A Neural Model of Empathic States in Attachment-Based Psychotherapy

David Cittern
Imperial College London, UK
Algorithmic Human Development Group
david.cittern10@imperial.ac.uk

Abbas Edalat
Imperial College London, UK
Algorithmic Human Development Group
a.edalat@imperial.ac.uk

Abstract

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 states of empathic concern involving an other that is perceived as being closely related to the self might enhance the motivation for self-directed bonding. We simulate our model computationally, and discuss the interplay between the bonding and empathy protocols of the therapy.

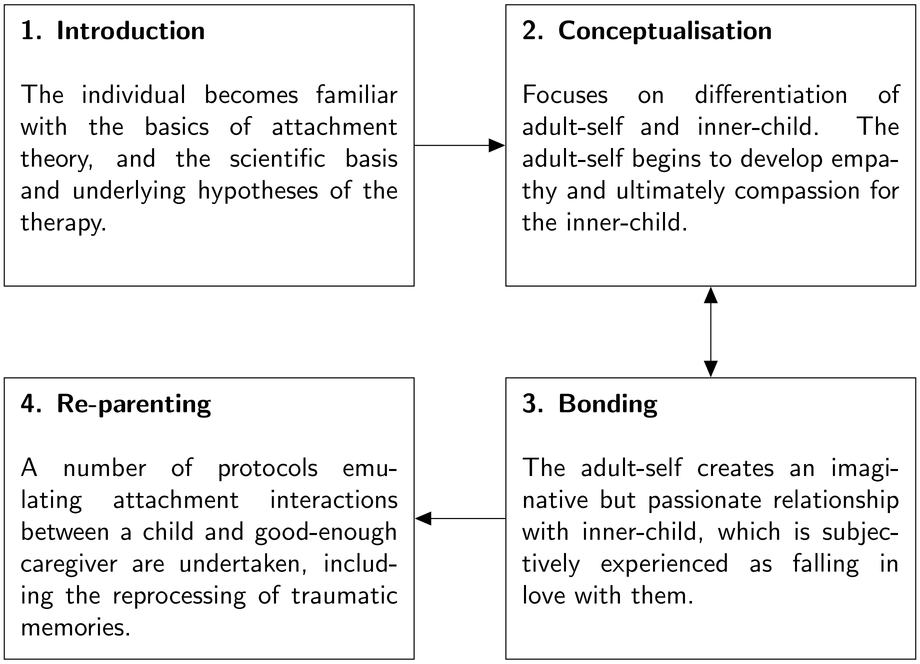
Introduction

Attachment theory grew out of the work of Bowlby (1969, 1973, 1980) who proposed that in order to fulfil their basic survival needs infants had evolved a genetic predisposition to form an attachment relationship with a primary caregiver, and that the nature of these early interactions is highly significant with respect to the formation of internal working models of self and other. While sensitive and timely caregiving in response to infant distress and requests for comfort facilitates secure-base exploration and optimal cognitive-emotional neural development and integration; neglectful, inconsistent and fear-inducing patterns of behaviour have been linked to the development of insecure (avoidant, anxious and disorganised, respectively) attachment schemas. Insecure attachment is thought to leave an individual vulnerable to a variety of psychopathologies (Mikulincer and Shaver, 2012); in the case of disorganisation these include serious disturbances such as dissociation (Liotti, 1995) and borderline personality disorder (BPD) (Fonagy et al., 2000; Carlson et al., 2009).

Self-Attachment (Edalat, 2015, 2017a,b) is a new, self-administrable, attachment-based psychotherapy which starts from the premise that at the root of many affect dysregulation, mood and anxiety disorders is a suboptimal attachment experience during early childhood (Mikulincer and Shaver, 2012). 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, whereas the adult-self corresponds to the more rational self dominant under times of calm and low perceived threat. The therapy aims to recreate the effects of early attachment-based interactions between an infant and good-enough primary caregiver using instead interactions that are fully internalised, in order to create a secure attachment schema within the individual. This is achieved by means of simulating (for example, using imagery techniques) the interactions between an infant and secure caregiver (from both perspectives). These interactions are proposed to naturally stimulate the release of hormones and neurotransmitters such as oxytocin (OXT) and dopamine (DA) in order to encourage neural plasticity and increasingly reduce suboptimal and pathological neural activity that inhibits abilities for self-agency. Since both the inner-child and adult-self are conceptualised as constituents of the self, the individual can be said to securely “self attach”.

The four stages of the Self-Attachment therapeutic process (Fig. 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 later on, 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.



1. Introduction

The individual becomes familiar with the basics of attachment theory, and the scientific basis and underlying hypotheses of the therapy.

