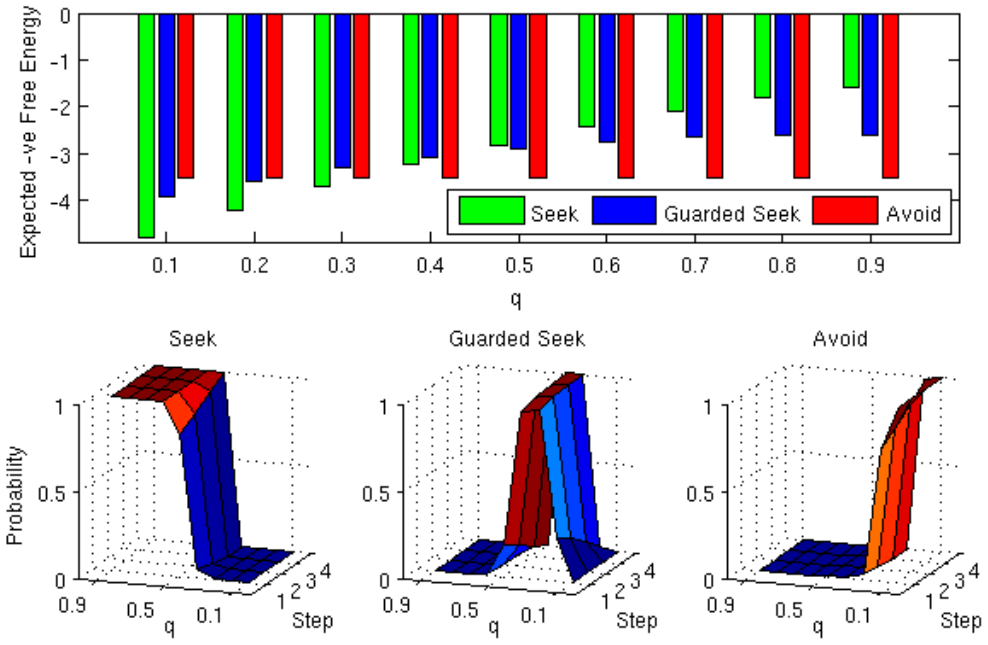

Figure 3.9: **Top**: mean (over repetitions) expected negative free energies for Seek, Guarded Seek and Avoid on final step of free energy minimisation (y-axis), for different values of q (x-axis). **Bottom**: mean action selection probabilities (z-axis) on each episode step of free energy minimisation (x-axis) for different values of q (y-axis) (**Left**: Seek, **Middle**: Guarded Seek, **Right**: Avoid).

3.4.2 Learning a Generative Model Encapsulating Caregiver Responsiveness

Up to now we have considered how infants behave when they minimise free energy over interoceptive observations according to a generative model that perfectly encapsulates the probability that the caregiver will attentively respond to their attachment needs. In reality, however, infants are not born knowing the particular responsiveness characteristics of

their caregiver, but must instead learn this over many repeated attachment interactions. Thus, we now consider an infant who learns the parameters of their generative model (with updates that minimise free energy). We pair the same infant with three different environments (i.e. types of caregivers): highly responsive ($q = 0.9$), which corresponds to an attentive caregiver that mostly attends to the infant's requests for attachment interaction; inconsistently responsive ($q = 0.4$), which corresponds to an inconsistent caregiver; and unresponsive ($q = 0.1$) which corresponds to a negligent caregiver that mostly ignores the infant's requests for attachment comfort, to see whether the distinct, organised forms of attachment emerge. In particular, we assume that each infant starts with a generative model with prior distributions that are uniform with respect to caregiving behaviour, in order to consider how the infant's model and preferred behavioural policies might adapt accordingly as experience accrues.

The prior distribution over hidden state transitions is given by the following matrix of Dirichlet prior concentration parameters [1]:

$$\phi(u_t = U_i \in U) = M(i, \epsilon, L, L) \otimes \{1\}^{J/L \times J/L} \tag{3.41}$$

where $\epsilon = 10^{-10}$ is a small positive number. Similarly, we assume that the initial hidden state distribution is also uniform with respect to caregiving behaviour (and also initial control state):

$$\xi = \{1\}^{J \times 1} \tag{3.42}$$

Since they are uniform with respect to caregiving behaviour, these parameters ϕ and ξ result in prior initial hidden state and hidden state transition distributions equivalent to the (uncertain) expectation of a caregiver with responsiveness $q = 0.5$ (Fig. 3.10), which (as we have seen previously) induces Seek behaviour in the infant for our chosen interoceptive observation preference and precision prior parameters. The fact that these priors (flat with respect to responsiveness) result in an initial tendency in the infant towards Seek behaviour is broadly consistent with the tenets of attachment theory, in that although infants are assumed to have no prior knowledge with respect to the effectiveness of their particular caregiver as an attachment figure, they are (according to Bowlby) nonetheless (genetically)

[1] We only consider here convergence to secure and avoidant forms of attachment for $q = 0.9$ and $q = 0.1$ respectively, for prior hidden state concentration parameters of 1. Convergence to these organised forms of attachment can also occur for lower/higher values of q (e.g. $q = 0.7$ for secure and $q = 0.2$ for avoidant, according to the regions outlined in the previous section), however we note that the convergence point is more organised for these less extreme values when the prior concentration parameters are increased (which decreases the implicit learning rate, and increases the amount of initial exploratory behaviour in the avoidant case).

predisposed to seek out an attachment relationship with them.

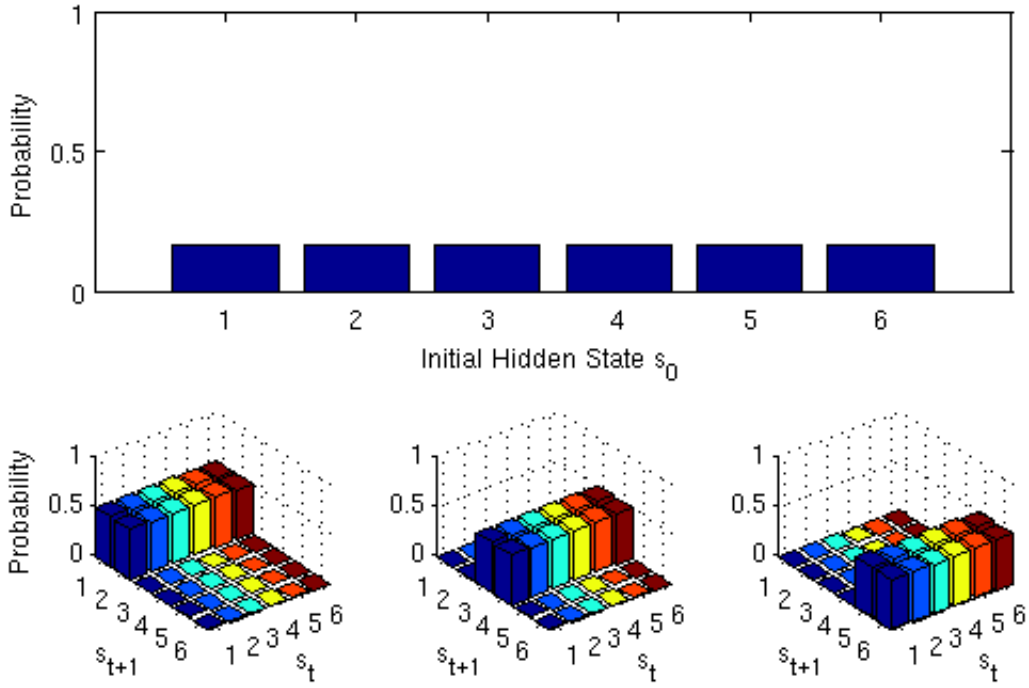

Figure 3.10: Prior initial hidden state distribution (**Top**), and prior hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ for control states Seek (**Bottom Left**), Guarded Seek (**Bottom Middle**) and Avoid (**Bottom Right**), for all agents. For the hidden state transition distributions, the x (lower right) axis gives current hidden state s_t, the y (lower left) axis gives the next hidden state s_{t+1}, and the z-axis gives the associated probability.

As in the previous section, we assume that the infant's generative model of outcomes given hidden states is a priori accurate (Eq. 3.39) so that, given a hidden state (which consists of a control state and caregiving behaviour), the infant knows the corresponding outcome (which consists of a control state and interoceptive observation) with certainty. These priors represent prior knowledge in the infant that, under states in which their attachment system is active, seeking out the caregiver might result in either an increase or decrease in their stress level relative to the previous timestep (and that this increase or decrease can be reduced in magnitude with guarded or resistant behaviour), whereas avoidance of the caregiver will result in no (externally induced) relative change to their stress level. We argue here that it is reasonable to assume the existence of such (perhaps evolutionarily ingrained) priors for our initial modelling efforts. Since we are primarily concerned with the learning of the hidden state transition distribution (which encapsulates caregiving

responsiveness), in the simulations that follow we do not perform learning on these observation model parameters (although the results presented here also apply to the case in which these observation model parameters are additionally updated, given sufficiently large concentration parameter priors).

In what follows we present simulation results for infants that all start with these same prior parameters, but differ with respect to the type of caregiver that they are exposed to (either highly responsive, inconsistent, or unresponsive). All results are averaged over 100 repetitions of 500 iterations (where each iteration consists of an individual episode of free energy minimisation and learning with respect to a process of depth 4 timesteps).

3.4.2.1 Highly Responsive Caregiver

We begin by considering an environment for which $q = 0.9$, i.e. an environment representing a caregiver who will (with relatively high probability) attend to the infant during high-stress states in which their attachment system is activated (Fig. 3.11). Since the infant's hidden state transition generative model is initially flat with respect to caregiver behaviour, they have an initial tendency to Seek out the caregiver during these high stress states,with this preference sustaining over iterations. The mean number of distinct actions chosen per iteration (a measure of organisation with respect to attachment behavioural strategy) drops slightly to approximately 1.3, suggesting that (on average) these infant prefer policies instructing sequentially-consistent Seek behaviour, while expected precision on the final step of each iteration rises as the infant comes to learn a more accurate generative model (which results in a higher probability for prediction of reaching states associated with the preferred interoceptive observation as a result of Seek behaviour).

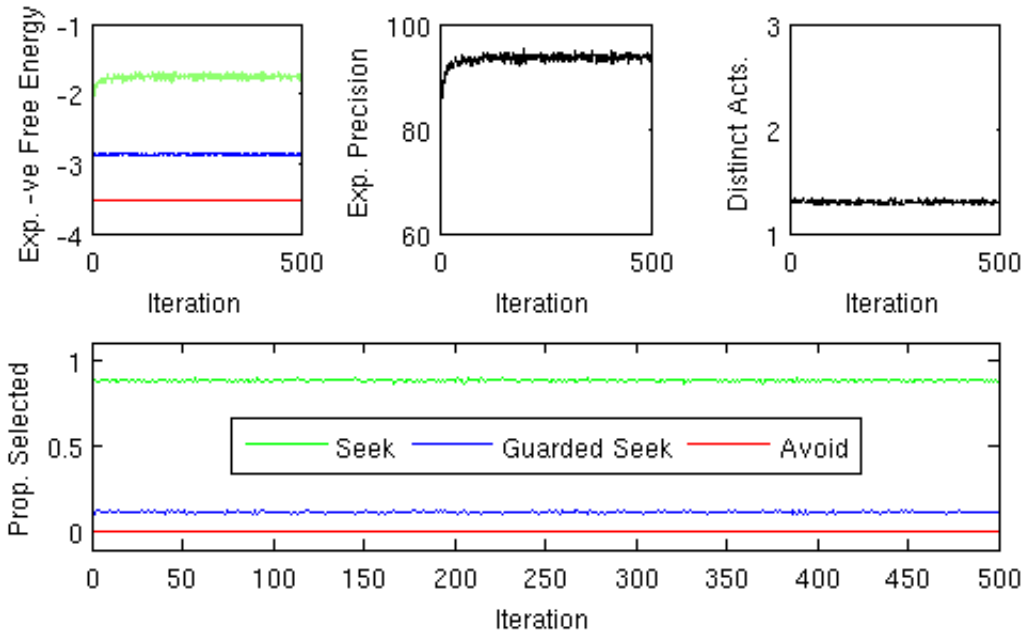

Figure 3.11: **Top Left**: Mean (over repetitions) expected negative free energies for the three actions on the final step of each iteration, for an infant paired with a high-q (responsive) caregiver. **Top Middle**: Mean (over repetitions) final-step expected precision. **Top Right**: Mean (over repetitions) number of distinct actions chosen per iteration. **Bottom**: Mean (over repetitions) proportion each action was chosen during each iteration.

Examining the final hidden state transition distribution in the infant's generative model (Fig. 3.12) we see that they have learned fairly accurate distributions for the Seek control state (particularly so for transitions from previous hidden states involving Seek or Avoid control states), and relatively accurate (compared to the prior distribution) transition probabilities for the less frequently selected Guarded Seek control state. It is worth noting that transition probabilities amongst hidden states involving Avoid have not been learned accurately, which is due to the fact that the interoceptive observation received is the same in hidden states involving the Avoid control state and both caregiver Attend and Ignore behaviour, however this is not a barrier to the formation of organised secure attachment.

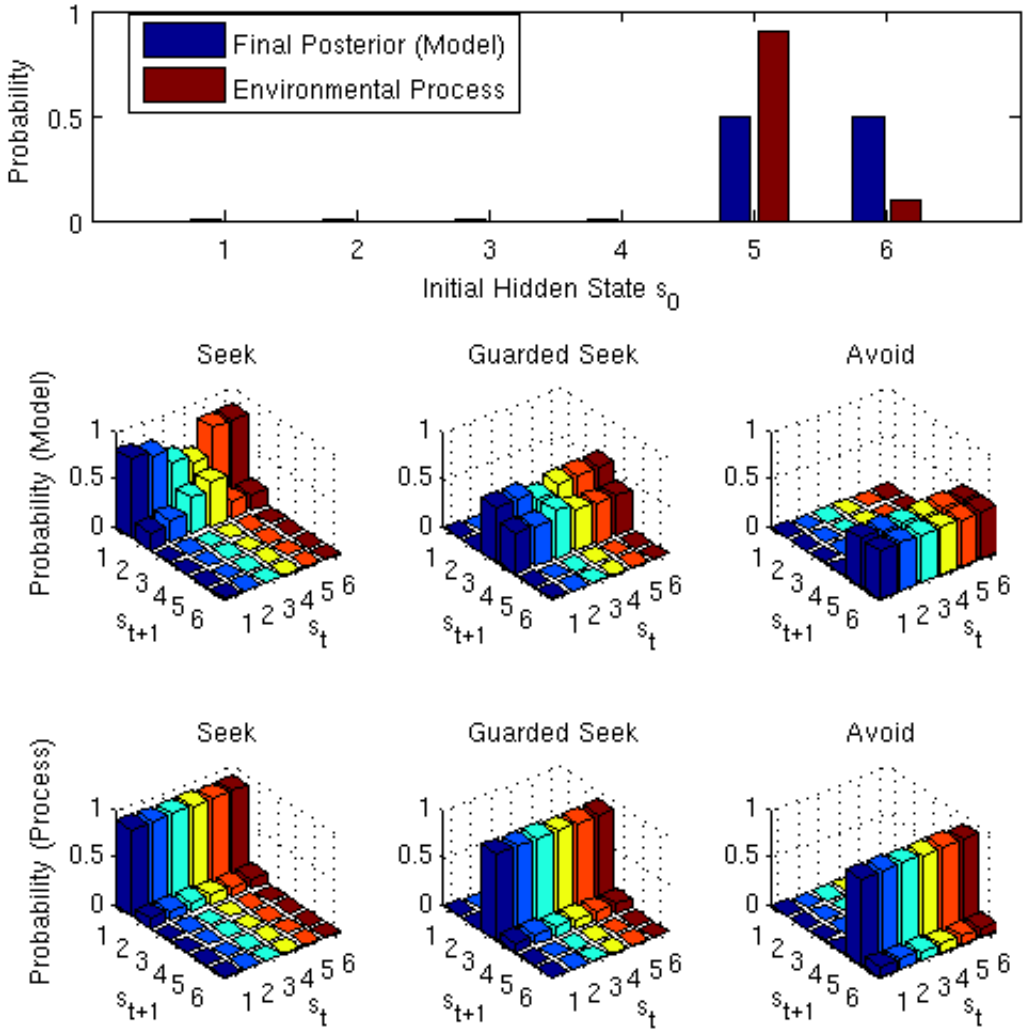

Figure 3.12: Mean (over repetitions) hidden state transition distributions for an agent interacting with a high-q (responsive) caregiver, for control states Seek (**Left**), Guarded Seek (**Middle**) and Avoid (**Right**). **Top**: Mean (over repetitions) final (posterior model) and actual (process) initial hidden state transition distributions. **Middle**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by the agent (generative model). **Bottom**: The actual hidden state transition distributions (generative process). The x (bottom right) axis gives current hidden state s_t, y (bottom left) axis the transitory hidden state s_{t+1}, and z-axis the probability.

3.4.2.2 Inconsistent Caregiver

We now consider the same infant paired instead with an inconsistent caregiver (with $q = 0.4$), for which we previously showed a strong tendency towards organised Guarded Seek (i.e. ambivalent) behaviour when the infant has a perfect generative model of hidden state transitions (i.e. caregiving behaviour). Since the infant's hidden state transition generative model is initially flat with respect to caregiver responsiveness, they have an initial tendency to Seek out the caregiver during these high stress states. However, this initial Seek behaviour leads to the preferred interoceptive observation (highest stress reduction) with lower probability than the prior model suggests, resulting in exploratory behaviour. As the transition model is gradually accordingly updated, the infant increasingly comes to prefer policies involving sequential Guarded Seek behaviour, with this ambivalent behaviour becoming relatively more organised as iterations progress (Fig. 3.13).

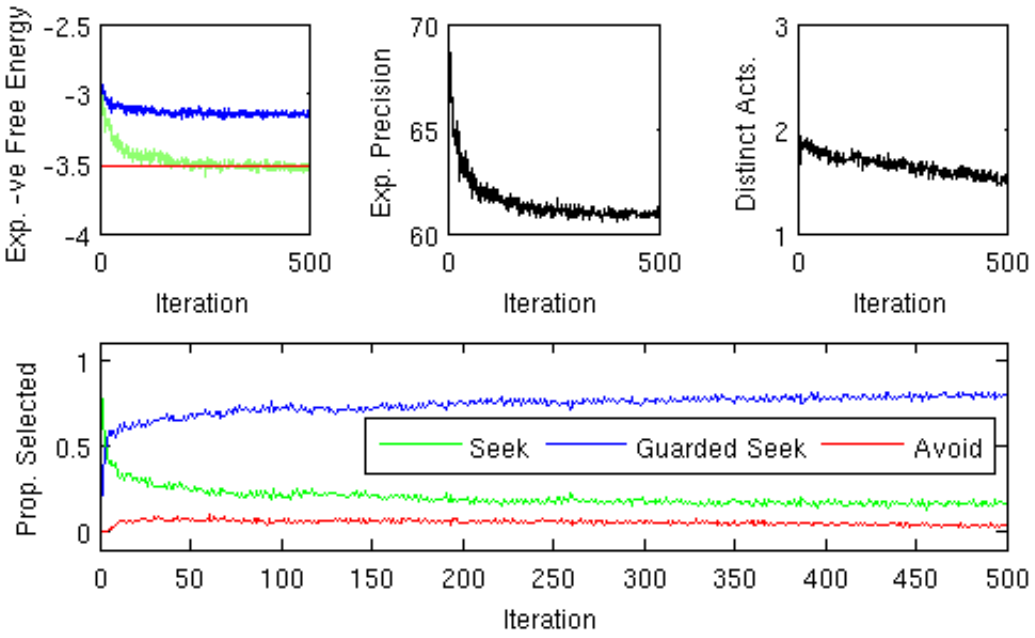

Figure 3.13: **Top Left**: Mean (over repetitions) expected negative free energies for the three actions on the final step of each iteration, for an infant paired with a mid-q (inconsistent) caregiver. **Top Middle**: Mean (over repetitions) final-step expected precision. **Top Right**: Mean (over repetitions) number of distinct actions chosen per iteration. **Bottom**: Mean (over repetitions) proportion each action was chosen during each iteration.

The final hidden state transition distribution in the infant's generative model (Fig. 3.14) has become relatively accurate for all transitions involving control states other than Avoid (as for the infant paired with the responsive caregiver, this is due to an inability to infer these hidden states accurately).

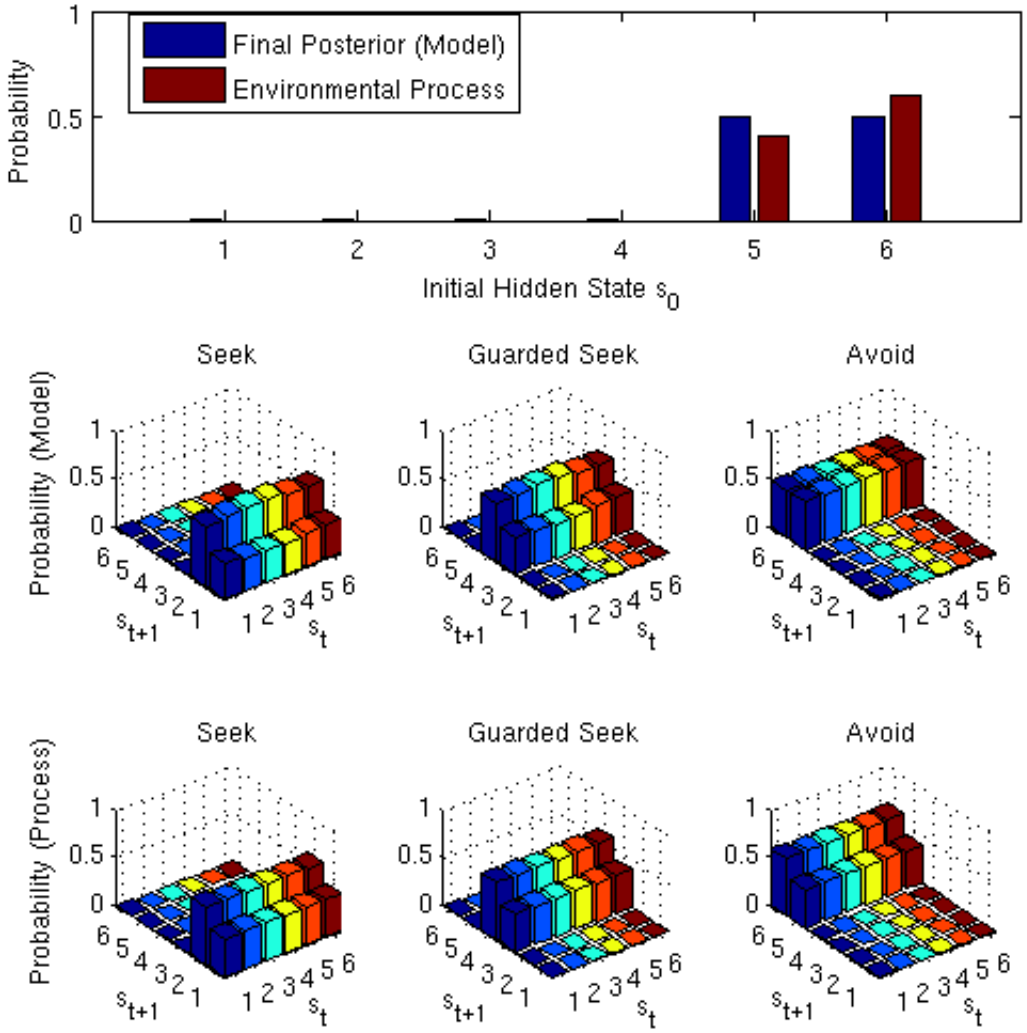

Figure 3.14: Mean (over repetitions) hidden state transition distributions for an agent interacting with a mid-q (inconsistent) caregiver, for control states Seek (**Left**), Guarded Seek (**Middle**) and Avoid (**Right**). **Top**: Mean (over repetitions) final (posterior model) and actual (process) initial hidden state transition distributions. **Middle**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by the agent (generative model). **Bottom**: The actual hidden state transition distributions (generative process). The x (bottom right) axis gives current hidden state s_t, y (bottom left) axis the transitory hidden state s_{t+1}, and z-axis the probability.

3.4.2.3 Unresponsive Caregiver

Finally we consider a consistently unresponsive caregiver (with $q = 0.1$), for which we previously showed a strong tendency towards organised Avoid behaviour when the infant has a perfect generative model of hidden state transitions (Fig. 3.15). Again, due to their flat prior hidden state distributions, the infant has an initial tendency to Seek out the caregiver, however they quickly come to prefer policies involving a majority of Avoid actions, with this avoidance behaviour becoming highly consistent and organised. We note that this transition from preference for policies involving Seek behaviour to policies involving Avoid behaviour tends to occur via a period during which Guarded Seek behaviour is preferred: this is a model prediction that can be tested empirically.

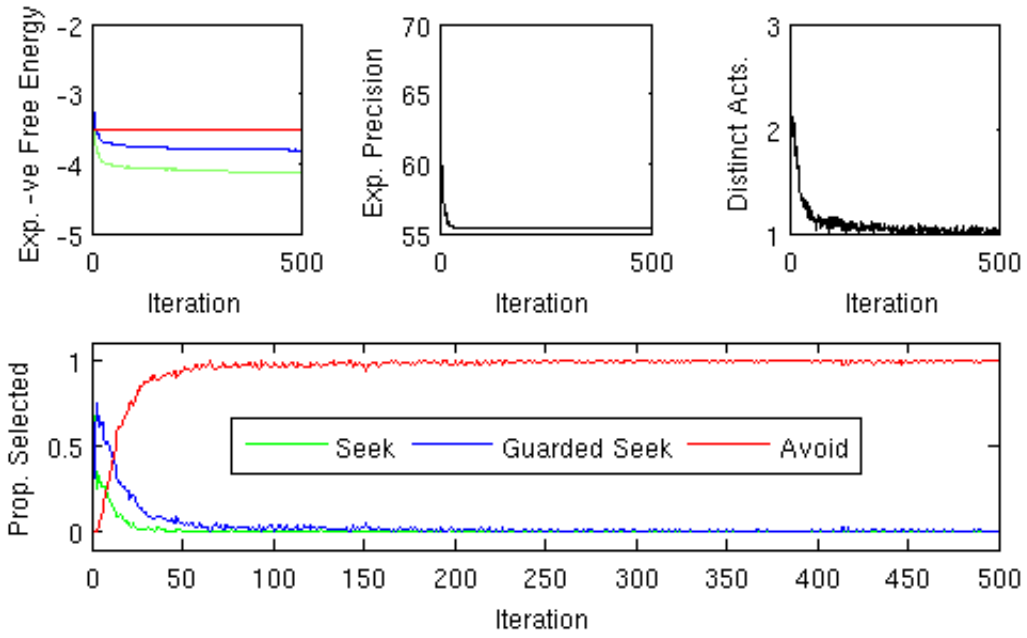

Figure 3.15: **Top Left**: Mean (over repetitions) expected negative free energies for the three actions on the final step of each iteration, for an infant paired with a low-q (unresponsive) caregiver. **Top Middle**: Mean (over repetitions) final-step expected precision. **Top Right**: Mean (over repetitions) number of distinct actions chosen per iteration. **Bottom**: Mean (over repetitions) proportion each action was chosen during each iteration.

The infant's final hidden state transition distribution has become relatively accurate for all hidden state transitions dependent on control states other than Avoid (Fig. 3.16).

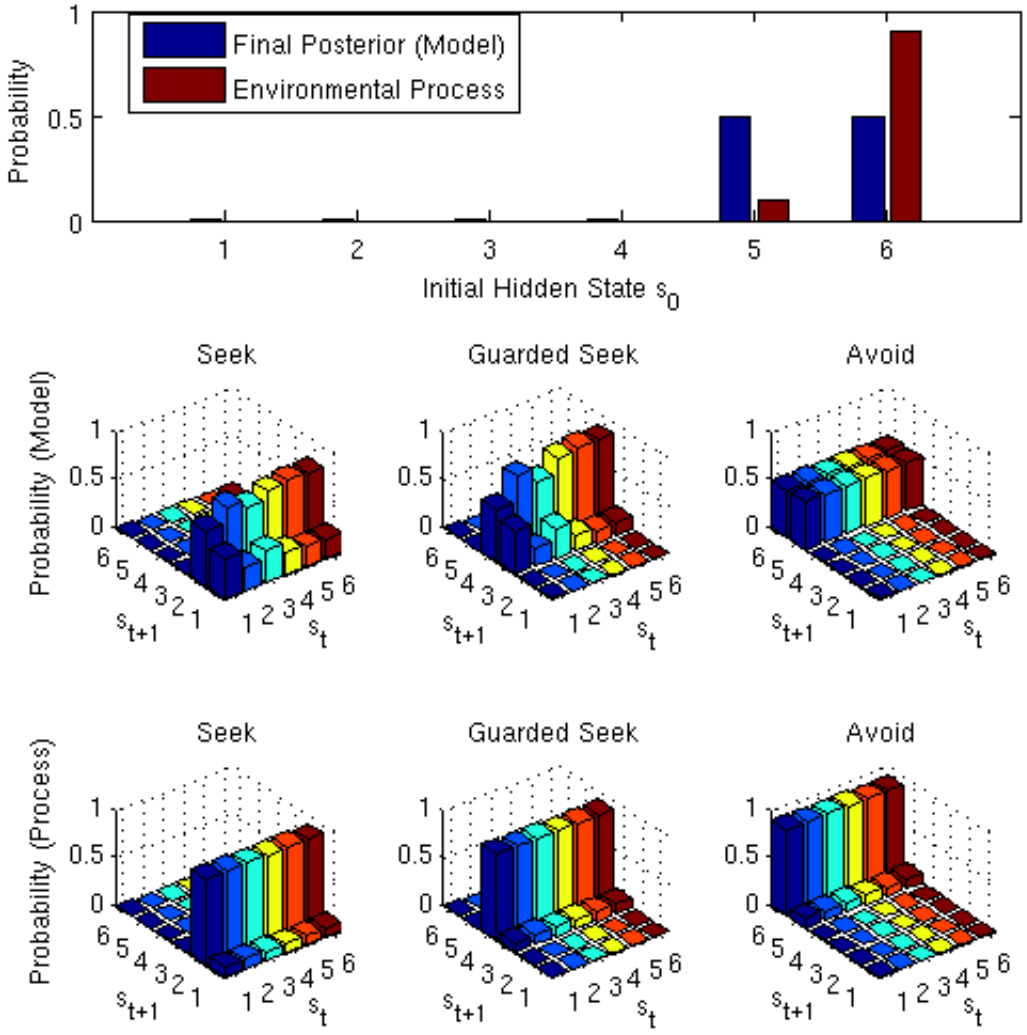

Figure 3.16: Mean (over repetitions) hidden state transition distributions for an agent interacting with a low-q (unresponsive) caregiver, for control states Seek (**Left**), Guarded Seek (**Middle**) and Avoid (**Right**). **Top**: Mean (over repetitions) final (posterior model) and actual (process) initial hidden state transition distributions. **Middle**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by the agent (generative model). **Bottom**: The actual hidden state transition distributions (generative process). The x (bottom right) axis gives current hidden state s_t, y (bottom left) axis the transitory hidden state s_{t+1}, and z-axis the probability.

3.4.3 Exteroceptive Observations, Ambivalence and Disorganisation

In the preceding simulations we have considered an infant who behaves, perceives and learns in order to minimise free energy based on interoceptive observations relating to changes in internal stress levels. In particular, we have shown that minimisation of free energy over these interoceptive observations is sufficient for the emergence of behaviour resembling the organised attachment types, in infants who differ only with respect to the environment (i.e. hidden state transition dynamics) that they interact with. However, as stated by Bowlby (1969), mother-infant attachment communications are:

> ...accompanied by the strongest of feelings and emotions, happy or the reverse, (and occur within the context of) facial expression, posture, tone of voice, physiological changes, tempo of movement, and incipient action

According to Bowlby (as quoted by Schore (Schore, 2012, p.168)):

> Emotion is nonverbal communication of basic but very powerful attitudes in mind and potential action

Thus, in addition to these interoceptive observations, we now also consider exteroceptive observations representing emotional cues from the caregiver. In particular, we consider how cues that conform to an infant's exteroceptive observation priors with respect to subsequent caregiving behaviour might have an organising effect for either highly responsive (leading to secure) or unresponsive (avoidant) caregivers. On the other hand, we show how caregivers who provide cues that are either ambiguous or misleading for the infant with respect to subsequent behaviour might lead to ambivalent attachment in the case where the caregiver is inconsistent, and how caregivers who consistently increase infant stress on approach but provide misleading cues might have a disorganising effect on infant attachment formation.

As detailed in Section 2.3.1.2, disrupted (atypical) affective communication according to the AMBIANCE scale (Bronfman et al., 1999; Safyer, 2013, Appendix G) has been linked to caregivers of both ambivalent and disorganised infants. For our initial modelling efforts we focus on the ACE dimension of this scale, which examines the quality of communication (encompassing verbal communication, along with emotional communication in the form of tone of voice, facial expressions, gestures and mood presentation) between the infant and caregiver. The ACE dimension has three components (see Appendix A and (Safyer, 2013, Appendix G) for further details):

1. Contradictory signalling to the infant. This category of communications includes the display of incongruent physical behaviours, and cases in which the caregiver's voice tone is incongruent

with physical responses, or the voice content or voice tone is incongruent with the accompanying facial expression.

2. Failure to initiate responsive behaviour to an infant cue. This category consists of absent responses from the caregiver to clear infant signals directed towards them.

3. Inappropriate responses to infant signals or needs.

In samples that did not include ambivalent infants: ACE was the only AMBIANCE dimension to individually differentiate organised and disorganised infants in Lyons-Ruth et al. (1999), and these behaviours were found at an elevated rate in caregivers of disorganised compared to organised infants in Madigan et al. (2006). In samples containing all four major attachment classifications: prevalence of ACEs was found to be elevated and highest in caregivers of ambivalent infants in Safyer (2013), and overall disrupted affective communication (including ACE behaviours) was found to be highest in caregivers of ambivalent and disorganised infants compared to secure and avoidant in Grienenberger et al. (2005). Elevated disrupted affective communication was also found in caregivers of disorganised infants in a sample categorised as either organised or disorganised (Goldberg et al., 2003).

In light of the above, we extend the hidden state transition model to incorporate both current and subsequent (proceeding timestep) caregiving behaviour, i.e. $X = \{(Attend, Attend), (Attend, Ignore), (Ignore, Attend), (Ignore, Ignore)\}$ so that we now have $J = 12$ total hidden states (i.e. all pairs of current and subsequent caregiving behaviour times all infant control states). We call these second-order hidden state representations. The hidden state transition generative process now becomes:

$$G(u_t = U_i \in U) = M(i, 0, L, L) \otimes \begin{bmatrix} q & 0 & q & 0 \\ (1-q) & 0 & (1-q) & 0 \\ 0 & q & 0 & q \\ 0 & (1-q) & 0 & (1-q) \end{bmatrix} \tag{3.43}$$

and, as before, the prior parameters for the transition generative model are uniform with respect to caregiving behaviour:

$$\phi(u_t = U_i \in U) = M(i, \epsilon, L, L) \otimes \{1\}^{J/L \times J/L} \tag{3.44}$$

Amongst other items, the ACE dimension of the AMBIANCE coding instrument includes both caregiver communications that are misleading or ambiguous (contradictory) with respect to subsequent infant-directed behaviour. In order to focus on such communication errors here, we consider 3 types of exteroceptive cues from the caregiver: cues that the

infant a priori associates with subsequent attention (which is given by the caregiver in corresponding hidden states with probability a), absence of a caregiving cue that the infant a priori associates with subsequent inattention (which the caregiver gives in corresponding hidden states with probability b), and an ambiguous cue that the infant a priori associates with *both* subsequent Attend and Ignore behaviour (given in these hidden states with probability c). The total observation set is now the product of the infant's interoceptive observations relating to internal stress levels (as before) and these exteroceptive cues from the caregiver, so that the distribution of outcomes given hidden states in the generative process now becomes:

$$O_1 = \begin{bmatrix} a & (1-b-c) & 0 & 0 \\ (1-a-c) & b & 0 & 0 \\ c & c & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & a & (1-b-c) \\ 0 & 0 & (1-a-c) & b \\ 0 & 0 & c & c \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.45}$$

$$O_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ a & (1-b-c) & 0 & 0 \\ (1-a-c) & b & 0 & 0 \\ c & c & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & a & (1-b-c) \\ 0 & 0 & (1-a-c) & b \\ 0 & 0 & c & c \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, O_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.46}$$

and the infant's distribution of observations given hidden states (in their generative model) is now given by:

$$\theta = \{\epsilon\}^{W \times J} + \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} \tag{3.47}$$

with:

$$\theta_1 = \begin{bmatrix} 900 & 0 & 0 & 0 \\ 0 & 900 & 0 & 0 \\ 100 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 900 & 0 \\ 0 & 0 & 0 & 900 \\ 0 & 0 & 100 & 100 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \theta_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 900 & 0 & 0 & 0 \\ 0 & 900 & 0 & 0 \\ 100 & 100 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 900 & 0 \\ 0 & 0 & 0 & 900 \\ 0 & 0 & 100 & 100 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \theta_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1000 & 1000 & 1000 & 1000 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.48}$$

In addition to the prior associations between interoceptive observations and hidden states discussed in the previous section, this model now additionally encodes prior beliefs associating a particular exteroceptive cue with subsequent caregiving Attend behaviour, another (lack of) exteroceptive cue with subsequent Ignore behaviour, and a third (ambiguous) cue which is expected under both hidden states in which the subsequent behaviour is Attend and Ignore. As for the previous case which considered only interoceptive observations, we assume for our initial modelling efforts here that the infant's distribution of observations given hidden states is static and not subject to learning. This is so that we can focus on the learning of the hidden state transition distribution (which encodes caregiving responsiveness q), but also because the misleading and ambiguous cues we consider here encompass many different emotional and verbal communications (rather than just a single cue). As such, we argue that effects of learning in the infant's model of exteroceptive observations given hidden states (in response to these wide variety of cues) are likely to be relatively small compared to learning in the hidden state transition model. The results that follow do, however, also apply to cases in which the observation model is additionally learned (given sufficiently large prior concentration parameters).

We begin by considering the three organised (secure, avoidant, ambivalent) forms of attachment. As discussed above, evidence suggests elevated rates of ACE in caregivers of ambivalent infants compared to caregivers of secure and avoidant infants. We thus ran simulations for an infant paired with four types of caregiver: a highly responsive caregiver

who signals subsequent behaviour appropriately and unambiguously ($q = 0.9$, $a = b = 1$, $c = 0$); a highly unresponsive caregiver who signals subsequent behaviour appropriately and unambiguously ($q = 0.05$, $a = b = 1$, $c = 0$); an inconsistent caregiver who signals subsequent behaviour appropriately and unambiguously ($q = 0.4$, $a = b = 1$, $c = 0$); and an inconsistent caregiver who performs various types of affective communication errors ($q = 0.4$, varying values for a, b, c). All results are averaged over 100 repetitions of 1000 iterations (where each iteration consists of an episode of free energy minimisation and learning with respect to a process of depth 4 timesteps).

Fig. 3.17 shows the mean proportion of action selections along with mean number of actions chosen per iteration for infants paired with these caregivers. Notice that a lack of ACEs in highly responsive and unresponsive caregivers results in organised secure and avoidant attachment, respectively. On the other hand, an infant paired with an inconsistent caregiver who signals subsequent behaviour appropriately is not led to an organised form of attachment, however when this inconsistent caregiver provides misleading ($a = b = 0.25$) and ambiguous ($c = 0.5$) cues, a relatively organised form of ambivalent attachment emerges. In other words, and consistent with the study outlined previously, ACEs can have an organising effect in infants of highly inconsistent caregivers.

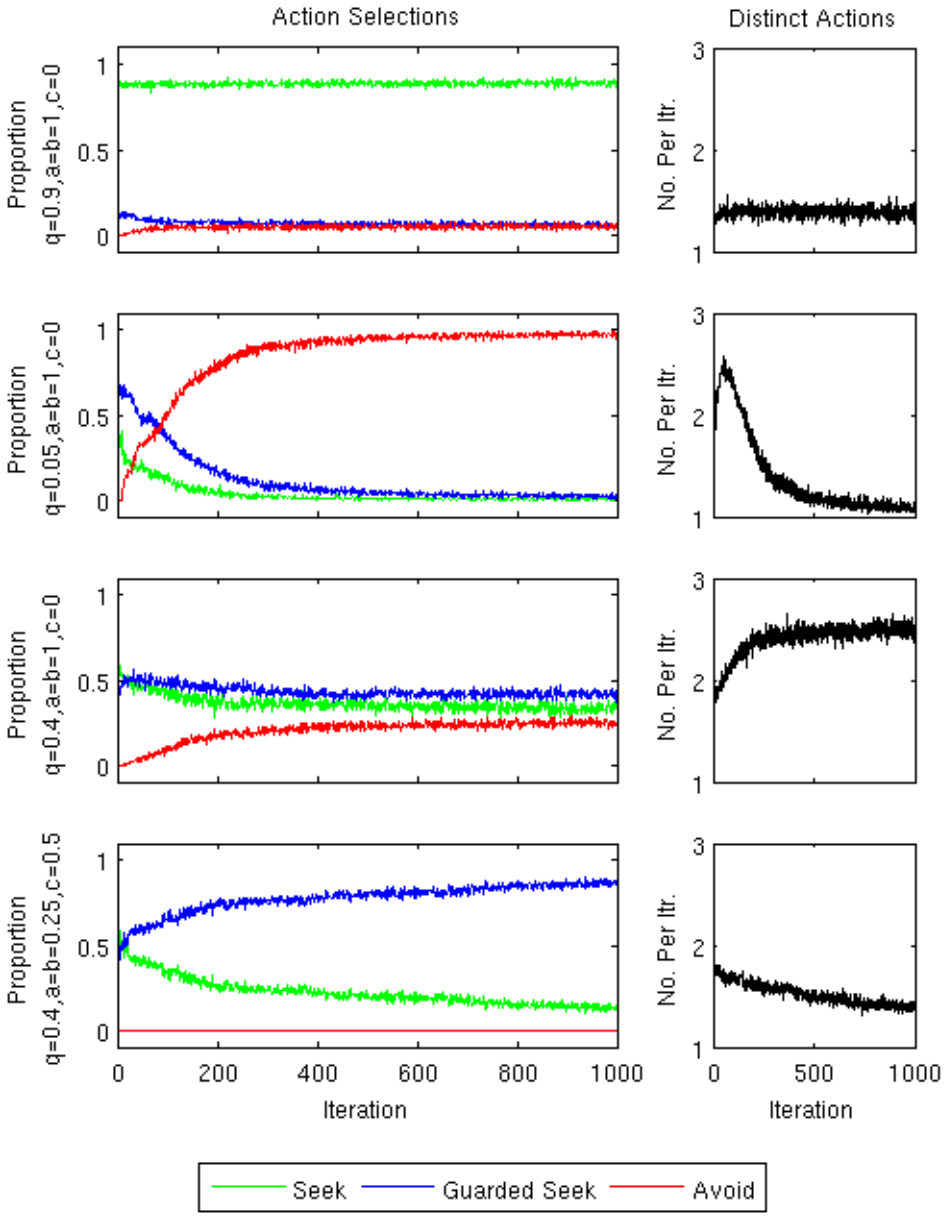

Figure 3.17: **Left Column**: Mean (over repetitions) proportion each action was chosen during each iteration. **Right Column**: Mean (over repetitions) number of distinct actions chosen per iteration. **Top Row**: Infant paired with highly responsive caregiver (q=0.9) with no ACEs (a=b=1, c=0). **Second Row**: Infant paired with highly unresponsive caregiver (q=0.05) with no ACEs (a=b=1, c=0). **Third Row**: Infant paired with inconsistent caregiver (q=0.4) with no ACEs (a=b=1, c=0). **Bottom Row**: Infant paired with inconsistent caregiver (q=0.4) with both ambiguous and misleading ACEs (a=b=0.25, c=0.5).

Fig. 3.18 shows the final posterior hidden state transition distributions learned by infants paired with these inconsistent caregivers in the cases where they either do or do not display these ACEs. The infant paired with the inconsistent caregiver who does not display ACEs is able to learn meaningful structure in their generative model with respect to transitions between second-order hidden states Seek and Guarded Seek actions (e.g. they learned that the hidden state representing current Attend and subsequent Ignore behaviour transitions to hidden states representing current Ignore and either subsequent Attend or Ignore behaviour). On the other hand, accurate second-order hidden state transition distributions have not been learned by the infant subjected to ACEs.

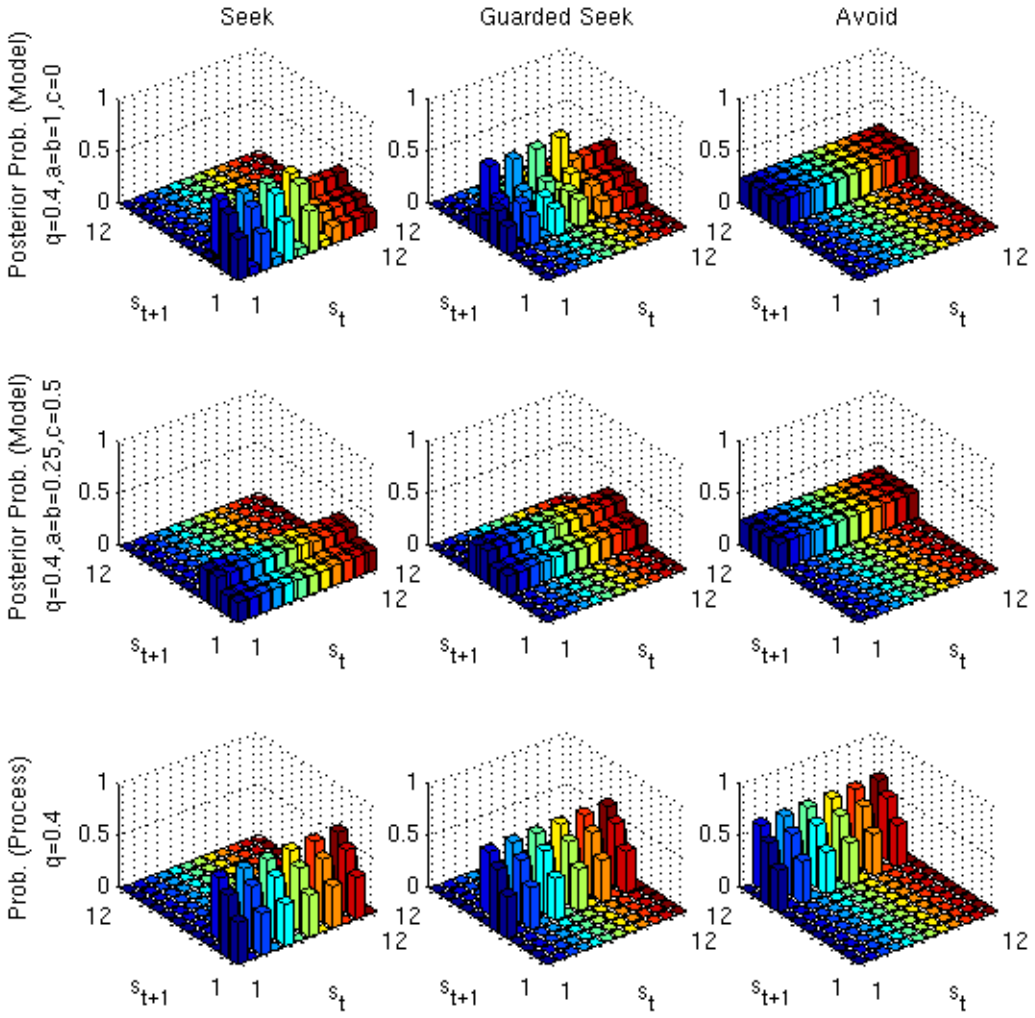

Figure 3.18: Mean (over repetitions) hidden state transition distributions for an agent interacting with a mid-q (inconsistent) caregiver, for control states Seek (**Left**), Guarded Seek (**Middle**) and Avoid (**Right**). **Top**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by an infant paired with an inconsistent caregiver (q=0.4) with no ACEs (a=b=1, c=0). **Middle**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by an infant paired with an inconsistent caregiver (q=0.4) with misleading and ambiguous ACEs (a=b=0.25,c=0.5). **Bottom**: The actual hidden state transition distributions (generative process). The x (bottom right) axis gives current hidden state s_t, y (bottom left) axis the transitory hidden state s_{t+1}, and z-axis the probability.

Fig. 3.19 shows the mean proportion each action was chosen during the last 10 iterations by an infant paired with an inconsistent caregiver ($q = 0.4$) for varying values of a, b, c. The

charts show how various combinations of misleading and ambiguous ACEs can results in relatively organised forms of ambivalent attachment compared to the case where cues are accurate and unambiguous ($a = b = 1$, $c = 0$). Ambivalent attachment is particularly organised for three of the configurations we have examined: when all cues are ambiguous ($a = b = 0$, $c = 1$), when cues are either misleading or accurate with equal probability ($a = b = 0.5$, $c = 0$), and for a mixture of misleading and ambiguous cues ($a = b = 0.25$, $c = 0.5$).

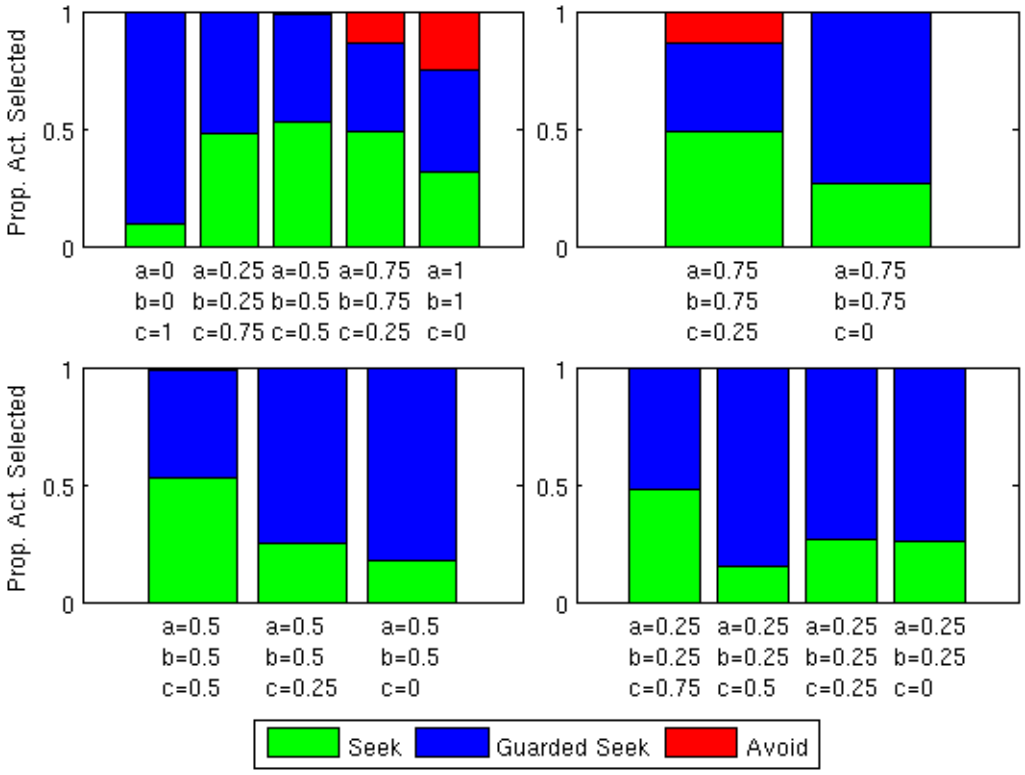

Figure 3.19: Mean (over repetitions) proportion each action was chosen by the infant during the last 10 iterations, when they were paired with an inconsistent caregiver (q=0.4) displaying varying rates and types of affective communication errors. **Top Left**: Proportions for a=b=1-c. **Top Right**: Proportions for a=b=0.75 and c decreasing from 0.25 to 0. **Bottom Left**: Proportions for a=b=0.5, and c decreasing from 0.5 to 0. **Bottom Right**: Proportions for a=b=0.25 and c decreasing from 0.75 to 0.

We now consider ACEs and the formation of disorganised attachment. Lyons-Ruth et al. (1999) highlighted particular ACE items that were found to be at least three times more prevalent in mothers of disorganised infants (see Appendix A). One of these communication errors (and the one that we focus on here) is "inviting approach verbally then distancing",

which corresponds to $b < 1$ in the generative process.

Although caregivers of disorganised infants (prone to withdrawal and frightening behaviours with high probability) are typically thought to increase infant stress levels at a greater magnitude than caregivers of avoidant infants (who instead simply ignore on approach), we note that for $a = b = 1$ and $c = 0$ the infants of such caregivers would learn policies consisting predominantly of Avoid behaviour (as in the case of the avoidant infant examined above). Thus, for reasons of simplicity we consider the same interoceptive preference distribution as before, and aim to show (in a general sense) how this particular ACE ($b < 1$) can have a disorganising effect in caregivers who (with high probability) increase the infant's stress on approach (which encompasses a form of disorganised caregiving).

Fig. 3.20 shows the mean proportion of action selections along with the mean number of actions chosen per iteration for infants paired with caregivers who, with high probability ($q = 0.05$), increase infant stress on approach. In accordance with the studies outlined previously, while caregivers who signal all subsequent behaviour accurately ($a = b = 1$, $c = 0$) lead infants towards organised avoidant attachment, caregivers that signal subsequent Attend behaviour accurately but subsequent Ignore behaviour inaccurately with probability 60% ($a = 1$, $b = 0.6$, $c = 0$) have a highly disorganising effect on infant behaviour. The mean proportion each action was chosen during the last 10 iterations by an infant paired with caregivers with $q = 0.05$ and varying values of b (with $a = 1$ and $c = 0$ in all cases) is shown in Fig. 3.21.

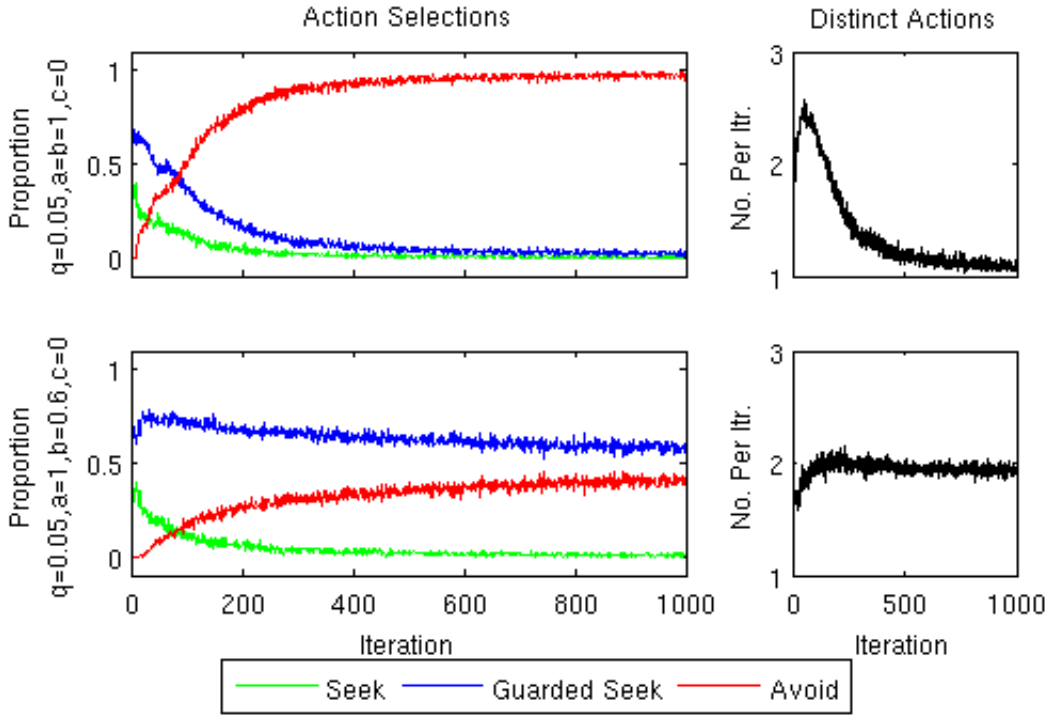

Figure 3.20: **Left Column**: Mean (over repetitions) proportion each action was chosen during each iteration. **Right Column**: Mean (over repetitions) number of distinct actions chosen per iteration. **Top Row**: Infant paired with highly unresponsive caregiver (q=0.05) with no ACEs (a=b=1, c=0). **Bottom Row**: Infant paired with highly unresponsive caregiver (q=0.05) with misleading ACEs on subsequent Ignore behaviour (a=1, b=0.6, c=0).

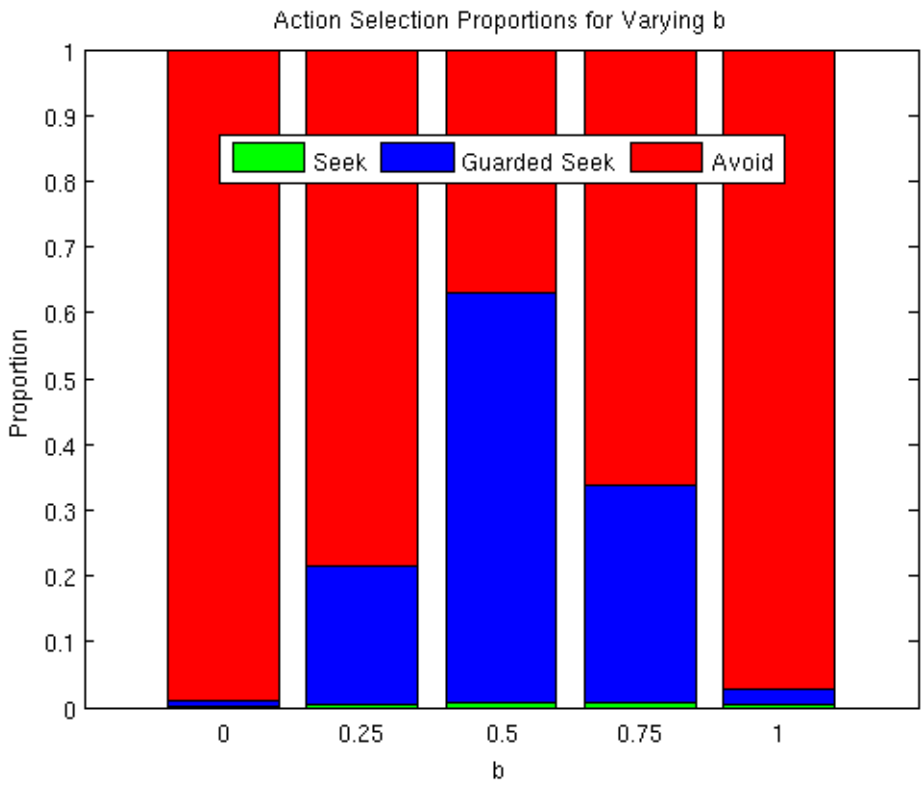

Figure 3.21: Mean (over repetitions) proportion each action was chosen by the infant during the last 10 iterations (y-axis), when they were paired with a highly unresponsive caregiver (q=0.05) displaying varying rates of misleading affective communication errors (values of $b \in \{0, 0.25, 0.5, 0.75, 1\}$ along the x-axis, for fixed $a = 1$ and $c = 0$).

Fig. 3.22 shows the final hidden state transition distributions learned by infants paired with these unresponsive caregivers in the cases where they either do or do not display this ACE with respect to subsequent inattention. The infant paired with the unresponsive caregiver who does not display ACEs is able to learn meaningful second-order hidden state transition distributions for Seek and Guarded Seek actions (although they have not learned a perfect model for these actions, since they very quickly come to display an organised form of avoidant attachment). On the other hand, the infant paired with an unresponsive caregiver who instead signals subsequent inattention inaccurately has learned distributions for Seek and Guarded seek which, for all current hidden states, predict subsequent hidden states overwhelmingly involving current Ignore behaviour. Thus, similarly to the case of the inconsistent caregiver considered above, ACEs prohibit the emergence of the second-order hidden state transition model.

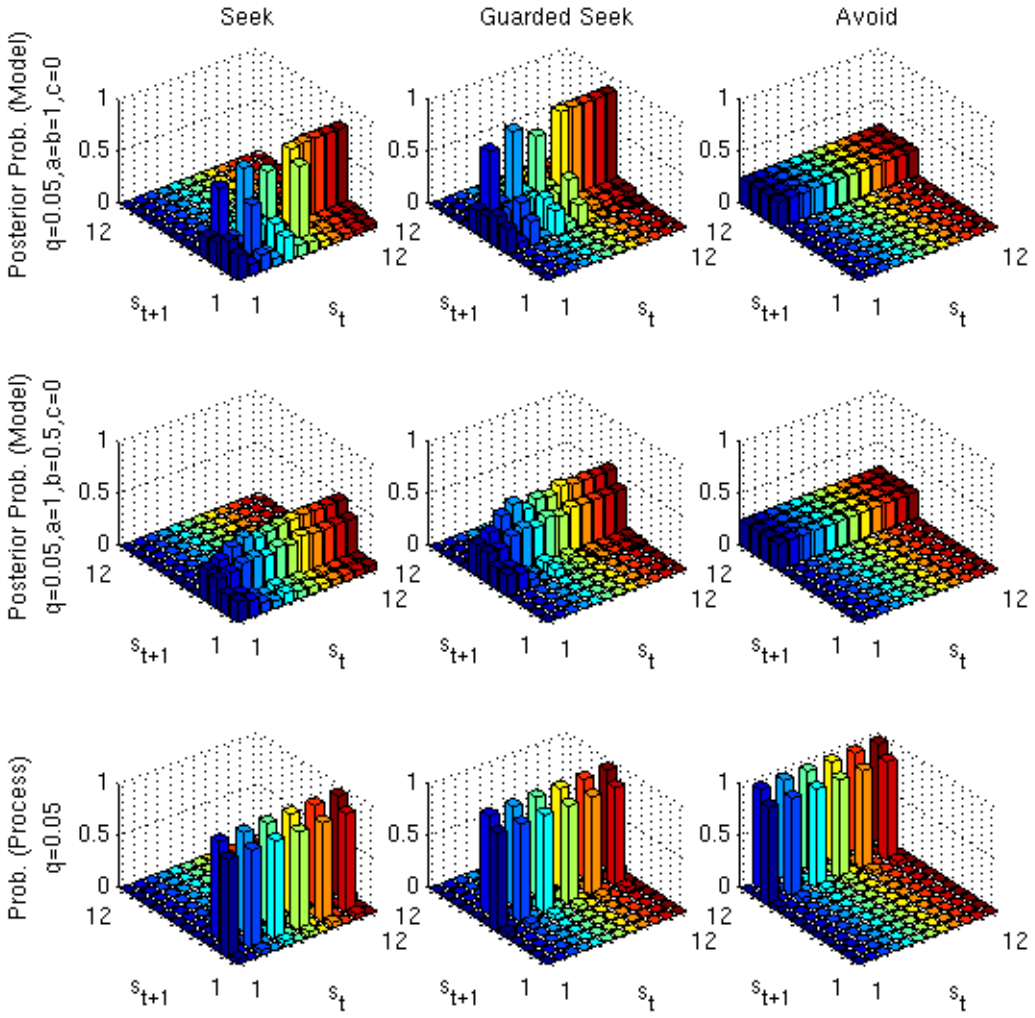

Figure 3.22: Mean (over repetitions) hidden state transition distributions for an agent interacting with a low-q (unresponsive) caregiver, for control states Seek (**Left**), Guarded Seek (**Middle**) and Avoid (**Right**). **Top**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by an infant paired with highly unresponsive caregiver (q=0.05) with no ACEs (a=b=1, c=0). **Middle**: Mean (over repetitions) hidden state transition distributions $P(s_{t+1}|s_t, u_t)$ learned by an infant paired with an unresponsive caregiver (q=0.05) with misleading ACEs on subsequent Ignore behaviour (a=1,b=0.5,c=0). **Bottom**: The actual hidden state transition distributions (generative process). The x (bottom right) axis gives current hidden state s_t, y (bottom left) axis the transitory hidden state s_{t+1}, and z-axis the probability.

In order to understand in more detail why the infant of the low-q caregiver who additionally displays ACEs exhibits roughly equal proportions of Avoid and Guarded Seek

behaviour within each episode, it is helpful to consider the expected negative free energy for policies containing each action. In Friston et al. (2015) it is shown that the expected negative free energy for some particular policy π at time $\tau > t$ consists of extrinsic and epistemic value terms:

$$\mathbf{Q}_\tau(\pi) = \mathbb{E}_{Q(o_\tau|\pi)}\left[\ln P(o_\tau) + KL[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)]\right] \tag{3.49}$$

The extrinsic value component $\mathbb{E}_{Q(o_\tau|\pi)}\left[\ln P(o_\tau)\right]$ of the expected negative free energy is the utility of the outcome o_τ (expected under the predictive posterior distribution) to the agent. In our model, extrinsic value corresponds to the preference the agent has for the interoceptive observation (stress increase or reduction) associated with that particular outcome, since they are assumed to be indifferent with respect to control states and exteroceptive observations. The epistemic value component $\mathbb{E}_{Q(o_\tau|\pi)}\left[KL[Q(s_\tau|o_\tau,\pi)||Q(s_\tau|\pi)]\right.$ reports the reduction in uncertainty about hidden states based on the outcome. Epistemic value can result in exploration, in the sense that an agent might select policies predicting outcomes with relatively low extrinsic value if these outcomes nonetheless help to reduce uncertainty with respect to hidden states.

Fig 3.23 shows the expected negative free energy, extrinsic and epistemic value for each action on the final step of each iteration (mean over repetitions), for both avoidant and disorganised infants. Both the avoidant and disorganised infants come to predict similar interoceptive observations for Seek and Guarded Seek behaviours (corresponding to the caregiver Ignoring them at τ) and thus both assign similar extrinsic value to these actions. However, epistemic value for Seek and Guarded Seek actions remains relatively high for the disorganised compared to the avoidant infant over iterations. The avoidant infant's posterior transition distribution has learned meaningful information with respect to second-order hidden state transitions, such that they tend to predict future hidden states at τ involving both current and subsequent Ignore behaviour, and outcomes at τ consisting of the interoceptive cue predicting current, and exteroceptive cue predicting subsequent, Ignore behaviour. On the other hand, as a result of previously experiencing ACEs, the disorganised infants have not learned these second-order hidden state transitions. These infants come to predict future hidden states at τ involving current Ignore, but either subsequent Attend or Ignore, behaviour (with exteroceptive cues a priori expected in each case to be able to disambiguate between these two hidden states). The additional epistemic value that results is enough to give rise to disorganisation.

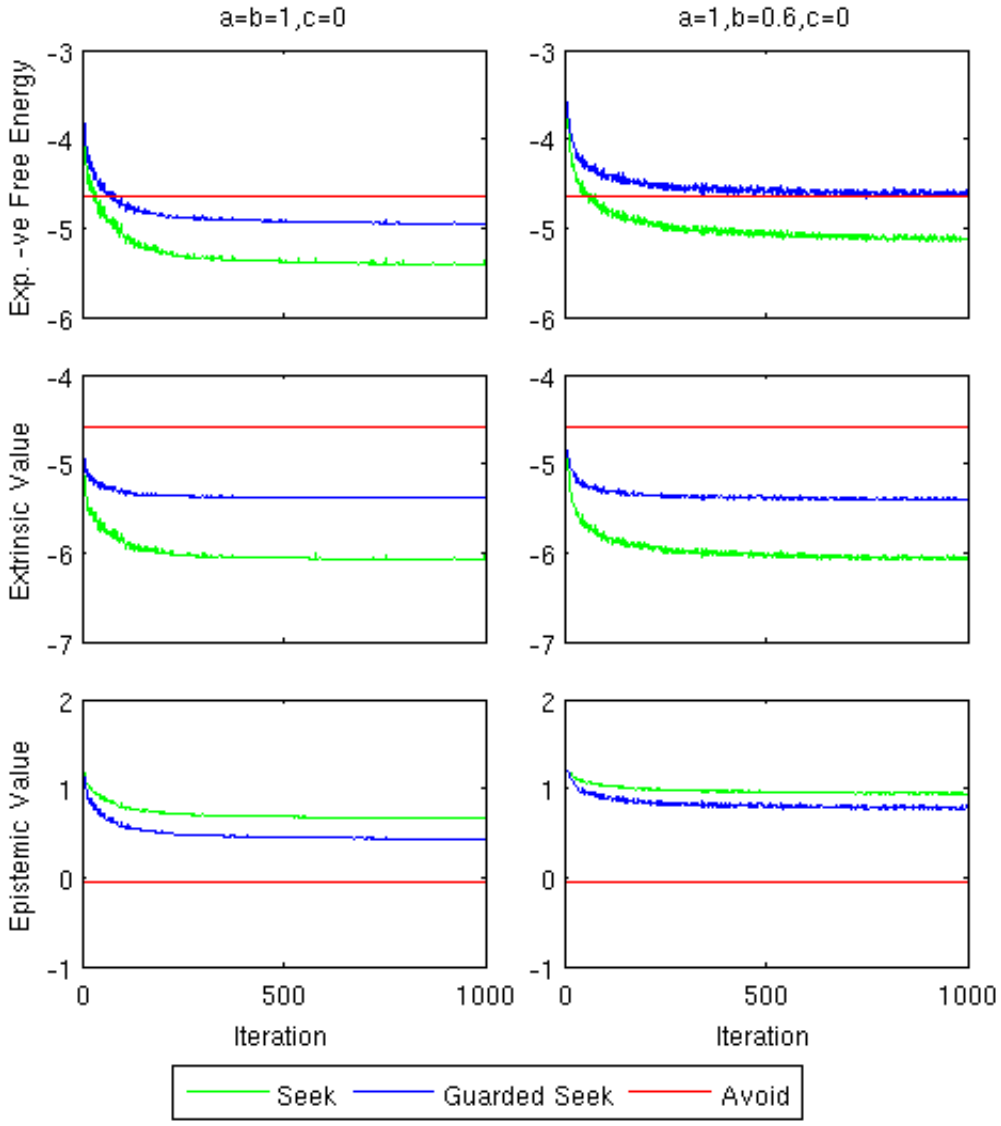

Figure 3.23: Mean (over repetitions) expected negative free energy, extrinsic value and epistemic value for Seek, Guarded Seek and Avoid actions on the final step of each iteration, for infants interacting with a low-q (unresponsive) caregiver displaying no ACEs, i.e. $a = b = 1$, $c = 0$ (**Left**) or ACEs with $a = 1$, $b = 0.6$, $c = 0$ (**Right**). **Top**: Mean (over repetitions) final step expected negative free energies. **Middle**: Mean (over repetitions) final step extrinsic values. **Bottom**: Mean (over repetitions) final step epistemic values.

3.5 Summary and Future Work

In this chapter we built on an existing decision theoretic model of organised attachment in order to propose a multi-step formulation of infant attachment formation in terms of free energy minimisation. In particular, we considered infant agents that minimise free energy over interoceptive observations related to changes in internal stress levels. Using this model, we showed how an infant with a prior hidden state transition distribution that is uniform with respect to caregiving responsiveness can come to acquire either a secure, avoidant or ambivalent form of organised attachment, with the particular type of attachment that emerges depending only on the responsiveness of the caregiver. Based on evidence relating to ACEs in caregivers of both ambivalent and disorganised infants, we then extended this model to consider also exteroceptive cues from the caregiver. We considered one particular ACE that has been found to be three times more prevalent in caregivers of disorganised infants, which is a cue that inaccurately implies subsequent attention. In order to capture the essence of disorganisation with this particular ACE, we showed how such a misleading cue might have a disorganising effect in infants of caregivers who, with high probability, increase infant stress on approach. Since (to the best of our knowledge) no particular ACE items on the AMBIANCE scale have as yet been associated with ambivalent forms of attachment, we explored the effect on the infant of both exteroceptive cues that are misleading and ambiguous with respect to subsequent caregiving behaviour. We showed how the introduction of various combinations of such ACEs might have an organising (towards ambivalence) effect in infants paired with these inconsistent caregivers. Our model makes a new prediction that can be tested empirically: namely that particular combinations (distributions) of misleading and ambiguous ACEs will lead to the most organised forms of ambivalent attachment.

The model (or atleast future versions of it) can potentially be used for phenotyping (Schwartenbeck and Friston, 2016), however there are a number of ways in which the work can be improved. For example, we considered free energy minimisation over fixed-length episodes in which all hidden states were taken to be states in which the infant's attachment system is active, i.e. we did not explicitly consider a return to baseline stress level for the infant. This was for reasons of simplicity, and also since evidence with respect to the relative time taken to return to baseline for avoidant, ambivalent and disorganised infants is currently either inconclusive (for interactions in which caregiver responsiveness is controlled) or unavailable (for uncontrolled interactions). As more empirical data becomes available, the model might be extended to include transitions from the current hidden states to an additional state (associated with a highest-preference interoceptive observation) representing deactivation of the infant's attachment system, with stress change parameters, transition probabilities and caregiving responsiveness set in order to model the data for each of these distinct attachment types. Future work could also extend the scope of the model to capture

the secure-base exploration paradigm more fully, to consider how undesirable stress states might arise during the course of exploration and how environment exploration might resume on transition to this baseline state.

In accordance with studies which found only the ACE dimension of the AMBIANCE scale to be a differentiator of disorganised compared to organised (secure and avoidant) attachment, and elevated ACEs in caregivers of ambivalent infants, we only considered the ACE dimension in our model. For disorganised attachment we considered only one particular type of ACE (cues that are misleading on subsequent caregiving inattention), whereas for ambivalent attachment we considered a number of distributions under which misleading and/or ambiguous cues were delivered with varying frequency. In particular, we considered these exteroceptive cues to be misleading or ambiguous to the infant a priori, captured using relatively large priors in the infant's model of observations given hidden state transitions. We chose to make this modelling assumption since the AMBIANCE scale describes ACEs in terms of broad groups of emotional and/or verbal cues, and so we decided to focus here on the learning of the infant's hidden state transition distribution (which encapsulates caregiving responsiveness). Future work could consider how ambiguity in exteroceptive cues might arise as a result of learning (and the two models could be selected between using, for example, a Bayesian model comparison). In addition (as discussed previously) elevated rates of other dimensions of the AMBIANCE disrupted affective communication scale have been associated with disorganised (withdrawal, disorientation, negative/intrusive and role confusion) and resistant (disorientation, negative/intrusive) forms of attachment. Future models can consider these other atypical caregiving behaviours, along with the other cues described by the ACE dimension, and might also attempt to differentiate between sub-categories of disorganisation (perhaps according to the Hostile/Helpless caregiving profiles associated with D-Insecure/D-Secure infants in Lyons-Ruth et al. (1999)). An attempt to capture additional aspects of disorganised (such as dissociative-like freezing) and ambivalent (e.g. hyperactivation) infant attachment behaviour can also be made: in the case of the ambivalent infant a self-induced increase in stress is believed to be a strategy to increase the likelihood of subsequently attentive caregiving, which can be captured relatively easily in the hidden state transition structure. As a further extension to the model, we might attempt to capture prolonged states of mind in the caregiver over each attachment episode (corresponding to contexts in the scenario modelled in Friston et al. (2015)) and an infant agent who learns a hierarchical generative model in which higher levels contextualise lower levels (Friston et al., 2015; Pezzulo et al., 2015).

Another possible extension to the model is to consider the subjective emotional experiences of infants paired with the different types of caregiver. A definition of emotion in terms of first and second-order time derivatives of free energy is proposed in Joffily and Coricelli (2013): according to that model, the valence of a state visited by an agent at time

t is defined as the negative first time-derivative of free energy at that state, so that free energy increasing over time ($F_t^{'} > 0$) implies negative valence, and free energy decreasing over time ($F_t^{'} < 0$) implies positive valence. Particular categorical emotions are then defined in terms of first and second time derivatives of free energy: for example, when $F_t^{'} < 0$ and $F_t^{''} < 0$ (i.e. free energy is decreasing at an increasing rate) the agent is hopeful of visiting a state with lower free energy in the future, whereas when $F_t^{'} > 0$ and $F_t^{''} > 0$ (free energy is increasing at an increasing rate) the agent is fearful of visiting a state of greater free energy in the future. Fear would be particularly desirable to capture in our model, since (according to Main's classical hypothesis) we would expect elevated experience of this emotion in cases of disorganisation.

Finally, a very recent update to the discrete free energy minimisation framework used here additionally accounts for habit learning (Friston et al., 2016) which would be interesting to consider within the context of attachment, and particularly Self-Attachment therapy (introduced in the next chapter). Briefly (and as we will see in more detail), Self-Attachment therapy aims to redress suboptimal early attachment experience by inducing DA and OXT-mediated neural plasticity in key attachment-related neural circuitry, which may involve the overcoming of deeply ingrained habits. It has recently been proposed that OXT plays a role in encoding the precision of interoceptive signals and therefore is involved in the association of interoceptive and exteroceptive observations within generative models of the self (Quattrocki and Friston, 2014). An interesting next step would thus be to formulate the hypothesised dynamics underlying a successful application of Self-Attachment therapy in terms of the free energy principle.

4. Self-Attachment Bonding Protocols

Having examined the formation of the organised (secure, ambivalent and avoidant) and disorganised forms of attachment in the previous chapter, we now consider a computational model of the neurobiological underpinnings of the Self-Attachment psychotherapy, which aims to re-train an individual's attachment schema. We begin by presenting a hypothesis relating to plasticity induced by the Self-Attachment bonding protocols, before considering the interplay between this and another aspect of the therapy (the empathy protocols) in the next chapter.

4.1 Introduction

As discussed in Section 2.5, 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 (Edalat, 2013b, 2015). 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 RH to RH interactions using instead internal LH to RH interactions within the individual. 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 OFC and limbic system, in order to increasingly contain pathological and suboptimal neural activity related to early insecure attachment experiences. While acknowledging the view of attachment theorists which states that attachment needs persist into adulthood (particularly in those with adverse childhood experience), the aim of Self-Attachment is to equip the individual with tools and techniques that will enable them to effectively regulate their own internal state.

In what follows we build on findings related to the neuroscience of attachment, bonding

and reward processing to present a computational model of fear counter-conditioning within neural circuits mediating stress and facilitative 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.

4.2 Neuroscience of Attachment and Bonding

As outlined previously, 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 (Schore, 2003a; Cozolino, 2014, p.53); areas known to be central to emotional processing, homeostatic regulation, social cognition and fear conditioning. Recent neuroimaging studies which have specifically investigated the neural correlates of attachment, bonding and parenting have begun to elaborate on this picture.

In a number of social and attachment-related contexts, insecure attachment has been found to correlate with relatively high activity in the amygdala and related stress circuitry. For example, Lemche et al. (2006) conducted a task in which individuals (rated as either secure or insecure for attachment according to the AAI) were shown a series of self-referential sentences describing either neutral situations or unpleasant attachment experiences (stress scenario), and were asked whether they agreed or disagreed with the statements while being scanned with fMRI and simultaneously having skin conductance (autonomic) levels measured. The study found that bilateral amygdala activity was positively correlated with attachment insecurity and autonomic response in the stress scenario. This seems to be the case in particular for ambivalent forms of insecure attachment. Vrtička et al. (2008) conducted a study in which faces with different expressions (either smiling or angry) were presented as social feedback signals about performance during a perceptual game (in which, after brief exposure, participants were asked to indicate which side of a screen contained more dots). The left amygdala selectively activated when angry faces were presented as negative feedback following incorrect responses (i.e. as social punishment), and the magnitude of this response was found to correlate with the degree of anxious (ambivalent) attachment (as measured by the relationship questionnaire) in participants, suggesting an increased sensitivity to social punishment and task failure. In Riem et al. (2012), women without children were exposed to the sounds of the crying of a non-related infant. Those women scoring high in ambivalent (preoccupied) or avoidant (dismissing) attachment (as measured by the AAI) showed elevated amygdala activation, reported more irritation, and applied more force on a handgrip dynamometer, relative to securely attached women. In another study (Kidd et al., 2013), salivary cortisol measurements were collected across an entire day (in a non-laboratory setting) from a sample of men and women whose attachment was

classified based on the relationship questionnaire. Anxious (preoccupied) attachment was found to be correlated with a relatively elevated and flat cortisol profile across the whole day, and these individuals also reported the highest subjective levels of stress.

The amygdala may also differentially activate in individuals with disorganised forms of attachment. Buchheim et al. (2006) conducted studies in which participants underwent fMRI while being rated for attachment type according to the adult attachment projective measure, in which the individual is asked to tell a story in order to explain what is happening in pictures that depict various attachment-related scenarios. The task was designed in such a way so that, as it progresses, it increasingly activates the attachment system (e.g. by depicting more traumatic scenes). The authors found increasing amygdala activation as a function of task progression only in individuals rated unresolved (disorganised). A recently reporting longitudinal study found increased left amygdala volume in adulthood in those classified as disorganised in infancy (Lyons-Ruth et al., 2016). A number of studies have also suggested hyperactive amygdala responses in individuals with BPD (a disorder closely related to disorganised attachment, see Section 2.5), for example in response to emotional pictures (Herpertz et al., 2001) and faces (Donegan et al., 2003); and in response to cues indicating threat in individuals who have experienced childhood maltreatment (Dannlowski et al., 2012; McCrory et al., 2011). However, the picture is perhaps not as entirely consistent as it seems to be for ambivalent forms of attachment. In the study of exposure to non-related infant crying outlined above (Riem et al., 2012), although there was a significant correlation between a measure of coherence of mind and amygdala activation, individuals with unresolved attachment did not show significant increased amygdala activation relative to secure individuals. In addition, mothers rated unresolved according to the AAI in Kim et al. (2014) showed blunted amygdala activity in response to the sad faces of their own infant compared to their happy faces (in contrast to resolved mothers, who showed the opposite effect).

Much evidence from rodent studies implicates the DA reward system in maternal and bonding behaviour (Numan and Stolzenberg, 2009; Numan, 2014), and a similar picture has started to emerge for humans. For example, Bartels and Zeki (2004) 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 VTA and 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 Minagawa-Kawai et al. (2009), anterior parts of the OFC were found to activate both in mothers viewing their own infants (with activation positively correlated with a rating of the pleasant mood of the infant), and in infants on viewing their mother's smile. As with activity in the amygdala and related stress circuitry, activity in the reward system appears to be modulated by attachment type. In Strathearn et al. (2009) relatively

low ventral striatum and OFC activity was found in avoidant relative to secure mothers (as assessed by the AAI) in response to seeing own-baby images. A comparison of own versus unknown infants (with either happy or sad emotional expression) showed significantly higher activation in reward circuitry when caregivers were considered irrespective of attachment type, suggesting a general increased motivation for interaction with an own as opposed to unknown infant. In Vrtička et al. (2008), increasing attachment avoidance (according to the adult attachment questionnaire) was found to correlate with decreasing activation 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, Kringelbach et al. (2008) found activation of the mOFC (a region implicated in stimulus-reward association) in response to images of unfamiliar infants but not adults.

Also important with regards to attachment is OXT. Elevated OXT has been linked to a range of pro-social behaviours, including parenting and bonding behaviours, and elevated trust. For example, Feldman et al. (2007) 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 (Gordon et al., 2010), 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 Feldman et al. (2007)), 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 development. For example, Heim et al. (2008) 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 (Declerck et al., 2010). Intranasal OXT has also been found to decrease the likelihood of cooperation during a social dilemma game in adults with BPD and high levels of attachment anxiety (according to the experience in close relationships self-report measure, (Bartz et al., 2010a)). In a recollection study (Bartz et al., 2010b), securely attached individuals (again according to the experience in close relationships measure) 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 (Olff et al., 2013), likely making exogenous OXT administration unsuitable as a treatment for many attachment-related disorders. This provides motivation for the development of therapies in which endogenous OXT secretion is naturally enhanced as a side effect of changes to the underlying attachment schema.

4.3 Self-Attachment Therapy

Self-Attachment (Edalat, 2015) is a new, self-administrable, attachment-based psychotherapy which aims to redress the insecure early attachment experiences that are believed to underpin many affect dysregulation disorders. Under the Self-Attachment paradigm, the self of the individual undergoing therapy is conceptualised as comprising two parts: the inner-child and the adult-self. The inner-child corresponds to the emotional self that becomes dominant under times of stress and perceived threat, thought to be rooted mainly in the RH of the brain. The adult-self corresponds to the more rational self dominant under times of calm and low perceived threat, 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, in order to create a secure attachment schema within the individual. These interactions are proposed to naturally stimulate the release of OXT and DA, in order to encourage neural plasticity and increasingly contain suboptimal and pathological neural activity that inhibits abilities for self-agency. This is achieved by means of simulating (for example, using mental imagery techniques) the interactions between an infant and secure caregiver, from both perspectives. Since both the inner-child and adult-self are conceptualised as constituents of the self, the individual can be said to securely "self attach".

4.3.1 Stages

The four stages of the Self-Attachment therapeutic process (Fig. 4.1) are outlined here (see Edalat (2015, 2017a,b) for further details). In the first (introductory) stage, the individual becomes familiar with the scientific basis and underlying hypotheses of the therapy, which includes an introduction to attachment theory, and the basics of the (developmental) neurobiology of attachment, love, bond making and emotion regulation. The aim of this preliminary phase is to provide initial motivation for undertaking the therapy, which requires dedication and self-discipline in terms of time and commitment.

Once this preliminary phase has been completed, the individual can begin to conceptualise the inner-child as an entity that is distinct from the adult-self, and the adult-self can begin to create a relationship with the inner-child with a view to establishing empathy and ultimately compassion with them. During the second (conceptualisation) phase of Self-

Attachment, the individual selects both a positive photograph of their childhood (which elicits emotions and memories such as happiness or contentment) and a negative photograph (for example associated with sadness). Several highly-structured exercises (termed protocols), focused towards these images, are then conducted in order to conceptualise the inner-child as concretely as possible. These protocols include (for example), with closed eyes, trying to visualise the two chosen childhood photos, and attempting to imagine that the child that they were is present and close to them and that they can touch and hold this child. Another protocol, aimed at strengthening the distinction between the adult-self and inner-child, involves projection of a negative internal state outwards onto an image associated with the inner-child. As we will discuss in the following chapter, we suggest that the undertaking of techniques from existing therapies (e.g. mentalization) might additionally be helpful in strengthening the self-other distinction during this phase.

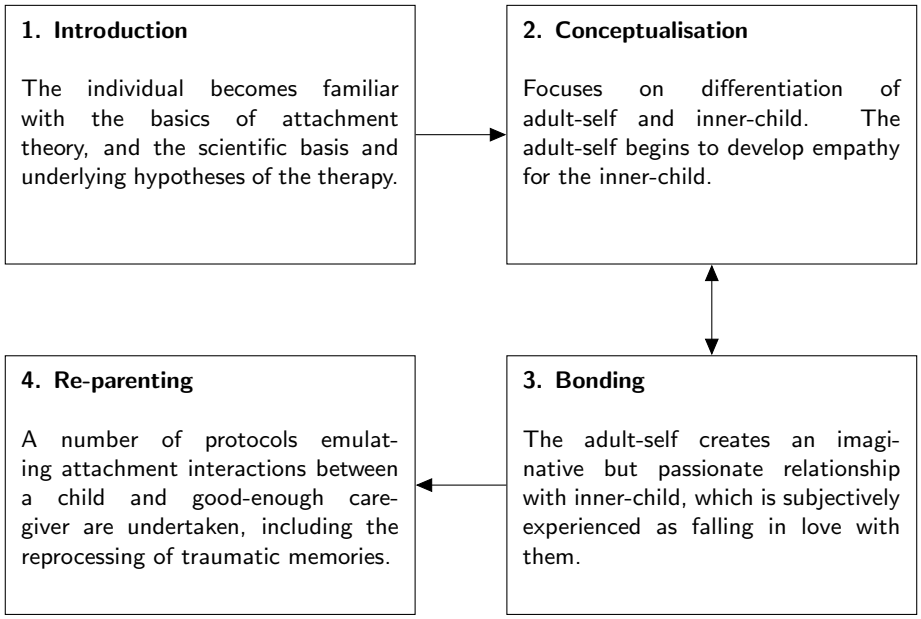

Figure 4.1: The four stages of the Self-Attachment therapeutic process. See text for further details.

The third stage of Self-Attachment (and the focus of this chapter) is concerned with building an imaginative but passionate affectional bond with the inner-child, which is subjectively experienced as falling in love with them. First, the adult-self adopts the inner-child and makes a vow to consistently support and love them. Then (from the perspective of the adult-self) the individual focuses on the images of the inner-child and attempts to bond with them, as a basis for creating the internalised attachment relationship. This bonding process

is enhanced with the use of activities such as self-massage and positive tactile stimulation (to simulate an embrace), and (overt and/or imagined) song and dance directed towards the inner-child which (as we will expand on) are hypothesised to assist in inducing neural plasticity in key attachment-related neural circuitry. Motivation to engage in this bonding may be enhanced with protocols involving empathic attunement with the emotional state of the inner-child: as we will explore in the following chapter, the protocols involved in the second (conceptualisation) and third (bonding) phases are closely intertwined, with application of each likely to drive progress in the other, and many of the protocols associated with these phases are carried out in a parallel rather than serial fashion.

The fourth phase of Self-Attachment therapy involves a number of protocols describing patterns of interaction between the adult-self and inner-child that emulate the function of a good enough parent interacting with a securely attached child, with the aim of minimising negative emotion and maximise positive affect. One example is a protocol that involves the reprocessing of painful and traumatic past events: first, the individual closes their eyes and recalls a traumatic childhood episode, remembering and re-experiencing in as much detail as possible the associated negative emotions (such as fear or helplessness). Once this state has been recalled, the individual imagines that the inner-adult quickly and competently intervenes in order to reduce distress in the inner-child, for example by embracing or vocally reassuring them. The aim is for these protocols to become habituated with repetition, so that the individual spontaneously engages with the inner-child in ways that alleviate their attachment needs.

4.3.2 Relationship to Existing Therapies

Self-Attachment can be regarded as an extension of attachment theory, but it is also related to and incorporates ideas from a range of existing psychotherapeutic methods including transactional analysis, behavioural therapy, mentalization, object-relations psychodynamic therapy, and cognitive behavioural therapy (Edalat, 2017a). Self-Attachment, which can be combined with any well-established therapeutic framework, integrates techniques from these therapies into its key and uniquely distinguishing focus of intervention, which is the creation of a fully internalised attachment relationship and affectional bond that emulates the characteristics of a secure infant-parent dyad. Although in its early stages, the results from a small number of initial (uncontrolled) preclinical trials (overseen by psychotherapists in private practice) have shown success in tackling chronic anxiety and depression in individuals who had previously (unsuccessfully) engaged with a number of other practices (including cognitive behavioural therapy, psychoanalytic therapy, mindfulness and neurofeedback) for long periods over several years (Edalat, 2015).

A closely related concept is security priming (Mikulincer and Shaver, 2007), which involves temporarily activating mental representations relating to the availability of a secure

attachment figure in order to reduce distress and restore positive mood. This is achieved using a variety of techniques involving subliminal (e.g. presentation of pictures suggesting attachment figure availability, or of the name of an individual perceived to be a secure attachment figure), visual (e.g. presentation of the face of a secure attachment figure) and imagery (e.g. guided imagery involving availability of an attachment figure) methods.

Also related is compassion-focused therapy (Gilbert, 2009), which involves activities designed to develop compassionate attributes and skills within the individual in order to improve capabilities for self-compassion and affect regulation. Techniques include the use of imagery, in which the individual imagines themselves receiving compassion from an external (not necessarily human) source (Rockliff et al., 2008). Recent studies have used virtual reality as a medium for switching an individual's perspective between an adult avatar (resembling themselves) and a generic child (Falconer et al., 2014, 2016). While embodied in the adult avatar the individual administered compassion to the distressed child, before switching to the perspective of the child in order to re-experience themselves administering the compassion from this alternative perspective; a practice which resulted in a reduction in measures of depression and self-criticism. In these experiments a non-related and non-self-resembling child avatar was used (although virtual embodiment during the recipient phase would have resulted in some sense of identification with the child). In contrast, rather than being a generic child, the recipient of attachment-based compassion in Self-Attachment (the inner-child) is conceptualised as comprising a part of the self, and there is a focus on the development of a dyadic attachment relationship between the inner-child and adult-self. It has been argued that using this inner-child representation, as opposed to a generic and/or non-related child, increases the efficacy of the therapy from the perspective of primary narcissism (Edalat, 2017a), and in Chapter 5 we will additionally argue that this inner-child representation should assist in inducing empathically-motivated caregiving (bonding) behaviour.

4.3.3 Bonding Protocols

In order to complete later stages of the therapy, the adult-self must first establish an imaginative (but passionate) loving relationship with the inner-child, which is subjectively experienced as falling in love with them. The individual begins by focusing on happy and unhappy images of their younger self, in order to form a conceptualisation of the inner-child. Then, from the perspective of the adult-self, the individual attempts to bond with the inner-child in order to create an attachment relationship with them, enhancing this process with the use of activities such as (overt and/or imagined) song and dance directed towards the inner-child. These activities can be experienced from the perspective of the inner-child in terms of attempts by the adult-self to engage and comfort during times of unease or distress.

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 (Edalat, 2017a). For example, it is common for children to form bonds with inanimate transitional objects that can serve as mother substitutes (Litt, 1986). In addition, throughout much of history, abstract bonds formed and reinforced through religious practice have been a predominant means of regulating emotion and social behaviour, and it has been argued that bonds between religious believers and God meet many of the criteria of an attachment relationship (Kirkpatrick, 2005). An fMRI study of Danish Christians (Schjødt et al., 2008) 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 a role in attachment-related approach behaviours (Villablanca, 2010).

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, along with positive tactile stimulation such as self massage, which are proposed to stimulate the dopaminergic reward system and areas involved in the representation of reward. Music 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. Koelsch et al. (2006)). Furthermore, recent fMRI studies have shown evidence for the involvement of the reward system during passive listening to self-reported pleasurable music (consistently implicating the Nucleus Accumbens (NAc), a major target for dopaminergic projections from the VTA). For example, in Salimpoor et al. (2011), fMRI and PET scans revealed striatal DA release in the CN and NAc in response to anticipation and peak emotional response to the music, respectively. Another study (Montag et al., 2011) found that self-reported favourite as opposed to least favourite music correlated with increased activity in the ventral striatum (which includes NAc), CN and insular. Similarly, in Koelsch et al. (2006) pleasant contrasted to unpleasant music correlated with activation of areas including the ventral striatum and anterior superior insular. In Mitterschiffthaler et al. (2007), subjects reported their emotional responses to classical music on a scale from 0 (sad) through 50 (neutral) to 100 (happy). Self-reported happy states correlated with activity in regions including ventral and dorsal striatum (CN), and sad with activity in regions including the amygdala. In a PET study (Brown et al., 2004), activations elicited by unfamiliar though pleasantly rated music resulted in activation of regions including AI and NAc compared to a rest condition. Song, too, has been shown to activate emotion and reward circuitry in both overt and imagined form. Jeffries et al. (2003) 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 mPFC and NAc. In another study (Kleber et al., 2007), professional classical singers were asked to imagine singing an aria (love song), resulting in intense activation in emotional areas (including amygdala, ACC, and mPFC) along with the

CN and putamen. Finally, there is evidence from human studies to suggest that pleasant touch (which is thought to be primarily reinforcing), of the sort that would be involved in self-massage, activates the OFC (with activation positively correlated to pleasantness (Rolls, 2013, p.72)), decreases cortisol (Field et al., 2005) and increases DA (Field et al., 2005) and OXT (Light et al., 2005) (in a recent study it was furthermore demonstrated that gentle stroking activates Paraventricular Nucleus of the Hypothalamus (PVN) OXT neurons in the rat (Okabe et al., 2015)).

As discussed in Section 4.2, 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 (Huang and Hessler, 2008). 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 chapter, 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 here 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.

4.4 Model of Hypothesised Self-Attachment Bonding Protocol Effects

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. 4.2.

Our starting point is Levine's neuroanatomical model in Levine (2008), which identified the OFC as the key region in mediating the relative strength of activity in fight/flight circuitry (focused on the amygdala, PVNp and 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 PVNm (which controls release of OXT and Vasopressin (VA)) and the PVNp (which controls release of CRH). Release of

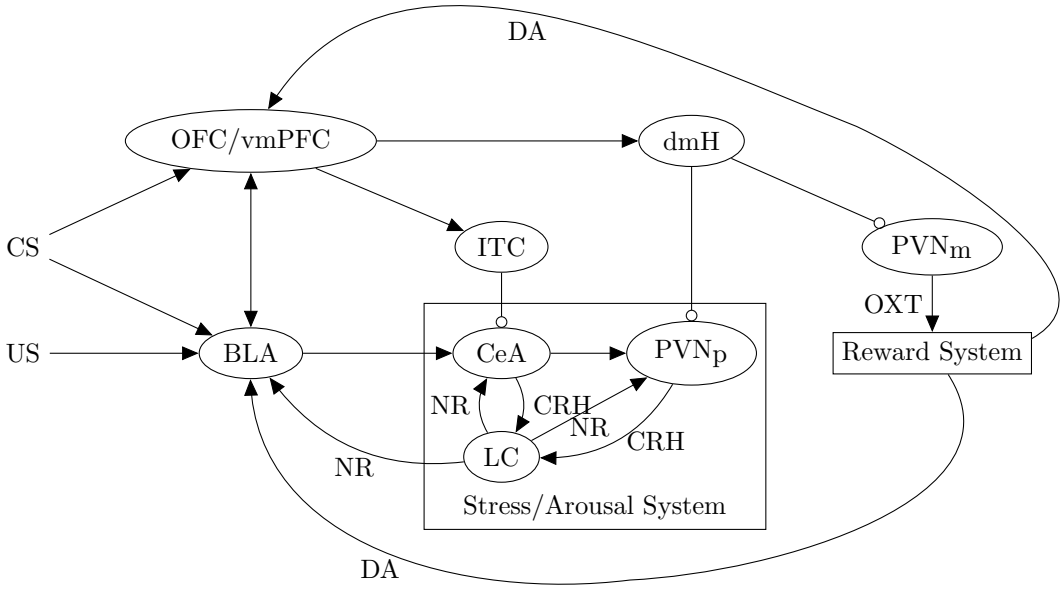

Figure 4.2: The overall neural architecture, based on both Levine (2008) and Moustafa et al. (2013). Conditioned (CS) and unconditioned (US) stimuli representations enter the basolateral amygdala (BLA), with CS representations also entering the orbitofrontal/ventromedial prefrontal cortex (OFC/vmPFC). 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.

OXT and VA to the reward system enhances the facilitation of the creation of social bonds, whilst CRH release serves to further stimulate activity in the amygdala-PVNp-LC stress loop, inhibiting social approach.

The OFC and BLA are known to have extensive bidirectional connections, which are implemented in our model within a Deep Belief Network (DBN) (see Appendix B.1) using the architecture shown in Fig. 4.3. We chose to use a DBN here primary due to its power in functioning as a hierarchical associative memory (and thus ability to capture the counter-conditioning processes hypothesised to underlie successful application of the bonding protocols). An RBM (i.e. a single layer DBN) has recently been used for modelling of the psychotherapeutic process (Edalat and Lin, 2014), while a DBN has been used to

model area V2 of the visual cortex (Lee et al., 2008). The Deep Boltzmann Machine (an architecture closely related to the DBN) has additionally been used to model both object-based attention in the visual cortex (Reichert et al., 2011) and Charles Bonnet syndrome (the common experience of complex visual hallucinations in those with blindness) (Series et al., 2010; Reichert et al., 2013).

An RBM is 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 $E(x, h)$, where:

$$E(x,h) = -\sum_{i,j} W_{ij} h_i x_j - \sum_j c_j x_j - \sum_i b_i h_i \tag{4.1}$$

giving joint-distribution over x and h:

$$p(x,h) = \frac{e^{-E(x,h)}}{\sum_{x,h} e^{-E(x,h)}} \tag{4.2}$$

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 (Hinton, 2002) 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 (Hinton et al., 2006), with successive layers of hidden units coming to learn increasingly higher-order features over the hidden unit activations of the previous layer.

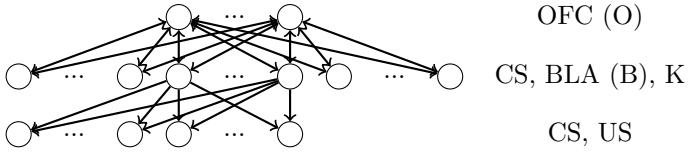

Figure 4.3: BLA-OFC DBN architecture

In both humans and animals, the amygdala has long been known to be crucially important in fear conditioning, in which a previously neutral CS is repeatedly paired with a noxious US that elicits a (fear related) Unconditioned Response (UR). Over time, the 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 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, Vrtička et al. (2012) 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. 4.4). 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. Note that we are modelling associations between CS and reward values in the OFC, but associations between CS and US features in the BLA, which is consistent with findings from reversal learning and devaluation experiments on primates highlighted by Rolls as suggesting that the OFC is more involved in the representation of reward value than the amygdala (Rolls, 2015, p.136).