2. Conceptualisation

Focuses on differentiation of adult-self and inner-child. The adult-self begins to develop empathy and ultimately compassion for the inner-child.

3. Bonding

The adult-self creates an imaginative but passionate relationship with inner-child, which is subjectively experienced as falling in love with them.

4. Re-parenting

A number of protocols emulating attachment interactions between a child and good-enough caregiver are undertaken, including the reprocessing of traumatic memories.

Figure 1: The four stages of the Self-Attachment therapeutic process. See text for further details.

The third stage of Self-Attachment 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 (Cittern and Edalat, 2015; Cittern, 2017). Motivation to engage in this bonding may be enhanced with protocols (described later on) involving empathic attunement with the emotional state of the inner-child. As we will explore, 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.

Relationship to Existing Therapies

Self-Attachment can be regarded as an extension of attachment theory (Edalat, 2017a), but it is also related to and incorporates ideas from a range of existing psychotherapeutic methods (Edalat, 2017b). These include the notion of an "inner child" from transactional analysis (Stewart and Joines, 1987), exposure as in behavioural therapy (Abramowitz et al., 2012), mentalization and the capacity to understand the emotions and mental states of others as well as those of oneself (Bateman and Fonagy, 2012), schema therapy and reparenting (Young et al., 2003), object-relations psychodynamic therapy (in a wider sense of the term in which objects can be impersonal as well as personal (Storr, 1988, p. 150)), and cognitive behavioural therapy (which includes techniques to identify the thoughts that induce negative affect and self-distress, challenge these thoughts as irrational or extreme, and replace them with more neutral or positive thoughts and behaviours (Hofmann, 2011)). Self-Attachment, which can be combined with any well-established therapeutic framework, integrates these techniques 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 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, yoga, mindfulness and neurofeedback) for long periods over several years (Edalat, 2015).

A particularly 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. We argue that using this inner-child representation, as opposed to a generic and/or non-related child, increases the efficacy of the therapy; both from the perspective of primary narcissism (Edalat, 2017a; Freud, 2011), and (as proposed in this paper) within the context of optimal representations for inducing empathically-motivated caregiving behaviour.

In addition, in Self-Attachment, the individual goes beyond expression of compassion for the inner-child by creating an internal affectional bond with that child, emulating the natural bonding between infants and their primary caregivers. This bonding, it has been hypothesised, activates the dopaminergic pathways of the reward system in the brain providing incentive, hope and energy for the successful conduct of the therapy (Edalat, 2015). Self-Attachment is also differentiated in its use of activities such as singing, dancing and self-massage that are known to increase DA and OXT and reduce cortisol levels (Jeffries et al., 2003; Kleber et al., 2007; Murcia et al., 2009; Field et al., 2005). By enhancing positive affects in these ways, individuals practicing Self-Attachment are better able to counter and contain their negative affects.

Is Self-Attachment a form of self-therapy or does it entail interaction with a therapist? The answer is that it can be both. Individuals who are able to conceptualise the inner-child and the adult-self in themselves may be able to undertake Self-Attachment as a self-help therapy on their own, in particular if they have already been exposed to some form of psychotherapy before. Others require a standard course of eight sessions in two to three months with a trained therapist in order to learn how to self-administer the Self-Attachment protocols. In any case, once the individual learns how to self-administer the protocols, they are required to practice them regularly by integrating them into their daily routine life style.

Empathy

As we have outlined above, the second (conceptualisation) stage of Self-Attachment involves protocols aiming to empathically attune with the emotional state of an inner-child that is in distress, and we have argued that this empathic attunement can assist in generating motivation for the protocols involved in the third stage, concerned with creating an internalised affectional bond between the adult-self and inner-child. In order to examine in detail how these empathic states might generate such motivation in Self-Attachment, we begin by defining the distinctions between various types of empathic experience and reviewing the role of empathy in attachment-based psychotherapies in general.

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 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, 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 the 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, however in contrast this mirroring is accompanied with a strong self-other distinction (i.e. 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 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).

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 in Attachment-Based Psychotherapy

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).

Empathy plays a particularly important role in attachment-based psychotherapies, which view the client-therapist relationship as an attachment bond and encourage the therapist 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 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 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.

Empathically-Motivated Bonding in Self-Attachment

In Self-Attachment, empathic resonance towards the inner-child is proposed as a method for increasing motivation to undertake the protocols related to self-directed bonding (stage 3), as previously described. The Self-Attachment 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 (representing the inner-child) and attempting to enter into an empathic state with them, in order to cultivate prosocial attitudes, feelings and states of mind geared towards alleviating their distress. Variations of the protocol can involve imagery or virtual reality techniques (Cittern et al., 2017), rather than the focus on an actual image of the younger self, in order to conceptualise the distressed and assistance-requiring inner-child.