The activation of the OFC's hidden layer O in turn serves as input to the ITC feed-forward network I (Eq. 4.5), whilst the activation state of the OFC's reward nodes K serves as input to the dmH-PVNp feed-forward network D (Eq. 4.6). 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 broadly consistent with a recent study that found a dominant directional influence from BLA to OFC during a choice task involving both appetitive and aversive food (i.e. primary reinforcer) stimuli (Jenison, 2014), and with evidence suggesting that projections from the amygdala are involved (but not crucially so) in stimulus value coding in the OFC (Rudebeck et al., 2013). This overall structure gives a reasonable (if basic) first approximation to dominant associations and information flows involved in the re-association of classes of social stimuli (previously conditioned as being threatening or stressful in nature) with new (appetitive) reward values.

4.4.1 Stress/Arousal System

The stress/arousal system is intended to encapsulate the basic dynamics of the CeA, PVNp and LC loop described in Levine (2008). 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 NR-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:

$$S(t) = \max(0, C(t) - I(t) - D(t)) \tag{4.3}$$

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:

$$C(t) = \sum_i \tanh\left(\sum_j W_{ij} B_j(t)\right) + \gamma S(t-1) \tag{4.4}$$

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:

$$I(t) = \sum_i \tanh\left(\sum_j W_{ij} O_j(t)\right) \tag{4.5}$$

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:

$$D(t) = \sum_i \tanh\left(\sum_j W_{ij} K_j(t)\right) \tag{4.6}$$

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:

$$\phi(t) = \max\left(1, q - g\left(\frac{q-1}{M_{max}}\right)\right) \tag{4.7}$$

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:

$$M(t) = \sum_i \tanh\left(\sum_j W_{ij} K_j(t)\right) \tag{4.8}$$

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).

4.4.2 Reward System

A prominent account of reinforcement learning in the brain suggests that phasic DA firing signals an appetitive reward prediction error, i.e. unexpected rewards (Niv, 2009). Midbrain DA neurons in the VTA project to the OFC and 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 Abraham et al. (2014) 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 (Milad and Quirk, 2012; Moustafa et al., 2013).

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 (Abraham et al., 2014). 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 (Bromberg-Martin et al., 2010; Hu, 2016). 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 NAc shell, 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). Salience signals are believed to be sent to areas involved in orienting, cognitive processing and general motivation, such as the NAc, dorsal striatum and dorsolateral prefrontal cortex.

In our model, we assume that reward predictions and prediction errors are computed by the OFC (Rolls, 2013, p.316) and signalled to the reward system (see Section 2.6.4). 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 increasing in uniform steps from $-n/2$ to $n/2$, exclusive of zero, then the predicted reward at time t is given by:

$$P(t) = \begin{cases} j - \frac{n}{2} - 1 & \text{if } j \leq \frac{n}{2} \\ j - \frac{n}{2} & \text{otherwise} \end{cases} \tag{4.9}$$

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:

$$F(t) = \phi(t)(R(t) + \delta P(t) - P(t-1)) \tag{4.10}$$

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 Love (2014), 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 Love (2014) for a discussion). Such OXT modulation of DA within the context of maternal and caregiving behaviour has been demonstrated directly in the rat (Shahrokh et al., 2010). In that study, mothers exhibiting higher rates of pup-directed caregiving behaviour expressed higher levels of OXT in the PVN and had increased projections from OXT cells in the PVN to the VTA, and in all mothers direct infusion of OXT

into the VTA increased the DA signal in the NAc. The authors found a positive correlation between increase in DA signal in the NAc and rate of maternal behaviour, but this relatively high increase was abolished with administration of OXT receptor antagonist directly into the VTA. Broadly in line with these findings, in our model OXT can modulate valance attribution by increasing phasic DA firing $F(t)$ for positive unmodulated reward prediction error $R(t) + \delta P(t) - P(t-1) > 0$. For simplicity, we also assume OXT serves to decrease $F(t)$ for unmodulated negative error. While we are primarily concerned with modelling the delivery (and modulation) of positive reward prediction errors (which are proposed to be involved in the counter-conditioning process driving the therapy), detail regarding the effect of OXT on signals representing the delivery of unexpected punishments can be refined in future, once a clearer picture emerges regarding the role for DA in conveying negative reward-prediction errors and the interplay between OXT and DA within this context.

4.4.3 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 (in particular music, and interactions such as singing directed towards inner-child representations) that naturally induce reward in order to drive a process of reinforcement learning. Supporting our proposal, based on the wealth of evidence showing activity in reward related brain areas and dopamine release during listening to music (outlined previously), Gold et al. (2013) investigated the effects of music listening on both accuracy and responsiveness in a learning task. In the training phase participants learned the reward contingencies for pairs of symbols, before a test phase in which participants were required to generalise what they had learned. Subjects heard either self-reported pleasurable or neutral music during the training and test phases of the task, and the accuracy (proportion of trials in which the stimulus with higher reward) and reaction times (amount of time between stimulus presentation and response) were measured. Pleasurable music generally accelerated reaction times during learning, and also accuracy for individuals reporting low prior musical experience. The authors argued for these results to be interpreted in terms of musically-elicited dopaminergic reward prediction errors directly enhancing reward-driven reinforcement learning. In particular, it was proposed that learning was improved by pleasurable music in individuals with low prior music experience as a result of amplified reward-prediction errors, since these individuals are less able to form accurate musical predictions.

Thus, we hypothesise that the inner-child directed singing and passive listening to pleasurable music involved in the Self-Attachment bonding protocols serve, 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 Levine (2008)) 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 (Milad and Quirk, 2012). 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. Raes and De Raedt (2012)).

As a first hypothesis, we consider here a similar vmPFC-ITC extinction inhibitory process to occur as a result of the counter conditioning-like aspects of the Self-Attachment bonding protocols. Our computational model of this process is based on the recent fear extinction model proposed in Moustafa et al. (2013). 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 Moustafa et al. (2013), 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:

$$\triangle W_{ij}(t) = \begin{cases} \alpha_{itc} F(t) O_i(t) I_j(t) & \text{if } F(t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.11}$$

and similarly, BLA-CeA weights according to

$$\triangle W_{ij}(t) = \begin{cases} -\alpha_{cea} F(t) B_i(t) C_j(t) & \text{if } F(t) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

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 protocols. 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 (Raes and De Raedt, 2012).

4.5 Simulations

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 $\alpha = 0.05$, a sparsity target of $d = 0.01$ (with gain $\eta = 0.25$ and decay $\gamma = 0.9$), an L2 weight decay penalty of $\lambda = 0.02$, and batch size 50, for 2000 epochs) using 1-step contrastive divergence, 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-. 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,

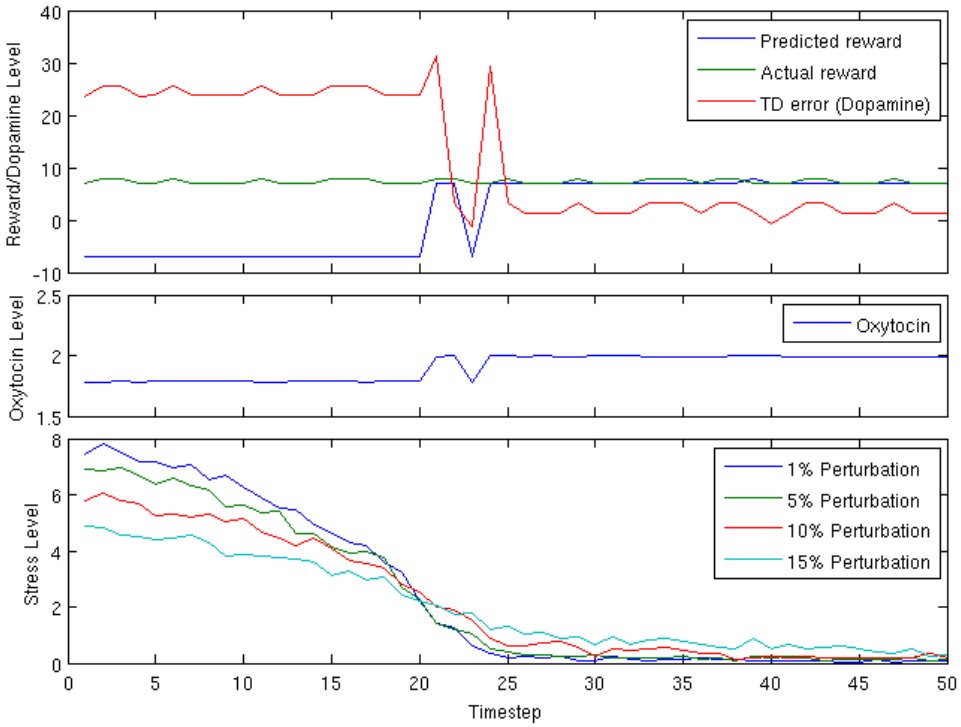

Figure 4.4: **Top**: Predicted/actual reward and dopamine at each time-step of the protocol (x-axis) for a typical simulation run. **Middle**: Oxytocin levels at each time-step (x-axis) of the protocol. **Bottom**: Stress levels at each time-step (x-axis) of a typical simulation (taken as an average over 100 activations) in response to a CS- permuted with 1,5,10 and 15% random noise.

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 high 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 therapy, for a particular CS- (see Fig. 4.5). 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: 4.9). 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. 4.10, and used this to update the weights in either the BLA-CeA or OFC-ITC feed forward networks (according to Eq. 4.11 and Eq. 4.12). 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. 4.4, top). This spike in DA coincides with a spike in OXT (Fig. 4.4, middle), and a large drop in average stress levels associated with the CS- randomly permuted at 1,5,10 and 15% levels (Fig. 4.4, bottom). The effect can be seen in Fig. 4.5, 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 (middle) 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.

4.6 Summary and Future Work

In this chapter we presented a neurobiological hypothesis with regards to the underlying effects of the Self-Attachment bonding protocols, based on a previous neuroanatomical model of OFC-mediated stress and bonding reactivity to social stimuli. We considered how this balance could be driven by both OFC-dmH inhibition on PVNp, and vmPFC-ITC inhibition on CeA. By way of a computational model, we showed how a counter-conditioning procedure (which we linked to application of the Self-Attachment bonding protocols) could be used to drive a re-balance of activity between these circuits, by creating new social stimuli-reward associations in the OFC that drive an increase in OXT-modulated DA release, and a decrease in CRH release (which is involved in the stress response).

In particular, we considered how the pairing of broad classes of social stimuli (that have previously been conditioned as threatening in nature) with additional, naturally induced reward (which is proposed to result from activities associated with the protocols) might result in reward prediction errors, originating in the OFC and encapsulated in DA firing. Building on a previous model of fear conditioning and extinction, DA-encapsulated reward prediction errors were proposed to drive strengthening of vmPFC-ITC inhibition on the CeA, whose output drives activity in autonomic and endocrine systems that have been found to be relatively overactive in individuals with insecure attachment classifications. Based on recent evidence we considered here a role for OXT in modulating DA (Love,

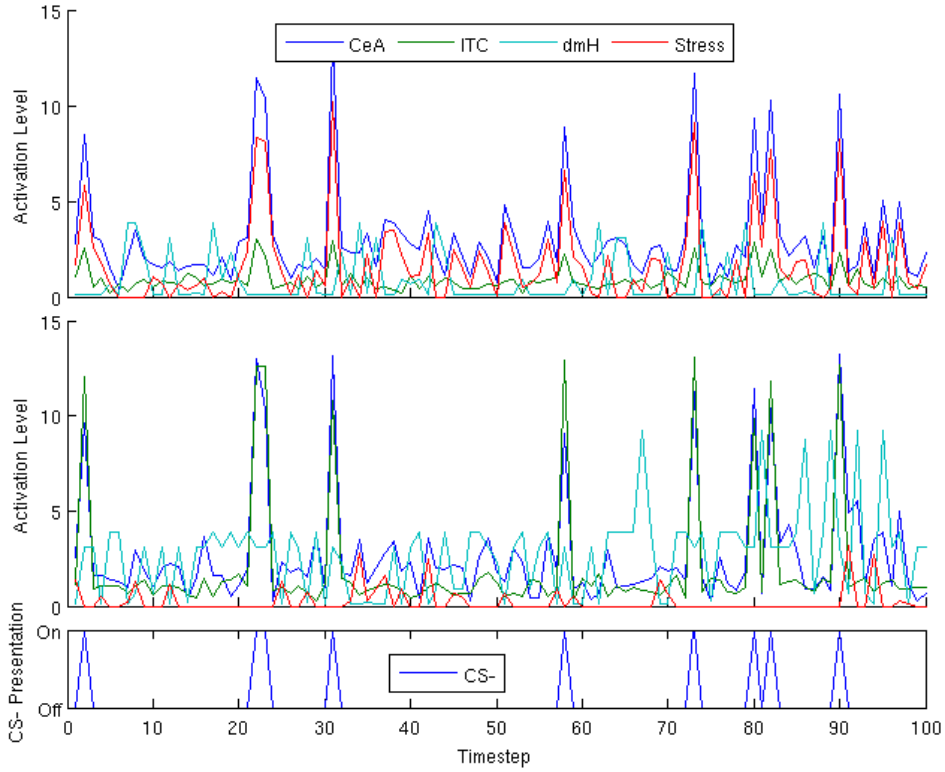

Figure 4.5: Stress circuit reactivity for the counter-conditioned CS- before and after bonding protocols have been applied. **Top**: Before application of the protocol, presentation of the the CS- at random discrete time-steps is associated with large CeA and stress spikes (red). **Middle**: After counter-conditioning, ITC and dmH inhibition reduces associated stress reactivity (red) in response to CS-. **Bottom**: Time-steps at which CS- was presented in both the before- and after-application testing conditions. In all other time-steps a random stimulus was presented.

2014), so that in our model the strengthening of the vmPFC-ITC pathway was driven by reward-prediction errors that were enhanced by OXT release, however we note that DA might also have a modulating effect on OXT (Love, 2014). As we discussed, the role of DA in signalling negative reward prediction errors remains somewhat inconclusive, and the role of DA encapsulated reward-prediction errors in the OFC/vmPFC has also been questioned (since the OFC is itself thought to compute and signal prediction errors, it may be that DA errors are used mostly for habit learning in the striatum (Rolls, 2013, p.316)). Further complicating the matter, there is also evidence to suggest that DA is released during the stress response (Pruessner et al., 2004) (possibly related to the firing of salience signalling DA neurons that we discussed previously). Future work can refine the precise role of DA in the model as functionality becomes clearer with further research.

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 (Moustafa et al., 2013) or short term memory effects of bonding (Levine, 2008). Neither did we consider VA, which is believed to play a role in selective attention along with memory and stress modulation (Levine, 2008; Bachner-Melman and Ebstein, 2014). Future work should consider such effects.

5. Self-Attachment Empathy Protocols

In this chapter we build on a neuroanatomical model of how empathic states can motivate caregiving behaviour, via empathy circuit-driven activation of regions in the hypothalamus and amygdala which in turn stimulate a mesolimbic-ventral pallidum pathway, by integrating findings related to the perception of pain in self and others. Based on this we propose a network to capture states of personal distress and empathic concern, which are particularly relevant for psychotherapists conducting attachment-based interventions. This model is then extended for the case of Self-Attachment therapy in which conceptualised components of the self serve as both the source of and target for empathic resonance, and we consider how such states might optimally generate motivation for self-directed bonding (as described in the previous chapter). We simulate our model computationally, and discuss the interplay between the bonding and empathy protocols of the therapy.

5.1 Introduction

As we have discussed, early insecure attachment experiences are 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, and create a predisposition for the future development of a range of psychopathological disturbances. On the other hand, the caregiver's mimicry and resonance of the emotional state of the infant in optimal early attachment experience is believed to underlie development of the ability to empathise with others. In attachment-based psychotherapies, which view the client-therapist relationship as an attachment bond, empathic resonance is used as a tool both to communicate attunement to the client, and to generate internal motivation for prosocial stances that guide the client towards states of positive emotion and internal regulation. While engaging in such resonance effectively and avoiding negative states of distress and burnout depends crucially on the ability to differentiate self from other, prosocial motivation resulting from an empathic

state is stronger the more closely related the other is deemed to be to the self.

In the previous chapter we hypothesised that Self-Attachment protocols involving the formation of a self-directed bond would stimulate neural plasticity in a network spanning the OFC, vmPFC and amygdala, increasingly containing stress circuitry and facilitating the release of hormones related to prosocial tendencies. In this chapter we build on this work by integrating neuroscientific findings relating to empathy, compassion and the self-other distinction, to present a computational model of how empathic states can provide additional motivation for the application of these bonding protocols. Within the context of findings relating to modulation of the empathic response, we argue that the particular conceptualisation that we use for the part of the self deemed as requiring attachment intervention aids in the generation of this motivation.

5.2 Empathy

Work in empathy distinguishes between a number of related yet distinct phenomena and states. We broadly follow the definitions set out in Gonzalez-Liencres et al. (2013), with states of "emotional contagion", "personal distress", "emotional empathy" and "sympathy" being relevant for our initial discussion here. A state of emotional contagion within the self involves a mirroring of the emotional state of another. Crucially, this mirroring occurs within a context of weak or absent self-other distinction, such that the emotional state of the other is perceived as belonging to the self, and is not necessarily attributed to the other. In cases in which the mirroring is of a negatively-valenced emotion belonging to the other, emotional contagion can result in emotional distress within the self ("personal distress"), driving egoistic withdrawal responses in which the self withdraws from its surrounding environment and stimuli triggering the state, in order to relieve the symptoms of the distress.

Similarly to emotional contagion, emotional empathy is a state that arises from a mirroring of the emotional state of another. In contrast, however, this mirroring is accompanied with a strong self-other distinction. In other words, there is both an experience of the other's emotional state within the self, and knowledge that this emotional state originates in the other rather than the self. Since there is this self-other distinction, empathic states involving negatively-valenced emotion can drive prosocial motivation aimed at relieving the perceived distress of the other (we discuss this in more detail in Section 5.2.2.1). The term "sympathy" is sometimes loosely used to describe prosocial motivation arising from such an empathic state (we prefer to use the phrase "empathic concern" here).

5.2.1 Empathy in Psychotherapy

The ability to empathise emotionally is recognised as an integral skill for professionals in a broad range of health-related occupations, including social work (Gerdes and Segal,

2011) and, increasingly, medicine (Halpern, 2003). The role of empathy in psychotherapy dates back to Carl Rogers, who proposed that a continuous effort to empathically attune with the client, along with an unconditional positive regard for them, are necessary in order for therapeutic change to occur (Rogers, 1957). Heinz Kohut's psychoanalytic self-psychology, developed from the 1960s onwards, holds that almost all psychopathology is rooted in empathic failure on the part of the parent in childhood, and the therapist serves as an empathic self-object in order to resume development towards maturity (Kohut, 1959; Baker and Baker, 1987). Empathy is now widely accepted as being important for the formation of an effective working relationship between the therapist and client. For example, the influential empathy cycle model defines a framework under which communication of the therapist's ever-strengthening empathic resonance helps to guide the client towards more accurate expression of their internal experience (Barrett-Lennard, 1981).

From an evolutionary perspective, empathic motivation has been argued to have its ultimate roots in selective pressures to care for offspring, identify with in-groups, and exclude out-groups (Zaki, 2014). The mechanisms driving empathic responses may have initially evolved in order to fulfil parental care (i.e. attachment) responsibilities, while only later being "co-opted" in order to increase survival prospects of in-groups (Gonzalez-Liencres et al., 2013), suggesting a primacy for attachment in the empathic experience. The infant's experiences of emotional mimicry and resonance within early attachment experiences have been proposed to underlie the later development of capabilities for empathy towards others (Decety and Meyer, 2008), and a number of studies have shown increased tendencies for self-reported compassionate states and prosocial behavioural responses with increasing attachment security (Mikulincer and Shaver, 2005).

Empathy plays a particularly important role in attachment-based psychotherapies, which are the main focus of this chapter. Attachment-based psychotherapies view the client-therapist relationship as an attachment bond, with the therapist aiming to provide a secure safe-haven for the relief of the client's distress (Obegi, 2008). The therapist uses empathic attunement and contingent communication as tools in order to help the client explore painful feelings and memories, and engages in interactive regulation of the client's emotion in order to guide them towards alternative ways of feeling and acting (Wallin, 2007, p.196). Thus, in order to cultivate secure attachment, the client must additionally experience the therapist as being both able and willing to help them cope with their difficult feelings, and the therapist must engage in a process of interactive regulation of the client's internal emotional state.

Empathic resonance is an effective tool for therapists and other health professionals to communicate attunement and understanding to the client, and generate prosocial motivation to guide interactions towards desirable outcomes. However, the demands of sharing the client's negative emotional state can potentially prove to be detrimental to the therapist's own mental health. Therapists and other health workers engaging in empathic reso-

nance without a sufficiently strong self-other distinction can become susceptible to secondary trauma (the second-hand exposure to traumatic events), burnout (overwhelming emotional exhaustion), alexithymia (dysfunction in emotional awareness), and states of personal emotional distress (Wagaman et al., 2015; Zenasni et al., 2012; Gleichgerrcht and Decety, 2013).

5.2.2 Neuroscience of Empathy

Imaging studies have uncovered a core empathy-for-pain network (Engen and Singer, 2013) involving the AI (an area involved in a range of emotion-related functions and experiences, including interoceptive awareness (Critchley et al., 2004; Zaki et al., 2012), emotional states induced by imagery or recall (Phan et al., 2002), and affective states that arise during social interaction, particularly relating to notions of fairness and cooperation (Lamm and Singer, 2010) and attachment functions including recognition of the mother's own infant (Noriuchi et al., 2008)) and Anterior Midcingulate Cortex (aMCC) (a part of the ACC, which is involved in a range of social and emotion-related functions including theory of mind cognition and the perception of fear (Baird et al., 2006) and the detection and appraisal of social exclusion (Kawamoto et al., 2015)), with lesion-based data furthermore suggesting that the AI is crucial for empathy (Gu et al., 2013). The considerable overlap in regions activated during the experiencing of pain oneself, and when perceiving others to be in pain (particularly in the AI and aMCC) has led to a shared-network hypothesis of empathy (Lamm et al., 2011). A number of factors, including perceived innocence (the degree to which the recipient of the empathic response is deemed to be responsible for their fate (Fehse et al., 2015)) [1], closeness (in terms of subjective similarity, or membership of an in-group (Hein et al., 2010)) and fairness (i.e. tendency to cooperate (Singer et al., 2006)) of the other have been found to modulate the strength of behavioural and neural empathic responses (Engen and Singer, 2013; Numan, 2014).

5.2.2.1 A Neuroanatomical Model of how Empathic States can Motivate Caregiving Behaviour

Numan has recently proposed a neuroanatomical model (Fig. 5.1)for how empathic states can give rise to caregiving behaviour (Numan, 2014, p.278). A large number of studies are cited by him as justification for this model: we overview the most important of those here, along with some additional studies that further support his architecture. We refer to Numan (2014) for a more comprehensive motivation of the model (see also Numan and Young (2016) and Numan (2017) for details on underlying circuits involved in parental caregiving behaviour).

[1] The authors used the term "compassion" to refer to a state more closely related to what we call "emotional empathy" and "sympathy"/"empathic concern".

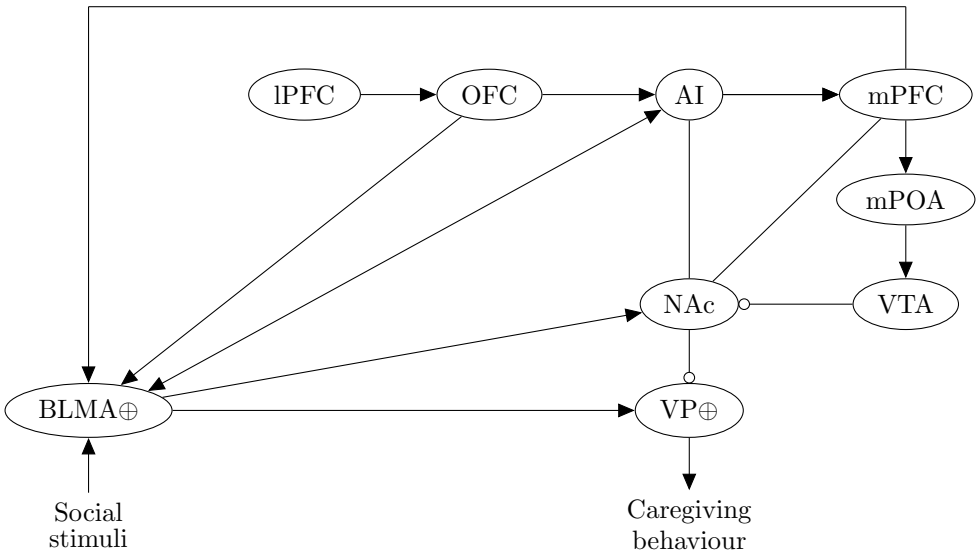

Figure 5.1: Numan's neuroanatomical model for how empathic states can motivate caregiving behaviour. Ellipses represent neural populations, arrows are excitatory synapses, and circles inhibitory (the nature of the connections between AI and NAc, and mPFC and NAc within this context is currently unknown). See text for details.

Under the model, projections from the Basolateral (Basal) and Basomedial (Accessory Basal) nuclei of the Amygdala (BLMA) to the AI are proposed to facilitate the creation of the shared empathic state (Hurlemann et al., 2010). It is proposed that, when AI activation levels are high enough, this region stimulates the mPFC (which includes the ACC, and thus aMCC, in Numan's definition) which in turn activates prosocial pathways resulting in caregiving behaviour. Supporting this claim, empathy-related activity in the AI and mPFC has been found in response to viewing an unfairly treated other (in the form of exclusion from a ball-tossing game) with activity correlating with subsequent (spontaneous) prosocial behaviour (Masten et al., 2011) (similar results were reported by Mathur et al. (2010) in response to viewing the suffering of another perceived as being in-group).

In particular, it is proposed by Numan that more ventral parts of the mPFC might be activated most strongly during the empathic concern response. There is evidence to suggest that the mPFC encodes a ventral-dorsal gradient for self-other reflection (Denny et al., 2012) that is also sensitive to self-relatedness of the other, with reflection on others perceived to have high degrees of self-relatedness (e.g. in terms of similarity, familiarity and closeness) correlating with more ventral mPFC activation, and reflection on non-self-related stimuli (e.g. a publicly known but personally unknown other) correlating with more dorsal mPFC activation, and that this gradient is sensitive within the context of empathy.

Using fMRI, Meyer et al. (2012) investigated the neural correlates of individuals viewing the social exclusion of either a friend or stranger, and found that observation of the friend was associated with relatively higher activation in the vmPFC. There is also evidence to suggest that more ventral parts of the mPFC activate in response to a perceived innocence of the other within the context of empathy: Fehse et al. (2015) found higher self-reported empathic concern, along with elevated activation of (more ventral parts of) the mPFC in response to stimuli depicting individuals that had experienced an unfortunate fate but were described as being innocent rather than responsible for their situation. These studies are consistent with Numan's proposal that more ventral parts of the mPFC are associated with relatively high subsequent activation in caregiving pathways in response to empathic resonance, according to a ventral-dorsal encoding of self and other representations in this area. An alternative perspective on mPFC functioning (Nicolle et al., 2012) proposes that activity in more ventral parts of the mPFC is associated with agent-independent choice value for executed behaviour (i.e. value for self/other when the self chooses on behalf of themselves/the other), whereas activity in more dorsal areas of the mPFC is associated with modelled value (i.e. value for the other when the self chooses on behalf of the self, or value for the self when the self chooses on behalf the other). According to this view, we might similarly expect value for empathic concern behaviour (towards an other perceived as being in distress) as executed by the self to be represented by activity in vmPFC.

The mPFC is proposed to initiate caregiving behaviour via activation of the Ventral Pallidum (VP) along both a stimulatory mPFC-BLMA-VP pathway, and a dis-inhibitory mPFC-mPOA-VTA-NAc-VP pathway (with inhibition of the NAc serving to release the VP from BLMA-mediated inhibition, thus potentiating caregiving behaviour). VP projections to midbrain locomotor regions have long been proposed to be involved in the translation of limbic motivation signals into motor output (Mogenson, 1987; Brudzynski et al., 1993; Jordan, 1998), and increasing activation in ventromedial parts of the VP (in response to NAc shell-mediated disinhibition) have been associated with goal-directed behaviour (Root, 2013; Numan, 2014, p.24-25).

The first of these pathways (mPOA-VTA-NAc-VP) has been identified as crucial for the onset and maintenance of maternal and caregiving behaviour based on a large body of animal (lesion) studies, and it is proposed that these same pathways also underlie the motivational aspects of empathic concern in humans (Numan, 2014). In particular, in animals, hormones and neurotransmitters (including OXT) act on the mPOA (a part of the anterior hypothalamus) resulting in projections from this region to the mesolimbic DA system in response to infant stimuli. The subsequent DA release from the VTA serves to inhibit the NAc, which releases the VP from inhibition and allows it to be responsive to infant stimuli-mediated projections from the BLMA. In animal studies, activation of this circuit has been found to result in motivated responses that attract a mother to her young,

and disruption of projections from the mPOA to the mesolimbic pathway halt this attraction and associated caregiving behaviour (Numan et al., 2005; Numan, 2014).

Since inactivation of mPFC projections to mPOA are also known to disrupt pup retrieval in rats (Numan, 2014, p.191), the mPFC is suggested as a crucial link between areas involved in the formation of empathic states and the dis-inhibitory mPOA-mesolimbic DA pathway identified as being crucially involved in caregiving behaviour. A number of studies are presented as evidence for involvement of this same pathway in prosocial and caregiving behaviours arising from empathy in humans, one of which is the investigation by Moll et al. (2006) that used fMRI to uncover the correlates of prosocial behaviour related to giving (in the form of a monetary donation) and receiving reward. Whilst both the receiving and giving of reward correlated with activity in the VTA and NAc, donation correlated additionally with activation in the preoptic area and Brodmann Area (BA)25 (a part of the vmPFC), and increased activation in the NAc. This suggests that the vmPFC and preoptic area might interact with the mesolimbic DA system within the context of prosocial acts that are interpreted as being rewarding to the recipient, as is the case in empathic concern responses. Interactions between the preoptic area and the mesolimbic DA system have furthermore been observed during the simulation of prosocial acts: in the fMRI study conducted by Decety and Porges (2011), participants viewed scenes including those involving individuals easing the pain of others, and were then asked to mentally simulate being the performer of such acts. In simulatory but not viewing scenarios, the authors found increased activation in the preoptic area and NAc, plus increased functional connectivity between the amygdala and NAc and VP. Using simulation theory (Hesslow, 2002), the authors argued that actual overt prosocial action would be expected to activate the same pathways as were found to be activated during these simulations. This suggests that activation of the same mPOA-mesolimbic DA pathway might drive both overt and imagined empathic concern.

The BLMA is anatomically positioned to relay the olfactory and somatic sensory inputs from infants (important for maternal and caregiving behaviour) to the NAc and VP, and suppression of BLMA activity and its input to the VP have been found to disrupt maternal and caregiving behaviour in rats (Numan et al., 2010). Within the context of empathic concern responses in humans, it is suggested in the model that projections from the mPFC to the BLMA are a significant pathway by which the type of social stimuli that can gain access to positively valent neurons in the BLMA and VP are regulated, allowing in-group members priority access to prosocial circuits. Finally, it is suggested that the Lateral Prefrontal Cortex (lPFC) and the OFC might also be involved in empathic concern responses in the form of cognitive modulatory influences via projections to the BLMA and AI. Later on we will consider in particular a possible role of the mOFC within this framework related to findings from the neuroscience of compassion and the hypothesised effects of the Self-Attachment bonding protocols.

5.2.2.2 Neural Correlates of Self and Other Pain Attribution

In order to extend Numan's model to additionally consider a state of personal distress (which behaviourally has been found to induce egoistic withdrawal as opposed to caregiving behaviour), we consider now the neural correlates of self and other pain attribution. Singer et al. (2004) used fMRI to asses brain activity while volunteers either experienced a painful stimulus (electrode) themselves, or received a cue indicating that their loved one (present in the same room) was receiving a similar painful stimulus. Their experiment followed on from a number of previous studies which had consistently shown activation in a pain network spanning the secondary somatosensory cortex (an area thought to be involved in the processing and integration of both painful and nonpainful somatosensory stimuli that are salient for higher-order functions such as memory and attention (Chen et al., 2008)), insular cortex, ACC, the cerebellum and supplementary motor areas (which are involved in movement, motor control and adaptation (Glickstein, 2007)), and (less consistently) the thalamus (which relays and controls the flow of information to the cortex (Sherman and Guillery, 2002)) and primary somatosensory cortex in response to painful noxious stimuli. The authors found that areas including the bilateral AI, rostral (perigenual) ACC, brainstem and cerebellum activated in both self- and other-pain conditions, whereas activity in the left Posterior Insular (PI)/secondary somatosensory cortex and right mid insular (areas otherwise implicated in the interoceptive (Craig, 2011) and sensory-discriminatory (Pavuluri and May, 2015) aspects of pain), caudal ACC and sensorimotor cortex (an area that is thought to be involved in both imagery and execution of motor function (Stippich et al., 2002)), comprising a pain network that has commonly been found to activate in response to painful noxious stimuli, was specific to the self-pain condition. The authors concluded that empathising with others in pain does not involve activation of the whole of the pain network, and that empathy is mediated by areas involved in representing the affective (but not sensory) aspects of pain.

Similarly, Zaki et al. (2007) conducted an fMRI study in order to determine the neural correlates of pain when experienced by the self, and contrasted this with the correlates of pain perceived to be experienced by another. During the self-pain condition, a thermal noxious stimulus was administered to the individual, whilst under the other-pain condition participants watched videos of other people receiving pain-inducing injuries. Based on this data, contrast (Ochsner et al., 2008) and functional connectivity (Zaki et al., 2007) analyses were performed.

In the contrast analysis (which looked at relative activation levels) a number of regions were found to have common activation during both self and other pain conditions. These regions include the aMCC and AI (regions previously implicated in the shared-network hypothesis of empathy), along with the middle frontal and premotor gyri, and the dorsal thalamus. The AI, PI and middle frontal gyrus were found to be more activated for self-

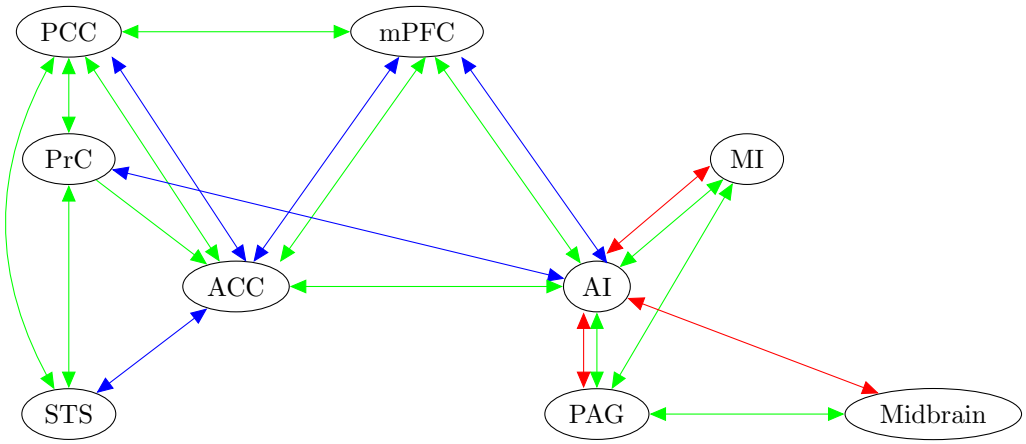

Figure 5.2: Distinct neural networks proposed for self and other pain (Zaki and Ochsner, 2011). Blue: more functionally connected during other pain, Red: more functionally connected during self pain, Green: anatomical connections. See text for details.

pain as opposed to other-pain, whereas in contrast greater activation was found in regions including the precuneus (an area that has been implicated in functions including the experience of a sense of agency (Cavanna and Trimble, 2006) and the recall of autobiographical memories involving familiar others (Maddock et al., 2001)), OFC and amygdala for other- as opposed to self-pain. The functional connectivity analysis revealed distinct circuits involved in the perception of pain in the self and other. Whilst the AI and aMCC were found to be functionally connected to each other during both self and other pain, these regions showed increased functional connectivity with more posterior (mid) areas of the insular during self-pain, along with the periaqueductal gray (which is involved in pain modulation and sensations associated with aversive emotions (Mai and Paxinos, 2011, p.367)) and areas in the midbrain. In contrast, during other-pain these regions were more functionally connected with a network comprising mPFC, precuneus, Posterior Cingulate Cortex (PCC) (a region that, similarly to the precuneus, is thought to be involved in the recall of autobiographical memories involving familiar others (Maddock et al., 2001)) and Superior Temporal Sulcus (STS) (which has been implicated in a variety of social processes including theory of mind (Beauchamp, 2015)).

5.3 A Model of Self-Other Representation-Mediated Personal Distress and Empathic Concern

Here, we have attempted to extract key findings from the studies by Zaki et al. (detailed in Section 5.2.2.2) with respect to self- and other-pain networks in order to model states of personal distress and empathic concern. In particular, we consider two points of current injection (Self and Other), with the self-pain condition involving current injection to neurons in the PI and AI, and other-pain involving current injection into neurons in the aMCC and mPFC.

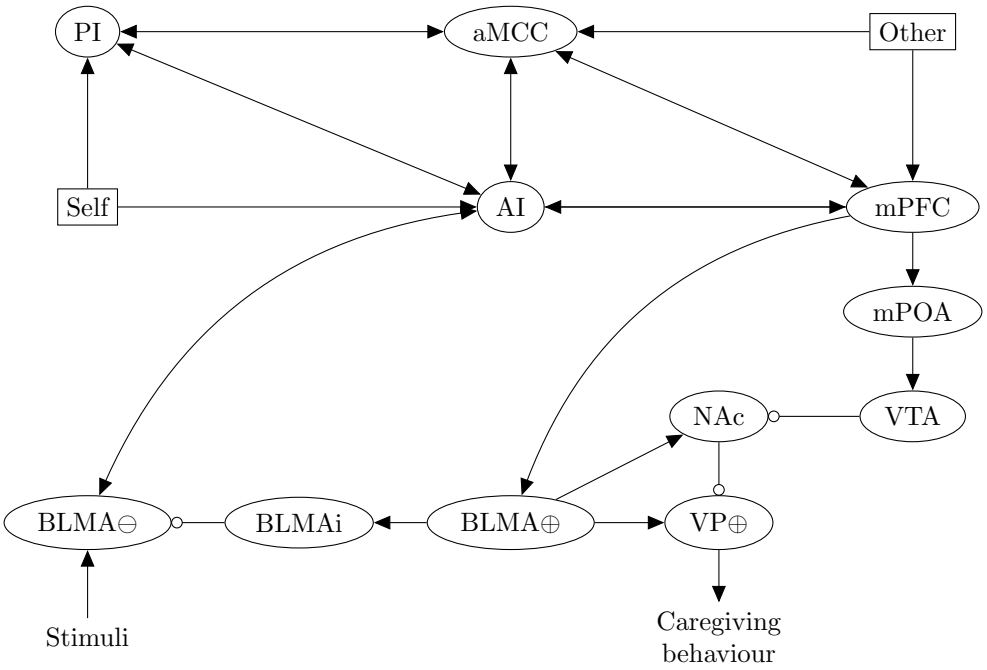

Figure 5.3: Model capturing how self-other distinction can mediate empathically-motivated caregiving behaviour in response to pain-inducing stimuli. Ellipses represent neural populations. Arrows are excitatory, and circles inhibitory (intra-group excitatory/inhibitory connections not shown). Rectangles (Self, Other) are points of current injection, proposed to represent activity in distinct neural networks involved in the perception of self and other pain. $\oplus$ represents positively-valent neurons (i.e. populations that subsequently activate prosocial/approach pathways) and $\ominus$ negatively-valent neurons (that subsequently activate withdrawal/avoidance pathways), which loosely correspond to positively and negatively valent emotional states, respectively. See text for details.

According to Zaki's model, self-pain involves increased activation in both the AI and PI. To capture this, during the self-pain condition, we inject current into neurons in both

the AI and PI. These injections to the insular might feasibly represent projections along a midbrain-periaqueductal gray pathway, since these regions are both anatomically and (during self-pain) functionally connected to this region, and/or via a pathway involving the thalamus and hypothalamus, which showed preferential activation for self- as opposed to other- referential processing in a separate meta-analysis (Zaki and Ochsner, 2011, p.21) and have strong connectivity with this region.

During the other-pain condition, we input current to neurons in the aMCC and mPFC. Input to the aMCC may represent inputs from the precuneus in Zaki's other-pain network, since this region was found to be both more active, and more connected to the aMCC, during other-pain, and there are known anatomical connections between these regions. Input to the mPFC during the other-pain condition represents activation of neurons preferentially encoding other-referential processing, and we propose that activation of a subset of these neurons encoding close-others will most strongly drive activation in the caregiving pathways described by Numan. This current injection might also represent increased activation in precuneus, which is anatomically connected with the mPFC via the PCC.

5.3.1 Implementation Details

We use Izhikevich neurons (Izhikevich et al., 2003) with current-based synapses (i.e. a postsynaptic current into a neuron is proportional to the weight between the presynaptic and postsynaptic neurons), with weights in the network tuned in order to capture the particular network states that we wish to model. We use the full 9-parameter Izhikevich model, which is able to imitate the firing properties of a large range of biological neuron types. The neuron membrane potential (voltage) v for a given current I is described by:

$$\frac{dv}{dt} = (1/C)(k(v - vr)(v - vt) - u + I) \tag{5.1}$$

where $I = I^{syn} + I^{ext}$ is the total current input, i.e. the sum of all synaptic and external currents to the neuron membrane. The synaptic currents in our model are proportional to synaptic weights, such that the total synaptic current I_j^{syn} at postsynaptic neuron j resulting from spikes at presynaptic neurons i at some particular point in time is given by:

$$I_j^{syn} = \sum_{i=1}^{N} s_i w_{ij} \tag{5.2}$$

where $s_i = 1$ if presynaptic neuron i is spiking, or 0 otherwise (neuron i spikes if its membrane voltage has exceeded the peak value, i.e. $v_i > vpeak$). In Eq. 5.2, w_{ij} is the synaptic weight between i and j, and N is the total number of presynaptic connections onto postsynaptic neuron j. External currents I^{ext} are additional currents injected into neurons as according to the phases of our model. The recovery variable u is described by:

	a	b	c	d	k	C	vr	vt	vpeak
RS	0.03	-2	-50	100	0.7	100	-60	-40	35
FS	0.15	8	-55	200	1	20	-55	-40	25
MSN	0.01	-20	-55	150	1	50	-80	-25	40
IB	0.01	5	-56	130	1.2	150	-75	-45	50

Figure 5.4: Izhikevich parameter values for the four types of neuron (RS, FS, MSN and IB) used in our model.

$$\frac{du}{dt} = a(b(v - vr) - u) \tag{5.3}$$

and there is an instantaneous reset of the membrane potential v, and a stepping of the recovery variable u, whenever v reaches a value $vpeak$:

$$v(v > vpeak) = c \tag{5.4}$$

$$u(v > vpeak) = u + d \tag{5.5}$$

The 9 open parameters of the model are thus a, b, c, d, k, C, vr, vt, $vpeak$, the setting of which define different types of neurons. The values we use in our simulation for our four neuron types (Regular Spiking (RS), Fast Spiking (FS), Medium Spiny Neuron (MSN) and Intrinsically Bursting Neuron (IB)) are given in Fig. 5.4. Parameters for RS, MSN and IB come from the models in Izhikevich and Moehlis (2008), and parameters for FS neurons come from the model in Izhikevich and Edelman (2008).

5.3.2 Regional Detail

Table 5.1 details the number of neurons, neuron types, and target connection probabilities and weights for all neural groups (regions) in our model. The number of neurons is derived from human non-clinical estimates, while neuron types and connection probabilities are based on animal (mainly rodent) data, due to a lack of human data. Weights describe the connection weight between any two neurons for a given presynaptic and postsynaptic group pair. Although this is just one of many sets of weights that would achieve our proposed network states, our focus here is on motivating the hypothesised neural dynamics underlying the therapeutic empathy, rather than on the particular weight magnitudes found in order to achieve this. Pre- and post-synaptic groups are connected randomly according to given probabilities (except for mPFC connections to mPOA and Positively valenced neurons of the Basolateral and Basomedial nuclei of the Amygdala (BLMA$\oplus$), which are connected according to the negative-binomial distribution with $r = 7$ and $p = 0.0025$ (Fig. C.1) intended

to capture a simple model of the ventral-dorsal gradient for self-other representations in the mPFC [18].

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
BLMA⊖		76 [2]	Excitatory (RS)			
				BLMA⊖	0.855 [3]	150
				BLMAi	5 [3]	350
				AI_exc	0.019 [3]	300
				AI_inh	0.008 [3]	25
BLMAi		13 [2]	Inhibitory (FS)			
				BLMAi	5 [3]	350
				BLMA⊖	10 [3]	1750
BLMA⊕		76 [2]	Excitatory (RS)			
				BLMA⊕	0.855 [3]	60
				BLMAi	5 [3]	225
				NAc	0.674 [3]	3000
				VP	0.674 [3]	550

[2]Total number of neurons taken as sum of estimates for basolateral and basomedial nuclei in García-Amado and Prensa (2012), with 15% assumed fast-spiking inhibitory interneurons (Inhibitory interneurons of the Basolateral and Basomedial nuclei of the Amygdala (BLMAi)) and the remaining regular-spiking excitatory principal (pyramidal) neurons (Sah et al., 2003; Zhang et al., 2013). Excitatory neurons are split evenly between groups with positive (BLMA⊕) and negative (Negatively valenced neurons of the Basolateral and Basomedial nuclei of the Amygdala (BLMA⊖)) valence. Total number of neurons in BLMAi is scaled by 0.5 (under crude assumption that interneurons target positively/negatively valenced excitatory cells in equal proportions).

[3]As in the neocortex (see footnote 14), we assume that 50% of targets of amygdala excitatory (exc) neurons are local, and 50% long-range, with local targets split evenly between excitatory and inhibitory interneurons (inh) neurons and long-range neocortical targets split so that 90% target exc neurons and 10% inh neurons (Lewis et al., 2002; Melchitzky et al., 2001). Furthermore, 100% of targets of amygdala inh neurons are taken to be local (i.e. within BLMA). Woodruff and Sah (2007) found basolateral amygdala intraconnectivity probabilities (for pairs within intersomatic distance $120\mu m$) of inh→inh=26%, inh→exc=50% and exc→inh=27.5%: assuming more dense local connectivity, we thus set exc→inh=inh→inh=5%, inh→exc=10%, from which the remaining probabilities are calculated.

VP		9 [4]	Inhibitory [5] (FS)			
NAc		184 [6]	Inhibitory [7] (MSN)			
				NAc	2.5 [8]	25
				VP	10 [9]	25
VTA		3 [10]	Dopamine (IB)			
				NAc	1.85 [11]	10
mPOA		20 [12]	Excitatory [5] (RS)			

[4]Total estimate of 350,000 VP neurons (Pakkenberg, 1990) divided by 2, to give a crude estimate of the total number of positively valent neurons. NAc-mediated γ-Aminobutyric Acid (GABA)-ergic VP projections to the mesencephalic locomotor region have long been proposed to be involved in the translation of limbic motivation signals into motor output (Jordan, 1998; Mogenson, 1987; Brudzynski et al., 1993). Both increasing and decreasing levels of activation in the VP have been associated with goal-directed behaviour (Numan, 2014, p.25-26) such that there are likely to be two distinct pathways, with inhibition of NAc shell acting to disinhibit (ventromedial) VP in the first, and stimulation of NAc core inhibiting (dorsolateral) VP in the second (Root, 2013). Since we are concerned primarily with the NAc shell-VP pathway here, we consider an increased firing rate to correspond to increased motivation for caregiving behaviour. We only consider GABAergic neurons, which constitute approximately 80% of all VP neurons (Gritti et al., 1993).

[5]We default to RS (for excitatory) and FS (for inhibitory) neurons when type is unknown (VP, mPOA).

[6]Based on total neuron estimate of 7285654 (control) from Wegiel et al. (2014).

[7]We only consider MSN cells, which are thought to account for 95% of NAc neurons (Shi and Rayport, 1994). Number of neurons is scaled accordingly.

[8]Tecuapetla et al. (2007) report a connection probability of 13% for MSN pairs with intersomatic distance within $100\mu m$. Taking into account BLMA connection probability scaling (see footnote 3), and assuming more dense local connectivity, we set the connection probability to 2.5%.

[9]Default connection probability of 10% (50% for mPOA-VTA) used when actual probabilities are unknown.

[10]Total number of human DA neurons estimated to be 450000 (German et al., 1983), of which 15% are in the VTA (Düzel et al., 2009) which gives 67500 VTA DA neurons. 55% of all VTA neurons are DA neurons (Margolis et al., 2006) giving 122727 total VTA neurons. We use the estimate of total number of neurons for the number of DA neurons, due to the otherwise very low number of DA neurons in the model.

[11]Based on the midpoint of the estimate in Bolam and Pissadaki (2012), but re-calculated using the more accurate rat nucleus accumbens volume of $6mm^3$ (McClure et al., 2004).

[12]Volume used is human anterior-superior hypothalamus (Makris et al., 2013), which includes preoptic area. Density is averaged across the four interstitial nuclei of the mPOA (Byne et al., 2000). A majority of mPOA neurons are GABAergic (Lonstein and De Vries,

				mPOA	10 [9]	25
				VTA	50 [9]	1750
AI						
	AI_exc	6212 [13]	Excitatory [17] (RS)			
				AI_exc	3.259 [14]	300
				AI_inh	13.036 [14]	50
				aMCC_exc	2.601 [14]	75
				aMCC_inh	1.157 [14]	25
				mPFC_exc	2.601 [14]	90
				mPFC_inh	1.156 [14]	100
				PI_exc	2.602 [14]	95
				PI_inh	1.155 [14]	70
				BLMA⊖	10 [9]	10
	AI_inh	1553 [13]	Inhibitory [17] (FS)			
				AI_exc	12.5 [15]	200

2000) and they might potentially be involved if projections to the VTA inhibit VTA GABAergic interneurons such that VTA DA neurons are released from local inhibition (Michael Numan, personal communication). Here we only consider glutamatergic neurons, since they form the majority in the central mPOA region that is thought to be particularly important for maternal behaviour (Tsuneoka et al., 2013). We use the total neuron number estimate due to a) relatively small size of mPOA and b) unknown precise overall proportion of glutamatergic neurons in mPOA.

[13]AI volume given in Makris et al. (2006). Density comes from BA13, in adjacent PI (Semendeferi et al., 1998) .

[14]We assume that 50% of targets of neocortical excitatory are local, and 50% long-range (actual proportions unknown). Of the local targets, we assume that these are split evenly between excitatory and inhibitory neurons (Lewis et al., 2002), whilst for long-range targets in other neocortical areas, we assume that 90% target excitatory neurons and 10% inhibitory neurons (Melchitzky et al., 2001). The average number of outgoing synapses per excitatory neocortical neuron is assumed to be the same as for inhibitory neocortical interneurons, from which connection probabilities are calculated.

[15]We assume that 100% of targets of neocortical inhibitory interneurons (inh) are local, with 95% targeting local excitatory (exc) neurons and 5% inh neurons (fast-spiking local inh-inh connections are relatively rare and sparse (Rudy et al., 2011; Markram et al., 2004)). Packer and Yuste (2011) found a local connectivity probability (for intersomatic distances less than $200\mu m$) of inh→exc of 62% (averaged across regions), with dense local connectivity that decreased as a function of intersomatic distance (probability became zero for distances greater than $450\mu m$). Based on both this and data for the basolateral amygdala [3] we set local connection probability inh→exc to 12.5%, from which other connection probabilities are calculated.

				AI_inh	2.632 [15]	750
mPFC						
	mPFC_exc	9574 [16]	Excitatory [17] (RS)			
				mPFC_exc	3.315 [14]	150
				mPFC_inh	13.261 [14]	50
				aMCC_exc	7.459 [14]	25
				aMCC_inh	3.318 [14]	50
				AI_exc	7.460 [14]	35
				AI_inh	3.316 [14]	45
				BLMA⊕	10 [9] [18]	12.5
				mPOA	10 [9] [18]	20
	mPFC_inh	2393 [16]	Inhibitory [17] (FS)			
				mPFC_exc	12.5 [15]	25
				mPFC_inh	2.632 [15]	300
PI						
	PI_exc	3106 [19]	Excitatory [17] (RS)			
				PI_exc	3.289 [14]	460

[16]Volume based on meta-analysis of MRI data on mPFC areas preferentially activated for self (in right vmPFC) and other (in left vmPFC and left Dorsomedial Prefrontal Cortex (dmPFC)) referential processing (contrasted with control data) (Murray et al., 2012). Dorsal areas activate during picture-based empathy-for-pain (Engen and Singer, 2013), while more ventral (possibly left-hemispheric) areas, thought to encode close-other representations (Murray et al., 2012), are proposed to project more strongly to BLMA⊕ and mPOA ((Numan, 2014, p.281), see footnote 18). Density is from BA10 (vmPFC) in Rabinowicz et al. (1999).

[17]mPFC and PI have 80% excitatory neurons and 20% inhibitory neurons, as per rough neocortex proportions. These are presumed to correspond to regular-spiking pyramidal and fast-spiking basket neurons, respectively (according to neocortex excitatory/inhibitory cell-type majorities) (see Rudy et al. (2011), Harris and Shepherd (2015) and references in Voges et al. (2010)). We also consider agranular AI and aMCC to have these same proportions.

[18]mPFC connected to BLMA⊕ and mPOA according to a negative-binomial distribution parametrised by r=7 and p=0.0025, intended to model a dorsal-ventral gradient for other-self representations (i.e. close-other representations, more ventral, have strongest connectivity, whilst neurons encoding self and other representations have sparser connectivity). The overall connection probabilities for each group pairing are as specified.

[19]Total neuron estimate based on PI control volume (Makris et al., 2006) and density (Semendeferi et al., 1998) estimates.

				PI_inh	13.149 [14]	100
				AI_exc	2.439 [14]	850
				AI_inh	1.084 [14]	25
				aMCC_exc	2.438 [14]	210
				aMCC_inh	1.085 [14]	80
	PI_inh	777 [19]	Inhibitory [17] (FS)			
				PI_exc	12.5 [15]	100
				PI_inh	2.630 [15]	200
aMCC						
	aMCC_exc	1329 [20]	Excitatory [17] (RS)			
				aMCC_exc	3.289 [14]	100
				aMCC_inh	13.168 [14]	100
				AI_exc	0.417 [14]	750
				AI_inh	0.185 [14]	300
				mPFC_exc	0.417 [14]	50
				mPFC_inh	0.185 [14]	50
				PI_exc	0.417 [14]	750
				PI_inh	0.185 [14]	500
	aMCC_inh	332 [20]	Inhibitory [17] (FS)			
				aMCC_exc	12.5 [15]	200
				aMCC_inh	2.634 [15]	400

Table 5.1: Number of neurons, neuron types, connection probabilities and connection weights for neuron groups/regions in the model. Estimates for total number of neurons in each region are calculated based on human bilateral volume and neuron density data (all control/non-clinical data, and averaged across age and gender when applicable), and then scaled to give 25660 total neurons in the model, and over 23 million synapses. Neuron types and connection probabilities are based on non-human (primarily rodent) data. Neurons in each group are connected randomly and uniformly (or according to a negative-binomial distribution in the case of mPFC to mPOA and BLMA) according to specified probabilities and with given weight.

[20]Volume is bilateral aMCC (dorsal ACC, labelled "whole" in Kitayama et al. (2006)). Density comes from chart for BA24a',b',c' (averaged across layers) in Höistad et al. (2013) .

5.3.3 Desired Network States and Simulation Phases

Here we describe three network states that are modelled in distinct phases of our simulation. The network is simulated at $1ms$ granularity, with discrete iterations $t = 1, 2, ..., 30000$ giving a total of $30s$ simulation time. We split our simulation into 3 phases of equal length $p = 10000$, where each phase corresponds to one of the three network states.

At each iteration, current is injected into neurons in the BLMA$\ominus$ (representing presence of a negatively valent social stimulus, for example an emotionally distressed client); the AI and PI (representing input from the self-pain network); and the aMCC and mPFC (representing input from the other-pain network). We define three values representing different proportional levels of stimulation: $L = 0.05$ ("low"), $M = 0.1$ ("medium") and $H = 0.15$ ("high"). For each of these five neural groups $G \in \{\text{BLMA}\ominus, \text{AI}, \text{PI}, \text{aMCC}, \text{mPFC}\}$, we now designate a subset of neurons $g(G) \subset G$ that can receive current injection, where $|g(G)| = |G| * H$. For BLMA$\ominus$, AI, PI and aMCC the neurons consisting this subset $g(G)$ are chosen randomly according to a uniform distribution $\mathcal{U}$.

Recall that the mPFC is believed to encode self-other representations along a ventral-dorsal gradient, and that we have connected neurons in this region to Numan's caregiving pathway accordingly. In particular, the mPFC has been connected to the mPOA and BLMA$\oplus$ according to a negative-binomial distribution parametrised by $r = 7$ and $p = 0.0025$, which is intended to model this ventral-dorsal gradient for self-other representations in the mPFC (i.e. neurons representing self or distant-other will be sampled with relatively low probability). In the simulations that follow, we consider two distinct target populations in the mPFC for neural activity in Zaki's other-pain network: a first population (encoding close-other representations) that is sampled according to this negative-binomial distribution (i.e. with $r = 7$ and $p = 0.0025$), and a second population (encoding more distant-other representations) that is sampled according to a negative-binomial distribution with $r = 14$ and $p = 0.035$ (see Fig. C.1).

We have a total of 5 subgroups g of neurons that can potentially receive external current on each iteration. At each time-step t, before each current injection, each of these neural subgroups g is perturbed by 10% (by random-uniformly switching 10% of neurons designated to receive current injection with neurons that previously were not). This perturbation results in $g_t(G)$, which defines the subset of neurons in neural group G that can potentially receive current injections at time-step t. This subset is then used to determine the final subset of neurons $c_t(G) \subseteq g_t(G)$ that actually receive current at time-step t, chosen according to a random-uniform distribution and varied according to which phase the simulation is currently in. For simplicity we use an amplitude $90mA$ for all current injections, and capture the different network states by varying $c_t(G)$ for all G across the phases.

At all iterations t (i.e. across all phases), we inject currents into a random subset $c_t(\text{BLMA}\ominus) \subseteq g_t(\text{BLMA}\ominus)$ of BLMA$\ominus$ neurons, where the size of this subset $|c_t(\text{BLMA}\ominus)| \sim$

$\mathcal{U}\{|\text{BLMA}\ominus|*M, |\text{BLMA}\ominus|*H\}$ is a uniformly distributed integer in the interval $[|\text{BLMA}\ominus|*M, |\text{BLMA}\ominus|*H]$. This current injection into the BLMA$\ominus$ represents high levels of negatively-valent input stimulation across the whole simulation.

5.3.3.1 Personal Distress

The first state that we want to capture is that of personal distress. In accordance with the findings of Zaki et al. described previously, AI and PI activations should be relatively high for this state. Since personal distress is associated with withdrawal rather than prosocial behaviour, we should have low activity in the VP (which drives caregiving behaviour). Relatively high activity in the BLMA$\ominus$-AI-PI network (and activation of the aMCC), along with low activity in the VP, thus defines a network state corresponding to personal distress.

In accordance with the above, during this phase of the simulation we inject current at a high level into both the AI and PI (in addition to current injections into the BLMA$\ominus$). This corresponds to injecting currents into a random subset $c_t(\text{AI}) \subseteq g_t(\text{AI})$ of AI neurons (with $|c_t(\text{AI})| \sim \mathcal{U}\{|\text{AI}| * M, |\text{AI}| * H\}$), and a random subset $c_t(\text{PI}) \subseteq g_t(\text{PI})$ of PI neurons (with $|c_t(\text{PI})| \sim \mathcal{U}\{|\text{PI}| * M, |\text{PI}| * H\}$. During this phase, we also inject current at a low level into the aMCC and mPFC (to represent low-levels of activity in the other-pain network, i.e. a weak self-other distinction). This corresponds to injecting currents into random neuron subsets $c_t(\text{aMCC}) \subseteq g_t(\text{aMCC})$ (with $|c_t(\text{aMCC})| \sim \mathcal{U}\{|\text{aMCC}| * L, |\text{aMCC}| * M\}$); and $c_t(\text{mPFC}) \subseteq g_t(\text{mPFC})$ (with $|c_t(\text{mPFC})| \sim \mathcal{U}\{|\text{mPFC}| * L, |\text{mPFC}| * M\}$). Current is injected into mPFC neural populations encoding a close-other representation (Fig. C.1).

5.3.3.2 Weak Empathic Concern

The second state that we want to capture (which we call "weak empathic concern") is that of an empathic state with a relatively strong self-other distinction (compared to the personal distress state), but with an other stimulus that is encoded as being a relatively distant-other. Despite this distant-other encoding, the weak empathy state should nonetheless potentially be sufficient for the motivation of caregiving behaviour. In this state we should again have activation in the BLMA$\ominus$ (stimulated by a focusing on the negative client stimulus), and AI and aMCC (which form core parts of empathy circuitry). In accordance with the findings by Zaki et al, AI and PI activations should be relatively low compared to a personal distress state, whilst aMCC should be roughly the same. Activity in mPFC neurons encoding distant-other representations should trigger activity in the BLMA$\oplus$-VP and mPOA-VTA-NAc-VP pathways that facilitate caregiving behaviour (at a level that is relatively high compared to the personal distress state, but relatively low compared to a strong empathy state considered next).

During the second phase of the simulation, in addition to current injections into the BLMA$\ominus$, we thus inject current at a low level into both the AI and PI. This corre-

sponds to injecting currents into a random subset $c_t(\text{AI}) \subseteq g_t(\text{AI})$ of AI neurons (with $|c_t(\text{AI})| \sim \mathcal{U}\{|\text{AI}| * L, |\text{AI}| * M\}$), and a random subset $c_t(\text{PI}) \subseteq g_t(\text{PI})$ of PI neurons (with $|c_t(\text{PI})| \sim \mathcal{U}\{|\text{PI}| * L, |\text{PI}| * M\}$. We also inject current at a high level into the aMCC and mPFC (to represent high-levels of activity in the other-pain network). This corresponds to injecting currents into random neuron subsets $c_t(\text{aMCC}) \subseteq g_t(\text{aMCC})$ (with $|c_t(\text{aMCC})| \sim \mathcal{U}\{|\text{aMCC}| * M, |\text{aMCC}| * H\}$); and $c_t(\text{mPFC}) \subseteq g_t(\text{mPFC})$ (with $|c_t(\text{mPFC})| \sim \mathcal{U}\{|\text{mPFC}| * M, |\text{mPFC}| * H\}$). Current is injected into mPFC neurons encoding more distant-other representations (Fig. C.1).

5.3.3.3 Strong Empathic Concern

The third state that we want to capture (which we call "strong empathic concern") is that of an empathic state with a strong self-other distinction, and a target that is now perceived to be a close-other. This close-other encoding should result in a relatively high level of motivation for caregiving behaviour compared to the weak empathy state. Activation levels across neural populations should thus be similar as for the weak empathy state, except that we should now expect relatively high activity in the BLMA⊕-VP and mPOA-VTA-NAc-VP pathways.

Current injections for the third phase of the simulation capturing this strong empathy state are the same as for the weak empathy state, except that current is now injected into mPFC neurons encoding close-other representations rather than distant-other representations (Fig. C.1).

5.3.4 Simulations

Here we describe results of simulations of our network over the three phases described in Section 5.3.3, implemented using the CARLsim 3.1 framework (Beyeler et al., 2015). In brief (and as covered previously), the first phase, for time-steps $t \in (0, 10]$ seconds, corresponds to a personal distress response, during which activation inputs from the self- and other-pain networks are high/low respectively (inputs from the other-pain network target close-other mPFC representations). The second phase ($t \in (10, 20]$ seconds) corresponds to a weak empathic concern state (i.e. a state in which the other's emotional state is mirrored, with the other being perceived as a relatively distant other). Input from the self-pain network is low, and from the other-pain network is high, and the other-pain network stimulates mPFC neurons encoding close-other representations leading to a relatively high level of caregiving behaviour compared to the personal distress state. The third phase ($t \in (20, 30]$ seconds) corresponds to a strong empathic concern state, with strong self-other distinction and the other perceived as a close-other. During this phase, input from the self-pain network is low, while input from the other-pain network is high (and targets mPFC neurons encoding close-other representations), such that relatively high amounts of caregiving behaviour result as

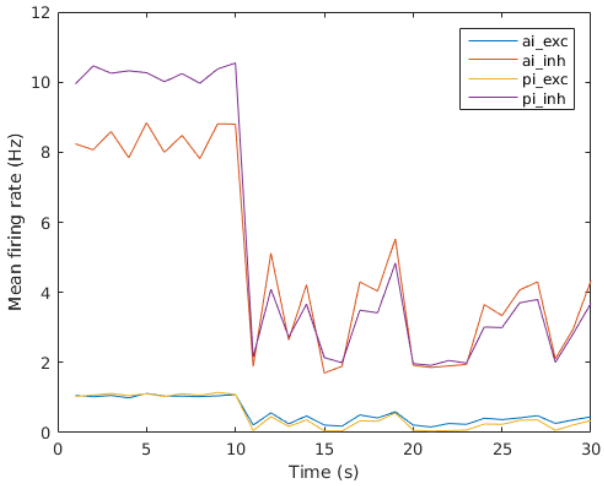

Figure 5.5: Mean firing rates for the AI and PI. See text for details.

compared to the weak empathy state.

The results presented here are representative of a typical simulation run using the configuration previously described. Fig. 5.5 gives the Mean Firing Rate (MFR) for neurons in the excitatory and inhibitory AI and PI neural groups. The chart shows that the MFR of AI and PI neurons is highest during the first phase (personal distress), and drop significantly relative to this during the final two phases (corresponding to weak and strong empathic concern). Excitatory/inhibitory AI neurons drop from a MFR of 1.08/8.80 Hz at 10s (end of the first phase and beginning of the second phase) to 0.22/1.92 Hz at 20s (the end of the second phase) and rise again slightly to 0.45/4.33 Hz at 30s (end of the third phase); while excitatory/inhibitory PI neurons drop from a MFR of 1.09/10.55 Hz at 10s to 0.06/1.97 Hz at 20s, and rise slightly to 0.34/3.69 Hz at 30s. These patterns correspond directly with the results of the experiments described in Section 5.2.2.2 regarding relative activation in these regions in self and other pain conditions.

MFR for neurons in the excitatory and inhibitory aMCC and mPFC groups are shown in Fig. 5.6. The aMCC MFR is relatively flat across the three phases: the MFR for excitatory/inhibitory aMCC is 1.09/5.97 Hz at 10s, 0.93/4.79 Hz at 20s, and 1.05/5.61 Hz at 30s. This relative stability is in accordance with the findings of the experiments described in Section 5.2.2.2, which did not find significant differences in activation of the aMCC across self and other pain paradigms. On the other hand, the MFR of the mPFC is slightly higher in the second and third phases compared to the first phase (0.84/5.25 Hz at 10s, 1.30/8.53 at 20s and 1.26/8.82 at 30s). This coincides with increased stimulation of neurons encoding distant- and close-other representations in this region by the other-pain network during the second and third phases (the mPFC is not directly stimulated with current injection during

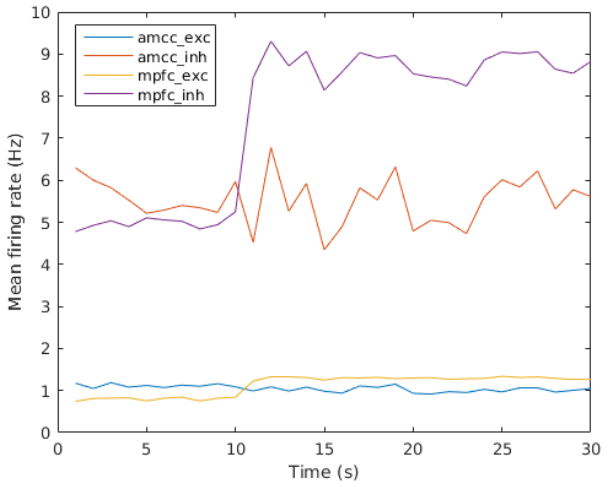

Figure 5.6: Mean firing rates for the aMCC and mPFC. See text for details.

the self-pain condition).

Fig. 5.7 shows the MFR for the three BLMA neural populations. The BLMA⊖, which receives a steady input current across all three phases (corresponding to client stimulus input) has a MFR that is relatively flat across all three phases, dropping slightly for the second and third phases relative to the first (0.92 Hz at 10s, 0.75 at 20s, 0.72 at 30s). This drop during the third phase coincides with an increase in MFR of the BLMA⊕ (from 0.21 Hz at 20s to 6.75 Hz at 30s), whose neurons are stimulated by the mPFC and excite the BLMAi.

Finally, MFR for the mPOA, VTA, NAc and VP are shown in Fig. 5.8. Activity in the mPOA, VTA and NAc is relatively low during the first phase, and rises significantly during the second phase (with a stronger self-other distinction, and perception of the other as a relatively distant-other) and further across the third phase (in which the other is instead encoded as a close-other). These firing patterns are reflected in firing in the VP, with relatively low MFR in the VP during the first phase but increased firing during the second phase, and higher firing yet during the third phase (with MFR 2.44 Hz at 30s). Firing of the VP (which represents facilitation of caregiving behaviour in response to an internal emotional state) is thus low during the personal distress state, increased for the weak empathy state, and highest for the strong empathy state.

5.4 Empathic Motivation in Self-Attachment Therapy

Up to this point we have considered states of personal distress and (weak and strong) empathic concern within an individual (e.g. a psychotherapist) resonating with the negatively-

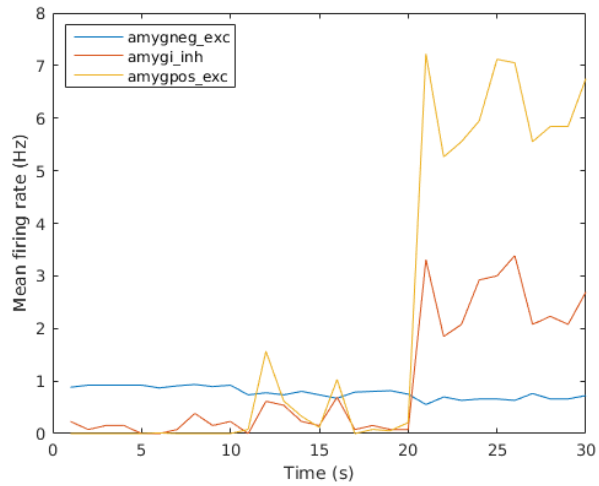

Figure 5.7: Mean firing rates for the three BLMA neural populations. See text for details.

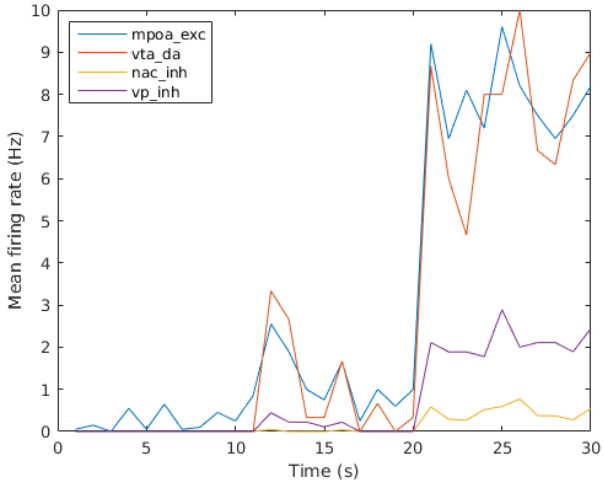

Figure 5.8: Mean firing rates for the mPOA, VTA, NAc and VP. See text for details.

valanced emotional state of another (e.g. the client). We have shown how these states are mediated by the relative strength of self-other emotional state attribution, and the perceived degree of self-relatedness of the other. We now extend this model to Self-Attachment: an attachment-based psychotherapy in which the individual takes the role of both psychotherapist and client.

As discussed in previous chapters, at a neural level 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 (Schore, 2003a; Cozolino, 2014, p.53); areas known to be central to, and crucially involved in, social cognition and emotional processing. Largely mature at birth (Ulfig et al., 2003), the amygdala is crucially involved in fear conditioning (Milad and Quirk, 2012), saliency, and stress-related processes that are likely to underpin many forms of insecure (particularly disorganised) attachment (Main and Hesse, 1990). High levels of attachment anxiety have been found to correlate with elevated cortisol profiles (Kidd et al., 2013), and a relatively over-active amygdala in response to angry faces conveying negative social feedback (Vrtička et al., 2008) and infant crying (Riem et al., 2012). The amygdala has strong bi-directional connectivity with the OFC, medial parts of which are (in the macaque) involved in the learning of stimulus-reward associations (Walton et al., 2010), and which is believed to mediate stress and facilitative reactivity to social stimuli via projections to dmH and the PVN. The PVNp releases CRH, which stimulates stress circuitry focused on the CeA and LC, while the PVNm releases OXT, one effect of which is thought to be a modulation of DA release (Love, 2014). Evidence implicates both DA (Bartels and Zeki, 2004; Strathearn et al., 2009; Vrtička et al., 2008) and OXT (Feldman et al., 2007; Gordon et al., 2010) as being crucial for a range of bonding and attachment-related behaviours. However, exogenous OXT administration has in certain cases been found to increase anti-social tendencies (Declerck et al., 2010; Bartz et al., 2011), and appears to amplify pre-existing interpersonal schemas (Olff et al., 2013; Bartz et al., 2010) making it unsuitable as a treatment for many attachment-related disorders.

In light of the above, in the previous chapter we hypothesised that a main effect of the Self-Attachment bonding protocols (in which the adult-self attempts to create an abstract bond with the inner-child) is to associate broad classes of social stimuli that have previously been conditioned as being fearful or threatening in nature with representations of additional, naturally-induced reward (which result from various interactions, for example directed singing (Salimpoor et al., 2011; Jeffries et al., 2003; Kleber et al., 2007) with inner-child imagery (Strathearn et al., 2008, 2009)). At a neural level, we proposed that this would result in a rebalancing in activation of stress and facilitative circuitry in response to such classes of social stimuli. As the protocols progress, the OFC should gradually come to learn new reward associations, increasingly facilitating natural OXT release and inhibiting

CRH release; while dopaminergic reward-prediction errors should drive a vmPFC-mediated inhibition of stress circuitry via strengthening of an ITC-CeA pathway, inhibiting activity in the amygdala and stress circuitry.

5.4.0.1 Bonding Protocols and Compassionate States

The Self-Attachment bonding protocols are closely related to the concept of compassion. In a compassionate state (Gonzalez-Liencres et al., 2013), as for an empathic state, there is a strong self-other distinction along with knowledge of the emotional state of another. Also in similarity with an empathic state, a compassionate state in the self can motivate prosocial behaviour aimed at relieving a perceived negative state in the other. The key distinction between empathic and compassionate states is that compassionate states do not necessarily involve the mirroring of the emotional state of the other. For example, one may perceive suffering in another but, rather than mirroring this suffering and experiencing this within the self (as in an empathic state), a compassionate state would instead involve positively-valenced emotion within the self. In a compassionate state, it is this positively valenced emotion within the self, rather than negatively valenced emotion that is mirrored from the other, that can motivate prosocial behaviour aimed at alleviating the suffering of the other. In this sense, a compassionate state can be seen as involving an attempt to project outwards a positive emotional state from the self to the other, whereas in contrast empathic states involve inward projections (mirroring) of the emotional state of the other. With respect to Self-Attachment, we propose that application of the bonding protocols towards the conceptualised inner-child serves to engender a more compassionate stance within the adult-self.

While empathy-for-pain states have been found to activate overlapping regions involved in self-pain, compassion instead seems to activate areas associated more with love, reward, and positive emotion. For example, Klimecki et al. (2013) investigated subjective emotional states and neural activations in response to another's distress, before and after both short-term empathy and compassion training. Empathy training increased empathic responses and negative affect, and was associated with activation in core AI-aMCC empathy-for-pain circuitry. Following compassion training (which involved watching videos of others in distress and cultivating feelings of benevolence towards them), negative affect response to others' pain returned to baseline, whilst activation in areas associated with positive affect (including mOFC, reward circuitry in the ventral striatum (which includes the NAc), and perigenual ACC) increased. A related study (Engen and Singer, 2015) examined neural activations involved in the formation of a compassionate state in response to someone in distress, as opposed to a re-appraisal (i.e. down-modulation of negative affect), with compassion compared to both re-appraisal and passive watching of the same negative stimuli resulting in increased activation in vmPFC, mOFC, perigenual ACC, and NAc.

5.4.1 Empathy Protocols

In this section we propose a model of the hypothesised neural underpinnings of another component of the Self-Attachment psychotherapy. The procedures that we consider, which we call the "empathy protocols", involve the individual undergoing the therapy taking the perspective of the "adult-self" while focusing on a negatively emotionally-valenced image of their younger self (the "inner-child") and attempting to enter into an empathic state with them. Variations of the protocol can involve imagery or virtual reality techniques, rather than the focus on an actual image of the younger self, in order to conceptualise the distressed and assistance-requiring inner-child. The purpose of the empathy protocols is to motivate caregiving behaviour towards the inner-child, as defined by the "bonding protocols" in the previous chapter. In this section, we use findings from the neuroscience of empathy, compassion, and the self-other distinction within the context of pain to explain how sufficient self-other distinction can lead to an empathic state which motivates bonding with the inner-child, and how application of these bonding protocols may in turn drive a progression from empathic to compassionate states within the adult-self.

The effectiveness of the Self-Attachment empathy protocols in motivating bonding behaviour may vary according to a number of dimensions, one of which is gender. As previously discussed, empathic motivation has been argued to have its ultimate roots in selective pressures for attachment bond formation. Evidence suggests that women (who have typically served as primary caregivers for young infants throughout history and across cultures) have more attuned empathic capabilities with respect to infants relative to men. These traits are believed to have roots more in biological than cultural factors, and are again possibly the result of evolutionary pressures (Christov-Moore et al., 2014).

Efficacy of these protocols may also vary according to prior attachment experience. In particular, in BPD (which, as discussed previously, is a condition linked to early disorganised attachment experience) empathic dysfunction may be experienced in the form of hyper-reactivity of reflexive systems involved in the sharing of others' mental states, along with impairments in more deliberative systems involved in perspective taking and the explicit attribution of empathic mental states to the other (Ripoll et al., 2013; Gonzalez-Liencres et al., 2013). Thus, in the case of BPD individuals, care may need to be taken so as to avoid overwhelming personal distress, and we suggest that the application of these protocols be supplemented with techniques that focus on forming clear mental distinctions between the adult-self and inner-child. One such therapy that may be particularly helpful in achieving this aim is Mentalization therapy (Bateman and Fonagy, 2012). Mentalization, defined as the ability to attribute mental states (including emotional) underlying overt behaviour to self and other, can be seen as inclusive of the notion of emotional empathy. The ability of an individual to mentalize is viewed as arising from contingently responsive mentalizing on the part of the caregiver in infancy, and therapies involve a joint focus on the client's

subjective inner states in order to strengthen their sense of self and ability to mentalize.

5.5 A Model of the Self-Attachment Empathy Protocols

Based on the evidence outlined above relating to the neuroscience of empathy and compassion; the model of how empathic states can motivate prosocial behaviour; the proposed networks for self- and other-pain; and findings relating to self-other representations in the mPFC, we propose a model of how the Self-Attachment empathy protocols can motivate application of the bonding protocols, and how repeated application of these bonding protocols can in turn facilitate a gradual shift towards inner-child directed caregiving behaviour that is mediated by a positive (compassionate) rather than negative (empathic) emotional state.

In particular, we propose that a sufficient self-other distinction is required in order for the mPFC to stimulate the mPOA-VTA-NAc-VP, BLMA-VP and BLMA-NAc-VP pathways (as described in Numan's model) that facilitate caregiving behaviour. Recall that there is evidence to suggest the vmPFC encodes "close-other" representations with high self-relatedness (Denny et al., 2012), and that ventral parts of the mPFC have been found to activate more strongly within the context of empathy with respect to the perceived closeness (Meyer et al., 2012) and innocence (Fehse et al., 2015) of the other. Thus, in line with Numan's proposal that more ventral parts of the mPFC might be involved in empathic concern, we suggest that stimulation of the caregiving pathways will be strongest for mPFC neural populations which encode a "close-other" that is perceived as having high self-relatedness/similarity and innocence (located in vmPFC), which are precisely the characteristics possessed by the conceptualised inner-child within the Self-Attachment framework. Within the context of empathically-motivated caregiving behaviour, this corresponds to the "strong empathic concern" state that we detailed above. Under the alternative view of mPFC function (proposing that ventral/dorsal areas encode executed/modelled value (Nicolle et al., 2012)) discussed previously, then as the adult-self and inner-child distinction is developed (and the idea of tending to the distressed inner-child is formulated) we might similarly expect progression towards patterns of activation (in more ventral areas) representing high value of caregiving behaviour (executed by the adult-self) for the inner-child.

As discussed previously, interactions between the mPOA and the mesolimbic DA system, along with increased functional connectivity between the amygdala and NAc and VP, have been observed during the simulation of prosocial acts. In terms of Self-Attachment therapy, this suggests that we should expect the mPOA-VP and BLMA-VP pathways to be activated both when the adult-self imagines, and overtly practices, caregiving and bonding behaviours with the inner-child.

With repetitive application of the bonding protocols, we hypothesise that the OFC should increasingly stimulate both inhibitory BLMAi (which inhibit the negatively valent BLMA⊖ neurons) and positively valent BLMA⊕, so that caregiving behaviour gradually comes to be facilitated via a positive (OFC-mediated) rather than negative (BLMA⊖-mediated) emotional state. As detailed above, we previously hypothesised that the Self-Attachment bonding protocols would lead to new stimulus-reward associations in the OFC, facilitating OXT release and inhibiting stress reactivity in response to social stimuli. Parts of the OFC have been found to preferentially activate in mothers viewing images of own vs other infants, with activation levels correlating with self-reported pleasant mood ratings (Nitschke et al., 2004; Minagawa-Kawai et al., 2009). This region is known more generally to increase firing in response to stimuli that predict rewards, with activation levels that are a function of current reward value (Gottfried et al., 2003). In particular we consider a neural population in the mOFC, lesions to which are known to disrupt stimulus-reward association learning in the macaque (Walton et al., 2010). Our proposal that increased activation of the mOFC should result from repeated application of the self-directed bonding protocols also mirrors the findings discussed above that found increased activation in this area following compassion training in human subjects.

5.5.1 Additional Regions and Connectivity

In his model of empathically-motivated caregiving behaviour, Numan proposes OFC projections to the AI and BLMA as additional pathways by which caregiving behaviour might be modulated by in-group preferences (Numan, 2014, p.279). Based on the previously proposed role for the OFC in the Self-Attachment bonding protocols, along with findings from the neuroscience of compassion, we extend our model to incorporate projections from mOFC to these regions (Fig. 5.9) which are proposed to increase in strength as the bonding protocols progress. Details of the number of neurons, neuron types and target connection probabilities and weights for this additional neural group and its synapses are given in Table 5.2.

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
mOFC						
	mOFC_exc	7472 [21]	Excitatory [22] (RS)			

[21] Based on control volume (Lacerda et al., 2004) and density (Rajkowska et al., 1999) estimates.

[22] mOFC follows standard neocortical neuron type proportions (see footnote 17).

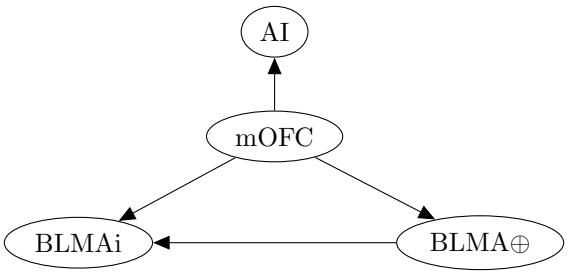

Figure 5.9: Additional connectivity emanating from the medial orbitofrontal cortex. The remaining architecture is as in Fig. 5.3

				mOFC_exc	3.319 [23]	175
				mOFC_inh	13.277 [23]	55
				AI_exc	7.058 [23]	40
				AI_inh	3.137 [23]	30
				BLMA⊕	10 [9]	10
				BLMAi	10 [9]	35
	mOFC_inh	1868 [21]	Inhibitory [22] (FS)			
				mOFC_exc	12.5 [24]	100
				mOFC_inh	2.632 [24]	500

Table 5.2: Details of the number of neurons, neuron types and connectivity for the additional mOFC neuron group. The remaining neuron groups in the model are as before (details given in Table 5.1). The total number of neurons in the extended model is 35000, with over 32 million synapses.

5.5.2 Desired Network States and Simulation Phases

As described previously, the empathy protocols involve focusing on an image of oneself (the inner-child) as an infant/child in distress, and empathising with the inner-child in order to motivate caregiving in the form of the bonding protocols. Here we describe three network states that are modelled in distinct phases of our simulation below. Based on our previously

[23]We assume that the mOFC excitatory neurons follow the same neocortical connectivity profile as in footnote 14.

[24]mOFC inhibitory neurons are assumed to follow the same neocortical connectivity profile as in footnote 15.

proposed model of personal distress and (weak and strong) empathic concern, we suggest that a typical individual undertaking Self-Attachment therapy might progress through these phases sequentially as the protocols are undertaken, and the simulations that follow thus provide predictions for patterns of activity that might be hypothesised to occur (under the assumption that the previously proposed model of personal distress and empathic concern states holds, and that the therapy induces the patterns of injected current that are to be described across each phase). Stimulation of the BLMA$\ominus$ (now corresponding to a focus on the stimulus representing the distressed inner-child) across the length of the simulation is as before (Section 5.3.3), but note that we now have an additional neural group (mOFC) that receives different amounts of current injection at each time-step (in order to stimulate different activation levels that are hypothesised to correspond to varying levels of progression with the bonding protocols).

5.5.2.1 Personal Distress

Initially, on application of the empathy protocols, it may be that there is only a weak distinction between the self (adult-self) and other (inner-child), i.e. high activity in the self-pain network and low activity in the other-pain network. Thus, the first state that we want to capture is that of personal distress within the adult-self. In this state, the adult-self focuses on the image of the inner-child and mirrors their negative emotional state, but does not have sufficient self-other distinction in order to distinguish the adult-self from the inner-child. This may correspond to early stages of the therapy, in which a sufficient self-other distinction has not yet been developed, and the reaction of the adult-self might thus be one of distress and withdrawal rather than caregiving. In accordance with the findings of Zaki et al. described previously, AI and PI activations should be relatively high for for this state, which involves a form of self-pain. Since personal distress is associated with withdrawal rather than prosocial behaviour, we should have low activity in the VP (which drives caregiving behaviour). Relatively high activity in the BLMA$\ominus$-AI-aMCC network, along with low activity in the VP, thus defines a network state corresponding to personal distress.

Current injections for the first simulation phase ($t \in (0, 10]$ seconds) are as previously defined for the personal distress state (Section 5.3.3.1), along with an additional stimulation of the mOFC (to represent low-levels of reward association with the inner-child stimulus, i.e. weak progress with respect to the bonding protocols). This corresponds to injecting currents into random neuron subsets $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ (with $|c_t(\text{mOFC})| \sim \mathcal{U}\{|\text{mOFC}| * L, |\text{mOFC}| * M\}$).

5.5.2.2 Strong Empathic Concern With Strengthening Self-Other Distinction

The second state that we want to capture is that of an empathic state with a sufficiently strong self-other distinction such that caregiving (i.e. the bonding protocols) results as motivated behaviour. We propose that once an individual has developed a self-other distinction between the adult-self and inner-child, they can progress from the first state to this second state, and commence effective application of the bonding protocols. We should again have activation in the BLMA$\ominus$ (stimulated by a focusing on the inner-child stimulus), and AI and aMCC (which form core parts of empathy circuitry). In accordance with the findings by Zaki et al, AI and PI activations should be relatively low compared to a personal distress state, whilst aMCC should be roughly the same. In this state, activity in mPFC neurons encoding other-referential representations should trigger activity in the BLMA$\oplus$-VP and mPOA-VTA-NAc-VP pathways that facilitate caregiving behaviour.

In particular, we want to capture a transition from personal distress (with weak self-other distinction) to strong empathic concern (with strong self-other distinction and a close-other representation). Thus, in addition to current injections into the BLMA$\ominus$, we also inject currents into a proportion of neurons in the AI and PI that linearly decreases from "high" to "low" as the second phase ($t \in (10, 20]$ seconds) progresses. This corresponds to injecting currents into a random subset $c_t(\text{AI}) \subseteq g_t(\text{AI})$ of AI neurons (with $|c_t(\text{AI})| = ((|\text{AI}| * (L - H)) * (t/p)) + |\text{AI}| * (2 * H - L)$, and a random subset $c_t(\text{PI}) \subseteq g_t(\text{PI})$ of PI neurons (with $|c_t(\text{PI})| = ((|\text{PI}| * (L - H)) * (t/p)) + |\text{PI}| * (2 * H - L)$. In order to capture increasing activation in the other-pain network, we inject currents into a proportion of neurons in the aMCC and mPFC that linearly increases from "low" to "high" as a function of phase progression. This corresponds to injecting currents into a random subset $c_t(\text{aMCC}) \subseteq g_t(\text{aMCC})$ of aMCC neurons (with $|c_t(\text{aMCC})| = ((|\text{aMCC}|*(H-L))*(t/p))+ |\text{aMCC}| * (2 * L - H)$; and a random subset $c_t(\text{mPFC}) \subseteq g_t(\text{mPFC})$ of mPFC neurons (with $|c_t(\text{mPFC})| = ((|\text{mPFC}| * (H - L)) * (t/p)) + |\text{mPFC}| * (2 * L - H)$.

The retrieval of autobiographical memories of familiar others involves activation of regions including the precuneus and PCC (Maddock et al., 2001), which are both regions in Zaki's other-pain network. It is therefore possible that conceptualisation protocols involving the individual actively attempting to associate an image of their younger self in distress with the conceptualised inner-child 'other' result in plasticity effects in circuits involving these regions, which might in turn account for the change in activation in self- and other-pain network activity that we have modelled here (in terms of a simple linear shift in current injection). As we will discuss in Section 5.6, OXT (hypothesised to be released as a result of the bonding protocols) may also play a role in facilitating this self-other shift. Thus, the neural mechanisms underlying this transformation are likely to be complex and multifaceted, and to differ according to the precise conceptualisation techniques that are employed, and we leave the further detail for future work.

As in the first phase, current is injected into the mOFC at low levels, corresponding to network states before which application of the bonding protocols have begun to take effect. We inject currents into a random subset $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ of mOFC neurons (with $|c_t(\text{mOFC})| \sim \mathcal{U}\{|\text{mOFC}| * L, |\text{mOFC}| * M\}$).

5.5.2.3 Compassion

The third network state that we want to capture corresponds to a more compassionate state, which occurs as a result of effective application of the bonding protocols. Now when the adult-self focuses on the distressed image of the inner-child, instead of mirroring their negative affective state, we propose that they will instead have a positive (compassionate) inner emotional state, and that this state will continue to motivate prosocial bonding behaviour towards the inner-child.

As discussed previously, we have hypothesised that one effect of the application of the Self-Attachment bonding protocols is to stimulate activity in neurons of the OFC and vmPFC representing positive reward and emotion, which in turn inhibit negatively valent representations in the amygdala. Consistent with our proposal, increased activation of the mOFC (along with the VTA) was a key finding in a study which looked at the effects of compassion training (Klimecki et al., 2013). In our model of the empathy protocols, this effect corresponds to stimulation of BLMA$\oplus$ by the mPFC (which occurs during both empathic concern and compassionate states) and mOFC (which occurs uniquely during compassionate states), and stimulation of BLMAi by mOFC which in turn inhibits BLMA$\ominus$ (which again occurs uniquely during a compassionate state). The mOFC also stimulates AI, which is consistent with evidence suggesting a role for the left AI in positive emotion and maternal behaviour (Craig, 2009, Fig.2). Stimulation of BLMA$\oplus$ and a different (presumed positively-valent) sub-population of AI neurons, along with inhibition of BLMA$\ominus$, is proposed to represent neurally the positively-valent emotional elements of a compassionate state within the adult-self. This state should also result in continued activation in Numan's caregiving pathways, facilitating bonding behavioural responses towards the inner-child.

During the third phase ($t \in (20, 30]$ seconds), in addition to current injections into the BLMA$\ominus$, we also inject currents into a "low" proportion of neurons in the AI and PI, to represent low levels of activity in the self-pain network. This corresponds to injecting currents into a random subset $c_t(\text{AI}) \subseteq g_t(\text{AI})$ of AI neurons (with $|c_t(\text{AI})| \sim \mathcal{U}\{|\text{AI}|*L, |\text{AI}|*M\}$), and a random subset $c_t(\text{PI}) \subseteq g_t(\text{PI})$ of PI neurons (with $|c_t(\text{PI})| \sim \mathcal{U}\{|\text{PI}|*L, |\text{PI}|*M\}$. To capture a strong self-other distinction, it is also assumed that activity in the other-pain network is sustained during this phase, so that we inject current into a "high" proportion of aMCC and mPFC neurons. This corresponds to injecting currents into a random subset $c_t(\text{aMCC}) \subseteq g_t(\text{aMCC})$ of aMCC neurons (with $|c_t(\text{aMCC})| \sim \mathcal{U}\{|\text{aMCC}| * M, |\text{aMCC}| * H\}$, and a random subset $c_t(\text{mPFC}) \subseteq g_t(\text{mPFC})$ of mPFC neurons (with $|c_t(\text{mPFC})| \sim$

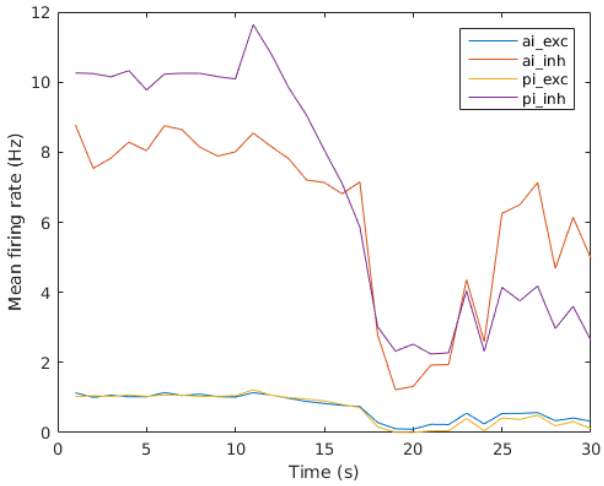

Figure 5.10: Mean firing rates for the AI and PI. See text for details.

$\mathcal{U}\{|\text{mPFC}| * M, |\text{mPFC}| * H\}$. To capture application of the bonding protocols, we linearly increase the proportion of mOFC neurons receiving input across the phase, by injecting current to a random subset $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ of PI neurons (with $|c_t(\text{mOFC})| = ((|\text{mOFC}| * (H - L)) * (t/p)) + |\text{mOFC}| * (3 * L - 2 * H)$. In accordance with the results in Klimecki et al. (2013), we might also expect elevated VTA activity during this phase.

5.5.3 Simulations

Fig. 5.10 gives the MFR for neurons in the excitatory and inhibitory AI and PI neural groups. The chart shows that the MFR of AI and PI neurons is highest for the first phase of the protocol, but drops significantly throughout the second phase (as self-pain inputs are decreased and other-pain inputs are increased), in accordance with the previously simulated personal distress and strong empathic concern states. Excitatory/inhibitory AI neurons drop from a MFR of 1.01/8.01 Hz at 10s (end of the first phase and beginning of the second phase) to 0.1/1.31 Hz at 20s (the end of the second phase), while excitatory/inhibitory PI neurons drop from a MFR of 1.05/10.09 at 10s to 0.01/2.52 at 20s. Firing rates for the AI rise again slightly during the third phase to 0.32/4.97 at 30s due to mOFC input, in line with a role for this region (particularly in the left hemisphere) in positive emotion and maternal behaviour (outlined previously).

MFR for neurons in the excitatory and inhibitory aMCC, mPFC and mOFC groups are shown in Fig. 5.11. The aMCC MFR is relatively flat across the three phases: the MFR for excitatory/inhibitory aMCC is 1.09/5.06 Hz at 10s, 1.11/6.23 Hz at 20s, and 1.01/4.97 Hz at 30s. Stability across the first and second phases mirrors the previously simulated

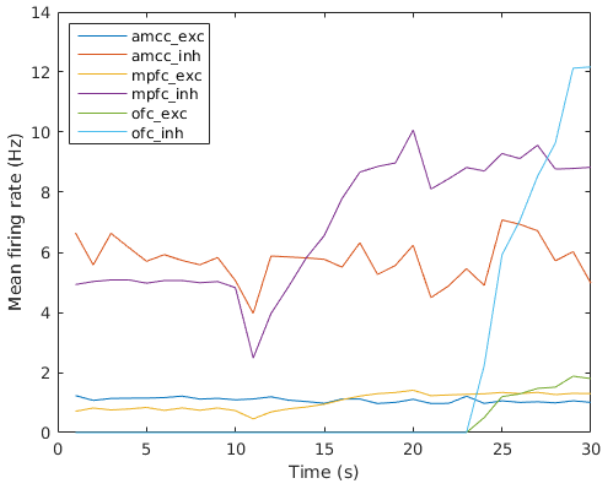

Figure 5.11: Mean firing rates for the aMCC, mPFC and mOFC. See text for details.

personal distress and strong empathic concern states, and is in accordance with the findings of the experiments described in Section 5.2.2.2 which did not find significant differences in activation of the aMCC across self and other pain paradigms. The relative decrease in firing rate for the aMCC during the third phase is consistent with evidence (discussed previously) that different (more perigenual) areas of the ACC may instead be involved in compassionate states.

The MFR of the mPFC rises steadily during phase 2 (from 0.73/4.82 Hz at 10s to 1.41/10.07 Hz at 20s) following relative stability in the first phase, and back toward relative stability in the third phase. The rise in mPFC activation coincides with a strengthening of the self-other distinction between adult-self and inner-child, which is captured by increasing activation of (more ventral) neurons encoding close-other representations proposed to be associated with the inner-child. For the mOFC, the MFR is relatively low during the first and second phases, but rises as the third phase progresses (with the MFR for excitatory/inhibitory mOFC neurons rising from near zero at 20s to 1.80/12.17 Hz at 30s). This rise during the third phase corresponds to progress in the application of the bonding protocols with respect to increasing expectations of reward associated with prosocial motivation towards the inner-child.

Fig. 5.12 shows the MFR for the three BLMA neural populations. The BLMA$\ominus$, which receives a steady input current across all three phases (corresponding to inner-child stimulus input) has a MFR that is relatively flat during the first phase (0.91 Hz at 10s), but drops as the self-other distinction becomes stronger during the second phase (to 0.61 Hz at 20s) and throughout progression of the third phase (to 0.51 Hz at 30s). This drop during the third phase coincides with a relatively large increase in MFR of the BLMAi (from 3.31 Hz at 20s

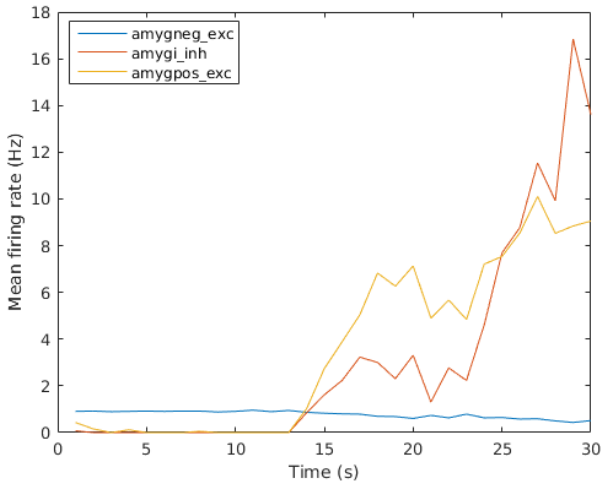

Figure 5.12: Mean firing rates for the three BLMA neural populations. See text for details.

to 13.62 Hz at 30s), whose neurons are stimulated by the mOFC. In contrast, the MFR of the BLMA⊕ is relatively low during the first phase, but rises for strong self-other distinction at the end of the second phase (to 7.13 Hz), and again as the third phase progresses (to 9.05 Hz at 30s). During the first and second phases, then, we have relatively low activation in the mOFC, BLMA⊕ and BLMAi, and relatively high activation in the BLMA⊖ which, along with activation in the AI and aMCC, is proposed to correspond to the negatively-valenced emotional state within the adult-self that is mirrored from the inner-child. During the third phase, the MFR in the BLMA⊖ falls significantly, yet rises in the BLMA⊕ and mOFC, and the mOFC stimulates distinct neurons in the AI. We propose that this pattern of activation corresponds to a positively-valenced compassionate state within the adult-self, in which the negatively-valenced, empathically-mirrored emotional state of the previous phases is (at least somewhat) suppressed.

MFR for the mPOA, VTA, NAc and VP are shown in Fig. 5.13. Activity in the mPOA, VTA and NAc is relatively low during the first phase, and rises significantly towards the end of the second phase (as the self-other distinction becomes stronger) and slightly further across the third phase (with relatively high MFR in the VTA during the third phase consistent with Klimecki et al. (2013)). These firing patterns are reflected in firing in the VP, with relatively low MFR in the VP during the first phase (and no firing at 10s); increased firing at the end of the second phase (with 2.56 Hz at 20s); and relatively high (and rising) MFR as the third phase progresses (to 3.56 Hz at 30s). An increasing MFR for the VP as the second phase progresses represents increasing facilitation of caregiving behaviour in response to a strong empathic concern state, as a result of strengthening self-other distinction. The MFR for the VP is highest towards the end of the third phase, which corresponds to an

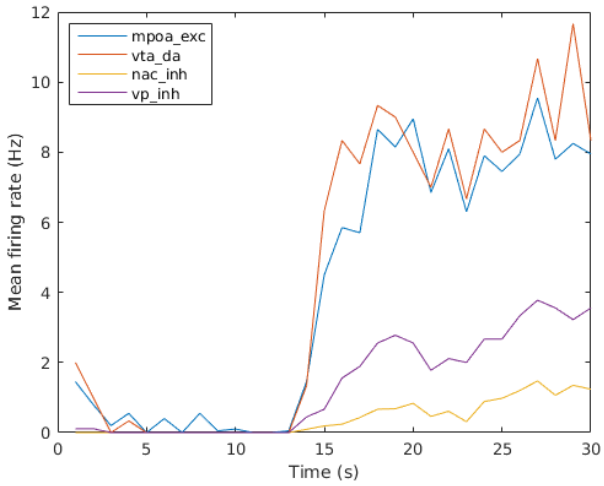

Figure 5.13: Mean firing rates for the mPOA, VTA, NAc and VP. See text for details.

increase in caregiving behaviour as the internal state of the adult-self transitions from an empathic state towards a more compassionate one.

5.6 Summary and Future Work

Based on an existing neuroanatomical model of how empathic states can motivate caregiving behaviour, fMRI data on self- and other-referential processing, and networks involved in the perception of pain in self and others, we presented a spiking neural model capable of describing three distinct empathy-related states: personal distress (involving emotional attunement within the context of a weak self-other distinction), weak empathic concern (prosocial motivation arising from emotional attunement, a strong self-other distinction, and perceptions of a relatively distant-other), and strong empathic concern (empathically-motivated prosocial behaviour arising from representations of a relatively close-other). We then extended this model to the case of Self-Attachment therapy: an attachment-based psychotherapy which involves a conceptualised adult-self empathising with an inner-child in order to motivate bonding behaviour towards them. We used this model to present a hypothesis as to how Self-Attachment might facilitate a transition within the adult-self from a state of personal distress, to one of strong empathic concern, to a compassionate stance towards the inner-child. We aimed to make our model biologically plausible with respect to connectivity and the relative number, and type, of neurons in each region, however there are a number of ways that its accuracy might be improved in future iterations.

The first point to note is that our model is highly compartmentalised. We didn't, for example, consider a full realisation of self- and other-pain networks, but rather argued

that current injection into the insular, aMCC and mPFC could feasibly represent activity in these two networks. Our model can thus be expanded in order to capture in more detail networks facilitating the perception of pain in self and other as this data becomes available, and also incorporate additional regions (such as those in the temporal poles, and the temporoparietal junction, which includes parts of the STS along with the inferior parietal lobule) that are commonly implicated in studies investigating cognitive empathy, theory of mind and mentalization (Walter, 2012; Abu-Akel and Shamay-Tsoory, 2011; Frith and Frith, 2006). Furthermore, we considered only anterior midcingulate parts of the ACC across empathy and compassion phases, whereas evidence (discussed above) suggests that more perigenual areas of the ACC are involved in compassionate states.

With respect to the regions that we did consider, our model is still highly simplified. Although all of our estimates for number of neurons were based on non-clinical human data, some were inaccurate due to lack of finer data (in particular the mPOA for which we used the volume of the whole encapsulating anterior-superior hypothalamic region, the PI where we used a neural density for adjacent AI, and the NAc for which we used an estimate for the total number of neurons in-spite of the caregiving pathway likely involving neurons more in the shell region). Furthermore, we defaulted to RS (for mPOA), FS (VP) and IB (VTA) neuron types in the absence of more detailed models, and typically only considered neuron types which either form a majority, or have been proposed as crucially important, in each region.

Due to a lack of human data, we used animal data in order to define connectivity, although this was not available in all cases. Accuracy of the model can thus be improved from this perspective as more connection data becomes available, and also by considering spatial connectivity (Voges et al., 2010) and potentially also laminar and columnar cortical structures. In addition, we didn't consider connections between AI and NAc, and mPFC and NAc. These connections were proposed in Numan's model to potentiate VP activation during empathic states, although details regarding axon terminals are for the time being unknown (Numan, 2014, p.278): future efforts can consider the nature of these additional connections. Finally, we assumed fixed connectivity weights with current-based synapses in order to demonstrate the three distinct states of the network, and so future work can consider conductance-based synapses and plasticity.

Studies that we highlighted have reported heightened AI activation with increasing perceptions of both closeness (Meyer et al., 2012) and in-group membership (Hein et al., 2010) of the other, and our strong (in contrast to weak) empathic concern state was broadly defined as involving representations of an other that was perceived as having these characteristics. Although our model replicated existing data on self and other perceptions of pain in the AI (with higher activation during the self-pain condition), along with appropriate activation in the mPFC during weak and strong empathic concern states and appropriate subsequent

activation in caregiving pathways in each case, activation in the AI across weak and strong empathic concern states did not significantly differ. Since (as we discussed) the mPFC is thought to be centrally involved in self-other representations (with more ventral parts encoding others with high self-relatedness, including closeness), and since Meyer et al. (2012) additionally report increased functional connectivity between the mPFC and AI for an empathic response directed towards a close as opposed to distant other, it might be that this effect is mediated by mPFC projections to AI that involve higher connectivity from ventral (encoding close-other representations) compared to more dorsal (encoding distant-others) areas. Future work can thus attempt to improve the model in this regard.

As discussed above, we previously hypothesised that one effect of the Self-Attachment bonding protocols is to stimulate OXT release from the PVNp, resulting in modulation of dopamine release in the VTA and enhanced vmPFC-ITC inhibition of the CeA and stress-related anti-social circuitry. In addition to this effect, we can predict that OXT release during the bonding protocols might enhance progress in the empathy protocols in a number of ways. As a result of known effects on receptors in the mPFC, mPOA, NAc and BLMA, OXT release should in general potentiate caregiving (and suppress withdrawal) motivation during application of the empathy protocols (Numan, 2014, p.287). In the case of the BLMA, we have considered the negatively-valent input stimulus as directly stimulating BLMA$\ominus$ in personal distress, empathic and compassionate states (with indirect stimulation of BLMA$\oplus$ occurring in empathic and compassionate states). However, OXT released during the bonding protocols might serve to facilitate additional and more direct stimulation of BLMA$\oplus$ (and suppression of BLMA$\ominus$ via stimulation of BLMAi). This means that we might expect the input stimulus to directly stimulate mostly BLMA$\ominus$ neurons during personal distress, but rather directly stimulate mostly BLMA$\oplus$ neurons during the empathic and compassionate states. Furthermore, OXT release might serve to strengthen the self-other distinction. In Colonnello et al. (2013), the ability of participants to differentiate their own identity was measured while they viewed a photo of themselves morphing into the photo of an unfamiliar face, with intranasal OXT shortening the time taken to differentiate self from other.

OXT released as a result of application of the bonding protocols might aid in the transition from an empathic to compassionately-motivated state in response to viewing the negatively-valenced inner-child stimulus. We previously hypothesised that OXT-modulated DA would increase reward predictions and firing rates in the mOFC, but OXT release might moreover be involved in the suppression of activity in AI areas crucially involved in the formation of negatively-valenced empathic states: Bos et al. (2015) found that empathy-related activation in the insular was strongly reduced after intranasal OXT in subjects observing others in pain. These studies highlight the strong interdependence between the bonding and empathy protocols in Self-Attachment, and the nature by which successful application

of each is likely to drive progress in the other. Future work can consider in more detail the effects of OXT release on empathically-motivated bonding, along with individual differences with regards to prior attachment experience. Since other types of human bonds (e.g. pair bonds) are thought to rely on overlapping circuitry between the amygdala, NAc and VP, and since OXT and DA release into the NAc is thought to result in plasticity that enhances activation in NAc-VP circuitry that promotes such attractions, future work can also consider the potential implications for other types of human bonds (Numan and Young, 2016).

6. Conclusions and Evaluation

Our first main aim was to expand the scope of computational and mathematical understandings of the different developmental trajectories that result in the distinct types of infant attachment. While a series of recent works have begun to explore attachment formation in the form of computational and mathematical models, the area remains relatively fertile ground - for example, our work is (to the best of our knowledge) the first to provide a computational account of disorganised attachment formation. Previous work has furthermore not tended to be grounded in computational neuroscience, and has not considered the large and influential body of work in attachment research concerned with disrupted patterns of affective communication.

In Chapter 3 we presented a model to account for the emergence of both organised and disorganised forms of infant attachment within the context of the free energy principle: a modern, holistic account of brain function. We argued that the fundamental principles of free energy minimisation and active inference conform to explanations of early infant attachment interactions (in terms of a balancing of exploration and caregiver interaction in order to achieve homeostatic regulation), and that this was an appropriate framework to use with respect to capturing a desire for stress reduction (due to links between stress and uncertainty, and the inherent drive towards the reduction of uncertainty under that framework). We built on an existing decision theoretic model which defined organised infant attachment in terms of optimal decisions (in response to caregiving attention or inattention) with respect to internal stress increases and decreases. We began by considering infant agents that act, perceive and learn in order to minimise free energy over interoceptive states related to stress, without any a priori knowledge about the responsiveness characteristics of their particular caregiver. Our model resulted in behaviour resembling the three classical organised (secure, avoidant and ambivalent) forms of attachment for this infant agent when they interacted with caregivers with different responsiveness characteristics. In particular, policies involving sequentially consistent and organised forms of secure (proximity seeking) behaviour emerged for infants who interacted with caregivers who, with high probability, reduced infant stress on approach. On the other hand, infants interacting with highly unresponsive caregivers (who with high probability increased infant stress on approach) came

to prefer policies consisting of sequentially consistent avoidance, whereas infants interacting with inconsistent caregivers came to prefer guarded (resistant) proximity seeking behaviour resembling an element of ambivalence. In accordance with evidence for heightened prevalence of ACEs in caregivers of both ambivalent and disorganised infants, we then expanded on this model to additionally consider exteroceptive cues from the caregiver. We showed how delivery of a particular exteroceptive cue (which has previously been associated with caregivers of disorganised infants) that is misleading for the infant with regards to subsequent caregiving behaviour could have a disorganising effects in infants of caregivers who, with high probability, increased infant stress on approach. While empirical evidence suggests links between ACEs and caregivers of ambivalent infants, to the best of our knowledge no particular ACE items on the AMBIANCE scale have as yet been associated with ambivalent infant attachment. Accordingly, we explored infant behavioural outcomes in response to various combinations of misleading and ambiguous ACE exteroceptive cues, showing how a number of combinations of such cues can indeed have an organising (towards ambivalence) effect in infants of these inconsistent caregivers.

In prior work in the field, the attachment secure-base exploration paradigm was considered within the context of a robot learning to balance exploration of their environment with proximity seeking for arousal reduction (Hiolle et al., 2012, 2014). Moreover, strong (i.e. multiply learned) patterns in a Hopfield network have also previously been proposed as a model of how an infant might come to conform to a particular pattern of attachment (Edalat and Mancinelli, 2013; Edalat, 2013a). However, in contrast to our work, these contributions did not address any specific differences in infant attachment types and their resulting emergent patterns of behaviour in response to caregiver interaction. Buono et al. (2006) and Stevens and Zhang (2009) considered an infant's internal state and external behaviour in response to interaction with a caregiver with varying impact on their physiology in the form of dynamical systems models. While these works were able to describe organised forms of attachment in terms of emergent differences in behaviour and internal state, in contrast to our work they did not consider the (clinically most important) disorganised form of attachment, nor were they grounded in computational neuroscience, or the influential body of work in the attachment literature on disrupted patterns of affective communication within insecure attachment dyads.

The most extensive prior work to consider the acquisition of attachment types in response to environmental impacts on internal state was Petters (2006a). Following Bowlby's control systems theory of attachment, the author designed a cognitive agent architecture that could account for the emergence of secure, avoidant and ambivalent infant attachment defined in terms of proximity seeking and signalling behaviour. The architecture differs notably from ours in its level of abstraction: while Petter's work described organised attachment formation by means of cognitive levels of processing which explicitly switched the infant's

goals (e.g. from exploration to security) based on the impact of external sensory data, we instead considered an agent with a single 'goal' (minimisation of surprise over interoceptive states related to stress, achieved by the minimisation of free energy) and the emergence of organised behaviour in response to this, during ISS reunion-type interactions (without considering the environmental exploration aspect). Although focusing on a broader scenario than our work here (i.e. capturing the secure-base exploration paradigm), similarly to the works above Petter's architecture also did not capture (the clinically most important) disorganised forms of infant attachment, nor did it consider the disrupted patterns of affective communication empirically linked to both disorganisation and ambivalence. Furthermore, while we captured resistant behaviours in ambivalent infants in terms of a guarding of caregiving behaviour in order to modulate the impact of this behaviour on internal state, these architectures instead accounted for ambivalent attachment in the form of prolonged signalling while in the presence of the caregiver (proposed to correspond to the heightened and prolonged distress observed in these infants). We instead proposed that this aspect (heightened and prolonged distress) of ambivalent attachment might be captured under our model in future work: in line with theories regarding this self-increase in stress in the infant as a strategy for eliciting caregiving attention, we might (for example) introduce an additional "hyperactivation" action that (with high probability) leads to hidden states associated with interoceptive observations involving increased internal stress, but which are also associated with relatively high probability of subsequently achieving caregiving attention. Our previous effort to synthesise the work of Petters (2006a) and Hiolle et al. (2012) in the form of an arousal-based neurocognitive architecture (Cittern and Edalat, 2014) did attempt to explain the acquisition of basic infant attachment behaviour and physiology for the organised and disorganised types within the secure-base exploration paradigm. Under that model, arousal levels during environment exploration are driven by a measure of novelty (as in Hiolle et al. (2012)) with an adaptive safe-range distance determining proximity seeking behaviour (similar to as in Petters (2006a)). While disorganised attachment behaviour was captured by us there in terms of the (presumed to be fear induced) freezing behaviour observed on ISS reunion, we instead captured a different aspect of disorganisation in our free energy based model described here: namely the inability to form a coherent and consistent behavioural policy. Unlike that previous work in which disorganised-freeze behaviour occurred as a result of overtly hostile caregiving behaviour, in our free energy model of attachment here disorganised policies resulted more specifically from exteroceptive affective cues that were misleading for the infant with respect to caregiving behaviour that subsequently would increase infant stress on approach.

As the first computational model of infant attachment formation to be grounded firmly in computational neuroscience, and additionally the first to consider both disorganisation (in terms of an inability to form a sequentially consistent/coherent behavioural policy) and

disrupted patterns of affective communication, our work has helped to bridge the attachment and computational modelling/neuroscience literature. Our model has furthermore made a new prediction that can be tested empirically: namely that particular combinations of misleading and ambiguous ACE items on the AMBIANCE scale will lead to the most organised forms of ambivalent attachment.

A better understanding of attachment formation, from a computational neuroscience perspective, was helpful groundwork in achieving our second aim, which was to motivate neurobiological hypotheses with respect to the effects of the recently introduced Self-Attachment psychotherapy. Self-Attachment therapy has proven successful in pre-clinical trials, but in order to move towards clinical trials (and to refine and improve the therapy going forwards) a clear and precise account of the neurobiological effects of the therapy on the individual's brain is desirable.

In Chapter 4 we presented a neurobiological hypothesis with regards to the underlying effects of the Self-Attachment bonding protocols, which are concerned with the creation of an abstract, self-directed bond. Based on a previous neuroanatomical model describing the control over balance in activation of neural circuitry involved in stress and facilitative reactivity to social stimuli, along with a computational model describing how reward-driven updates to representations can reduce activity in circuitry involved in the stress reaction, we showed how a counter-conditioning procedure (linked to application of the Self-Attachment bonding protocols) might serve to drive a rebalancing of activity in these circuits. In accordance with neurobiological evidence on parenting and bonding, we considered in particular how OXT-modulated DA might drive this rebalancing in activation, with additional reward proposed to be introduced as a result of the representations and activities involved in the therapy and resulting in DA-encapsulated reward prediction errors that served to strengthen stress inhibitory pathways. Our model was simulated computationally, and (in alignment with existing empirical data) we suggested that the resulting rebalancing in activation corresponds to the emergence of increasingly secure forms of attachment.

We followed this up in Chapter 5 by considering how empathic states directed towards the conceptualised inner-child might be used in order to generate motivation for application of these bonding protocols. Based on an existing neuroanatomical model of how empathic states can motivate caregiving behaviour, along with data on the neural underpinnings of self- and other- referential processing, we presented a spiking neural model capable of describing three distinct empathy-related states: personal distress (involving emotional attunement within the context of a weak self-other distinction), weak empathic concern (prosocial motivation arising from emotional attunement, a strong self-other distinction, and perceptions of a relatively distant-other), and strong empathic concern (empathically-motivated prosocial behaviour arising from representations of a relatively close-other). This model was then extended and applied to the case of Self-Attachment therapy, for which we hypoth-

esised a transition within the adult-self from personal distress, through strong empathic concern, to a compassionate stance towards the inner-child. We aimed to make our model as biologically accurate as possible with respect to connectivity and the relative number, and type, of neurons in each brain region considered.

A couple of previous works had attempted to model psychotherapy within the context of attachment theory. Edalat and Lin (2014) presented a computational neural model of mentalization-based psychotherapy from an attachment-developmental perspective, in order to explain how a rebalancing in activation of neural circuitry might occur in response to the introduction of additional reward, which was proposed to drive a tendency towards more deliberative and cognitive forms of decision making. Although focusing on key attachment-related neural circuitry, in contrast to our work here that model does not consider Self-Attachment therapy and its protocols directly, but rather how the introduction of additional reward might encourage more cognitive forms of decision making in individuals who initially show a tendency towards emotional decision making as a result of adverse early attachment experience. Edalat (2017a) describes a reinforcement learning procedure in a multi-agent environment, in which additional reward is supplied in order to incentivise secure behaviour in an attachment game played by the adult-self and inner-child. This gives a model of the reward-driven Self-Attachment intervention, however (in contrast to the work here) that model did not consider in detail the neurobiological regions or processes underpinning the bonding and empathy protocols involved in the therapy.

We hope that the work undertaken here can be instructive as we progress towards clinical trials; assisting in the clarification of issues such as (for example) along what dimensions the therapy should be customised for different individuals, and helping to frame clear descriptions of the (hypothesised) effects of the treatment for participants. From a broader perspective, our work is unique in the sense that it is the first to attempt to combine neurobiological data on the self-other distinction, bonding, parenting and empathy in order to provide a hypothesised model for the creation of fully internalised self-directed attachments, rudimentary forms of which might also manifest outside of Self-Attachment therapy (Edalat, 2017a).

6.1 Limitations

With respect to our free energy model of infant attachment formation, we were able to explain basic patterns relating to the emergence of organised attachment in infant agents who minimise free energy with respect only to their interoceptive states. We furthermore explained disorganised attachment by disruptions in patterns of affective communication conveyed by exteroceptive cues. However (as discussed in more depth in Section 3.5) there are aspects of ambivalent (hyperactivation) and disorganised (e.g. dissociation) attachment

behaviour that we did not attempt to capture, and there are furthermore many patterns of disrupted affective communication that have been observed empirically in caregivers but were not considered here. We also did not attempt to capture processes giving rise to the intergenerational transmission of attachment types (i.e. the computational mechanisms dictating how and why a disorganised infant themself tends to become a caregiver of a disorganised infant later in life), which is an important concern for future work.

We focused on one aspect of the Self-Attachment therapy, which is the creation of an internalised attachment bond, along with how particular self-other representations might optimally motivate its creation. Self-Attachment therapy describes a number of other protocols not considered here, for example the recalling from memory of a previously traumatic experience which is re-imagined as having an alternative (positive) outcome. Furthermore, our models with respect to self-directed bonding were highly compartmentalised, focusing only on brain regions which (according to current research) are thought to be most significantly (but by no means solely) involved in this process. Although a large amount of the data used to motivate our models was from human studies, some was based on animal research in the absence of research on human brains. Since we are attempting to motivate a therapy for humans this is clearly a shortcoming, although one that can hopefully be rectified over the coming years as more data from human studies becomes available. Perhaps most significantly, although we did consider differences in activation in brain regions within the model according to attachment types and related disorders (such as BPD), we did not make any detailed proposals with regards to how the Self-Attachment therapeutic process might be tailored accordingly (although, in the case of BPD, we did suggest that the empathy protocols be supplemented with mentalization-based therapies and techniques in order to assist in the strengthening of self-other distinctions between the adult-self and inner-child). This is an important consideration moving forwards.

6.2 Future Directions

Our neurobiological understanding of attachment, parenting, bonding, trauma and early development is increasing rapidly, with the amount of new data continuing to grow at a fast pace, and so there is much scope for refinement and expansion of the models moving forwards. In addition, as we discussed in detail in the final sections of each chapter, there are a number of ways in which the models can be expanded in scope according to currently available data. With respect to our account of infant attachment formation according to the free energy principle, we could for example consider additional aspects of ambivalent (hyperactivation of stress and attachment system) and disorganised (e.g. freezing/dissociative) behaviours seen empirically in these infants. As well as considering the remaining types of exteroceptive cues defined by the ACE dimension of the AMBIANCE scale, we might

additionally attempt to capture other dimensions of disrupted affective communication on this scale that have also been associated with the emergence of disorganised and ambivalent forms of attachment. In the case of disorganised attachment, a more comprehensive account of patterns of disrupted affective communication may facilitate a distinction in the model between those infants subclassified as D-Secure and D-Insecure with respect to organised behavioural tendencies. The scope of the model might also be extended to account for secure-base environment exploration, hierarchical generative models and contextualisation, or an account of Self-Attachment driven changes to an individual's suboptimal attachment schema. Our models of Self-Attachment therapy, too, might be expanded in scope with respect to existing data: for example, in our model of the Self-Attachment bonding protocols we did not consider effects of OXT on direct modulation of regions (such as the amygdala) involved in stress reactivity. Furthermore, evidence suggests a role for VA, a neuropeptide with a similar chemical structure to OXT which is also released from the PVN, in bonding and social behaviour more generally. In particular, VA might play roles in selective attention and the modulation of memory and stress in social situations, and has (under certain situations) been associated with reduced capacity for forms of empathy (Bachner-Melman and Ebstein, 2014), all of which might be important considerations within this context. Particularly interesting would be further consideration of the interplay between application of the bonding and empathy protocols. According to known effects on receptors in the mPFC, mPOA, NAc and BLMA, OXT release during self-directed bonding should potentiate caregiving motivation (and might additionally enhance capacities for differentiation of the adult-self from the inner-child). Evidence suggests that OXT release might also assist in transition from empathic to compassionate states, via suppression of activity in AI and BLMA areas involved in formation of negatively-valenced empathic states.

Although we have made suggestions for future enhancements, a key aim now is to test the hypotheses and predictions presented in this thesis. While the new predictions made by our free energy model of infant attachment are behavioural in nature, the hypotheses relating to Self-Attachment therapy involve neurobiological effects within the individual's brain which can be tested with neuroimaging. Relevant to this, Barsaglini et al. (2014) conducted a review of imaging studies on the neurobiological effects on adults of psychotherapy in comparison to pharmacotherapy. The individual studies included PET and fMRI investigations into a range of psychotherapeutic treatments (including cognitive-behavioural and virtual reality exposure therapies) on individuals with disorders such as obsessive-compulsive, unipolar major depressive and post-traumatic stress. In many cases, the authors found evidence to suggest that the effects of psychotherapy were similar to pharmacotherapy in terms of changes in both activation in, and functional integration of, brain regions otherwise associated with each disorder. These areas included the amygdala, OFC, mPFC, cingulate cortex, insular and basal ganglia, all of which are specific regions of interest for

the Self-Attachment interventions explored in this thesis. Two predictions in particular that we have made which can be examined empirically in similar neuroimaging studies are the preferential activation of motivational circuitry in response to empathising with the inner-child representation (as opposed to alternative representations), and the effects of the inner-child directed bonding protocols on activation and integration in networks spanning the OFC/vmPFC and amygdala.

Another important focus for future work, underway now, is the creation of technological tools to assist in administration of the protocols and potentially enhance their effectiveness. These include a mobile software application to provide information on the therapy, instructions for its administration, and tracking of the individual's progress; and an immersive virtual reality environment in which the individual can experience the Self-Attachment protocols from the perspectives of both the adult-self and inner-child. We hope that the development and functioning of these tools will be informed and aided by the work undertaken here.

Bibliography

Abraham, A. D., K. A. Neve, and K. M. Lattal (2014). Dopamine and extinction: A convergence of theory with fear and reward circuitry. *Neurobiology of learning and memory 108*, 65–77.

Abrams, K. Y., A. Rifkin, and E. Hesse (2006). Examining the role of parental frightened/frightening subtypes in predicting disorganized attachment within a brief observational procedure. *Development and Psychopathology 18*(02), 345–361.

Abu-Akel, A. and S. Shamay-Tsoory (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia 49*(11), 2971–2984.

Ainsworth, M. D. S. (1963). The development of infant-mother interaction among the Ganda. *Determinants of infant behavior 2*, 67–112.

Ainsworth, M. D. S. (1967). *Infancy in Uganda: Infant care and the growth of love.* Johns Hopkins Press.

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

Al-Hasani, R., J. Foster, A. Metaxas, C. Ledent, S. Hourani, I. Kitchen, and Y. Chen (2011). Increased desensitization of dopamine D 2 receptor-mediated response in the ventral tegmental area in the absence of adenosine A 2A receptors. *Neuroscience 190*, 103–111.

Allen, J. P., S. T. Hauser, and E. Borman-Spurrell (1996). Attachment theory as a framework for understanding sequelae of severe adolescent psychopathology: an 11-year follow-up study. *Journal of Consulting and Clinical psychology 64*(2), 254.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub.

Anderson, K. J. (1990). Arousal and the inverted-u hypothesis: A critique of Neiss's "Reconceptualizing arousal.". *Psychological bulletin 107*(1), 96–100.

Bachner-Melman, R. and R. Ebstein (2014). The role of oxytocin and vasopressin in emotional and social behaviors. In E. Fliers, M. Korbonits, and J. A. Romijn (Eds.), *Handbook of clinical neurology*, Volume 124, pp. 53. Elsevier.

Baird, A., B.-K. Dewar, H. Critchley, R. Dolan, T. Shallice, and L. Cipolotti (2006). Social and emotional functions in three patients with medial frontal lobe damage including the anterior cingulate cortex. *Cognitive Neuropsychiatry 11*(4), 369–388.

Baker, H. S. and M. N. Baker (1987). Heinz Kohut's self psychology: An overview. *American Journal of Psychiatry 144*(1), 1–9.

Bakermans-Kranenburg, M. J., M. H. Van Ijzendoorn, and F. Juffer (2003). Less is more: meta-analyses of sensitivity and attachment interventions in early childhood. *Psychological bulletin 129*(2), 195.

Baldi, E. and C. Bucherelli (2005). The inverted "u-shaped" dose-effect relationships in learning and memory: modulation of arousal and consolidation. *Nonlinearity in Biology, Toxicology, and Medicine 3*(1), 9–21.

Barbas, H. (2007). Flow of information for emotions through temporal and orbitofrontal pathways. *Journal of Anatomy 211*(2), 237–249.

Barbey, A. K., M. Koenigs, and J. Grafman (2013). Dorsolateral prefrontal contributions to human working memory. *Cortex 49*(5), 1195–1205.

Barrett-Lennard, G. T. (1981). The empathy cycle: Refinement of a nuclear concept. *Journal of counseling psychology 28*(2), 91.

Barsaglini, A., G. Sartori, S. Benetti, W. Pettersson-Yeo, and A. Mechelli (2014). The effects of psychotherapy on brain function: A systematic and critical review. *Progress in neurobiology 114*, 1–14.

Bartels, A. and S. Zeki (2004). The neural correlates of maternal and romantic love. *Neuroimage 21*(3), 1155–1166.

Bartholomew, K. and L. M. Horowitz (1991). Attachment styles among young adults: a test of a four-category model. *Journal of personality and social psychology 61*(2), 226.

Bartz, J., D. Simeon, H. Hamilton, S. Kim, S. Crystal, A. Braun, V. Vicens, and E. Hollander (2011, a). Oxytocin can hinder trust and cooperation in borderline personality disorder. *Social Cognitive and Affective Neuroscience 6*(5), 556.

Bartz, J., J. Zaki, K. N. Ochsner, N. Bolger, A. Kolevzon, N. Ludwig, and J. E. Lydon (2010, b). Effects of oxytocin on recollections of maternal care and closeness. *Proceedings of the National Academy of Sciences 107*(50), 21371–21375.

Bateman, A. W. and P. Fonagy (2012). *Handbook of mentalizing in mental health practice*. American Psychiatric Pub.

Beauchamp, M. S. (2015). The social mysteries of the superior temporal sulcus. *Trends in cognitive sciences 19*(9), 489–490.

Belsky, J. and M. Rovine (1987). Temperament and attachment security in the strange situation: An empirical rapprochement. *Child Development*, 787–795.

Beyeler, M., K. D. Carlson, T.-S. Chou, N. Dutt, and J. L. Krichmar (2015). CARLsim 3: A user-friendly and highly optimized library for the creation of neurobiologically detailed spiking neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE.

Bolam, J. P. and E. K. Pissadaki (2012). Living on the edge with too many mouths to feed: why dopamine neurons die. *Movement Disorders 27*(12), 1478–1483.

Bos, P. A., E. R. Montoya, E. J. Hermans, C. Keysers, and J. van Honk (2015). Oxytocin reduces neural activity in the pain circuitry when seeing pain in others. *NeuroImage 113*, 217–224.

Bowlby, J. (1944). Forty-four juvenile thieves: Their characters and home life. *International Journal of Psychoanalysis 25*(19-52), 107–127.

Bowlby, J. (1958). The nature of the child's tie to his mother. *The International journal of psycho-analysis 39*(5), 350.

Bowlby, J. (1960a). Grief and mourning in infancy and early childhood. *Psychoanalytic study of the child 15*(9), 92.

Bowlby, J. (1960b). Separation anxiety. *The International Journal of Psychoanalysis 41*, 89.

Bowlby, J. (1969). *Attachment and loss: Attachment (Vol 1)*. Basic Books (New York).

Bowlby, J. (1973). *Attachment and loss: Separation, anxiety and anger (Vol 2)*. Basic Books (New York).

Bowlby, J. (1980). *Attachment and loss: Loss, sadness and depression (Vol. 3)*. Basic Books (New York).

Bowlby, J. (1982). Attachment and loss: retrospect and prospect. *American journal of Orthopsychiatry 52*(4), 664.

Bowlby, J. (1988). *A Secure Base*. Basic Books (New York).

Bowlby, J. et al. (1951). *Maternal care and mental health.* World Health Organization Geneva.

Bowlby, J. and J. Robertson (1953). A two-year-old goes to hospital. *Proceedings of the Royal Society of Medicine 46*(6), 425.

Braun, C. M., R. Daigneault, S. Gaudelet, and A. Guimond (2008). Diagnostic and statistical manual of mental disorders, fourth edition symptoms of mania: which one(s) result(s) more often from right than left hemisphere lesions? *Comprehensive psychiatry 49*(5), 441–459.

Brennan, K. A., C. L. Clark, and P. R. Shaver (1998). Self-report measurement of adult attachment: An integrative overview. In J. A. E. Simpson and W. S. E. Rholes (Eds.), *Attachment theory and close relationships*, pp. 46–76. New York, NY, US: Guilford Press.

Bretherton, I. (1985). Attachment theory: Retrospect and prospect. *Monographs of the society for research in child development*, 3–35.

Bromberg-Martin, E. S., M. Matsumoto, and O. Hikosaka (2010). Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron 68*(5), 815–834.

Bronfman, E., E. Parsons, and K. Lyons-Ruth (1999). Atypical Maternal Behavior Instrument for Assessment and Classification (AMBIANCE): Manual for coding disrupted affective communication. *Unpublished manuscript. Cambridge, MA: Harvard Medical School*.

Brown, S., M. J. Martinez, and L. M. Parsons (2004). Passive music listening spontaneously engages limbic and paralimbic systems. *Neuroreport 15*(13), 2033–2037.

Brudzynski, S. M., M. Wu, and G. J. Mogenson (1993). Decreases in rat locomotor activity as a result of changes in synaptic transmission to neurons within the mesencephalic locomotor region. *Canadian journal of physiology and pharmacology 71*(5-6), 394–406.

Buchheim, A., S. Erk, C. George, H. Kächele, M. Ruchsow, M. Spitzer, T. Kircher, and H. Walter (2006). Measuring attachment representation in an fMRI environment: A pilot study. *Psychopathology 39*(3), 144–152.

Buono, L., R. Chau, G. Lewis, N. Madras, M. Pugh, L. Rossi, and T. Witelski (2006). Mathematical models of mother/child attachment. *Fields-MITACS Industrial Problem Solving Workshop August 2006*.

Byne, W., M. S. Lasco, E. Kemether, A. Shinwari, M. A. Edgar, S. Morgello, L. B. Jones, and S. Tobet (2000). The interstitial nuclei of the human anterior hypothalamus: an investigation of sexual variation in volume and cell size, number and density. *Brain research 856*(1), 254–258.

Carlson, E. A., B. Egeland, and L. A. Sroufe (2009). A prospective investigation of the development of borderline personality symptoms. *Development and psychopathology 21*(04), 1311–1334.

Carlson, V., D. Cicchetti, D. Barnett, and K. Braunwald (1989). Disorganized/disoriented attachment relationships in maltreated infants. *Developmental psychology 25*(4), 525.

Carran, M., C. Kohler, M. O'Connor, W. Bilker, and M. Sperling (2003). Mania following temporal lobectomy. *Neurology 61*(6), 770–774.

Cavanna, A. E. and M. R. Trimble (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain 129*(3), 564–583.

Chan, R. C., D. Shum, T. Toulopoulou, and E. Y. Chen (2008). Assessment of executive functions: Review of instruments and identification of critical issues. *Archives of Clinical Neuropsychology 23*(2), 201–216.

Chen, T. L., C. Babiloni, A. Ferretti, M. G. Perrucci, G. L. Romani, P. M. Rossini, A. Tartaro, and C. Del Gratta (2008). Human secondary somatosensory cortex is involved in the processing of somatosensory rare stimuli: an fmri study. *Neuroimage 40*(4), 1765–1771.

Chiron, C., I. Jambaque, R. Nabbout, R. Lounes, A. Syrota, and O. Dulac (1997). The right brain hemisphere is dominant in human infants. *Brain 120*(6), 1057–1065.

Christov-Moore, L., E. A. Simpson, G. Coudé, K. Grigaityte, M. Iacoboni, and P. F. Ferrari (2014). Empathy: Gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews 46*, 604–627.

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

Cittern, D. and A. Edalat (2015a). Reinforcement learning for Nash equilibrium generation. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1727–1728. International Foundation for Autonomous Agents and Multiagent Systems.

Cittern, D. and A. Edalat (2015b). Reinforcement learning for Nash equilibrium generation: Extended version. `http://humandevelopment.doc.ic.ac.uk/papers/rlneg.pdf`.

Cittern, D. and A. Edalat (2015c). Towards a neural model of bonding in self-attachment. In *2015 International Joint Conference on Neural Networks (IJCNN)*.

Cohen, M. and P. Shaver (2004). Avoidant attachment and hemispheric lateralisation of the processing of attachment-and emotion-related words. *Cognition and Emotion 18*(6), 799–813.

Colonnello, V., F. S. Chen, J. Panksepp, and M. Heinrichs (2013). Oxytocin sharpens self-other perceptual boundary. *Psychoneuroendocrinology 38*(12), 2996–3002.

Connors, B. W. and M. J. Gutnick (1990). Intrinsic firing patterns of diverse neocortical neurons. *Trends in neurosciences 13*(3), 99–104.

Cozolino, L. (2010). *The Neuroscience of Psychotherapy: Healing the Social Brain (Norton Series on Interpersonal Neurobiology)*. WW Norton & Company.

Cozolino, L. (2014). *The Neuroscience of Human Relationships: Attachment and the Developing Social Brain (Norton Series on Interpersonal Neurobiology)*. WW Norton & Company.

Craig, A. (2011). Significance of the insula for the evolution of human awareness of feelings from the body. *Annals of the New York Academy of Sciences 1225*, 72–82.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature reviews neuroscience 10*(1).

Creed, M. C., N. R. Ntamati, and K. R. Tan (2014). VTA GABA neurons modulate specific learning behaviors through the control of dopamine and cholinergic systems. *Frontiers in behavioral neuroscience 8*.

Critchley, H. D., S. Wiens, P. Rotshtein, A. Öhman, and R. J. Dolan (2004). Neural systems supporting interoceptive awareness. *Nature Neuroscience 7*(2), 189–195.

Damasio, A. R., B. J. Everitt, and D. Bishop (1996). The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences 351*(1346), 1413–1420.

Dannlowski, U., A. Stuhrmann, V. Beutelmann, P. Zwanzger, T. Lenzen, D. Grotegerd, K. Domschke, C. Hohoff, P. Ohrmann, J. Bauer, et al. (2012). Limbic scars: long-term consequences of childhood maltreatment revealed by functional and structural magnetic resonance imaging. *Biological psychiatry 71*(4), 286–293.

Davidson, R. J. and N. A. Fox (1989). Frontal brain asymmetry predicts infants' response to maternal separation. *Journal of abnormal psychology 98*(2), 127.

Davis, M. (1992). The role of the amygdala in fear and anxiety. *Annual review of neuroscience 15*(1), 353–375.

Dawson, G., S. B. Ashman, D. Hessl, S. Spieker, K. Frey, H. Panagiotides, and L. Embry (2001). Autonomic and brain electrical activity in securely-and insecurely-attached infants of depressed mothers. *Infant Behavior and Development 24*(2), 135–149.

de Berker, A. O., R. B. Rutledge, C. Mathys, L. Marshall, G. F. Cross, R. J. Dolan, and S. Bestmann (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature communications 7*.

Decety, J. and M. Meyer (2008). From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and psychopathology 20*(04), 1053–1080.

Decety, J. and E. C. Porges (2011). Imagining being the agent of actions that carry different moral consequences: an fMRI study. *Neuropsychologia 49*(11), 2994–3001.

Declerck, C. H., C. Boone, and T. Kiyonari (2010). Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. *Hormones and Behavior 57*(3), 368–374.

Dégenètais, E., A.-M. Thierry, J. Glowinski, and Y. Gioanni (2002). Electrophysiological properties of pyramidal neurons in the rat prefrontal cortex: an in vivo intracellular recording study. *Cerebral Cortex 12*(1), 1–16.

Denny, B. T., H. Kober, T. D. Wager, and K. N. Ochsner (2012). A meta-analysis of functional neuroimaging studies of self-and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of cognitive Neuroscience 24*(8), 1742–1752.

Donegan, N. H., C. A. Sanislow, H. P. Blumberg, R. K. Fulbright, C. Lacadie, P. Skudlarski, J. C. Gore, I. R. Olson, T. H. McGlashan, and B. E. Wexler (2003). Amygdala hyperreactivity in borderline personality disorder: implications for emotional dysregulation. *Biological psychiatry 54*(11), 1284–1293.

Donovan, W. L. and L. A. Leavitt (1985). Physiologic assessment of mother-infant attachment. *Journal of the American Academy of Child Psychiatry 24*(1), 65–70.

Düzel, E., N. Bunzeck, M. Guitart-Masip, B. Wittmann, B. H. Schott, and P. N. Tobler (2009). Functional imaging of the human dopaminergic midbrain. *Trends in neurosciences 32*(6), 321–328.

Edalat, A. (2013a). Capacity of strong attractor patterns to model behavioural and cognitive prototypes. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2661–2669.

Edalat, A. (2013b). Self-attachment: A new and integrative psychotherapy (presented at the Institute of Psychiatry, Kings College London).

Edalat, A. (2015). Introduction to self-attachment and its neural basis. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE.

Edalat, A. (2017a). Self-attachment: A holistic approach to computational psychiatry. In A. C. Peter Erdi, Basabdatta Sen Bhattacharya (Ed.), *Computational Neurology and Psychiatry*, Springer series of Bio/Neuroinformatics. Springer.

Edalat, A. (2017b). Self-attachment: A self-administrable intervention for chronic anxiety and depression. Technical report, Department of Computing, Imperial College London.

Edalat, A. and Z. Lin (2014). A neural model of mentalization/mindfulness based psychotherapy. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pp. 2743–2751.

Edalat, A. and F. Mancinelli (2013). Strong attractors of Hopfield neural networks to model attachment types and behavioural patterns. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–10. IEEE.

Engen, H. G. and T. Singer (2013). Empathy circuits. *Current Opinion in Neurobiology 23*(2), 275–282.

Engen, H. G. and T. Singer (2015). Compassion-based emotion regulation up-regulates experienced positive affect and associated neural networks. *Social cognitive and affective neuroscience*.

Falconer, C. J., A. Rovira, J. A. King, P. Gilbert, A. Antley, P. Fearon, N. Ralph, M. Slater, and C. R. Brewin (2016). Embodying self-compassion within virtual reality and its effects on patients with depression. *British Journal of Psychiatry Open 2*(1), 74–80.

Falconer, C. J., M. Slater, A. Rovira, J. A. King, P. Gilbert, A. Antley, and C. R. Brewin (2014). Embodying compassion: A virtual reality paradigm for overcoming excessive self-criticism. *PloS one 9*(11).

Fehse, K., S. Silveira, K. Elvers, and J. Blautzik (2015). Compassion, guilt and innocence: an fmri study of responses to victims who are responsible for their fate. *Social neuroscience 10*(3), 243–252.

Feinstein, J. S., R. Adolphs, A. Damasio, and D. Tranel (2011). The human amygdala and the induction and experience of fear. *Current biology 21*(1), 34–38.

Feldman, R., A. Weller, O. Zagoory-Sharon, and A. Levine (2007). Evidence for a neuroendocrinological foundation of human affiliation plasma oxytocin levels across pregnancy and the postpartum period predict mother-infant bonding. *Psychological Science 18*(11), 965–970.

Field, T., M. Hernandez-Reif, M. Diego, S. Schanberg, and C. Kuhn (2005). Cortisol decreases and serotonin and dopamine increase following massage therapy. *International Journal of Neuroscience 115*(10), 1397–1413.

Fischer, A. and C. Igel (2014). Training restricted Boltzmann machines: An introduction. *Pattern Recognition 47*(1), 25–39.

FitzGerald, T. H., R. J. Dolan, and K. Friston (2015). Dopamine, reward learning, and active inference. *Frontiers in computational neuroscience 9*, 136.

Flor-Henry, P., J. C. Lind, and Z. J. Koles (2004). A source-imaging (low-resolution electromagnetic tomography) study of the eegs from unmedicated males with depression. *Psychiatry Research: Neuroimaging 130*(2), 191–207.

Fonagy, P. (2000). Attachment and borderline personality disorder. *Journal of the american psychoanalytic association 48*(4), 1129–1146.

Fonagy, P., M. Target, and G. Gergely (2000). Attachment and borderline personality disorder: A theory and some evidence. *Psychiatric Clinics of North America 23*(1), 103–122.

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience 11*(2), 127–138.

Friston, K., T. FitzGerald, F. Rigoli, P. Schwartenbeck, J. O'Doherty, and G. Pezzulo (2016). Active inference and learning. *Neuroscience & Biobehavioral Reviews 68*, 862–879.

Friston, K., F. Rigoli, D. Ognibene, C. Mathys, T. Fitzgerald, and G. Pezzulo (2015). Active inference and epistemic value. *Cognitive neuroscience 6*(4), 187–214.

Friston, K. J., J. Daunizeau, J. Kilner, and S. J. Kiebel (2010). Action and behavior: a free-energy formulation. *Biological cybernetics 102*(3), 227–260.

Frith, C. D. and U. Frith (2006). The neural basis of mentalizing. *Neuron 50*(4), 531–534.

Frodl, T., E. Reinhold, N. Koutsouleris, M. Reiser, and E. M. Meisenzahl (2010). Interaction of childhood stress with hippocampus and prefrontal cortex volume reduction in major depression. *Journal of psychiatric research 44*(13), 799–807.

García-Amado, M. and L. Prensa (2012). Stereological analysis of neuron, glial and endothelial cell numbers in the human amygdaloid complex. *PloS one 7*(6).

George, C., M. West, and O. Pettem (1999). *The Adult Attachment Projective: Disorganization of adult attachment at the level of representation.* Guilford Press.

Gerdes, K. E. and E. Segal (2011). Importance of empathy for social work practice: Integrating new science. *Social Work 56*(2), 141–148.

German, D., D. Schlusselberg, and D. Woodward (1983). Three-dimensional computer reconstruction of midbrain dopaminergic neuronal populations: from mouse to man. *Journal of neural transmission 57*(4), 243–254.

Ghaziri, J., A. Tucholka, G. Girard, J.-C. Houde, O. Boucher, G. Gilbert, M. Descoteaux, S. Lippé, P. Rainville, and D. K. Nguyen (2015). The corticocortical structural connectivity of the human insula. *Cerebral Cortex*, 13.

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in psychiatric treatment 15*(3), 199–208.

Giudice, S., R. Thatcher, and R. Walker (1987). Human cerebral hemispheres develop at different rates and ages. *Science 236*, 1110.

Gleichgerrcht, E. and J. Decety (2013). Empathy in clinical practice: how individual dispositions, gender, and experience moderate empathic concern, burnout, and emotional distress in physicians. *PLoS One 8*(4).

Glickstein, M. (2007). What does the cerebellum really do? *Current Biology 17*(19), R824–R827.

Gold, B. P., M. J. Frank, B. Bogert, and E. Brattico (2013). Pleasurable music affects reinforcement learning according to the listener. *Frontiers in psychology 4*.

Goldberg, S. (2000). *Attachment and development*. Oxford University Press.

Goldberg, S., D. Benoit, K. Blokland, and S. Madigan (2003). Atypical maternal behavior, maternal representations, and infant disorganized attachment. *Development and psychopathology 15*(02), 239–257.

Goldman-Rakic, P. S. (1987). Development of cortical circuitry and cognitive function. *Child development*, 601–622.

Golkar, A., T. B. Lonsdorf, A. Olsson, K. M. Lindstrom, J. Berrebi, P. Fransson, M. Schalling, M. Ingvar, and A. Öhman (2012). Distinct contributions of the dorsolateral prefrontal and orbitofrontal cortex during emotion regulation. *PLoS One 7*(11).

Gonzalez-Liencres, C., S. G. Shamay-Tsoory, and M. Brüne (2013). Towards a neuroscience of empathy: ontogeny, phylogeny, brain mechanisms, context and psychopathology. *Neuroscience & Biobehavioral Reviews 37*(8), 1537–1548.

Gordon, I., O. Zagoory-Sharon, J. F. Leckman, and R. Feldman (2010). Oxytocin and the development of parenting in humans. *Biological psychiatry 68*(4), 377–382.

Gottfried, J. A., J. O'Doherty, and R. J. Dolan (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science 301*(5636), 1104–1107.

Grienenberger, J. F., K. Kelly, and A. Slade (2005). Maternal reflective functioning, mother–infant affective communication, and infant attachment: Exploring the link between mental states and observed caregiving behavior in the intergenerational transmission of attachment. *Attachment & human development 7*(3), 299–311.

Gritti, I., L. Mainville, and B. E. Jones (1993). Codistribution of GABA- with acetylcholine-synthesizing neurons in the basal forebrain of the rat. *The Journal of comparative neurology 329*(4), 438–457.

Grossmann, T. (2013). Mapping prefrontal cortex functions in human infancy. *Infancy 18*(3), 303–324.

Gu, X., P. R. Hof, K. J. Friston, and J. Fan (2013). Anterior insular cortex and emotional awareness. *Journal of Comparative Neurology 521*(15), 3371–3388.

Halpern, J. (2003). What is clinical empathy? *Journal of general internal medicine 18*(8), 670–674.

Hamilton, C. E. (2000). Continuity and discontinuity of attachment from infancy through adolescence. *Child development 71*(3), 690–694.

Hane, A. A., N. A. Fox, H. A. Henderson, and P. J. Marshall (2008). Behavioral reactivity and approach-withdrawal bias in infancy. *Developmental Psychology 44*(5), 1491.

Harmon-Jones, E., C. K. Peterson, and C. Harmon-Jones (2010). Anger, motivation, and asymmetrical frontal cortical activations. In *International handbook of anger*, pp. 61–78. Springer.

Harris, K. D. and G. M. Shepherd (2015). The neocortical circuit: themes and variations. *Nature neuroscience 18*(2), 170–181.

Heim, C., L. Young, D. Newport, T. Mletzko, A. Miller, and C. Nemeroff (2008). Lower CSF oxytocin concentrations in women with a history of childhood abuse. *Molecular psychiatry 14*(10), 954–958.

Hein, G., G. Silani, K. Preuschoff, C. D. Batson, and T. Singer (2010). Neural responses to ingroup and outgroup members' suffering predict individual differences in costly helping. *Neuron 68*(1), 149–160.

Henriques, J. B. and R. J. Davidson (1991). Left frontal hypoactivation in depression. *Journal of abnormal psychology 100*(4), 535.

Henssen, A., K. Zilles, N. Palomero-Gallagher, A. Schleicher, H. Mohlberg, F. Gerboga, S. B. Eickhoff, S. Bludau, and K. Amunts (2016). Cytoarchitecture and probability maps of the human medial orbitofrontal cortex. *Cortex 75*, 87–112.

Herpertz, S. C., T. M. Dietrich, B. Wenning, T. Krings, S. G. Erberich, K. Willmes, A. Thron, and H. Sass (2001). Evidence of abnormal amygdala functioning in borderline personality disorder: a functional MRI study. *Biological psychiatry 50*(4), 292–298.

Herr, N. R., C. Hammen, and P. A. Brennan (2008). Maternal borderline personality disorder symptoms and adolescent psychosocial functioning. *Journal of Personality Disorders 22*(5), 451–465.

Hertsgaard, L., M. Gunnar, M. F. Erickson, and M. Nachmias (1995). Adrenocortical responses to the strange situation in infants with disorganized/disoriented attachment relationships. *Child development 66*(4), 1100–1106.

Hesse, E. (2008). *Handbook of attachment: Theory, research, and clinical applications*, Chapter The Adult Attachment Interview: Protocol, method of analysis, and empirical studies., pp. 395–433. Guilford Press.

Hesslow, G. (2002). Conscious thought as simulation of behaviour and perception. *Trends in cognitive sciences 6*(6), 242–247.

Hill-Soderlund, A. L., W. R. Mills-Koonce, C. Propper, S. D. Calkins, D. A. Granger, G. A. Moore, J.-L. Gariepy, and M. J. Cox (2008). Parasympathetic and sympathetic responses to the strange situation in infants and mothers from avoidant and securely attached dyads. *Developmental Psychobiology 50*(4), 361–376.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation 14*(8), 1771–1800.

Hinton, G. E. (2010). A practical guide to training restricted Boltzmann machines. *Momentum 9*(1), 926.

Hinton, G. E., S. Osindero, and Y.-W. Teh (2006). A fast learning algorithm for deep belief nets. *Neural computation 18*(7), 1527–1554.

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

Hiolle, A., L. Cañamero, M. Davila-Ross, and K. A. Bard (2012). Eliciting caregiving behavior in dyadic human-robot attachment-like interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS) 2*(1), 3.

Hiolle, A., M. Lewis, and L. Canamero (2014). A robot that uses arousal to detect learning challenges and seek help. In *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, Volume 14, pp. 726–733.

Hobson, R. P., M. Patrick, L. Crandell, R. GarcÍa-pÉrez, and A. Lee (2005). Personal relatedness and attachment in infants of mothers with borderline personality disorder. *Development and Psychopathology 17*(2), 329–347.

Hobson, R. P., M. P. Patrick, J. A. Hobson, L. Crandell, E. Bronfman, and K. Lyons-Ruth (2009). How mothers with borderline personality disorder relate to their year-old infants. *The British Journal of Psychiatry 195*(4), 325–330.

Höistad, M., H. Heinsen, B. Wicinski, C. Schmitz, and P. R. Hof (2013). Stereological assessment of the dorsal anterior cingulate cortex in schizophrenia: absence of changes in neuronal and glial densities. *Neuropathology and applied neurobiology 39*(4), 348–361.

Hooker, C. I. and R. T. Knight (2006). The role of the lateral orbitofrontal cortex in the inhibitory control of emotion. In *The Orbitofrontal Cortex*, Chapter 12, pp. 307. Oxford University Press.

Hu, H. (2016). Reward and aversion. *Annual review of neuroscience 39*, 297–324.

Huang, Y.-C. and N. A. Hessler (2008). Social modulation during songbird courtship potentiates midbrain dopaminergic neurons. *PloS one 3*(10), e3281.

Humphrey, T. (1968). The development of the human amygdala during early embryonic life. *Journal of Comparative Neurology 132*(1), 135–165.

Hurlemann, R., A. Patin, O. A. Onur, M. X. Cohen, T. Baumgartner, S. Metzler, I. Dziobek, J. Gallinat, M. Wagner, W. Maier, et al. (2010). Oxytocin enhances amygdala-dependent, socially reinforced learning and emotional empathy in humans. *The Journal of Neuroscience 30*(14), 4999–5007.

Izhikevich, E. M. et al. (2003). Simple model of spiking neurons. *IEEE Transactions on neural networks 14*(6), 1569–1572.

Izhikevich, E. M. and G. M. Edelman (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the national academy of sciences 105*(9), 3593–3598.

Izhikevich, E. M. and J. Moehlis (2008). Dynamical systems in neuroscience: The geometry of excitability and bursting. *SIAM review 50*(2), 397.

Jacobvitz, D., N. Hazen, and S. Riggs (1997). Disorganized mental processes in mothers, frightening/frightened caregiving, and disoriented/disorganized behavior in infancy. In *biennial meeting of the Society for Research in Child Development, Washington, DC*.

Jacobvitz, D., K. Leon, and N. Hazen (2006). Does expectant mothers' unresolved trauma predict frightened/frightening maternal behavior? risk and protective factors. *Development and Psychopathology 18*(02), 363–379.

Jeffries, K., J. Fritz, and A. Braun (2003). Words in melody: an H215O PET study of brain activation during singing and speaking. *Neuroreport 14*(5), 749–754.

Jenison, R. L. (2014). Directional influence between the human amygdala and orbitofrontal cortex at the time of decision-making. *PloS one 9*(10).

Joffily, M. and G. Coricelli (2013). Emotional valence and the free-energy principle. *PLoS Comput Biol 9*(6).

Jordan, L. M. (1998). Initiation of locomotion in mammals. *Annals of the New York Academy of Sciences 860*(1), 83–93.

Kalló, I., C. S. Molnár, S. Szöke, C. Fekete, E. Hrabovszky, and Z. Liposits (2015). Area-specific analysis of the distribution of hypothalamic neurons projecting to the rat ventral tegmental area, with special reference to the GABAergic and glutamatergic efferents. *Frontiers in neuroanatomy 9*.

Kandel, E. R., J. H. Schwartz, T. M. Jessell, et al. (2012). *Principles of neural science* (Fifth ed.). McGraw-hill New York.

Kawamoto, T., M. Ura, and H. Nittono (2015). Intrapersonal and interpersonal processes of social exclusion. *Frontiers in Neuroscience 9*, 62.

Kidd, T., M. Hamer, and A. Steptoe (2013). Adult attachment style and cortisol responses across the day in older adults. *Psychophysiology 50*(9), 841–847.

Kiel, E. J., K. L. Gratz, S. A. Moore, R. D. Latzman, and M. T. Tull (2011). The impact of borderline personality pathology on mothers' responses to infant distress. *Journal of Family Psychology 25*(6), 907.

Kim, S., P. Fonagy, J. Allen, and L. Strathearn (2014). Mothers' unresolved trauma blunts amygdala response to infant distress. *Social neuroscience*, 1–12.

Kirkpatrick, L. A. (2005). *Attachment, evolution, and the psychology of religion.* Guilford Press.

Kitayama, N., S. Quinn, and J. D. Bremner (2006). Smaller volume of anterior cingulate cortex in abuse-related posttraumatic stress disorder. *Journal of affective disorders 90*(2), 171–174.

Kleber, B., N. Birbaumer, R. Veit, T. Trevorrow, and M. Lotze (2007). Overt and imagined singing of an italian aria. *Neuroimage 36*(3), 889–900.

Klimecki, O. M., S. Leiberg, M. Ricard, and T. Singer (2013). Differential pattern of functional brain plasticity after compassion and empathy training. *Social cognitive and affective neuroscience 9*(6), 873–879.

Kocsis, K., J. Kiss, A. Csaki, and B. Halasz (2003). Location of putative glutamatergic neurons projecting to the medial preoptic area of the rat hypothalamus. *Brain research bulletin 61*(4), 459–468.

Koelsch, S., T. Fritz, K. Müller, A. D. Friederici, et al. (2006). Investigating emotion with music: an fMRI study. *Human brain mapping 27*(3), 239–250.

Kohut, H. (1959). Introspection, empathy, and psychoanalysis: An examination of the relationship between mode of observation and theory. *Journal of the American Psychoanalytic Association*.

Kringelbach, M. L., A. Lehtonen, S. Squire, A. G. Harvey, M. G. Craske, I. E. Holliday, A. L. Green, T. Z. Aziz, P. C. Hansen, P. L. Cornelissen, et al. (2008). A specific and rapid neural signature for parental instinct. *PLoS One 3*(2).

Kringelbach, M. L. and E. T. Rolls (2004). The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in neurobiology 72*(5), 341–372.

Krugers, H., P. Goltstein, S. Van der Linden, and M. Joels (2006). Blockade of glucocorticoid receptors rapidly restores hippocampal CA1 synaptic plasticity after exposure to chronic stress. *European Journal of Neuroscience 23*(11), 3051–3055.

Kupchik, Y. M. and P. W. Kalivas (2013). The rostral subcommissural ventral pallidum is a mix of ventral pallidal neurons and neurons from adjacent areas: an electrophysiological study. *Brain Structure and Function 218*(6), 1487–1500.

Kurth, F., S. B. Eickhoff, A. Schleicher, L. Hoemke, K. Zilles, and K. Amunts (2010). Cytoarchitecture and probabilistic maps of the human posterior insular cortex. *Cerebral Cortex 20*(6), 1448–1461.

Lacerda, A. L., M. S. Keshavan, A. Y. Hardan, O. Yorbik, P. Brambilla, R. B. Sassi, M. Nicoletti, A. G. Mallinger, E. Frank, D. J. Kupfer, et al. (2004). Anatomic evaluation of the orbitofrontal cortex in major depressive disorder. *Biological psychiatry 55*(4), 353–358.

Lamm, C., J. Decety, and T. Singer (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *Neuroimage 54*(3), 2492–2502.

Lamm, C. and T. Singer (2010). The role of anterior insular cortex in social emotions. *Brain Structure and Function 214*(5-6), 579–591.

LeDoux, J. (2003). The emotional brain, fear, and the amygdala. *Cellular and molecular neurobiology 23*(4-5), 727–738.

LeDoux, J. E. (1997). Emotion, memory and the brain. *Scientific American 7*(1), 68–76.

LeDoux, J. E., J. Iwata, P. Cicchetti, and D. Reis (1988). Different projections of the central amygdaloid nucleus mediate autonomic and behavioral correlates of conditioned fear. *The Journal of Neuroscience 8*(7), 2517–2529.

LeDoux, J. E. and E. A. Phelps (1993). Emotional networks in the brain. *Handbook of emotions*, 109–118.

Lee, G. P., K. J. Meador, D. W. Loring, J. D. Allison, W. S. Brown, L. K. Paul, J. J. Pillai, and T. B. Lavin (2004). Neural substrates of emotion as revealed by functional magnetic resonance imaging. *Cognitive and Behavioral Neurology 17*(1), 9–17.

Lee, H., C. Ekanadham, and A. Y. Ng (2008). Sparse deep belief net model for visual area V2. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 873–880.

Lemche, E., V. P. Giampietro, S. A. Surguladze, E. J. Amaro, C. M. Andrew, S. C. Williams, M. J. Brammer, N. Lawrence, M. A. Maier, T. A. Russell, et al. (2006). Human attachment security is mediated by the amygdala: Evidence from combined fMRI and psychophysiological measures. *Human brain mapping 27*(8), 623–635.

Leuner, B. and T. Shors (2013). Stress, anxiety, and dendritic spines: what are the connections? *Neuroscience 251*, 108–119.

Levine, D. S. (2008). Neural networks of human nature and nurture. *Avances en psicología latinoamericana 26*(1), 82–98.

Lewis, D. A., D. S. Melchitzky, and G.-G. Burgos (2002). Specificity in the functional architecture of primate prefrontal cortex. *Journal of neurocytology 31*(3-5), 265–276.

Light, K. C., K. M. Grewen, and J. A. Amico (2005). More frequent partner hugs and higher oxytocin levels are linked to lower blood pressure and heart rate in premenopausal women. *Biological psychology 69*(1), 5–21.

Liotti, G. (1992). Disorganized/disoriented attachment in the etiology of the dissociative disorders. *Dissociation: Progress in the Dissociative Disorders*.

Liotti, G. (2004). Trauma, dissociation, and disorganized attachment: Three strands of a single braid. *Psychotherapy: Theory, research, practice, training 41*(4), 472.

Litt, C. J. (1986). Theories of transitional object attachment: An overview. *International Journal of Behavioral Development 9*(3).

Lonstein, J. and G. De Vries (2000). Maternal behaviour in lactating rats stimulates c-fos in glutamate decarboxylase-synthesizing neurons of the medial preoptic area, ventral bed nucleus of the stria terminalis, and ventrocaudal periaqueductal gray. *Neuroscience 100*(3), 557–568.

Love, T. M. (2014). Oxytocin, motivation and the role of dopamine. *Pharmacology Biochemistry and Behavior 119*.

Luna, B., 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 (2001). Maturation of widely distributed brain function subserves cognitive development. *Neuroimage 13*(5), 786–793.

Lyons-Ruth, K. (1996). Attachment relationships among children with aggressive behavior problems: The role of disorganized early attachment patterns. *Journal of consulting and clinical psychology 64*(1), 64.

Lyons-Ruth, K., L. Alpern, and B. Repacholi (1993). Disorganized infant attachment classification and maternal psychosocial problems as predictors of hostile-aggressive behavior in the preschool classroom. *Child development 64*(2), 572–585.

Lyons-Ruth, K., E. Bronfman, and E. Parsons (1999). Maternal frightened, frightening, or atypical behavior and disorganized infant attachment patterns. *Monographs of the Society for Research in Child Development*, 67–96.

Lyons-Ruth, K., S. Melnick, M. Patrick, and R. P. Hobson (2007). A controlled study of hostile-helpless states of mind among borderline and dysthymic women. *Attachment & human development 9*(1), 1–16.

Lyons-Ruth, K., P. Pechtel, S. Yoon, C. Anderson, and M. Teicher (2016). Disorganized attachment in infancy predicts greater amygdala volume in adulthood. *Behavioural brain research 308*, 83–93.

Lyons-Ruth, K., C. Yellin, S. Melnick, and G. Atwood (2005). Expanding the concept of unresolved mental states: Hostile/helpless states of mind on the adult attachment interview are associated with disrupted mother–infant communication and infant disorganization. *Development and psychopathology 17*(01), 1–23.

Macfie, J. and S. A. Swan (2009). Representations of the caregiver–child relationship and of the self, and emotion regulation in the narratives of young children whose mothers have borderline personality disorder. *Development and Psychopathology 21*(03), 993–1011.

Macfie, J., S. A. Swan, K. L. Fitzpatrick, C. D. Watkins, and E. M. Rivas (2014). Mothers with borderline personality and their young children: Adult attachment interviews, mother–child interactions, and children's narrative representations. *Development and psychopathology 26*(02), 539–551.

Maddock, R. J., A. S. Garrett, and M. H. Buonocore (2001). Remembering familiar people: the posterior cingulate cortex and autobiographical memory retrieval. *Neuroscience 104*(3), 667–676.

Madigan, S., G. Moran, and D. R. Pederson (2006). Unresolved states of mind, disorganized attachment relationships, and disrupted interactions of adolescent mothers and their infants. *Developmental Psychology 42*(2), 293.

Mai, J. K. and G. Paxinos (2011). *The human nervous system*. Academic Press.

Main, M. and E. Hesse (1990). Parents' unresolved traumatic experiences are related to infant disorganized attachment status: Is frightened and/or frightening parental behavior the linking mechanism? In *Attachment in the pre-school years: theory, research and intervention*. University of Chicago Press.

Main, M. and E. Hesse (1992). Frightening, frightened, dissociated, or disorganized behavior on the part of the parent: A coding system for parent–infant interactions. *Unpublished manuscript. University of California at Berkeley*.

Main, M., N. Kaplan, and J. Cassidy (1985). Security in infancy, childhood, and adulthood: A move to the level of representation. *Monographs of the society for research in child development*, 66–104.

Main, M. and H. Morgan (1996). Disorganization and disorientation in infant strange situation behavior. In *Handbook of dissociation*, pp. 107–138. Springer.

Main, M. and J. Solomon (1986). Discovery of an insecure-disorganized/disoriented attachment pattern: Procedures, findings, and implications for the classification of behavior. In T. B. Brazelton and M. Yogman (Eds.), *Affective development in infancy*, pp. 95–124. Ablex Publishing.

Main, M. and J. Solomon (1990). Procedures for identifying infants as disorganized/disoriented during the Ainsworth strange situation. *Attachment in the preschool years: Theory, research, and intervention 1*, 121–160.

Makris, N., J. M. Goldstein, D. Kennedy, S. M. Hodge, V. S. Caviness, S. V. Faraone, M. T. Tsuang, and L. J. Seidman (2006). Decreased volume of left and total anterior insular lobule in schizophrenia. *Schizophrenia research 83*(2), 155–171.

Makris, N., D. F. Swaab, A. van der Kouwe, B. Abbs, D. Boriel, R. J. Handa, S. Tobet, and J. M. Goldstein (2013). Volumetric parcellation methodology of the human hypothalamus in neuroimaging: Normative data and sex differences. *NeuroImage 69*, 1–10.

Margolis, E. B., H. Lock, G. O. Hjelmstad, and H. L. Fields (2006). The ventral tegmental area revisited: is there an electrophysiological marker for dopaminergic neurons? *The Journal of physiology 577*(3), 907–924.

Markram, H., M. Toledo-Rodriguez, Y. Wang, A. Gupta, G. Silberberg, and C. Wu (2004). Interneurons of the neocortical inhibitory system. *Nature Reviews Neuroscience 5*(10), 793–807.

Masten, C. L., S. A. Morelli, and N. I. Eisenberger (2011). An fMRI investigation of empathy for social pain and subsequent prosocial behavior. *Neuroimage 55*(1), 381–388.

Mathur, V. A., T. Harada, T. Lipke, and J. Y. Chiao (2010). Neural basis of extraordinary empathy and altruistic motivation. *Neuroimage 51*(4), 1468–1475.

McClure, W. O., A. Ishtoyan, and M. Lyon (2004). Very mild stress of pregnant rats reduces volume and cell number in nucleus accumbens of adult offspring: some parallels to schizophrenia. *Developmental brain research 149*(1), 21–28.

McCrory, E. J., S. A. De Brito, C. L. Sebastian, A. Mechelli, G. Bird, P. A. Kelly, and E. Viding (2011). Heightened neural reactivity to threat in child victims of family violence. *Current Biology 21*(23), R947–R948.

McDonald, A. J. (1992). Projection neurons of the basolateral amygdala: a correlative golgi and retrograde tract tracing study. *Brain research bulletin 28*(2), 179–185.

McMahan True, M., L. Pisani, and F. Oumar (2001). Infant–mother attachment among the Dogon of Mali. *Child development 72*(5), 1451–1466.

Melchitzky, D. S., G. González-Burgos, G. Barrionuevo, and D. A. Lewis (2001). Synaptic targets of the intrinsic axon collaterals of supragranular pyramidal neurons in monkey prefrontal cortex. *Journal of Comparative Neurology 430*(2), 209–221.

Meyer, M. L., C. L. Masten, Y. Ma, C. Wang, Z. Shi, N. I. Eisenberger, and S. Han (2012). Empathy for the social suffering of friends and strangers recruits distinct patterns of brain activation. *Social cognitive and affective neuroscience 8*(4), 446–454.

Mikulincer, M. and P. R. Shaver (2005). Attachment security, compassion, and altruism. *Current directions in psychological science 14*(1), 34–38.

Mikulincer, M. and P. R. Shaver (2007). Boosting attachment security to promote mental health, prosocial values, and inter-group tolerance. *Psychological Inquiry 18*(3), 139–156.

Mikulincer, M. and P. R. Shaver (2012). An attachment perspective on psychopathology. *World Psychiatry 11*(1), 11–15.

Milad, M. R. and G. J. Quirk (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annual review of psychology 63*, 129–151.

Minagawa-Kawai, Y., S. Matsuoka, I. Dan, N. Naoi, K. Nakamura, and S. Kojima (2009). Prefrontal activation associated with social attachment: facial-emotion recognition in mothers and infants. *Cerebral Cortex 19*(2), 284–292.

Mitterschiffthaler, M. T., C. H. Fu, J. A. Dalton, C. M. Andrew, and S. C. Williams (2007). A functional MRI study of happy and sad affective states induced by classical music. *Human brain mapping 28*(11), 1150–1162.

Mogenson, G. J. (1987). Limbic-motor integration. *Progress in psychobiology and physiological psychology 12*, 117–170.

Moll, J., F. Krueger, R. Zahn, M. Pardini, R. de Oliveira-Souza, and J. Grafman (2006). Human fronto–mesolimbic networks guide decisions about charitable donation. *Proceedings of the National Academy of Sciences 103*(42), 15623–15628.

Montag, C., M. Reuter, and N. Axmacher (2011). How one's favorite song activates the reward circuitry of the brain: Personality matters! *Behavioural brain research 225*(2), 511–514.

Moustafa, A. A., M. W. Gilbertson, S. P. Orr, M. M. Herzallah, R. J. Servatius, and C. E. Myers (2013). A model of amygdala–hippocampal–prefrontal interaction in fear conditioning and extinction in animals. *Brain and cognition 81*(1), 29–43.

Murray, R. J., M. Schaer, and M. Debbané (2012). Degrees of separation: a quantitative neuroimaging meta-analysis investigating self-specificity and shared neural activation between self-and other-reflection. *Neuroscience & Biobehavioral Reviews 36*(3), 1043–1059.

Mushiake, H., N. Saito, K. Sakamoto, Y. Itoyama, and J. Tanji (2006). Activity in the lateral prefrontal cortex reflects multiple steps of future events in action plans. *Neuron 50*(4), 631–641.

Nair-Roberts, R., S. Chatelain-Badie, E. Benson, H. White-Cooper, J. Bolam, and M. Ungless (2008). Stereological estimates of dopaminergic, GABAergic and glutamatergic neurons in the ventral tegmental area, substantia nigra and retrorubral field in the rat. *Neuroscience 152*(4), 1024–1031.

Newman, L. K., C. S. Stevenson, L. R. Bergman, and P. Boyce (2007). Borderline personality disorder, mother–infant interaction and parenting perceptions: Preliminary findings. *Australian and New Zealand Journal of Psychiatry 41*(7), 598–605.

Nicolle, A., M. C. Klein-Flügge, L. T. Hunt, I. Vlaev, R. J. Dolan, and T. E. Behrens (2012). An agent independent axis for executed and modeled choice in medial prefrontal cortex. *Neuron 75*(6), 1114–1121.

Nikolić, I. and I. Kostović (1986). Development of the lateral amygdaloid nucleus in the human fetus: transient presence of discrete cytoarchitectonic units. *Anatomy and embryology 174*(3), 355–360.

Nitschke, J. B., E. E. Nelson, B. D. Rusch, A. S. Fox, T. R. Oakes, and R. J. Davidson (2004). Orbitofrontal cortex tracks positive mood in mothers viewing pictures of their newborn infants. *Neuroimage 21*(2), 583–592.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology 53*(3), 139–154.

Noriuchi, M., Y. Kikuchi, and A. Senoo (2008). The functional neuroanatomy of maternal love: mother's response to infant's attachment behaviors. *Biological Psychiatry 63*(4), 415–423.

Numan, M. (2014). *Neurobiology of Social Behavior: Toward an Understanding of the Prosocial and Antisocial Brain*. Academic Press.

Numan, M. (2017). Parental behavior. In *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier.

Numan, M., J. A. Bress, L. R. Ranker, A. J. Gary, A. L. DeNicola, J. K. Bettis, and S. E. Knapp (2010). The importance of the basolateral/basomedial amygdala for goal-directed maternal responses in postpartum rats. *Behavioural Brain Research 214*(2), 368–376.

Numan, M. and T. R. Insel (2003). *The neurobiology of parental behavior*, Volume 1. Springer Science & Business Media.

Numan, M., M. J. Numan, J. M. Schwarz, C. M. Neuner, T. F. Flood, and C. D. Smith (2005). Medial preoptic area interactions with the nucleus accumbens–ventral pallidum circuit and maternal behavior in rats. *Behavioural Brain Research 158*(1), 53–68.

Numan, M. and D. S. Stolzenberg (2009). Medial preoptic area interactions with dopamine neural systems in the control of the onset and maintenance of maternal behavior in rats. *Frontiers in neuroendocrinology 30*(1), 46–64.

Numan, M. and L. J. Young (2016). Neural mechanisms of mother–infant bonding and pair bonding: Similarities, differences, and broader implications. *Hormones and behavior 77*, 98–112.

Obegi, J. H. (2008). The development of the client-therapist bond through the lens of attachment theory. *Psychotherapy: Theory, Research, Practice, Training 45*(4), 431.

Ochsner, K. N., S. A. Bunge, J. J. Gross, and J. D. Gabrieli (2002). Rethinking feelings: an fMRI study of the cognitive regulation of emotion. *Journal of cognitive neuroscience 14*(8), 1215–1229.

Ochsner, K. N., R. D. Ray, J. C. Cooper, E. R. Robertson, S. Chopra, J. D. Gabrieli, and J. J. Gross (2004). For better or for worse: neural systems supporting the cognitive down-and up-regulation of negative emotion. *Neuroimage 23*(2), 483–499.

Ochsner, K. N., J. Zaki, J. Hanelin, D. H. Ludlow, K. Knierim, T. Ramachandran, G. H. Glover, and S. C. Mackey (2008). Your pain or mine? Common and distinct neural systems supporting the perception of pain in self and other. *Social Cognitive and Affective Neuroscience 3*(2), 144–160.

Okabe, S., M. Yoshida, Y. Takayanagi, and T. Onaka (2015). Activation of hypothalamic oxytocin neurons following tactile stimuli in rats. *Neuroscience letters 600*, 22–27.

Olff, M., J. L. Frijling, L. D. Kubzansky, B. Bradley, M. A. Ellenbogen, C. Cardoso, J. A. Bartz, J. R. Yee, and M. van Zuiden (2013). The role of oxytocin in social bonding, stress regulation and mental health: An update on the moderating effects of context and interindividual differences. *Psychoneuroendocrinology 38*(9), 1883–1894.

Olijslagers, J., T. Werkman, A. McCreary, C. Kruse, and W. Wadman (2006). Modulation of midbrain dopamine neurotransmission by serotonin, a versatile interaction between neurotransmitters and significance for antipsychotic drug action. *Current neuropharmacology 4*(1), 59.

Öngür, D., A. T. Ferry, and J. L. Price (2003). Architectonic subdivision of the human orbital and medial prefrontal cortex. *Journal of Comparative Neurology 460*(3), 425–449.

Overton, P. and D. Clark (1997). Burst firing in midbrain dopaminergic neurons. *Brain Research Reviews 25*(3), 312–334.

Packer, A. M. and R. Yuste (2011). Dense, unspecific connectivity of neocortical parvalbumin-positive interneurons: a canonical microcircuit for inhibition? *The Journal of Neuroscience 31*(37), 13260–13271.

Pakkenberg, B. (1990). Pronounced reduction of total neuron number in mediodorsal thalamic nucleus and nucleus accumbens in schizophrenics. *Archives of General Psychiatry 47*(11), 1023–1028.

Papp, E., Z. Borhegyi, R. Tomioka, K. S. Rockland, I. Mody, and T. F. Freund (2012). Glutamatergic input from specific sources influences the nucleus accumbens-ventral pallidum information flow. *Brain Structure and Function 217*(1), 37–48.

Pavuluri, M. and A. May (2015). I feel, therefore, i am: the insula and its role in human emotion, cognition and the sensory-motor system. *AIMS Neurosci 2*(1), 18–27.

Petrovic, P., C. J. Ekman, J. Klahr, L. Tigerström, G. Rydén, A. G. Johansson, C. Sellgren, A. Golkar, A. Olsson, A. Öhman, et al. (2015). Significant gray matter changes in a region of the orbitofrontal cortex in healthy participants predicts emotional dysregulation. *Social cognitive and affective neuroscience 1*, 1–9.

Petters, D. (2006a). *Designing agents to understand infants.* Ph. D. thesis, School of Computer Science, The University of Birmingham.

Petters, D. (2006b). Implementing a theory of attachment: A simulation of the strange situation with autonomous agents. In *Proceedings of the Seventh International Conference on Cognitive Modelling*, Volume 7, pp. 226–231.

Peyrache, A., N. Dehghani, E. N. Eskandar, J. R. Madsen, W. S. Anderson, J. A. Donoghue, L. R. Hochberg, E. Halgren, S. S. Cash, and A. Destexhe (2012). Spatiotemporal dynamics of neocortical excitation and inhibition during human sleep. *Proceedings of the National Academy of Sciences 109*(5), 1731–1736.

Pezzulo, G., F. Rigoli, and K. Friston (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology 134*, 17–35.

Phan, K. L., D. A. Fitzgerald, P. J. Nathan, G. J. Moore, T. W. Uhde, and M. E. Tancer (2005). Neural substrates for voluntary suppression of negative affect: a functional magnetic resonance imaging study. *Biological psychiatry 57*(3), 210–219.

Phan, K. L., T. Wager, S. F. Taylor, and I. Liberzon (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage 16*(2), 331–348.

Planalp, E. M. and J. M. Braungart-Rieker (2013). Temperamental precursors of infant attachment with mothers and fathers. *Infant Behavior and Development 36*(4), 796–808.

Porges, S. W. (2011). *The Polyvagal Theory: Neurophysiological Foundations of Emotions, Attachment, Communication, and Self-regulation (Norton Series on Interpersonal Neurobiology)*. WW Norton & Company.

Pruessner, J. C., F. Champagne, M. J. Meaney, and A. Dagher (2004). Dopamine release in response to a psychological stress in humans and its relationship to early life maternal care: a positron emission tomography study using [11c] raclopride. *The Journal of Neuroscience 24*(11), 2825–2831.

Quattrocki, E. and K. Friston (2014). Autism, oxytocin and interoception. *Neuroscience & Biobehavioral Reviews 47*, 410–430.

Rabinowicz, T., D. E. Dean, J. M.-C. Petetot, and G. M. de Courten-Myers (1999). Gender differences in the human cerebral cortex: more neurons in males; more processes in females. *Journal of Child Neurology 14*(2), 98–107.

Radley, J. J., A. B. Rocher, W. G. Janssen, P. R. Hof, B. S. McEwen, and J. H. Morrison (2005). Reversibility of apical dendritic retraction in the rat medial prefrontal cortex following repeated stress. *Experimental neurology 196*(1), 199–203.

Radley, J. J., A. B. Rocher, M. Miller, W. G. Janssen, C. Liston, P. R. Hof, B. S. McEwen, and J. H. Morrison (2006). Repeated stress induces dendritic spine loss in the rat medial prefrontal cortex. *Cerebral Cortex 16*(3), 313–320.

Raes, A. K. and R. De Raedt (2012). The effect of counterconditioning on evaluative responses and harm expectancy in a fear conditioning paradigm. *Behavior therapy 43*(4), 757–767.

Rajkowska, G., J. J. Miguel-Hidalgo, J. Wei, G. Dilley, S. D. Pittman, H. Y. Meltzer, J. C. Overholser, B. L. Roth, and C. A. Stockmeier (1999). Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. *Biological psychiatry 45*(9), 1085–1098.

Reichert, D. P., P. Series, and A. J. Storkey (2011). A hierarchical generative model of recurrent object-based attention in the visual cortex. In *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 18–25. Springer.

Reichert, D. P., P. Seriès, and A. J. Storkey (2013). Charles Bonnet syndrome: evidence for a generative model in the cortex? *PLoS computational biology 9*(7).

Riem, M. M., M. J. Bakermans-Kranenburg, M. H. van IJzendoorn, D. Out, and S. A. Rombouts (2012). Attachment in the brain: adult attachment representations predict amygdala and behavioral responses to infant crying. *Attachment & human development 14*(6), 533–551.

Ripoll, L. H., R. Snyder, H. Steele, and L. J. Siever (2013). The neurobiology of empathy in borderline personality disorder. *Current psychiatry reports 15*(3), 1–11.

Robison, A. J. and E. J. Nestler (2011). Transcriptional and epigenetic mechanisms of addiction. *Nature reviews neuroscience 12*(11), 623–637.

Rockliff, H., P. Gilbert, K. McEwan, S. Lightman, and D. Glover (2008). A pilot exploration of heart rate variability and salivary cortisol responses to compassion-focused imagery. *Journal of Clinical Neuropsychiatry 5*, 132–139.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology 21*(2), 95.

Rognoni, E., D. Galati, T. Costa, and M. Crini (2008). Relationship between adult attachment patterns, emotional experience and eeg frontal asymmetry. *Personality and Individual Differences 44*(4), 909–920.

Rolls, E. T. (1990). A theory of emotion, and its application to understanding the neural basis of emotion. *Cognition & Emotion 4*(3), 161–190.

Rolls, E. T. (2013). *Emotion and decision making explained*. Oxford University Press.

Rolls, E. T. (2015). Limbic systems for emotion and for memory, but no single limbic system. *Cortex 62*, 119–157.

Root, D. H. (2013). The ventromedial ventral pallidum subregion is necessary for outcome-specific pavlovian-instrumental transfer. *The Journal of Neuroscience 33*(48), 18707–18709.

Root, D. H., R. I. Melendez, L. Zaborszky, and T. C. Napier (2015). The ventral pallidum: Subregion-specific functional anatomy and roles in motivated behaviors. *Progress in neurobiology 130*, 29–70.

Rudebeck, P. H., A. R. Mitz, R. V. Chacko, and E. A. Murray (2013). Effects of amygdala lesions on reward-value coding in orbital and medial prefrontal cortex. *Neuron 80*(6), 1519–1531.

Rudy, B., G. Fishell, S. Lee, and J. Hjerling-Leffler (2011). Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental neurobiology 71*(1), 45–61.

Russo, S. J. and E. J. Nestler (2013). The brain reward circuitry in mood disorders. *Nature Reviews Neuroscience 14*(9), 609–625.

Safyer, M. P. (2013). *When good enough mothering is not good enough: a study of mothers' secure base scripts, atypical and disrupted caregiving and the transmission of infant attachment quality.* Ph. D. thesis, University of Michigan.

Sah, P., E. L. Faber, M. L. De Armentia, and J. Power (2003). The amygdaloid complex: anatomy and physiology. *Physiological reviews 83*(3), 803–834.

Salimpoor, V. N., M. Benovoy, K. Larcher, A. Dagher, and R. J. Zatorre (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature neuroscience 14*(2), 257–262.

Sandler, J. and J. Bowlby (1989). *Dimensions of psychoanalysis.* Karnac Books.

Saper, C. B. and B. B. Lowell (2014). The hypothalamus. *Current Biology 24*(23), R1111–R1116.

Schjødt, U., H. Stødkilde-Jørgensen, A. W. Geertz, and A. Roepstorff (2008). Rewarding prayers. *Neuroscience letters 443*(3), 165–168.

Schore, A. N. (2001). Effects of a secure attachment relationship on right brain development, affect regulation, and infant mental health. *Infant mental health journal 22*(1-2), 7–66.

Schore, A. N. (2002). Advances in neuropsychoanalysis, attachment theory, and trauma research: Implications for self psychology. *Psychoanalytic Inquiry 22*(3), 433–484.

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

Schore, A. N. (2003b). *Affect Regulation and the Repair of the Self (Norton Series on Interpersonal Neurobiology)*, Volume 2. WW Norton & Company.

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

Schuck, N. W., M. B. Cai, R. C. Wilson, and Y. Niv (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron 91*(6), 1402–1412.

Schuengel, C., M. J. Bakermans-Kranenburg, and M. H. Van IJzendoorn (1999). Frightening maternal behavior linking unresolved loss and disorganized infant attachment. *Journal of consulting and clinical psychology 67*(1), 54.

Schultz, W., P. Dayan, and P. R. Montague (1997). A neural substrate of prediction and reward. *Science 275*(5306), 1593–1599.

Schwartenbeck, P. and K. Friston (2016). Computational phenotyping in psychiatry: a worked example. *eNeuro 3*(4).

Semendeferi, K., E. Armstrong, A. Schleicher, K. Zilles, and G. W. Van Hoesen (1998). Limbic frontal cortex in hominoids: a comparative study of area 13. *American Journal of Physical Anthropology 106*(2), 129–155.

Series, P., D. P. Reichert, and A. J. Storkey (2010). Hallucinations in Charles Bonnet syndrome induced by homeostasis: a deep Boltzmann machine model. In *Advances in Neural Information Processing Systems*, pp. 2020–2028.

Shahrokh, D. K., T.-Y. Zhang, J. Diorio, A. Gratton, and M. J. Meaney (2010). Oxytocin-dopamine interactions mediate variations in maternal behavior in the rat. *Endocrinology 151*(5), 2276–2286.

Sherman, S. M. and R. Guillery (2002). The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences 357*(1428), 1695–1708.

Shi, W. and S. Rayport (1994). GABA synapses formed in vitro by local axon collaterals of nucleus accumbens neurons. *The Journal of neuroscience 14*(7), 4548–4560.

Siegel, D. J. (2012). *The developing mind: How relationships and the brain interact to shape who we are*. Guilford Press.

Singer, T., B. Seymour, J. O'doherty, H. Kaube, R. J. Dolan, and C. D. Frith (2004). Empathy for pain involves the affective but not sensory components of pain. *Science 303*(5661), 1157–1162.

Singer, T., B. Seymour, J. P. O'Doherty, K. E. Stephan, R. J. Dolan, and C. D. Frith (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature 439*(7075), 466–469.

Smith, J. D., S. S. Woodhouse, C. A. Clark, and E. A. Skowron (2016). Attachment status and mother–preschooler parasympathetic response to the strange situation procedure. *Biological psychology 114*, 39–48.

Smith, K. S., A. J. Tindell, J. W. Aldridge, and K. C. Berridge (2009). Ventral pallidum roles in reward and motivation. *Behavioural brain research 196*(2), 155–167.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing*, Volume 1, Chapter 6, pp. 194–281. Cambridge: MIT Press.

Spangler, G. (1998). Emotional and adrenocortical responses of infants to the strange situation: The differential function of emotional expression. *International Journal of Behavioral Development 22*(4), 681–706.

Spangler, G. and K. E. Grossmann (1993). Biobehavioral organization in securely and insecurely attached infants. *Child development 64*(5), 1439–1450.

SPM12 (2014). `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`.

Sroufe, L. A., B. Egeland, and E. A. Carlson (1999). One social world: The integrated development of parent-child and peer relationships. In *Relationships as developmental contexts. The Minnesota symposia on child psychology*, Volume 30, pp. 241–261. Lawrence Erlbaum Associates, Inc.

Steele, H., M. Steele, and P. Fonagy (1996). Associations among attachment classifications of mothers, fathers, and their infants. *Child development*, 541–555.

Stepp, S. D., D. J. Whalen, P. A. Pilkonis, A. E. Hipwell, and M. D. Levine (2012). Children of mothers with borderline personality disorder: identifying parenting behaviors as potential targets for intervention. *Personality Disorders: Theory, Research, and Treatment 3*(1), 76.

Stevens, G. T. and J. Zhang (2009). A dynamic systems model of infant attachment. *IEEE Transactions on Autonomous Mental Development 1*(3), 196–207.

Stippich, C., H. Ochmann, and K. Sartor (2002). Somatotopic mapping of the human primary sensorimotor cortex during motor imagery and motor execution by functional magnetic resonance imaging. *Neuroscience Letters 331*(1), 50–54.

Strathearn, L., P. Fonagy, J. Amico, and P. R. Montague (2009). Adult attachment predicts maternal brain and oxytocin response to infant cues. *Neuropsychopharmacology 34*(13), 2655–2666.

Strathearn, L., J. Li, P. Fonagy, and P. R. Montague (2008). What's in a smile? maternal brain responses to infant facial cues. *Pediatrics 122*(1), 40–51.

Symonds, L. L., N. S. Gordon, J. C. Bixby, and M. M. Mande (2006). Right-lateralized pain processing in the human cortex: an fMRI study. *Journal of Neurophysiology 95*(6), 3823–3830.

Tabarean, I., B. Conti, M. Behrens, H. Korn, and T. Bartfai (2005). Electrophysiological properties and thermosensitivity of mouse preoptic and anterior hypothalamic neurons in culture. *Neuroscience 135*(2), 433–449.

Tachibana, Y. and O. Hikosaka (2012). The primate ventral pallidum encodes expected reward value and regulates motor action. *Neuron 76*(4), 826–837.

Tecuapetla, F., L. Carrillo-Reid, J. Bargas, and E. Galarraga (2007). Dopaminergic modulation of short-term synaptic plasticity at striatal inhibitory synapses. *Proceedings of the National Academy of Sciences 104*(24), 10258–10263.

Tsuneoka, Y., T. Maruyama, S. Yoshida, K. Nishimori, T. Kato, M. Numan, and K. O. Kuroda (2013). Functional, anatomical, and neurochemical differentiation of medial preoptic area subregions in relation to maternal behavior in the mouse. *Journal of Comparative Neurology 521*(7), 1633–1663.

Ulfig, N., M. Setzer, and J. Bohl (2003). Ontogeny of the human amygdala. *Annals of the New York Academy of Sciences 985*(1), 22–33.

van IJzendoorn, M. H. (1995). Adult attachment representations, parental responsiveness, and infant attachment: a meta-analysis on the predictive validity of the adult attachment interview. *Psychological bulletin 117*(3), 387.

van IJzendoorn, M. H. and M. J. Bakermans-Kranenburg (2004). Maternal sensitivity and infant temperament in the formation of attachment. *Theories of infant development*, 231–257.

van Ijzendoorn, M. H., C. Schuengel, and M. J. Bakermans-Kranenburg (1999a). Disorganized attachment in early childhood: Meta-analysis of precursors, concomitants, and sequelae. *Development and psychopathology 11*(02), 225–250.

van Ijzendoorn, M. H., C. Schuengel, and M. J. Bakermans-Kranenburg (1999b). Disorganized attachment in early childhood: Meta-analysis of precursors, concomitants, and sequelae. *Development and psychopathology 11*(02), 225–250.

van IJzendoorn, M. H., C. M. Vereijken, M. J. Bakermans-Kranenburg, and J. Marianne Riksen-Walraven (2004). Assessing attachment security with the attachment Q sort: Meta-analytic evidence for the validity of the observer AQS. *Child development 75*(4), 1188–1213.

Varga, Z., D. Csabai, A. Miseta, O. Wiborg, and B. Czéh (2017). Chronic stress affects the number of GABAergic neurons in the orbitofrontal cortex of rats. *Behavioural Brain Research 316*, 104–114.

Villablanca, J. R. (2010). Why do we have a caudate nucleus. *Acta Neurobiologiae Experimentalis 70*(1).

Vlachos, I., C. Herry, A. Luthi, A. Aertsen, and A. Kumar (2011). Context-dependent encoding of fear and extinction memories in a large-scale network model of the basal amygdala. *PLoS Comput Biol 7*(3).

Voeller, K. K., J. A. Hanson, and R. N. Wendt (1988). Facial affect recognition in children a comparison of the performance of children with right and left hemisphere lesions. *Neurology 38*(11), 1744–1744.

Voges, N., A. Schüz, A. Aertsen, and S. Rotter (2010). A modeler's view on the spatial structure of intrinsic horizontal connectivity in the neocortex. *Progress in neurobiology 92*(3), 277–292.

Vrtička, P., F. Andersson, D. Grandjean, D. Sander, and P. Vuilleumier (2008). Individual attachment style modulates human amygdala and striatum activation during social appraisal. *PLoS One 3*(8).

Vrtička, P., D. Sander, and P. Vuilleumier (2012). Lateralized interactive social content and valence processing within the human amygdala. *Frontiers in human neuroscience 6*.

Wagaman, M. A., J. M. Geiger, C. Shockley, and E. A. Segal (2015). The role of empathy in burnout, compassion satisfaction, and secondary traumatic stress among social workers. *Social work 60*(3), 201–209.

Wallin, D. J. (2007). *Attachment in psychotherapy*. Guilford Press.

Wallis, J. D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nature neuroscience 15*(1), 13–19.

Walter, H. (2012). Social cognitive neuroscience of empathy: concepts, circuits, and genes. *Emotion Review 4*(1), 9–17.

Walton, M. E., T. E. Behrens, M. J. Buckley, P. H. Rudebeck, and M. F. Rushworth (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron 65*(6), 927–939.

Waters, E., C. E. Hamilton, and N. S. Weinfield (2000). The stability of attachment security from infancy to adolescence and early adulthood: General introduction. *Child development 71*(3), 678–683.

Wegiel, J., M. Flory, I. Kuchna, K. Nowicki, S. Ma, H. Imaki, J. Wegiel, I. L. Cohen, E. London, T. Wisniewski, et al. (2014). Stereological study of the neuronal number and volume of 38 brain subdivisions of subjects diagnosed with autism reveals significant alterations restricted to the striatum, amygdala and cerebellum. *Acta Neuropathol. Commun 2*, 141.

White, H., T. J. Flanagan, A. Martin, and D. Silvermann (2011). Mother–infant interactions in women with borderline personality disorder, major depressive disorder, their co-occurrence, and healthy controls. *Journal of Reproductive and Infant Psychology 29*(3), 223–235.

Wilson, R. C., Y. K. Takahashi, G. Schoenbaum, and Y. Niv (2014). Orbitofrontal cortex as a cognitive map of task space. *Neuron 81*(2), 267–279.

Wittling, W. and M. Pflüger (1990). Neuroendocrine hemisphere asymmetries: salivary cortisol secretion during lateralized viewing of emotion-related and neutral films. *Brain and Cognition 14*(2), 243–265.

Wolff, M. S. and M. H. Ijzendoorn (1997). Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment. *Child development 68*(4), 571–591.

Woodruff, A. R. and P. Sah (2007). Networks of parvalbumin-positive interneurons in the basolateral amygdala. *The Journal of neuroscience 27*(3), 553–563.

Woon, F. L., S. Sood, and D. W. Hedges (2010). Hippocampal volume deficits associated with exposure to psychological trauma and posttraumatic stress disorder in adults: a meta-analysis. *Progress in Neuro-Psychopharmacology and Biological Psychiatry 34*(7), 1181–1188.

World Health Organization (1992). *The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines*. Geneva: World Health Organization.

Yang, Q., R. Gu, P. Tang, and Y.-J. Luo (2013). How does cognitive reappraisal affect the response to gains and losses? *Psychophysiology 50*(11), 1094–1103.

Yang, Q., P. Tang, R. Gu, W. Luo, and Y.-j. Luo (2014). Implicit emotion regulation affects outcome evaluation. *Social cognitive and affective neuroscience 10*(6), 824–31.

Zaki, J. (2014). Empathy: a motivated account. *Psychological bulletin 140*(6), 1608.

Zaki, J., J. I. Davis, and K. N. Ochsner (2012). Overlapping activity in anterior insula during interoception and emotional experience. *Neuroimage 62*(1), 493–499.

Zaki, J. and K. Ochsner (2011). You, me, and my brain: Self and other representation in social cognitive neuroscience. *Social neuroscience: Toward understanding the underpinnings of the social mind*, 14–39.

Zaki, J., K. N. Ochsner, J. Hanelin, T. D. Wager, and S. C. Mackey (2007). Different circuits for different pain: patterns of functional connectivity reveal distinct networks for processing pain in self and others. *Social neuroscience 2*(3-4), 276–291.

Zelazo, P. D. and U. Müller (2002). Executive function in typical and atypical development. In U. Goswami (Ed.), *Handbook of childhood cognitive development*, pp. 445 – 469. Oxford, UK: Blackwell.

Zelenko, M., H. Kraemer, L. Huffman, M. Gschwendt, N. Pageler, and H. Steiner (2005). Heart rate correlates of attachment status in young mothers and their infants. *Journal of the American Academy of Child & Adolescent Psychiatry 44*(5), 470–476.

Zenasni, F., E. Boujut, A. Woerner, and S. Sultan (2012). Burnout and empathy in primary care: three hypotheses. *British Journal of General Practice 62*(600), 346–347.

Zhang, J., J. Muller, and A. McDonald (2013). Noradrenergic innervation of pyramidal cells in the rat basolateral amygdala. *Neuroscience 228*, 395–408.

A. Appendix For Chapter 3

In this appendix we give details on the ACE dimension of the AMBIANCE scale, and the discrete formulation of the free energy principle and active inference, both of which were used in Chapter 3.

A.1 Affective Communication Errors (AMBIANCE)

Here we give further details on the behaviour codes for the three categories (1A,1B,1C) of the ACE dimension of the AMBIANCE scale (Bronfman et al., 1999; Safyer, 2013, Appendix G). Behaviours in italics were found to be displayed 3 times more often in mothers of disorganised infants in Lyons-Ruth et al. (1999).

- **1A: Contradictory signalling to infant**
  - **Voice tone incongruent with message:**
    - Stern voice, but permissive message
    - Sweet voice with derogatory, demanding, or impatient message
  - **Verbal content or voice tone incongruent with physical response:**
    - *Invites approach verbally then distances*
    - *Uses friendly tone while maintaining threatening posture*
    - Says something positive about infant while simultaneously indicating aversion
  - **Verbal content or voice tone incongruent with facial expression:**
    - Smiles while using stern voice
    - Exhibits angry facial expression but speaks pleasantly
  - **Incongruent physical behaviours:**
    - *Directs infant to do something then not to do it*
    - Offers then withdraws toy

    - ∗ Holds affectionately, while simultaneously withdrawing or threatening infant

- **1B: Failure to initiate responsive behaviour to infant cue**
    - *Does not attempt to soothe infant when distressed*
    - *Does not offer comfort when infant falls*
    - *Fails to set appropriate limits around safety*
    - Ignores cues for pick up
    - Does not intervene when infant engages in dangerous behaviour
    - Does not respond to clear infant cue (e.g. infant looks at mother and vocalises, but mother fails to respond; infant shows mother a toy and mother does not respond)

- **1C: Inappropriate responding to infant signals or needs**
    - *Laughs while infant crying or distressed*
    - *Directs inauthentic affect towards infant (fake over bright affect)*
    - Ignores infant cue for distance
    - Ignores infant's "no"
    - Mother smiles when infant is angry, upset, afraid, or sad
    - Attempts to minimise or discount infant's display of distress or asking for help (e.g. "what's the matter" or "oh, you're okay")

A.2 Discrete Markovian Formulation of the Free Energy Principle

We follow the mathematical models outlined in Friston et al. (2015) and FitzGerald et al. (2015) based on a finite discrete partially observable Markov decision process (the derivations here are given in more detail in those papers). We thus have a finite set of observations (or observable outcomes) $\widetilde{o} \in O = \{1, ..., W\}$, a finite set of J discrete hidden states $\widetilde{s} \in S = \{1, ..., J\}$ and a finite set of L discrete actions actions $\widetilde{a} \in \Omega = \{1, ..., L\}$, where $\sim$ denotes a sequence of variables over time. The generative process R defining the environment dynamics up to current time t is then:

$$R(\widetilde{o}, \widetilde{s}, \widetilde{a}) = Pr(\{o_0, ..., o_t\} = \widetilde{o}, \{s_0, ..., s_t\} = \widetilde{s}, \{a_0, ..., a_t\} = \widetilde{a}) \tag{A.1}$$

The agent's internal model of this process (their generative model) over observations $\widetilde{o} \in O$, hidden states $\widetilde{s} \in S$ and control states $\widetilde{u} \in U$ is:

$$P(\widetilde{o}, \widetilde{s}, \widetilde{u}) = Pr(\{o_0, ..., o_T\} = \widetilde{o}, \{s_0, ..., s_T\} = \widetilde{s}, \{u_0, ..., u_T\} = \widetilde{u}) \tag{A.2}$$

A policy $\pi \in \{1, ..., K\}$ indexes a sequence of future control states $(\widetilde{u} \mid \pi) = (u_t, ..., u_T|\pi)$, and it is assumed that the agent has an approximate posterior distribution Q over hidden and control states:

$$Q(\widetilde{s}, \widetilde{u}) = Pr(\{s_0, ..., s_T\} = \widetilde{s}, \{u_0, ..., u_T\} = \widetilde{u}) \tag{A.3}$$

which is parametrised by $(\widehat{s}, \widehat{\pi})$, where $\widehat{s} \in [0, 1]^J$ is a $J \times 1$ vector of hidden state expectations, and $\widehat{\pi} \in [0, 1]^K$ is a $K \times 1$ vector of policy expectations.

The variational free energy F is [1]:

$$\begin{aligned} F(\widetilde{o}, \widehat{s}, \widehat{\pi}) &= \mathbb{E}_Q[-\ln P(\widetilde{o}, \widetilde{s}, \widetilde{u}) - H[Q(\widetilde{s}, \widetilde{u})] \\ &= -\ln P(\widetilde{o}) + KL[Q(\widetilde{s}, \widetilde{u})||P(\widetilde{s}, \widetilde{u} \mid \widetilde{o})] \end{aligned} \tag{A.4}$$

where:

$$H[P(x)] = \mathbb{E}_{P(x)}[-\ln P(x)] \tag{A.5}$$

is the Shannon entropy, and:

$$KL[Q(x)||P(x)] = \mathbb{E}_{Q(x)}[\ln Q(x) - \ln P(x)] \tag{A.6}$$

is the Kullback-Leibler divergence, such that minimising free energy minimises an upper bound on surprise $-\ln P(\widetilde{o})$ (see second rearrangement in Eq. A.4, which upper-bounds surprise since the KL divergence term cannot be less than zero).

A.2.1 Factorising the Generative Model

The generative model used to model the finite horizon Markovian process up to time $t \in (0, ..., T)$ can be expressed in terms of the following factorisation:

$$P(\widetilde{o}, \widetilde{s}, \widetilde{u}, \gamma| \widetilde{a}) = P(\widetilde{o} \mid \widetilde{s})P(\widetilde{s} \mid \widetilde{a})P(\widetilde{u} \mid \gamma)P(\gamma|\alpha, \beta) \tag{A.7}$$

The first factor $P(\widetilde{o} \mid \widetilde{s})$ implies that observations only depend on the current hidden state:

$$P(\widetilde{o} \mid \widetilde{s}) = P(o_0|s_0)P(o_1|s_1)...P(o_t|s_t) \tag{A.8}$$

[1] See (Friston et al., 2010, Appendix 2) for derivation.

which is encoded in matrix form as:

$$\forall t : P(o_t|s_t) = A \tag{A.9}$$

such that $P(o_t = i|s_t = j) = A_{ij}$.

The second factor $P(\widetilde{s} \mid \widetilde{a})$ expresses prior beliefs about state transitions (under the assumption that the agent knows their past actions):

$$P(\widetilde{s} \mid \widetilde{a}) = P(s_t|s_{t-1}, a_t)...P(s_1|s_0, a_1)P(s_0) \tag{A.10}$$

with transition probabilities from one state to the next encoded in matrix form as:

$$\forall t : P(s_{t+1}|s_t, u_t) = B(u_t) \tag{A.11}$$

and the prior distribution over initial states (i.e. the agent's prior beliefs with regards to the environmental hidden state that they will start in) encoded as:

$$P(s_0) = D \tag{A.12}$$

The third factor $P(\widetilde{u} \mid \gamma)$ expresses beliefs about sequences of control states (policies). It is assumed that the agent has the prior belief that control states will minimise free energy. The expected negative free energy of a policy $\mathbf{Q}(\pi)$ is defined as the free energy expected under that policy:

$$\mathbf{Q}(\pi) = \sum_{\tau=t+1}^{T} \mathbf{Q}_\tau(\pi) \tag{A.13}$$

with:

$$\begin{aligned} \mathbf{Q}_\tau(\pi) &= \mathbb{E}_{Q(o_\tau, s_\tau|\pi)}[\ln P(o_\tau, s_\tau)] + H[Q(s_\tau|\pi)] \\ &= \{1\}^{W\times 1} \cdot (A \circ \ln A)\,\widehat{s}_\tau(\pi) - (\ln \widehat{o}_\tau(\pi) - \ln C_\tau) \cdot \widehat{o}_\tau(\pi) \end{aligned} \tag{A.14}$$

where $\{1\}^{W\times 1}$ is a column vector of ones, $A \cdot B$ is $A^\top B$, $A \circ B$ is the Hadamard (element-wise) matrix product (i.e. $(A \circ B)_{ij} = A_{ij}B_{ij}$), $\ln A$ denotes the element-wise logarithm of matrix A, and:

$$\begin{aligned} Q(o_\tau, s_\tau|\pi) &= P(o_\tau|s_\tau)Q(s_\tau|\pi) \\ &= P(o_\tau|s_\tau)\mathbb{E}_{Q(s_t)}[P(s_\tau|s_t, \pi)] \\ &= \mathbb{E}_{Q(s_t)}[P(o_\tau, s_\tau|s_t, \pi)] \end{aligned} \tag{A.15}$$

is a posterior predictive distribution over future states and outcomes. The expected states at time τ under policy π are given by:

$$\widehat{s}_\tau(\pi) = B(u_\tau|\pi)...B(u_t|\pi)\widehat{s}_t \tag{A.16}$$

and the expected observations at time τ under policy π are:

$$\widehat{o}_\tau(\pi) = A\widehat{s}_\tau(\pi) \tag{A.17}$$

The agent also encodes a prior distribution over future outcomes:

$$P(o_\tau) = C_\tau \tag{A.18}$$

which specifies the utility (preference) of each outcome to the agent. We can now associate the prior probability over control states with the expected negative free energy of that policy:

$$\ln P(\widetilde{u}\,|\gamma) = \gamma \cdot \mathbf{Q}(\pi) \tag{A.19}$$

where γ is called the "precision" and encodes the confidence in prior beliefs. This expression states that a policy is a-priori more likely if its expected free energy is small. For σ a softmax function, we then have:

$$P(\widetilde{u}\,|\gamma) = \sigma(\gamma \cdot \mathbf{Q}) \tag{A.20}$$

where $\mathbf{Q}$ is a $K \times 1$ vector containing the expected negative free energy of each policy at the current time.

The final factor $P(\gamma|\alpha, \beta)$ expresses a prior over precision γ, which is assumed to have a gamma distribution with shape and rate parameters α and β:

$$P(\gamma|\alpha, \beta) = Gamma(\alpha, \beta) \tag{A.21}$$

A.2.2 Minimising Variational Free Energy

If the following factorisation for the approximate posterior distribution is assumed:

$$\begin{aligned} Q(\widetilde{x}\mid\widehat{x}) &= Q(s_0|\widehat{s}_0)...Q(s_T|\widehat{s}_T)Q(u_t,...,u_T|\,\widehat{\pi})Q(\gamma|\widehat{\gamma}) \\ Q(\gamma|\widehat{\gamma}) &= Gamma(\alpha, \widehat{\beta}{=}\,\alpha/\,\widehat{\gamma}) \end{aligned} \tag{A.22}$$

for $\widetilde{x}{=}\widetilde{s}, \widetilde{u}, \gamma$ and $\widehat{x}{=}\widehat{s}, \widehat{\pi}, \widehat{\gamma}$, then:

$$\begin{aligned}\mathbb{E}_Q[\ln P(\widetilde{o},\widetilde{x})] &= \mathbb{E}_Q[\ln P(o_t|s_t) + \ln P(s_t|s_{t-1},a_t) + \ln P(\widetilde{u}\,|\gamma) + \ln P(\gamma|\alpha,\beta)] \\ &= \widehat{s}_t \cdot \ln A \cdot o_t \\ &\quad + \widehat{s}_t \cdot \ln(B(a_{t-1})\ \widehat{s}_{t-1}) \\ &\quad + \widehat{\gamma}\ \mathbf{Q}\cdot\ \widehat{\pi} \\ &\quad + (\alpha-1)(\psi(\alpha) - \ln\ \widehat{\beta}) - \beta\ \widehat{\gamma}\ + \alpha\ln\beta - \ln\Gamma(\alpha)\end{aligned} \tag{A.23}$$

with $B(a_0)\ \widehat{s}_0 = D$, and where Γ is the gamma function, ψ is the digamma function, and o_t is a $W \times 1$ binary column vector with the index of the non-zero element corresponding to the agent's observation at time t; since for $\gamma \sim Gamma(\alpha,\beta)$ we have probability density function:

$$P(\gamma|\alpha,\beta) = Gamma(\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\gamma^{\alpha-1}e^{-\beta\gamma} \tag{A.24}$$

and so:

$$\begin{aligned}\mathbb{E}_Q[\ln P(\gamma|\alpha,\beta)] &= \mathbb{E}_Q\left[\ln\left(\frac{\beta^\alpha}{\Gamma(\alpha)}\gamma^{\alpha-1}e^{-\beta\gamma}\right)\right] \\ &= \alpha\ln\beta - \ln\Gamma(\alpha) + \mathbb{E}_Q[(\alpha-1)\ln(\gamma) - \beta\gamma] \\ &= \alpha\ln\beta - \ln\Gamma(\alpha) + (\alpha-1)(\psi(\alpha) - \ln\ \widehat{\beta}) - \beta\ \widehat{\gamma}\end{aligned} \tag{A.25}$$

The approximate posterior distribution has entropy:

$$\begin{aligned}H[Q(\widetilde{x}\mid\widehat{x})] &= \mathbb{E}_Q[-\ln Q(\widetilde{x}\mid\widehat{x})] \\ &= \mathbb{E}_Q[-\ln Q(s_t|\ \widehat{s}_t) - \ln Q(\widetilde{u}\mid\widehat{\pi}) - \ln Q(\gamma|\ \widehat{\gamma})] \\ &= -\ \widehat{s}_t \cdot \ln\ \widehat{s}_t - \widehat{\pi}\cdot\ln\ \widehat{\pi} + \ln\Gamma(\alpha) + (1-\alpha)\psi(\alpha) - \ln(\widehat{\beta}) + \alpha\end{aligned} \tag{A.26}$$

since:

$$\begin{aligned}\mathbb{E}_Q[-\ln Q(\gamma|\,\widehat{\gamma})] &= \mathbb{E}_Q\left[-\ln\left(\frac{\widehat{\beta}^{\alpha}}{\Gamma(\alpha)}\gamma^{\alpha-1}e^{-\widehat{\beta}\gamma}\right)\right] \\ &\quad -\alpha\ln\widehat{\beta}+\ln\Gamma(\alpha)-\mathbb{E}_Q[(\alpha-1)\ln(\gamma)-\widehat{\beta}\,\gamma] \\ &= \ln\Gamma(\alpha)+(1-\alpha)\psi(\alpha)-\ln(\widehat{\beta})+\alpha\end{aligned} \tag{A.27}$$

$$\mathbb{E}_Q[\ln Q(\gamma|\,\widehat{\gamma})] = \mathbb{E}_Q\left[\ln\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\gamma^{\alpha-1}e^{-\beta\gamma}\right)\right] \tag{A.28}$$

Thus, the variational free energy is given by:

$$\begin{aligned}F(\widetilde{o},\widetilde{x},\widehat{x}) &= -\mathbb{E}_Q[\ln P(\widetilde{o},\widetilde{x})-H(Q(\widetilde{x}\mid\widehat{x})] \\ &= \widehat{s}_t\cdot(\ln\widehat{s}_t-\ln A\cdot o_t-\ln(B(a_{t-1})\,\widehat{s}_{t-1})) \\ &\quad +\widehat{\pi}\cdot(\ln\widehat{\pi}-\widehat{\gamma}\,\mathbf{Q})+\beta\,\widehat{\gamma}+\alpha(\ln\alpha-\ln\widehat{\gamma}-\ln\beta-1)\end{aligned} \tag{A.29}$$

The partial derivatives of the free energy F with respect to $\widehat{s}_t$, $\widehat{\pi}$ and $\widehat{\gamma}$ are:

$$\frac{\partial F}{\partial\,\widehat{s}_t} = \{1\}^{W\times 1}+\ln\widehat{s}_t-\ln A\cdot o_t-\ln(B(a_{t-1})\,\widehat{s}_{t-1})-\widehat{\gamma}\cdot\bigtriangledown_{\widehat{s}}\,\mathbf{Q}\cdot\widehat{\pi} \tag{A.30}$$

$$\frac{\partial F}{\partial\,\widehat{\pi}} = \{1\}^{K\times 1}+\ln\widehat{\pi}-\widehat{\gamma}\cdot\mathbf{Q} \tag{A.31}$$

$$\frac{\partial F}{\partial\,\widehat{\gamma}} = \beta-\mathbf{Q}\cdot\widehat{\pi}-\widehat{\beta} \tag{A.32}$$

Setting these equal to zero (ignoring the final term in the derivative with respect to hidden states, for simplicity) and re-arranging gives the variational updates that minimise free energy:

$$\widehat{s}_t = \sigma(\ln A\cdot o_t+\ln(B(a_{t-1})\widehat{s}_{t-1})) \tag{A.33}$$

$$\widehat{\pi} = \sigma(\widehat{\gamma}\cdot\mathbf{Q}) \tag{A.34}$$

$$\widehat{\gamma} = \alpha/(\beta-\mathbf{Q}\cdot\widehat{\pi}) \tag{A.35}$$

The updates in equations A.34 and A.35 are iterated until convergence (or, and as here, for a fixed number of N steps).

A.2.3 Learning the Generative Model

Now we consider how the generative model (i.e. the distributions defined by A, B and D) might be learned. Recall that $A_{\bullet j}$ (i.e. column j of A) encodes the likelihood of observations given hidden state j. Following the method and derivation in FitzGerald et al. (2015), we place a Dirichlet prior over each of these multinomial distributions, with concentration parameters θ:

$$P(A_{\bullet j} \mid \theta) = Dirichlet(\theta_{\bullet j}) \tag{A.36}$$

Recall also that $B(u)_{\bullet j}$ encodes the likelihood of hidden states at $t+1$ given that the hidden state at time t is j. We similarly place a Dirichlet prior over these distributions, with concentration parameters ϕ:

$$P(B(u)_{\bullet j} \mid \phi(u)) = Dirichlet(\phi(u)_{\bullet j}) \tag{A.37}$$

The following factorisation is assumed for the generative model:

$$P(\tilde{o}, \tilde{x} \mid \tilde{a}) = P(\tilde{o} \mid \tilde{s}, A)P(\tilde{s} \mid \tilde{a}, B)P(\tilde{u} \mid \gamma)P(\gamma|\alpha, \beta)P(A|\theta)P(B|\phi) \tag{A.38}$$

for $\tilde{x}{=}\tilde{s}, \tilde{u}, \gamma, A, B$; and for the approximate posterior:

$$\begin{aligned} Q(\tilde{x} \mid \widehat{x}) &= Q(s_0|\widehat{s}_0)...Q(s_T|\widehat{s}_T)Q(u_t, ..., u_T|\ \widehat{\pi})Q(\gamma|\widehat{\gamma})Q(A|\widehat{\theta})Q(B|\widehat{\phi}) \\ Q(\gamma|\widehat{\gamma}) &= Gamma(\alpha, \widehat{\beta}{=}\ \alpha/\ \widehat{\gamma}) \\ Q(A|\widehat{\theta}) &= Dirichlet(\widehat{\theta}) \\ Q(B|\widehat{\phi}) &= Dirichlet(\widehat{\phi}) \end{aligned} \tag{A.39}$$

with $\widehat{x}{=}\widehat{s}, \widehat{\pi}, \widehat{\gamma}, \widehat{\theta}, \widehat{\phi}$. Recall that the probability density function of $P(A_{\bullet j} \mid \theta) = Dirichlet(\theta_{\bullet j})$ is given by:

$$P(A_{\bullet j} \mid \theta) = Dirichlet(\theta_{\bullet j}) = \frac{\Gamma(\sum_i \theta_{ij})}{\prod_i \Gamma(\theta_{ij})} \prod_i A_{ij}^{\theta_{ij}-1} \tag{A.40}$$

and thus we have that:

$$\begin{aligned}
\mathbb{E}_Q[\ln P(A_{\bullet j} \mid \theta)] &= \mathbb{E}_Q\left[\ln\left(\frac{\Gamma(\sum_i \theta_{ij})}{\prod_i \Gamma(\theta_{ij})} \prod_i A_{ij}^{\theta_{ij}-1}\right)\right] \\
&= \mathbb{E}_Q\left[\ln(\Gamma(\sum_i \theta_{ij})) - \sum_i \ln(\Gamma(\theta_{ij})) + \sum_i (\theta_{ij}-1)\ln(A_{ij})\right] \\
&= \ln(\Gamma(\sum_i \theta_{ij})) - \sum_i \ln(\Gamma(\theta_{ij})) \\
&\quad + \sum_i (\theta_{ij}-1)[\psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})]
\end{aligned} \tag{A.41}$$

and:

$$\begin{aligned}
\mathbb{E}_Q[\ln Q(A_{\bullet j} \mid \widehat{\theta})] &= \ln(\Gamma(\sum_i \widehat{\theta}_{ij})) - \sum_i \ln(\Gamma(\widehat{\theta}_{ij})) \\
&\quad + \sum_i (\widehat{\theta}_{ij} - 1)[\psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})]
\end{aligned} \tag{A.42}$$

(and similarly for $\mathbb{E}_Q[\ln P(B_{\bullet j} \mid \phi)]$ and $\mathbb{E}_Q[\ln Q(B_{\bullet j} \mid \widehat{\phi})]$). Therefore, for our generative model P we have that:

$$
\begin{aligned}
\mathbb{E}_Q[\ln P(\widetilde{o},\widetilde{x})] &= \mathbb{E}_Q[\ln P(o_t|s_t,A) + \ln P(s_t|s_{t-1},a_t,B) + \ln P(\widetilde{u}\,|\gamma) \\
&\quad + \ln P(\gamma|\alpha,\beta) + \ln P(A|\theta) + \ln P(B|\phi)] \\
&= \widehat{s}_t \cdot \widehat{A} \cdot o_t \\
&\quad + \widehat{s}_t \cdot \widehat{B}\,(a_{t-1})\ \widehat{s}_{t-1} \\
&\quad + \widehat{\gamma}\ \mathbf{Q}\cdot \widehat{\pi} \\
&\quad + (\alpha-1)(\psi(\alpha) - \ln\widehat{\beta}) - \beta\ \widehat{\gamma} + \alpha\ln\beta - \ln\Gamma(\alpha) \\
&\quad + \sum_j \left(\ln(\Gamma(\sum_i \theta_{ij})) - \sum_i \ln(\Gamma(\theta_{ij})) + \sum_i (\theta_{ij}-1)[\psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})] \right) \\
&\quad + \sum_j \left(\ln(\Gamma(\sum_i \phi_{ij})) - \sum_i \ln(\Gamma(\phi_{ij})) + \sum_i (\phi_{ij}-1)[\psi(\widehat{\phi}_{ij}) - \psi(\sum_i \widehat{\phi}_{ij})] \right)
\end{aligned}
\tag{A.43}
$$

for $\widehat{A}_{ij} = \mathbb{E}_Q[\ln A_{ij}] = \psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})$ and $\widehat{B}_{ij} = \mathbb{E}_Q[\ln B_{ij}] = \psi(\widehat{\phi}_{ij}) - \psi(\sum_i \widehat{\phi}_{ij})$, and:

$$
\mathbf{Q}(\pi) = \sum_{\tau=t+1}^{T} \{1\}^{W\times 1} \cdot (\overline{A} \circ \widehat{A})\widehat{s}_\tau(\pi) - (\ln \widehat{o}_\tau(\pi) - \ln C_\tau) \cdot \widehat{o}_\tau(\pi) \tag{A.44}
$$

with:

$$
\widehat{s}_\tau(\pi) = \overline{B}(u_\tau|\pi)...\overline{B}(u_t|\pi)\ \widehat{s}_t \tag{A.45}
$$

and:

$$
\widehat{o}_\tau(\pi) = \overline{A}\ \widehat{s}_\tau(\pi) \tag{A.46}
$$

for $\overline{A}_{ij} = \mathbb{E}_Q[A_{ij}] = \widehat{\theta}_{ij}\ /\sum_i \widehat{\theta}_{ij}$ and $\overline{B}_{ij} = \mathbb{E}_Q[B_{ij}] = \widehat{\phi}_{ij}\ /\sum_i \widehat{\phi}_{ij}$. The approximate posterior distribution has entropy:

$$
\begin{aligned}
H[Q(\widetilde{x} \mid \widehat{x})] &= \mathbb{E}_Q[-\ln Q(\widetilde{x} \mid \widehat{x})] \\
&= \mathbb{E}_Q[-\ln Q(s_t|\ \widehat{s}_t) - \ln Q(\widetilde{u} \mid \widehat{\pi}) - \ln Q(\widetilde{\gamma} \mid \widehat{\gamma}) \\
&\quad - \ln Q(A|\ \widehat{\theta}) - \ln Q(B|\ \widehat{\phi})] \\
&= -\ \widehat{s}_t \cdot \ln\ \widehat{s}_t - \ \widehat{\pi} \cdot \ln\ \widehat{\pi} + \ln\Gamma(\alpha) + (1-\alpha)\psi(\alpha) - \ln(\widehat{\beta}) + \alpha \\
&\quad - \sum_j \left(\ln(\Gamma(\sum_i \widehat{\theta}_{ij})) - \sum_i \ln(\Gamma(\widehat{\theta}_{ij})) + \sum_i (\widehat{\theta}_{ij} - 1)[\psi(\widehat{\theta}_{ij}) - \psi(\sum_i \widehat{\theta}_{ij})] \right) \\
&\quad - \sum_j \left(\ln(\Gamma(\sum_{i=1}^{n} \widehat{\phi}_{ij})) - \sum_i \ln(\Gamma(\widehat{\phi}_{ij})) + \sum_i (\widehat{\phi}_{ij} - 1)[\psi(\widehat{\phi}_{ij}) - \psi(\sum_i \widehat{\phi}_{ij})] \right)
\end{aligned} \tag{A.47}
$$

Thus, the variational free energy is given by:

$$
\begin{aligned}
F(\widetilde{o}, \widetilde{x}, \widehat{x}) &= -\mathbb{E}_Q[\ln P(\widetilde{o}, \widetilde{s}, \widetilde{u}, \gamma, A, B)] - H[Q(\widetilde{x} \mid \widehat{x})] \\
&= \widehat{s}_t \cdot (\ln\ \widehat{s}_t - \widehat{A} \cdot o_t - \widehat{B}\ (a_{t-1})\ \widehat{s}_{t-1}) \\
&+ \widehat{\pi} \cdot (\ln\ \widehat{\pi} - \widehat{\gamma}\ \mathbf{Q}) + \beta\ \widehat{\gamma} + \alpha(\ln\alpha - \ln\ \widehat{\gamma} - \ln\beta - 1) \\
&+ \sum_{ij} \left((\widehat{\theta}_{ij} - \theta_{ij})\ \widehat{A}_{ij} + \ln\Gamma(\theta_{ij}) - \ln\Gamma(\widehat{\theta}_{ij}) \right) + \sum_j \left(\ln\Gamma(\sum_i \widehat{\theta}_{ij}) - \ln\Gamma(\sum_i \theta_{ij}) \right) \\
&+ \sum_{ij} \left((\widehat{\phi}_{ij} - \phi_{ij})\ \widehat{B}_{ij} + \ln\Gamma(\phi_{ij}) - \ln\Gamma(\widehat{\phi}_{ij}) \right) + \sum_j \left(\ln\Gamma(\sum_i \widehat{\phi}_{ij}) - \ln\Gamma(\sum_i \phi_{ij}) \right)
\end{aligned} \tag{A.48}
$$

The partial derivatives of the free energy F with respect to $\widehat{s}_t$, $\widehat{\pi}$, $\widehat{\gamma}$, $\widehat{\theta}_{ij}$ and $\widehat{\phi}_{ij}$ are:

$$
\frac{\partial F}{\partial\ \widehat{s}_t} = \{1\}^{W\times 1} + \ln\ \widehat{s}_t - \widehat{A} \cdot o_t - \widehat{B}\ (a_{t-1})\ \widehat{s}_{t-1} - \widehat{\gamma} \cdot \bigtriangledown_{\widehat{s}}\ \mathbf{Q} \cdot\ \widehat{\pi} \tag{A.49}
$$

$$
\frac{\partial F}{\partial\ \widehat{\pi}} = \{1\}^{K\times 1} + \ln\ \widehat{\pi} - \widehat{\gamma} \cdot \mathbf{Q} \tag{A.50}
$$

$$
\frac{\partial F}{\partial\ \widehat{\gamma}} = \beta - \mathbf{Q} \cdot\ \widehat{\pi} - \widehat{\beta} \tag{A.51}
$$

$$
\frac{\partial F}{\partial\ \widehat{\theta}_{ij}} = \frac{\partial\ \widehat{A}_{ij}}{\partial\ \widehat{\theta}_{ij}} (\widehat{\theta}_{ij} - \theta_{ij} - o_{ti}\, \widehat{s}_{tj}) \tag{A.52}
$$

$$\frac{\partial F}{\partial \widehat{\phi}_{ij}} = \frac{\partial \widehat{B}_{ij}}{\partial \widehat{\phi}_{ij}}(\widehat{\phi}_{ij} - \phi_{ij} - \widehat{s}_{ti}\widehat{s}_{t-1j}) \tag{A.53}$$

Setting these equal to zero (ignoring the final term in the derivative with respect to hidden states, for simplicity) and re-arranging gives the variational updates that minimise free energy (FitzGerald et al., 2015; Friston et al., 2016):

$$\widehat{s}_t = \sigma(\widehat{A} \cdot o_t + \widehat{B}\,(a_{t-1})\widehat{s}_{t-1}) \tag{A.54}$$

$$\widehat{\pi} = \sigma(\widehat{\gamma} \cdot \mathbf{Q}) \tag{A.55}$$

$$\widehat{\gamma} = \alpha/(\beta - \mathbf{Q} \cdot \widehat{\pi}) \tag{A.56}$$

$$\widehat{\theta}_{ij} = \theta_{ij} + \sum_t o_{ti}\widehat{s}_{tj} \tag{A.57}$$

$$\widehat{\phi}\,(u)_{ij} = \phi(u)_{ij} + \sum_t [u = a_{t-1}] \cdot \widehat{s}_{ti}\widehat{s}_{t-1j} \tag{A.58}$$

where the Iverson brackets $[\cdot]$ returns one if the expression is true and zero otherwise.

Here we consider learning only of the hidden state initial and transition distributions (i.e. we assume that the generative model of observations given hidden states is known and fixed). We assume Y iterations of the learning algorithm, where the initial step of each iteration returns the agent to a hidden state sampled from the generative process initial hidden state distribution. Similarly to the hidden state transition model, initial state beliefs D are assumed to be captured by a Dirichlet distribution with prior concentration parameters ξ (i.e. $P(D \mid \xi) = Dirichlet(\xi)$ with $\widehat{D}_i = \mathbb{E}_Q[\ln D_i] = \psi(\widehat{\xi}_i) - \psi(\sum_i \widehat{\xi}_i)$) which are incrementally updated based on initial hidden state beliefs at each iteration (Friston et al., 2016). The full process is shown concisely in Algorithm A1.

Algorithm A1 Free energy minimisation

input: generative process R
input: generative model prior parameters $\theta, \phi, C, \xi, \alpha, \beta$
initialise: $\widehat{\gamma} = \alpha/\beta$, $\widehat{\theta} = \theta$, $\widehat{\phi} = \phi$, $\widehat{\xi} = \xi$
for all $y = 1, ..., Y$ **do**
  sample: $s_1, o_1 \sim R$
  for $t = 1, ..., T$ **do**
    if t is 1 **then**
      $\widehat{s}_t \leftarrow \sigma(\widehat{A} \cdot o_t + \widehat{D})$
      $\widehat{\pi} \leftarrow \sigma(\{1\}^{K \times 1})$
    else
      $\widehat{s}_t \leftarrow \sigma(\widehat{A} \cdot o_t + \widehat{B}\,(a_{t-1})\widehat{s}_{t-1})$
    end if
    $\mathbf{Q} \leftarrow \{0\}^{K \times 1}$
    for all π **do**
      for $\tau = t, ..., T$ **do**
        $\widehat{s}_\tau(\pi) \leftarrow \overline{B}(u_\tau|\pi)...\overline{B}(u_t|\pi)\widehat{s}_t$
        $\widehat{o}_\tau(\pi) \leftarrow \overline{A}\ \widehat{s}_\tau(\pi)$
        $\mathbf{Q}_\pi \leftarrow \mathbf{Q}_\pi + \{1\}^{W \times 1} \cdot (\overline{A} \circ \widehat{A})\widehat{s}_\tau(\pi) - (\ln \widehat{o}_\tau(\pi) - \ln C_\tau) \cdot \widehat{o}_\tau(\pi)$
      end for
    end for
    for $n = 1, ..., N$ **do**
      $\widehat{\pi} \leftarrow \sigma(\widehat{\gamma} \cdot \mathbf{Q})$
      $\widehat{\gamma} \leftarrow \alpha/(\beta - \mathbf{Q} \cdot \widehat{\pi})$
    end for
    sample: $a_t \sim \widehat{\pi}$
    sample: $s_{t+1}, o_{t+1} \sim R(a_t, s_t)$
  end for
  if learn hidden state transition model **then**
    for all u, i, j **do**
      $\widehat{\phi}\ (u)_{ij} \leftarrow \widehat{\phi}\ (u)_{ij} + \sum_{\tau=2}^{T} [u = a_{\tau-1}] \cdot \widehat{s}_{\tau i}\widehat{s}_{\tau-1 j}$
    end for
  end if
  if learn observation model **then**
    for all i, j **do**
      $\widehat{\theta}_{ij} \leftarrow \widehat{\theta}_{ij} + \sum_{\tau=1}^{T} o_{\tau i}\widehat{s}_{\tau j}$
    end for
  end if
  if learn initial hidden state model **then**
    $\widehat{\xi} \leftarrow \widehat{\xi} + \widehat{s}_1$
  end if
end for

B. Appendix For Chapter 4

In this appendix we give details on the algorithms used to train the DBN used in Chapter 4.

B.1 Restricted Boltzmann Machines

An RBM (Smolensky, 1986) is a stochastic generative neural network with X visible units x and H hidden units h. In the typical case (and as considered here) both $x \in \{0,1\}^X$ and $h \in \{0,1\}^H$ are binary. The network has parameters $\theta = \{b, c, W\}$. The weights W between each visible and hidden unit are symmetric and bidirectional (i.e. $W_{ij} = W_{ji}$), and there are no connections between hidden units, and no connections between visible units. Hidden units have biases b, and visible units have biases c.

The aim is to model a distribution over the input data the network sees. In order to achieve this, we first define an energy function:

$$E(x,h) = -h^\top W x - c^\top x - b^\top h = -\sum_j \sum_k W_{j,k} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \tag{B.1}$$

i.e. each configuration of x and h is assigned a scalar energy. The joint-distribution over x and h is then:

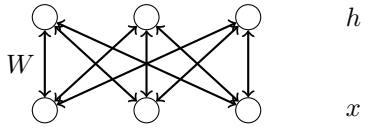

Figure B.1: Restricted Boltzmann machine with 3 hidden units and 3 visible units

$$p(x,h) = \frac{e^{-E(x,h)}}{Z} \tag{B.2}$$

Thus, the probability of some particular configuration is the exponential of the negative energy for that configuration, over the partition function Z. The partition function (sometimes called a "normalisation constant") is the sum of the numerator over all values of x and h:

$$Z = \sum_{x,h} e^{-E(x,h)} \tag{B.3}$$

Given these properties, it can be shown that (Fischer and Igel, 2014):

$$p(h_j = 1|x) = \sigma(b_j + \sum_k W_{jk} x_k) \tag{B.4}$$

and:

$$p(x_j = 1|h) = \sigma(c_j + \sum_i W_{ij} h_i) \tag{B.5}$$

where σ is the logistic function:

$$\sigma(t) = \frac{1}{1 + e^{-t}} \tag{B.6}$$

B.1.1 Training with Contrastive Divergence

We want to adjust the parameters θ (i.e. the weights and biases) in order to maximise the probability of the training data $x \in D$. This is equivalent to minimising the average negative log-likelihood:

$$\min_\theta \frac{1}{|D|} \sum_{x \in D} - \ \log \ p(x;\theta) \tag{B.7}$$

for $|D|$ total training data examples. We use stochastic gradient descent to iteratively update θ based on the gradient of $log\ p(x)$ so that the model comes to assign high probability to those inputs which it has seen. Parameters $\vartheta \in \theta$ are updated over mini-batches $T \subset D$ according to:

$$\triangle\vartheta = \frac{\alpha}{|T|}\frac{\partial}{\partial\vartheta}\left(\sum_{x\in T}\log\ p(x;\theta)\right) - \lambda\vartheta \tag{B.8}$$

for learning rate parameter $\alpha \in \mathbb{R}^+$. In order to achieve models with parameters having small absolute values, we minus half of the sum of squared parameters (i.e. $\frac{1}{2}||\theta||^2$) weighted by $\lambda \in \mathbb{R}_0^+$ (i.e. the set of all positive real numbers along with zero) from the objective function (i.e. negative log likelihood of the training data) being minimised, which results in the $-\lambda\vartheta$ term above. The log-likelihood of any particular data point x under the model parametrised by θ is:

$$\log\ p(x;\theta) = log\sum_h p(x,h;\theta) = \log\frac{1}{Z}\sum_h e^{-E(x,h)} = \log\sum_h e^{-E(x,h)} - \log\sum_{x,h} e^{-E(x,h)} \tag{B.9}$$

and so for the gradient we have:

$$\begin{aligned}\frac{\partial}{\partial\vartheta}\left(-\log\ p(x;\theta)\right) &= -\frac{\partial}{\partial\vartheta}\left(\log\sum_h e^{-E(x,h)}\right) + \frac{\partial}{\partial\vartheta}\left(\log\sum_{x,h} e^{-E(x,h)}\right) \\ &= \sum_h p(h|x)\frac{\partial E(x,h)}{\partial\vartheta} - \sum_x p(x)\sum_h p(h|x)\frac{\partial E(x,h)}{\partial\vartheta}\end{aligned} \tag{B.10}$$

Over mini-batch $T \subset D$ with $|T|$ examples we have that the mean of this derivative is:

$$\begin{aligned}&\frac{1}{|T|}\sum_{x^{(t)}\in T}\frac{\partial}{\partial\vartheta}\ \left(-\ \log\ p(x^{(t)};\theta)\right) \\ &\quad= \frac{1}{|T|}\sum_{x^{(t)}\in T}\left(\sum_h p(h|x^{(t)})\frac{\partial E(x^{(t)},h)}{\partial\vartheta} - \sum_x p(x)\sum_h p(h|x)\frac{\partial E(x,h)}{\partial\vartheta}\right) \\ &\quad= \frac{1}{|T|}\sum_{x^{(t)}\in T}\left(\mathbb{E}_h\left[\frac{\partial E(x^{(t)},h)}{\partial\vartheta}\ \middle|\ x^{(t)}\right] - \mathbb{E}_{h,x}\left[\frac{\partial E(x,h)}{\partial\vartheta}\right]\right)\end{aligned} \tag{B.11}$$

where the first term $\mathbb{E}_h\left[\frac{\partial E(x^{(t)},h)}{\partial\vartheta}\ \middle|\ x^{(t)}\right]$ is called the "positive phase", and the second term $\mathbb{E}_{h,x}\left[\frac{\partial E(x,h)}{\partial\vartheta}\right]$ is called the "negative phase". We can compute $\frac{\partial E(x,h)}{\partial\vartheta}$ for the weights:

$$\frac{\partial E(x,h)}{\partial W_{jk}} = \frac{\partial}{\partial W_{jk}} \left(-\sum_{j,k} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \right) = -h_j x_k \tag{B.12}$$

For the hidden biases b we have that:

$$\frac{\partial E(x,h)}{\partial b_j} = \frac{\partial}{\partial b_j} \left(-\sum_{j,k} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \right) = -h_j \tag{B.13}$$

and for the visible biases c:

$$\frac{\partial E(x,h)}{\partial c_k} = \frac{\partial}{\partial c_k} \left(-\sum_{j,k} W_{jk} h_j x_k - \sum_k c_k x_k - \sum_j b_j h_j \right) = -x_k \tag{B.14}$$

B.1.1.1 Positive phase

The positive phase is the expectation (over the hidden layer values) of the partial derivative of the energy E with respect to the parameter $\vartheta \in \theta$, given the observation $x^{(t)}$. From Eq (B.12), it follows that:

$$\mathbb{E}_h \left[\frac{\partial E(x^{(t)}, h)}{\partial W_{jk}} \;\middle|\; x^{(t)} \right] = \mathbb{E}_h \left[-h_j x_k \;\middle|\; x^{(t)} \right] = \sum_{h_j \in \{0,1\}} -h_j x_k \; p(h_j | x^{(t)}; \theta) = -x_k \; p(h_j = 1 | x^{(t)}; \theta) \tag{B.15}$$

which can be written in matrix form as:

$$\mathbb{E}_h[\bigtriangledown_W E(x,h) | x^{(t)}] = -h(x^{(t)}; \theta) {x^{(t)}}^\top \tag{B.16}$$

where $h(x;\theta)$ is a vector containing the probability that each hidden unit h_j is equal to 1, given x:

$$h(x;\theta) = \begin{pmatrix} p(h_1 = 1 | x; \theta) \\ ... \\ p(h_H = 1 | x; \theta) \end{pmatrix} = \sigma(b + Wx) \tag{B.17}$$

so that the entry at row j, column k in $-h(x;\theta)x^\top$ will be $-x_k\ p(h_j = 1|x;\theta)$.

For the hidden biases b, from Eq. (B.13) it follows that:

$$\mathbb{E}_h\left[\frac{\partial E(x^{(t)},h)}{\partial b_j}\ \middle|\ x^{(t)}\right] = \mathbb{E}_h\left[-h_j\ \middle|\ x^{(t)}\right] = \sum_{h_j\in\{0,1\}} -h_j\ p(h_j|x^{(t)};\theta) = -\ p(h_j = 1|x^{(t)};\theta) \tag{B.18}$$

which can be written in matrix form:

$$\mathbb{E}_h[\bigtriangledown_b E(x,h)|x^{(t)}] = -h(x^{(t)};\theta) \tag{B.19}$$

Finally, for the visible biases c, from Eq. (B.14) it follows that:

$$\mathbb{E}_h\left[\frac{\partial E(x^{(t)},h)}{\partial c_k}\ \middle|\ x^{(t)}\right] = \mathbb{E}_h\left[-x_k\ \middle|\ x^{(t)}\right] = -x_k \tag{B.20}$$

which can be written in matrix form:

$$\mathbb{E}_h[\bigtriangledown_c E(x,h)|x^{(t)}] = -x^{(t)} \tag{B.21}$$

B.1.1.2 Negative phase

The negative phase is the expectation (over the hidden and visible layer values) of the partial derivative of the energy E with respect to the parameter $\vartheta \in \theta$, i.e. it is an expectation under our model's distribution. Thus, the negative phase is hard to compute (generally intractable), since it involves a sum over both h and x which grows exponentially in the number of these units. Therefore this term is typically approximated using a technique called k-step Contrastive Divergence (CD-k) (Hinton, 2002).

B.1.1.3 Contrastive Divergence

Since the negative phase is generally intractable, we replace the expectation over x with a point estimate at a sample $\tilde{x}$ from the RBM's model distribution. We could use Gibbs sampling (Markov chain Monte Carlo) to obtain unbiased samples $\tilde{x}$ from the RBM distribution, and then use these samples to obtain an unbiased estimate of the log-likelihood gradient, since it is easy to perform Gibbs sampling with a RBM given that all units in a layer are conditionally independent given the other layer. However, this would require lots

of sampling steps, since we would have to run the Markov chain until reaching the stationary distribution of the RBM. Instead, we can use an algorithm called CD-k which involves a variation on Gibbs sampling. In CD-k we start at the training vector $x^{(t)}$ and run the Markov chain for only k steps, setting:

$$x^{(t,0)} := x^{(t)} \tag{B.22}$$

and:

$$\forall\ 0 \leq i < k : h^{(t,i)} \sim p(h|x^{(t,i)}) \text{ and } x^{(t,i+1)} \sim p(x|h = h^{(t,i)}) \tag{B.23}$$

In other words, we initialise the visible units with a training sample $x^{(t)}$, giving the first visible unit configuration of the chain $x^{(t,0)}$. Then we sample all of the hidden units in parallel, from $p(h|x = x^{(t,0)})$, to give $h^{(t,0)}$. Then we sample $x^{(t,1)}$ from $p(x|h = h^{(t,0)})$. This repeats for k steps. Then, using $x^{(t,k)}$ as our sample $\tilde{x}^{(t)}$, our estimate for the negative phase part of the gradient is:

$$\mathbb{E}_{h,x}\left[\frac{\partial E(x,h)}{\partial \vartheta}\right] \approx \mathbb{E}_h\left[\frac{\partial E(\tilde{x}^{(t)},h)}{\partial \vartheta}\ \middle|\ \tilde{x}^{(t)}\right] \tag{B.24}$$

so that the gradient of the negative log-likelihood for one training vector $x^{(t)}$ with respect to ϑ is approximated by:

$$\frac{\partial}{\partial \vartheta}\ \Big(-\ \log\ p(x^{(t)};\theta)\Big) \approx \mathbb{E}_h\left[\frac{\partial E(x^{(t)},h)}{\partial \vartheta}\ \middle|\ x^{(t)}\right] - \mathbb{E}_h\left[\frac{\partial E(\tilde{x}^{(t)},h)}{\partial \vartheta}\ \middle|\ \tilde{x}^{(t)}\right] \tag{B.25}$$

The gradient descent update rule is decreasing the energy at the training observation $x^{(t)}$ and its associated hidden layer (positive phase), and increasing the energy at the sample $\tilde{x}^{(t)}$ and its associated hidden layer. Since low energy corresponds to high probability, we are increasing the probability of observing $x^{(t)}$ with its hidden layer under the model's distribution, and at the same time decreasing the probability that $\tilde{x}^{(t)}$ will be observed. Since $\tilde{x}^{(t)}$ is not a sample from the stationary distribution, this approximation (Eq. B.25) is biased. As $k \to \infty$ we are guaranteed to get a sample from the equilibrium distribution, however in practice $k = 1$ works surprisingly well.

Given a training vector $x^{(t)} \in D$ and a contrastive divergence sample $\tilde{x}^{(t)}$, the estimate for the difference between the positive and negative phases for the weights is thus:

$$\begin{aligned}\nabla_W \left(-\log\left(p\left(x^{(t)}\right)\right)\right) &= \mathbb{E}_h\left[\nabla_W E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_{h,x}\left[\nabla_W E(x,h)\right] \\ &\approx \mathbb{E}_h\left[\nabla_W E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_h\left[\nabla_W E(\tilde{x}^{(t)},h)\middle|\tilde{x}^{(t)}\right] \\ &= h(x^{(t)}){x^{(t)}}^\top - h(\tilde{x}^{(t)}){\tilde{x}^{(t)}}^\top\end{aligned} \tag{B.26}$$

so that the stochastic gradient descent update, for some particular mini-batch of training examples $T \subset D$, is:

$$\triangle W = \frac{\alpha}{|T|}\sum_{x^{(t)}\in T}\left(h(x^{(t)}){x^{(t)}}^\top - h(\tilde{x}^{(t)}){\tilde{x}^{(t)}}^\top\right) - \lambda W \tag{B.27}$$

For the hidden biases b we have:

$$\begin{aligned}\nabla_b \left(-\log\left(p\left(x^{(t)}\right)\right)\right) &= \mathbb{E}_h\left[\nabla_b E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_{h,x}\left[\nabla_b E(x,h)\right] \\ &\approx \mathbb{E}_h\left[\nabla_b E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_h\left[\nabla_b E(\tilde{x}^{(t)},h)\middle|\tilde{x}^{(t)}\right] \\ &= h(x^{(t)}) - h(\tilde{x}^{(t)})\end{aligned} \tag{B.28}$$

so that:

$$\triangle b = \frac{\alpha}{|T|}\sum_{x^{(t)}\in T}\left(h(x^{(t)}) - h(\tilde{x}^{(t)})\right) - \lambda b \tag{B.29}$$

and for the visible biases c we have:

$$\begin{aligned}\nabla_c \left(-\log\left(p\left(x^{(t)}\right)\right)\right) &= \mathbb{E}_h\left[\nabla_c E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_{h,x}\left[\nabla_c E(x,h)\right] \\ &\approx \mathbb{E}_h\left[\nabla_c E(x^{(t)},h)\middle|x^{(t)}\right] - \mathbb{E}_h\left[\nabla_c E(\tilde{x}^{(t)},h)\middle|\tilde{x}^{(t)}\right] \\ &= x^{(t)} - \tilde{x}^{(t)}\end{aligned} \tag{B.30}$$

so that:

$$\triangle c = \frac{\alpha}{|T|}\sum_{x^{(t)}\in T}\left(x^{(t)} - \tilde{x}^{(t)}\right) - \lambda c \tag{B.31}$$

This gives us the following general algorithm for the gradient descent update of the weights and biases, using k-step CD for estimation of the gradient and mini-batches:

Algorithm B1 RBM Training algorithm using CD-k with mini-batches

initialise: W, b, c
initialise: $q = 0$
for all training epochs **do**
  for all training mini-batches T **do**
    initialise: $\triangle W = \{0\}^{X \times H}$, $\triangle b = \{0\}^{H \times 1}$, $\triangle c = \{0\}^{X \times 1}$, $A = \{0\}^{|T| \times H}$
    for all $x \in T$ **do**
      $x^{(0)} \leftarrow x$
      for $i = 0, ..., k-1$ **do**
        sample: $h^{(i)} \sim p(h|x^{(i)})$
        sample: $x^{(i+1)} \sim p(x|h^{(i)})$
      end for
      $\tilde{x} \leftarrow x^{(k)}$
      $\triangle W \leftarrow \triangle W + h(x)x^\top - h(\tilde{x})\tilde{x}^\top$
      $\triangle b \leftarrow \triangle b + h(x) - h(\tilde{x})$
      $\triangle c \leftarrow \triangle c + x - \tilde{x}$
      $A_x \leftarrow p(h|x)$
    end for
    $W \leftarrow W + (\alpha/|T|) \triangle W - \lambda W$
    $b \leftarrow b + (\alpha/|T|) \triangle b - \lambda b$
    $c \leftarrow c + (\alpha/|T|) \triangle c - \lambda c$
    $q \leftarrow \gamma q + (1-\gamma)(1/|T|)\{1\}^{1 \times |T|} A$
  end for
  $b \leftarrow b - \eta(q - d)$
end for

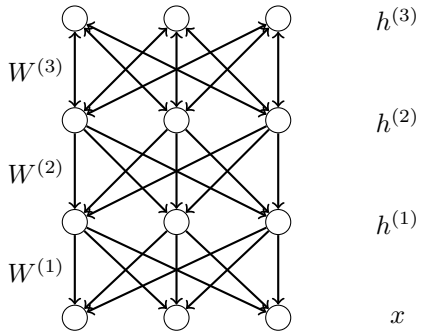

Figure B.2: A 3 layered deep belief network (with 4 layers of units)

for $\{0\}^{i\times j}$ the $i \times j$ matrix of zeros, and $\{1\}^{i\times j}$ the $i \times j$ matrix of ones. Note that for the hidden biases b we have additionally minused another term $q - d$ at the end of each epoch, weighted by $\eta \in \mathbb{R}_0^+$ (this term is sometimes also minused from the weights). This is in order to encourage sparse hidden unit activations. $q \in \mathbb{R}^H$ is a moving average (parametrised by γ) of the hidden unit activation probabilities over batches, and $d \in \mathbb{R}^H$ is the desired (target) probability of each hidden unit being active, with $0 < p \in d \ll 1$ (see Hinton (2010) for further details).

B.2 Deep Belief Networks

A DBN is a multi-layered stochastic generative model that mixes undirected and directed connections between the units. For a DBN with J layers of hidden units, the network contains undirected connections $W^{(J)}$ between the top two hidden unit layers $h^{(J)}$ and $h^{(J-1)}$. There are directed connections $W^{(j)}$ between the remaining hidden unit layers $h^{(j)}$ and $h^{(j-1)}$, $\forall 1 < j < J$, and directed connections $W^{(1)}$ between the first hidden layer $h^{(1)}$ and the visible layer x. Additionally, $\forall 1 \leq j \leq J$, hidden layer $h^{(j)}$ has biases $b^{(j)}$, and the visible layer x has biases c. The distribution $p(h^{(3)}, h^{(2)})$ over the top two layers of units forms an undirected RBM (associative memory), and the layers below form a directed sigmoid belief network.

An example of a 3-layered DBN (with 4 layers of units) is given in B.2). The full distribution of this DBN is:

$$p(x, h^{(1)}, h^{(2)}, h^{(3)}; \theta) = p(h^{(2)}, h^{(3)})p(h^{(1)}|h^{(2)})p(x|h^{(1)}) \tag{B.32}$$

for parameters $\theta = \{W^{(1)}, W^{(2)}, W^{(3)}, b^{(1)}, b^{(2)}, b^{(3)}, c\}$, where:

$$p(h^{(2)}, h^{(3)}) = exp\left({h^{(2)}}^\top W^{(3)} h^{(3)} + {b^{(2)}}^\top h^{(2)} + {b^{(3)}}^\top h^{(3)}\right)/Z \tag{B.33}$$

The conditional distribution of layer $h^{(1)}$ given the layer above, $h^{(2)}$, takes the following form:

$$\begin{aligned} p(h^{(1)}|h^{(2)}) &= \prod_j p(h_j^{(1)}|h^{(2)}) \\ p(h_j^{(1)} = 1|h^{(2)}) &= \sigma(b_j^{(1)} + W_j^{(2)} h^{(2)}) \end{aligned} \tag{B.34}$$

and the conditional distribution of layer x given the layer above, $h^{(1)}$, is:

$$\begin{aligned} p(x|h^{(1)}) &= \prod_i p(x_i|h^{(1)}) \\ p(x_i = 1|h^{(1)}) &= \sigma(c_i + W_i^{(1)} h^{(1)}) \end{aligned} \tag{B.35}$$

DBNs are typically trained in a layer-wise fashion, starting with an RBM at the first level. Starting with the input layer x and the first hidden layer $h^{(1)}$, we firstly train a RBM on the full set of data. Once the RBM has been trained, then by holding the weights $W^{(1)}$ and visible biases c fixed, we can now add another hidden layer $h^{(2)}$ and train the weights $W^{(2)}$ between $h^{(1)}$ and $h^{(2)}$, and biases $b^{(1)}$, as another RBM (with inputs propagated up to $h^{(1)}$ via feed-forward connections from x). A third (and arbitrarily many) number of RBM layer(s) can be added in this way, and it has been shown that a variational lower bound on $log\ p(x)$ can always be increased with each additional new layer (Hinton et al., 2006).

B.2.1 Sampling

To generate a visible layer sample $\tilde{x} \sim p(x, h^{(1)}, h^{(2)}, h^{(3)})$ from the DBN, we first perform Gibbs sampling at the top-level RBM to get a sample at layer $h^{(2)}$ from the prior distribution $p(h^{(3)}, h^{(2)})$, and then activate $h^{(1)}$ based on $h^{(2)}$. Based on this activation of $h^{(1)}$, we activate x. The result is a data vector $\tilde{x}$, which we refer to as a "generative sample" from the DBN's model.

It is possible to generate samples in an auto-associative-like manner by holding some subset of the lower-level units of the top RBM constant during Gibbs sampling. One option is to append a class label y to the input to the top level RBM during training B.3, and then present and hold constant the desired class label during Gibbs sampling to generate a sample associated with this class (Hinton et al., 2006), which is the approach taken in this work.

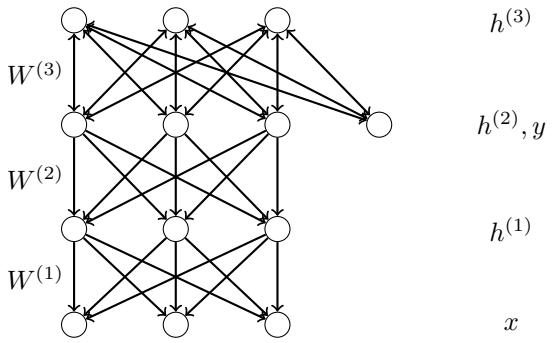

Figure B.3: Auto-associative sampling in a DBN with class label y

C. Appendix For Chapter 5

In this appendix we give further details on how the neuron numbers and connection probabilities were calculated in Chapter 5.

C.1 Regional Detail

Below we calculate a number of neurons for each region of our model which is relative to the proportion of actual neurons that this region has within the human brain, and connectivity probabilities between pairs of neurons both within and across regions, as well as detailing the types of neuron (in terms of electrophysiology and firing patterns) that are found in each region. Neuron numbers are calculated using region volume and neuron density estimates, which all come from non-clinical human data. Data on connectivity and neuron types are typically not available for the human brain, and thus our data mostly comes from animal subjects.

In order to calculate the number of neurons for each region of our model, we start by estimating the total number of neurons in each human brain region of interest for our model (estimated as a total of 1,313,268,048 neurons across all regions: details for each individual region are given in the sections below). Then, based on these estimates and assuming a total of 35000 neurons in our model, we calculated the relative proportion of neurons for each region in our model. Finally, these relative region neuron numbers were rounded to the nearest integer, giving a total of 34997 neurons in our model.

C.1.1 Neocortex

The neocortex is the phylogenetically newest part of the human brain. It has a well-defined six-layered structure that distinguishes it from the evolutionarily older allocortex (which has between three and five layers), with layers classified according to the types of neuron that occupy it and their afferent/efferent connectivity profiles (Kandel et al., 2012, p.345-346). Although relative differences in layer thickness amongst regions do exist, this laminar structure is largely homogeneous across the whole neocortex.

The AI and aMCC are defined as agranular cortex, which means that they lack a layer IV, whereas PI has both granular (i.e. with a full layer IV) and dysgranular (with a rudimentary layer IV) cortical segments (see Sections C.1.2, C.1.3 and C.1.9 for further details). Given their structure, the AI and aMCC are often considered to be transitory areas between neocortex and allocortex, however (for simplicity) we consider these regions (along with the PI, mPFC and mOFC) to be parts of the neocortex, and do not consider in detail differences in cell composition and connectivity amongst these regions.

C.1.1.1 Neocortical neuron types

The neocortex comprises a large number of different types of neuron, roughly 80% of which are excitatory and 20% inhibitory (Harris and Shepherd, 2015; Peyrache et al., 2012). The most common type of excitatory neuron in the neocortex are pyramidal cells, which use glutamate as their neurotransmitter. They occupy layers II-VI and make up 60-70% of all neocortical neurons (Connors and Gutnick, 1990). The majority of these pyramidal cells (70% in the rat study in Dégenètais et al. (2002)) exhibit RS firing characteristics; in response to steady input current they fire single spikes at a rate that initially slows (adapts) before stabilising.

In the mouse, the majority (approximately 40%) of inhibitory neocortical neurons are known to be Parvalbumin-expressing basket or chandelier cells, with similar proportions believed to apply to humans (Rudy et al., 2011). These cells use GABA as their neurotransmitter and, as with pyramidal cells, are present in layers II-VI (Markram et al., 2004). They exhibit FS dynamics (Rudy et al., 2011), with a frequency of firing in response to stable input current that is non-adaptive and relatively high frequency compared to pyramidal cells. Thus, for simplicity, we assume that in all neocortical regions, 80% of neurons are excitatory RS cells, and the remaining 20% are inhibitory FS cells, as per these majority types.

C.1.1.2 Neocortical connectivity

We now consider connectivity within and between neocortical regions. For excitatory neocortical neurons targeting neocortical regions, we assume that 50% target distant (i.e. other region) neurons, and the remaining 50% target local (i.e. same region) neurons (the actual proportions are unknown). For distant neocortical targets of excitatory neocortical neurons, we assume that 90% target distant excitatory neocortical neurons, and the remaining 10% target distant inhibitory neocortical neurons (Melchitzky et al., 2001). Local targets of excitatory neocortical neurons are split evenly between excitatory and inhibitory neurons (Lewis et al., 2002).

For inhibitory neocortical neurons, we assume that 100% of outgoing connections target local neurons, with 95% of these targeting local excitatory neurons and 5% targeting local

inhibitory neurons (since interneurons are rarely and sparsely connected (Markram et al., 2004)). Packer and Yuste (2011) found a local connection probability (for intersomatic distances less than $200\mu m$) for Parvalbumin basket cells to pyramidal cells of 62% (averaged across regions), with dense local connectivity that decreased as a function of intersomatic distance (such that the probability of connection became zero for distances greater than $450\mu m$). Similar data for the BLA suggests a connection probability of 50% for inhibitory interneurons to local excitatory pyramidal-like neurons (within $120\mu m$), which we roughly scaled to 10% to take dense local connectivity into account (see Section C.1.4). Based on this, we use a slightly higher connection probability of 12.5% for inhibitory neocortical neurons to local excitatory neurons (although we note that this is a very crude estimate).

These connection proportions and probabilities, along with the total number of neurons within each neocortical region within our model (which gives the total number of possible connections), allow us to calculate connection probabilities within and between all neocortical regions. These are detailed below for each specific neocortical region, along with details of connectivity between between neocortical and non-neocortical regions.

C.1.1.3 Neocortical connectivity example: Anterior Insular

As an example of the method we used to calculate connection probabilities for pairs of neurons with a neocortical source, we use the AI which we have considered to be part of the neocortex (see Section C.1.2 for further discussion). We calculated 7765 AI neurons for our model, of which 80% (6212) are regular-spiking excitatory neurons, and 20% (1553) are fast-spiking inhibitory neurons. The connection probability of AI inhibitory to AI excitatory neurons is 12.5%: since the total number of possible connections is $6212 * 1553 = 9647236$ this gives $9647236 * 0.125 = 1205904.5$ total connections. For AI inhibitory to AI inhibitory, the total number of possible connections is $1553^2 = 2411809$, and the number of connections is thus $0.05 * (1205904.5 * 100/95) = 63468.657894737$, giving a connection probability of $63468.657894737/2411809 = 2.6315789474\%$ (rounded to 2.632% in the model). The average number of outgoing connections per inhibitory AI neuron is $(9647236 + 63468.657894737)/1553 = 817.3684210526$.

AI excitatory neurons project to the AI (which has a total of 6212 excitatory and 1553 inhibitory neurons), aMCC (1329 excitatory and 332 inhibitory), mPFC (9547 excitatory and 2393 inhibitory) and PI (3106 excitatory and 777 inhibitory), along with excitatory (negatively valanced) neurons in the BLMA$\ominus$ (total 76 neurons). We assume that the average number of outgoing connections per excitatory AI neuron is 817.3684210526 (i.e. the same as for inhibitory AI neurons, above), giving a total number of outgoing connections from AI excitatory neurons of $817.3684210526 * 6212 = 5077492.631578751$.

The total number of possible connections from AI excitatory neurons to BLMA$\ominus$ neurons is $6212 * 76 = 472112$ which, given connection probability of 10%, gives a total of 47211.2

connections from AI excitatory to BLMA$\ominus$ neurons. Thus, the total number of connections from AI excitatory neurons to the other neocortical regions (AI, aMCC, mPFC and PI) is $5077492.631578751 - 47211.2 = 5030281.431578751$, with 50% targeting local neurons in the AI ($5030281.431578751 * 0.5 = 2515140.71578947$) and the remaining 50% targeting distant neurons in other neocortical areas (i.e. the aMCC, mPFC or PI). We assume that the number of connections to each of the aMCC, mPFC and PI is proportional to the total number of neurons in that region (giving 238572.824448993 connections from AI excitatory to aMCC, 1718844.66597297 from AI excitatory to mPFC, and 557723.225367513 from AI excitatory to PI).

The total number of possible connections from AI excitatory to AI excitatory neurons is $6212^2 = 38588944$, and the total number of possible connections from AI excitatory neurons to AI inhibitory neurons is $6212 * 1553 = 9647236$. The total number of connections from AI excitatory to other local AI excitatory neurons is $2515140.71578947 * 0.5 = 1257570.357894735$, and the total number of connections from AI excitatory to local AI inhibitory neurons is $2515140.71578947 * 0.5 = 1257570.357894735$, corresponding to connection probabilities of 3.259% and 13.036% respectively.

The total number of possible connections from AI excitatory neurons to aMCC excitatory neurons is $6212 * 1329 = 8255748$, and the total number of possible connections from AI excitatory neurons to aMCC inhibitory neurons is $6212 * 332 = 2062384$. The total number of connections from AI excitatory to aMCC excitatory neurons is $238572.824448993 * 0.9 = 214715.542004094$, and the total number of connections from AI excitatory to aMCC inhibitory neurons is $238572.824448993 * 0.1 = 23857.282444899$, corresponding to connection probabilities of 2.601% and 1.157% respectively.

The total number of possible connections from AI excitatory neurons to mPFC excitatory neurons is $6212 * 9574 = 59473688$, and the total number of possible connections from AI excitatory neurons to mPFC inhibitory neurons is $6212 * 2393 = 14865316$. The total number of connections from AI excitatory to mPFC excitatory neurons is $1718844.66597297 * 0.9 = 1546960.199375673$, and the total number of connections from AI excitatory to mPFC inhibitory neurons is $1718844.66597297 * 0.1 = 171884.466597297$, corresponding to connection probabilities of 2.601% and 1.156% respectively.

The total number of possible connections from AI excitatory neurons to PI excitatory neurons is $6212 * 3106 = 19294472$, and the total number of possible connections from AI excitatory neurons to PI inhibitory neurons is $6212 * 777 = 4826724$. The total number of connections from AI excitatory to PI excitatory neurons is $557723.225367513 * 0.9 = 501950.902830762$, and the total number of connections from AI excitatory to PI inhibitory neurons is $557723.225367513 * 0.1 = 55772.322536751$, corresponding to connection probabilities of 2.602% and 1.155% respectively. The final connection probabilities are all rounded to 3 decimal places.

C.1.2 Anterior Insular

C.1.2.1 Number of neurons

An MRI-based estimate for gray matter volume of the human anterior insular for both schizophrenic and control subjects is given in Makris et al. (2006): we take the control subject data, which estimated a mean bilateral AI volume of $9.6cm^3 = 9600mm^3$ across 40 male and female subjects aged between 23 and 66 (mean age 40). Our neuron density estimate comes from BA13 (which is an area in the adjacent PI): Semendeferi et al. (1998) estimated a density of $30351/mm^3$ in a postmortem biopsy of a single 75 year old male subject who did not die of a neurological disease. This gives a total estimate for the number of neurons in the human AI of $9600 * 30351 = 291,369,600$. This is 22.1866% of the total number of biological neurons in our regions of interest, and thus the corresponding number of AI neurons in our model is set to 7765.

C.1.2.2 Neuron types

The AI is agranular and typically considered transitory cortex (Öngür et al., 2003; Wallis, 2012, Fig. 1), however for simplicity we assume standard neocortical cell composition and connectivity. For total number of AI neurons in our model 7765, we set 6212 to be excitatory RS and 1553 inhibitory FS, according to standard neocortical proportions described above in Section C.1.1.1.

C.1.2.3 Connection probabilities

In our model, in addition to recurrent AI connections, excitatory AI neurons project to PI, aMCC, mPFC and BLMA⊖. AI excitatory neurons project locally to AI excitatory and inhibitory neurons with probabilities 3.259% and 13.036%, respectively. For the distant neocortical projections, we connect AI excitatory neurons to excitatory and inhibitory PI neurons with probabilities 2.602% and 1.155%; to excitatory and inhibitory aMCC neurons with probabilities 2.601% and 1.157%; and to excitatory and inhibitory mPFC neurons with probabilities 2.601% and 1.156%. We additionally connect AI excitatory neurons to subcortical BLMA⊖ neurons with probability 10% (used as a default value since actual connection probability is unknown).

Inhibitory AI neurons are assumed to project only to local excitatory and inhibitory AI neurons: we take connection probabilities of 12.5% and 2.632% to local excitatory and inhibitory neurons, respectively.

The full calculations underlying these connection probabilities is given in the example calculation in Section C.1.1.3 (the same method was used to calculate connection probabilities for neurons projecting from other neocortical regions).

C.1.3 Anterior Midcingulate Cortex

C.1.3.1 Number of neurons

We use the aMCC gray matter volume in Kitayama et al. (2006) (labelled "whole", and comprising BA24 and 32): they used MRI to estimate a total bilateral volume of $2710.9mm^3$ based on a control sample of 13 men and women. We assume a neuronal density of $23000/mm^3$, based on the findings in Höistad et al. (2013) who present estimates for BA24a',b',c' following a postmortem biopsy on 13 male control subjects aged 25 to 65 (mean age 51.9). This gives a total estimate of $62,350,700$ neurons in the bilateral human aMCC which, at 4.746% of the total, gives 1661 neurons in the model.

C.1.3.2 Neuron types

The aMCC, as a subpart of the ACC, is agranular and typically considered to be transitionary cortex (Öngür et al., 2003; Wallis, 2012, Fig. 1). However, as for the AI, we assume standard neocortical cell composition and connectivity. For total number of aMCC neurons in our model 1661, we set 1329 to be excitatory RS and 332 inhibitory FS, according to standard neocortical proportions described above in Section C.1.1.1.

C.1.3.3 Connection probabilities

In addition to recurrent aMCC connections, excitatory aMCC neurons project to PI, AI and mPFC (which are all neocortical targets, see Section C.1.1.2). In our model, aMCC excitatory neurons project locally to aMCC excitatory and inhibitory neurons with probabilities 3.289% and 13.168%, respectively. For the distant neocortical projections, we connect aMCC excitatory neurons to excitatory and inhibitory PI, AI and mPFC neurons with probabilities 0.417% and 0.185% (for all three regions).

Inhibitory aMCC neurons are assumed to project only to local excitatory and inhibitory aMCC neurons: we use connection probabilities of 12.5% and 2.634% to local excitatory and inhibitory neurons, respectively.

C.1.4 Amygdala

C.1.4.1 Number of neurons

García-Amado and Prensa (2012) estimated a total of $5.02 * 10^6$ neurons in the basolateral (basal) nucleus, and $1.73 * 10^6$ in the basomedial (accessory basal) nucleus, based on postmortem biopsy of 7 control male and female subjects aged from 20 to 75 (mean age 52.1). This gives a total of $6,750,000$ neurons in our regions of interest within the human amygdala. We assume that 85% of these are excitatory and the remaining 15% inhibitory (see Section

C.1.4.2), and that there are the same number of positively and negatively valanced excitatory neurons, giving an estimate of $2,868,750$ for both positively and negatively valanced excitatory neurons in human amygdala. Since we only consider a population of BLMA inhibitory neurons that inhibit only BLMA$\ominus$, we use an estimate of $506,250$ inhibitory BLMA neurons. Thus, we have estimated that positively valanced excitatory, negatively valanced excitatory, and inhibitory neurons correspond to 0.218%, 0.218% and 0.0385% of the total number of neurons in our regions in interest, respectively. This results in 76 BLMA$\oplus$, 76 BLMA$\ominus$ and 13 BLMAi neurons in our model.

C.1.4.2 Neuron types

We assume that 85% of neurons in the BLMA are glutamatergic excitatory pyramidal-like neurons with regular-spiking characteristics, and the remaining 15% of BLMA neurons are fast-spiking GABAergic inhibitory interneurons (McDonald, 1992; Sah et al., 2003). In our model, pyramidal neurons constitute those in BLMA$\oplus$ and BLMA$\ominus$, whilst cells in BLMAi are interneurons.

C.1.4.3 Connection probabilities

Pyramidal neurons in the basolateral complex of the amygdala (lateral, basolateral and accessory basal nuclei) are believed to have substantial local projections, mostly to interneurons (IN) but also local pyramidal neurons (PN), whilst interneurons innervate both local pyramidal cells and other local interneurons (Sah et al., 2003). Vlachos et al. (2011), in their model of the basolateral nuclei of the amygdala, connect inhibitory PN-PN with probability 0.01, PN-IN with probability 0.15, IN-PN with probability 0.15, and IN-IN with probability 0.1, although this is not justified with any neuroscientific data.

Woodruff and Sah (2007) studied connectivity (within and between Parvalbumin IN, which account for 50% of BLA interneurons; and PN neurons) in the BLA of the mouse. They found that 47% of the Parvalbumin IN were fast-spiking. In particular, they made recordings from 160 IN-PIN neuron pairs, with neurons within somatic distance $120\mu m$ of each other. 99 of these pairs were synaptically connected, with 34% (55 of 160) forming a unidirectional IN-PN inhibitory connection; 12% (19 of 160) having an unreciprocated PN-IN excitatory connection; and 16% (25 of 160) reciprocally connected. The remaining 38% (61 of 160) of IN–PN pairs were not connected in any way. This gives connection probabilities of IN$\rightarrow$IN: 26%, IN$\rightarrow$PN: 50% and PN$\rightarrow$IN: 27.5%. Since they only considered pairs within somatic distance $120\mu m$, we scale these probabilities down slightly (while roughly keeping their relative proportions): we connect BLMAi to BLMA$\ominus$ with probability 10%; BLMA$\ominus$ to BLMA$\ominus$ and BLMA$\oplus$ to BLMA$\oplus$ with probability 5%; and BLMAi to BLMAi also with probability 5%.

In order to calculate the remaining connection probabilities, we assume that the BLMA follows neocortical proportions for long-range target proportions, and proportions of excitatory-inhibitory neocortical targets (see Section C.1.1.2). The BLMA$\ominus$ group consists of excitatory cells which project to local excitatory cells (BLMA$\ominus$), local inhibitory cells (BLMAi), and excitatory and inhibitory cells of the AI. Connection probability from BLMA$\ominus$ to BLMAi is given at 5% (see above). Since the total number of possible connections from BLMA$\ominus$ to BLMAi is $76 * 13 = 988$, this gives a total of $988 * 0.05 = 49.4$. Following neocortical target proportions, we assume that there will also be 49.4 total connections from BLMA$\ominus$ to BLMA$\ominus$: for total number of possible connection $76^2 = 5776$, this gives connection probability $(49.4/5776) * 100 = 0.855\%$. Now, assuming the same local-distant target proportions as for excitatory neocortex pyramidal cells, we can assume a total of $49.4 * 2 = 98.8$ connections to AI cells. Again assuming the same target proportions as in the neocortex (i.e. of distant neocortex targets, 10% are inhibitory neocortical cells and 90% distant neocortical excitatory cells), this gives a connection probability of 0.008% to inhibitory AI cells, and 0.019% to excitatory AI cells.

The BLMA$\oplus$ group consists of excitatory cells which project to local excitatory cells (BLMA$\oplus$), local inhibitory cells (BLMAi), and inhibitory cells in the VP and NAc. As for the BLMA$\ominus$, we use connection probability of 5% for BLMA$\oplus$ to BLMAi, and use this (along with neocortical local target proportions) to calculate a probability of connection 0.855% for BLMA$\oplus$ to BLMA$\oplus$. Again, assuming neocortical local-distant target proportions (and assuming that total connections from BLMA$\oplus$ to VP and NAc are proportional to the number of targets in each area), we calculate connection probability of 0.674% from BLMA$\oplus$ to both VP and NAc.

C.1.5 Medial Prefrontal Cortex

C.1.5.1 Number of neurons

In our model, we are concerned specifically with a subpart of the mPFC that preferentially activates for other referential processing; activation that is presumed to be required in order to stimulate the mPFC-mPOA-VTA-NAc-VP, mPFC-BLMA$\oplus$-VP and mPFC-BLMA$\oplus$-NAc-VP caregiving pathways. Murray et al. (2012) conducted a meta-analysis of studies concerning self- and other-referential processing, and found evidence for a ventral-dorsal gradient for such processing in the mPFC. Although other vs self comparisons did not reveal any mPFC clusters exhibiting preferential activation, self vs control, self vs other and self vs public other comparisons all revealed preferential activation in the right vmPFC. Meanwhile, other vs control comparisons revealed preferential activations in three non-overlapping clusters (one each in the left and right vmPFC, and one in the left dmPFC), whilst close-other vs control showed increased activation in left vmPFC. Based on this

analysis, we use a volume estimate of $4528mm^3$ which comes from the sum of the volumes of the three non-overlapping clusters found in the other vs control (left dmPFC and left vmPFC) and self vs control (right vmPFC) contrasts. We tentatively suggest that the area encoding an optimal close-other representation (with strongest connectivity to our caregiving pathways) would be located mainly in the left vmPFC (which was preferentially activated in both the other and close-other contrasts with the control condition).

Rabinowicz et al. (1999) report neural densities in two subparts of BA10 (a part of the vmPFC) for both left and right hemispheres, in a postmortem biopsy study with 11 male and female control subjects aged 12 to 24 (mean age 16.7). We average these eight results to give a density of $2380/0.024mm^3$ (i.e. $99167/mm^3$). This, along with our volume estimate, gives a total of $449,028,176$ neurons in the human mPFC, which is 34.19% of the total in our regions of interest, corresponding to 11967 neurons in our model.

C.1.5.2 Neuron types

For total number of mPFC neurons in our model 11967, we set 9574 to be excitatory RS and 2393 inhibitory FS, according to standard neocortical proportions described above in Section C.1.1.1.

C.1.5.3 Connection probabilities

mPFC excitatory neurons have local projections to mPFC excitatory and inhibitory neurons, plus long range projections to neocortical regions aMCC and AI and subcortical regions BLMA$\oplus$ and mPOA.

Numan also considered projections from mPFC to NAc, although the nature of these projections was left undefined in his model. In the rat study in Papp et al. (2012), it is claimed that the mPFC does not directly innervate MSN NAc neurons that directly project to the VP, whereas the BLA does. Instead, it is proposed that the mPFC might innervate other MSN in the NAc (i.e. ones that do not directly project to the VP), or alternatively NAc interneurons. Thus, since the nature of the connectivity was undefined in Numan's original model, and since the evidence above seems to suggest more direct connectivity along the BLMA-NAc-VP pathway compared to the mPFC-NAc-VP pathway, we do not consider direct projections from the mPFC to the NAc in our initial model.

We connect mPFC excitatory neurons to local mPFC excitatory and inhibitory neurons with probabilities 3.315% and 13.261% respectively; for long-range neocortical targets, we connect mPFC excitatory neurons to excitatory and inhibitory neurons in the aMCC with probabilities 7.459% and 3.318% respectively, and to excitatory and inhibitory neurons in the AI with probabilities 7.460% and 3.316% (all according to standard neocortical connectivity patterns defined in Section C.1.1.2). Finally, mPFC excitatory neurons are connected with subcortical targets in the mPOA and BLMA$\oplus$. Pre- and post-synaptic neurons for these

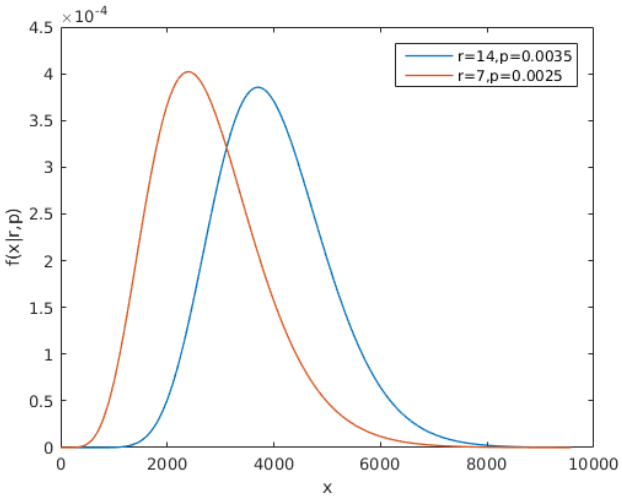

Figure C.1: Negative-binomial distributions used to model close- and distant-other representations in the mPFC. The close-other distribution ($r = 7$, $p = 0.0025$) is used to connect the mPFC to the mPOA and BLMA, and to sample mPFC targets from the other-pain network encoding close-other representations. The distant-other distribution ($r = 14$, $p = 0.0035$) is used to sample mPFC targets from the other-pain network encoding distant-other representations.

projections are connected according to a negative-binomial distribution, with parameters $r = 7$ (number of successes) and $p = 0.0025$ (probability of success in a single trial) (Figs. C.1, C.2, C.3). This is a discrete distribution intended to model a ventral-dorsal gradient for self-other representations in the mPFC (i.e. close-other representations, located more ventrally, have strongest connectivity, whilst neurons encoding self and other representations have sparser connectivity). Neurons are connected in this manner such that the overall connection probabilities for each group pairing is 10% (we use this default value in the absence of connectivity data).

Inhibitory mPFC neurons are assumed to project only to local excitatory and inhibitory mPFC neurons: we use connection probabilities of 12.5% and 2.632% to local excitatory and inhibitory neurons, respectively.

C.1.6 Medial Preoptic Area

C.1.6.1 Number of neurons

We use as an estimate of the volume of the mPOA volume estimates for the anterior-superior hypothalamus (which includes the mPOA along with the diagonal band of broca, sexually dimorphic nucleus of the preoptic area, and the paraventricular nucleus): Makris et al. (2013) reports left and right anterior-superior hypothalamus volumes in an MRI study of

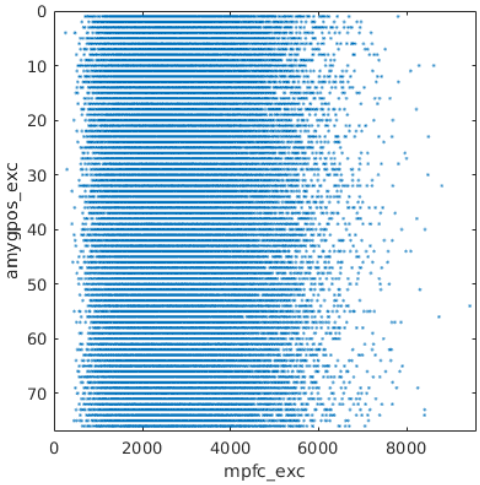

Figure C.2: Connectivity between mPFC and positively valent BLMA neurons: mPFC neurons close to id 0 (encoding self) and 9574 (encoding other) are connected to positive BLMA neurons with more sparsity than remaining neurons (encoding close-other).

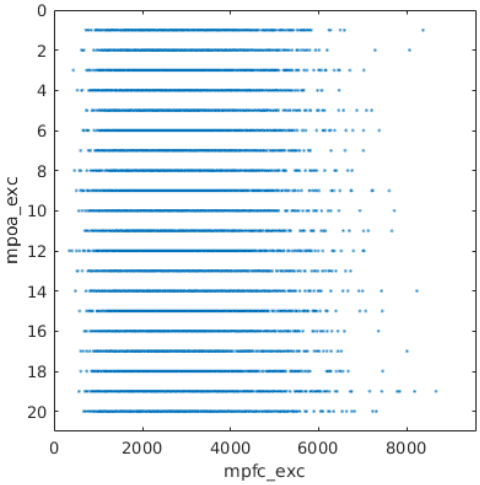

Figure C.3: Connectivity between mPFC and mPOA neurons: mPFC neurons close to id 0 (encoding self) and 9574 (encoding other) are connected to positive BLMA neurons with more sparsity than remaining neurons (encoding close-other).

26 men and 18 women (control, average age 40), giving total bilateral volume $57.5mm^3$ (averaged over hemispheres and genders). We use neural density $13144/mm^3$, which is an average of the densities of the four interstitial nuclei of the anterior hypothalamus (INAH1-4) reported in Byne et al. (2000) (averaged over nuclei and genders). This gives as an estimate for the total number of biological neurons $755,780$, which corresponds to 0.0575% of the total neurons in our regions of interest, thus corresponding to 20 neurons in our model.

C.1.6.2 Neuron types

Very little is known about the types of neuron in the human mPOA, and so we rely here primarily on rodent data. The rodent mPOA contains a number of different types of neuron, including those expressing GABA, glutamate and galanin. It has been estimated that $50-95\%$ of all mPOA neurons are GABAergic (Lonstein and De Vries, 2000; Tabarean et al., 2005), although precise proportions for other neuron types are (to the best of our knowledge) currently unknown.

Despite the believed predominance of GABAergic neurons throughout the mPOA, we only consider glutamatergic neurons here for the following reasons. Firstly, glutamatergic neurons are believed to form the majority of neurons in the central mPOA, an area that is thought to be particularly important in maternal behaviour (Tsuneoka et al., 2013) (we do also note, however, that this study reports that only 6% of central mPOA cells activated during maternal behaviour were glutamatergic whereas 75% were GABAergic, and that 2.5% of all glutamatergic cells in this area were activated whereas 17.5% of GABAergic cells were). Secondly, Numan has argued that GABAergic mPOA neurons, although found to be active during maternal behaviour, are crucially uninvolved in the mPOA-VTA-NAc pathway in our model [p.188](Numan and Insel, 2003). In particular, he argues that GABAergic mPOA neurons serve primarily to inhibit both incoming inhibitory synapses, and areas outside of this pathway, during maternal behaviour, and that this pathway instead primarily involves glutamatergic projections from mPOA to VTA. A recent study found evidence for GABAergic projections from mPOA to VTA in addition to the glutamatergic projections described by Numan although, consistent with Numan, it was reported that a majority of VTA projecting neurons in the mPOA were glutamatergic (36.55%) rather than GABAergic (18.26%) (Kalló et al., 2015).

Thus, because of the relatively small volume of the mPOA; the unknown overall proportion of glutamatergic neurons; the unknown electrophysiological properties of these excitatory mPOA neurons; and because it is believed that mPOA-VTA projections are primarily glutamatergic during maternal behaviour, we assume that all 20 mPOA neurons are excitatory RS neurons in our model.

C.1.6.3 Connection probabilities

Evidence suggests that glutamatergic neurons within the mPOA are self-innervate (Kocsis et al., 2003). Thus, in addition to projections from mPOA to the VTA, we also consider local mPOA-mPOA projections. Precise connection probabilities are unknown in both cases, and so we default to 10% for local projections, and 50% for projections to the VTA (which is required in order to achieve the desired network dynamics, due to the very low relative number of neurons in the VTA in our model).

C.1.7 Nucleus Accumbens

C.1.7.1 Number of neurons

We use the estimate of $7,285,654$ for total number of neurons in the human NAc (control) given in Wegiel et al. (2014), which is based on a postmortem biopsy of 14 individuals. We scale this value by 0.95, which is the believed proportion of MSN in the rodent NAc (Shi and Rayport, 1994). This gives a total of 6921371 MSN NAc neurons, which accounts for 0.527034143% of the total number of biological neurons in our regions of interest, corresponding to 184 NAc neurons in our model.

C.1.7.2 Neuron types

As described above, in the rodent NAc, 95% of neurons are believed to be GABAergic MSN neurons, with the remaining neurons mainly composed of local-circuit GABAergic and cholinergic interneurons (Shi and Rayport, 1994; Robison and Nestler, 2011). We only consider the predominant MSN cells in our initial model, which receive dopaminergic inputs from the VTA along with glutamatergic input from the BLMA (see Fig. 2 in Russo and Nestler (2013) for a good diagram of major NAc MSN input sources).

C.1.7.3 Connection probabilities

Tecuapetla et al. (2007) examined 106 MSN-MSN cell pairs (all within distance $100\mu m$) in Wistar rats: they report that 13% of MSN cells received input from nearby MSN cells. For the BLMA it was found that IN-IN neurons connect with a probability of 26% (within 120 μm), which we scaled to 5% for our model: assuming again denser local connectivity, this suggests that a connection probability of about half this (2.5%) is roughly reasonable for MSN-MSN pairs in the NAc. We use a default connection probability of 10% for NAc to VP.

C.1.8 Medial Orbitofrontal Cortex

C.1.8.1 Number of neurons

We use as an estimate for the volume of mOFC $5890mm^3$, which comes from an MRI study of 34 male and female subjects (control, mean age 37.03) (Lacerda et al., 2004). Our neural density estimate of $59500/mm^3$ comes from Rajkowska et al. (1999) (averaged across rostral and caudal estimates for all layers), which was a postmortem biopsy study of 12 men and women (control, mean age 43.3). This gives an estimate of $350,455,000$ for the total number of neurons in the human mOFC, which is 26.686% of the total number of neurons in our regions of interest, corresponding to 9340 neurons in the model.

C.1.8.2 Neuron types

Although the OFC does contain agranular and dysgranular cortical sections, medial regions are exclusively granular (Henssen et al., 2016). For total number of mOFC neurons in our model 9340, we set 7472 to be excitatory RS and 1868 inhibitory FS, according to standard neocortical proportions described above in Section C.1.1.1.

C.1.8.3 Connection probabilities

mOFC excitatory neurons have local projections to mOFC excitatory and inhibitory neurons, plus long range projections to the neocortical region AI and subcortical regions BLMA$\oplus$ and BLMAi. We connect mOFC excitatory neurons to local mOFC excitatory and inhibitory neurons with probabilities 3.319% and 13.277% respectively; and mOFC excitatory neurons project to long-range neocortical targets in AI with probabilities 7.058% (for excitatory targets) and 3.137% (for inhibitory targets), according to standard neocortical connectivity patterns defined in Section C.1.1.2. Connection probabilities between mOFC and BLMA are unknown: we thus connect mOFC excitatory neurons to both BLMA$\oplus$ and BLMAi neurons with a default probability of 10%.

Inhibitory mOFC neurons are assumed to project only to local excitatory and inhibitory mOFC neurons: we use connection probabilities of 12.5% and 2.632% to these local excitatory and inhibitory neurons, respectively.

C.1.9 Posterior Insular

C.1.9.1 Number of neurons

We use the bilateral PI gray matter volume of $4800mm^3$ reported in Makris et al. (2006) (which is based an average across 40 male and female control subjects aged between 23 and 66, with mean age 40, measured by MRI). Our neuron density estimate comes from BA13 (a part of the PI): Semendeferi et al. (1998) estimated a density of $30351/mm^3$ in a

postmortem biopsy of a single 75 year old male subject who did not die of a neurological disease. This gives a total estimate of 145, 684, 800 neurons in the bilateral human PI which, at 11.093% of the total, results in 3883 neurons in our model.

C.1.9.2 Neuron types

In contrast to the agranular AI, the PI contains granular and dysgranular sections of cortex (Kurth et al., 2010), and is typically considered part of the neocortex. For total number of PI neurons in our model 3883, we set 3106 to be excitatory RS and 777 inhibitory FS, according to standard neocortical proportions described above in Section C.1.1.1.

C.1.9.3 Connection probabilities

In addition to recurrent PI connections, excitatory PI neurons project to AI and aMCC (which are both neocortical targets, see Section C.1.1.2). Posterior parts of the human Insular do have structural connectivity with the ACC region, particularly to posterior midcingulate cortex (which is adjacent to aMCC) (Ghaziri et al., 2015). In our initial model, for simplicity, we only consider anterior midcingulate regions of the ACC and assume structural connectivity (with a more fully realised and compartmentalised model of the ACC suggested as work for future models).

In our model, PI excitatory neurons project locally to PI excitatory and inhibitory neurons with probabilities 3.289% and 13.149%, respectively. For the distant neocortical projections, we connect PI excitatory neurons to excitatory and inhibitory AI excitatory neurons with probability 2.439% and to inhibitory neurons with probability 1.084%; and to aMCC excitatory/inhibitory neurons with probabilities 2.438% and 1.085.

Inhibitory PI neurons are assumed to project only to local excitatory and inhibitory PI neurons: we use connection probabilities of 12.5% and 2.630% to local excitatory and inhibitory neurons, respectively.

C.1.10 Ventral Tegmental Area

C.1.10.1 Number of neurons

The total number of DA neurons in the human brain is estimated to be 450, 000 (German et al., 1983), of which 15% are thought to be in the VTA (Düzel et al., 2009), giving 67, 500 VTA DA neurons in total. Since 55% of all VTA neurons are thought to be DA neurons (Margolis et al., 2006), this gives a total of 122, 727 neurons in the human VTA, which is 0.00829% of all neurons in our regions of interest, resulting in 3 neurons in our model's VTA.

C.1.10.2 Neuron types

The human VTA consists of approximately 55% DA neurons (Margolis et al., 2006), with the remaining neurons mainly GABAergic (approximately 35% in the rat) and glutamatergic (approximately 3% in the rat) (Nair-Roberts et al., 2008). Numan identified DA projections from VTA to NAc as crucial in driving caregiving behaviour. It is known that GABAergic neurons in the VTA project to cholinergic neurons in the NAc, and that projections from MSN in the NAc target VTA GABAergic interneurons which regulate VTA DA neuron activity (Creed et al., 2014). However, for simplicity, and given the relatively small number of neurons in the region, we only consider DA neurons in the VTA in our initial model.

Above we estimated $67,500$ total DA neurons in the human VTA, which is 0.005% of all neurons in our regions of interest. Using only this estimate, then for $35,000$ total model neurons, we would have 2 VTA DA neurons in our model: we instead assume that all 3 VTA model neurons are DA neurons, in order to (slightly) ease our simulatory efforts. DA neurons in the VTA are known to burst in response to stimuli predicting primary reward (Overton and Clark, 1997): we use intrinsically bursting neurons as a very basic model of this.

C.1.10.3 Connection probabilities

The VTA is believed to contain DA receptors whose activation inhibits the firing activity of dopamine neurons, thus carrying out an auto-inhibitory function (Al-Hasani et al., 2011; Olijslagers et al., 2006). However, local connection density is largely unknown, and we thus only consider long-range projections from VTA DA neurons to NAc MSN neurons.

In Bolam and Pissadaki (2012) it is estimated that, in the rat, each VTA DA neuron synapses onto between 12351 and 29644 MSN neurons in the ventral striatum. The total number of MSN cells in the striatum is given as $2,780,000$, and the ventral striatum is taken to be one fifth of the volume of the total striatum. Based on this, we can crudely estimate that $2780000/5 = 556000$ MSN cells are in the ventral striatum, which means that each DA VTA neuron synapses onto between 2.22% and 5.33% of ventral striatal MSN neurons.

McClure et al. (2004) reports a control ("unhandled rat") NAc volume of $6mm^3$. If we re-perform the calculations in Bolam and Pissadaki (2012) with this more accurate volume of $6mm^3$, we get the volume of NAc per VTA DA neuron of $6/20000 = 0.0003mm^3$. Then, the ratio of NAc versus dorsal striatum volume per DA neuron is $0.0003/0.001658 = 0.180940893$, and the number of synapses formed by one DA neuron in the NAc is $(0.181 * 102,165) = 18492 to (0.181 * 245,103) = 44364$. The total number of MSN neurons in the striatum (one hemisphere) is given as $2,780,000$, and the total volume of striatum is given as 19.9. Thus, we crudely estimate number of MSN cells in NAc as $(2 * 2780000) * (6/19.9) = 1676382$. This means that each VTA DA neuron synapses onto approximately $(18492 * 100/1676382)\% = 1.1\%$ to $(44364 * 100/1676382)\% = 2.6\%$ of NAc MSN neurons in the rat

brain. In our model, we take the midpoint of this estimate, which is 1.85%.

C.1.11 Ventral Pallidum

C.1.11.1 Number of neurons

Pakkenberg (1990) estimate 350,000 total neurons in the human VP, based on a postmortem control sample of 10 men and women aged 41 to 86 (mean age 60.4). This is 0.0267% of the total number of neurons in our regions of interest, corresponding to 9 neurons in our model.

C.1.11.2 Neuron types

In the rat VP, it is believed that approximately 80% of neurons are GABAergic, and smaller populations of cholinergic (approximately 20%) and glutamatergic neurons (Root et al., 2015; Kupchik and Kalivas, 2013). NAc-mediated GABAergic VP projections to the mesencephalic locomotor region have long been proposed to be involved in the translation of limbic motivation signals into motor output (Jordan, 1998; Mogenson, 1987; Brudzynski et al., 1993). Both increasing and decreasing levels of activation in the VP have been associated with goal-directed behaviour (Numan, 2014, p.25-26) such that there are likely to be two distinct pathways, with inhibition of NAc shell acting to disinhibit (ventromedial) VP in the first, and stimulation of NAc core inhibiting (dorsolateral) VP in the second (Root, 2013). Since we are concerned primarily with the NAc shell-VP pathway here, we consider an increased firing rate to correspond to increased motivation for caregiving behaviour. We only consider GABAergic neurons, which constitute approximately 80% of all VP neurons (Gritti et al., 1993).

C.1.11.3 Connection probabilities

As detailed above, increased activity in the (ventromedial) VP has been shown to correlate with goal-directed behaviour. Another perspective for VP activity is put forward by the experiments on primates in Tachibana and Hikosaka (2012), who found that VP neurons encoded the expected value associated with upcoming actions, such that the VP provides expected reward signals that facilitate or inhibit motor output. In our model, we can thus consider activity in the VP to represent an increase in expected reward associated with the initiation of caregiving behaviour (which in turn leads to an increase in caregiving behavioural output which drives application of the bonding protocols), without explicitly considering the nature of the onward projections that facilitate this behavioural output in any further detail. We additionally do not consider here projections back to input regions such as the VTA and NAc (Smith et al., 2009).

Acronyms

AAI Adult Attachment Interview. 12, 13, 16, 17, 74, 75

ACC Anterior Cingulate Cortex. 18, 23, 25, 79, 96, 97, 107, 115, 116, 122, 126, 190, 199

ACE Affective Communication Errors. 11, 12, 16, 57, 58, 61–71, 129, 161

AI Anterior Insular. 18, 23, 79, 96–100, 105, 106, 108–111, 115, 118, 120–124, 126, 127, 186–190, 192, 193, 198, 199

AMBIANCE Atypical Maternal Behaviour Instrument for Assessment and Classification. 11, 12, 57, 58, 70, 71, 129, 161

aMCC Anterior Midcingulate Cortex. 96–100, 106–111, 115, 120–123, 126, 186–190, 193, 199

ANS Autonomic Nervous System. 18, 20, 22

BA Brodmann Area. 98, 105–107, 189, 190, 193, 198

BLA Basolateral Amygdala Complex, consisting of the Lateral, Basolateral (Basal) and Basomedial (Accessory Basal) nuclei. 22, 81–85, 87–89, 187, 191, 193

BLMA Basolateral (Basal) and Basomedial (Accessory Basal) nuclei of the Amygdala. 96–98, 103, 104, 112, 117, 118, 123, 127, 191–193, 197, 198

BLMA⊖ Negatively valenced neurons of the Basolateral and Basomedial nuclei of the Amygdala. 103, 108, 109, 112, 117, 119–124, 127, 187–189, 191, 192

BLMA⊕ Positively valenced neurons of the Basolateral and Basomedial nuclei of the Amygdala. 102, 103, 106, 108–110, 112, 117, 120, 121, 123, 124, 127, 191–193, 198

BLMAi Inhibitory interneurons of the Basolateral and Basomedial nuclei of the Amygdala. 102, 103, 112, 117, 121, 123, 127, 191, 192, 198

BPD Borderline Personality Disorder. 16, 17, 75, 116

CD-k k-step Contrastive Divergence. 178, 179

CeA Central Nucleus of the Amygdala. 22, 24, 82–84, 87–89, 91, 92, 114, 115, 127

CN Caudate Nucleus. 75, 78, 79

CRH Corticotropin-releasing Hormone. 19, 22, 81, 91, 114, 115

CS Conditioned Stimulus. 22, 82, 85, 87–89

DA Dopamine. 23, 25, 72, 75, 77, 79, 84–86, 89, 91, 92, 98, 104, 105, 114, 117, 127, 128, 199, 200

DBN Deep Belief Network. 81–84, 88, 89, 174, 182, 183

dmH Dorsomedial Hypothalamus. 81, 83, 84, 87, 89, 91, 114

dmPFC Dorsomedial Prefrontal Cortex. 96, 106, 192, 193

EEG Electroencephalography. 20, 21

fMRI Functional Magnetic Resonance Imaging. 20, 21, 74, 75, 78, 98, 130

FS Fast Spiking. 101, 102, 104, 126, 186, 189, 190, 193, 198, 199

GABA γ-Aminobutyric acid. 104, 105, 186, 191, 196, 197, 200, 201

IB Intrinsically Bursting Neuron. 101, 102, 104, 126

ISS Infant Strange Situation. 6–9, 11–13, 16, 20, 21, 26, 28, 31

ITC Intercalated Cells of the Amygdala. 24, 83, 84, 87–89, 91, 92, 115, 127

LC Locus Coeruleus. 18, 22, 81, 83, 114

LH Left Hemisphere. 20, 21, 23, 73, 77

MFR Mean Firing Rate. 111, 112, 122–124

mOFC Medial Orbitofrontal Cortex. 24, 115, 116, 118–124, 127, 186, 198

mPFC Medial Prefrontal Cortex. 19, 23, 79, 96–102, 106, 108–112, 117, 120–123, 126, 127, 186–190, 192–194

mPOA Medial Preoptic Area. 23, 97, 98, 102, 104–106, 108–110, 112, 117, 120, 124, 126, 127, 192–194, 196, 197

MSN Medium Spiny Neuron. 101, 104, 193, 197, 200

NAc Nucleus Accumbens. 78, 79, 85, 86, 97, 98, 104, 109, 110, 112, 115–117, 120, 124, 126–128, 192, 193, 196, 197, 200, 201

NR Norepinephrine. 18, 22, 83

OFC Orbitofrontal Cortex. 18, 19, 22–25, 73–75, 79, 81–85, 87, 89, 91, 92, 94, 98, 99, 114, 115, 117, 118, 121, 198

OXT Oxytocin. 23, 72, 75–77, 79–81, 84, 86, 89, 91, 92, 114, 115, 118, 127, 128

PCC Posterior Cingulate Cortex. 99, 101

PI Posterior Insular. 98–100, 105, 106, 108–111, 120–122, 126, 186–190, 198, 199

PSNS Parasympathetic Nervous System. 18, 19

PVN Paraventricular Nucleus of the Hypothalamus. 79, 86, 114

PVNm Magnocellular part of the Paraventricular Nucleus of the Hypothalamus. 23, 81, 84, 89, 114

PVNp Parvocellular part of the Paraventricular Nucleus of the Hypothalamus. 22, 81, 83, 84, 87, 89, 91, 114, 127

RBM Restricted Boltzmann Machine. 81, 174, 178, 179, 182, 183

RH Right Hemisphere. 1, 18, 20, 21, 23, 73, 74, 77, 114

RS Regular Spiking. 101, 104, 126, 186, 189, 190, 193, 196, 198, 199

SNc Substantia Nigra pars Compacta. 37, 75, 85

SNS Sympathetic Nervous System. 18, 19, 22

STS Superior Temporal Sulcus. 99, 126

UR Unconditioned Response. 82, 87

US Unconditioned Stimulus. 22, 82, 87–89

VA Vasopressin. 81, 92

vmPFC Ventromedial Prefrontal Cortex. 24, 84, 85, 87–89, 91, 92, 94, 98, 106, 115, 116, 121, 127, 192, 193

VP Ventral Pallidum. 97, 98, 104, 109, 110, 112, 117, 120, 124, 126–128, 192, 193, 197, 201

VTA Ventral Tegmental Area. 25, 37, 75, 78, 79, 84–86, 97, 98, 104, 105, 109, 110, 112, 117, 120, 124, 126, 127, 192, 196, 197, 199, 200