The effectiveness of the Self-Attachment empathy protocols in motivating bonding behaviour may vary according to a number of dimensions. We briefly discuss two key factors here but do not consider them in the model that follows (these points are left as considerations for future work, and discussed further in Section 5). Firstly, as we have highlighted, 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), which involves a joint focus on the client's subjective inner states in order to strengthen their sense of self and ability to mentalize (i.e. attribute mental and emotional states underlying overt behaviour to self and other).

Neuroscience of Empathy and Self-Other Pain

In this section we overview evidence relating to the neuroscience of empathic states and self and other perceptions of pain, in order to lay the groundwork for our computational neural model of personal distress and empathic concern. Imaging studies have uncovered a core empathy-for-pain network (Engen and Singer, 2013) involving the anterior insular (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 anterior cingulate cortex (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 lead 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).

Empathically-Motivated Caregiving

Numan has recently proposed a neuroanatomical model 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).

Under the model, projections from the basolateral and basomedial nuclei of the amygdala (BLMA) (a major input region for social stimuli, which has long been associated with Pavlovian fear conditioning (Feinstein et al., 2011) but is now also known to be involved in goal-directed behaviour and to respond learned or innate rewarding stimuli (Jenison et al., 2011; Morrison and Salzman, 2010)) 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 medial prefrontal cortex (mPFC) (which includes the ACC, and thus aMCC, in Numan's definition) which in turn activates prosocial pathways. 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

[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".

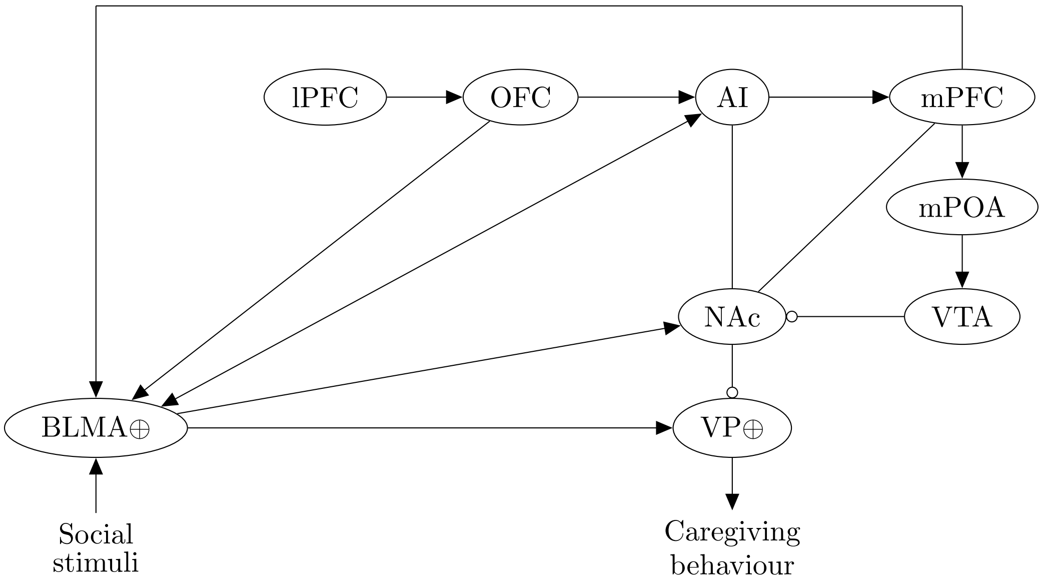


Figure 2: Numan's neuroanatomical model for how empathic states can motivate caregiving behaviour. Stimuli enter the basolateral and basomedial amygdala (BLMA), which projects to the anterior insular (AI) to create an empathic state, and to the nucleus accumbens (NAc) and ventral pallidum (VP) in order to drive caregiving behaviour. The AI projects to the medial prefrontal cortex (including the anterior cingulate) which is proposed to activate a mesolimbic caregiving pathway from the medial preoptic area of the hypothalamus (mPOA) to the ventral tegmental area (VTA) and the NAc, inhibition of which releases the VP from NAc inhibition mediated via BLMA inputs. Projections from lateral prefrontal (lPFC) and orbitofrontal (OFC) cortices to the AI and BLMA are additional pathways that can potentiate the caregiving response, as are projections from AI and mPFC to the NAc. 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). ⊕ denotes positively-valent neurons (i.e. populations that subsequently activate prosocial/approach pathways).

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 ventromedial prefrontal cortex (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-medial preoptic area (mPOA)-ventral tegmental area (VTA)-nucleus accumbens (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 neuropeptides (including OXT) act on the mPOA (a part of the hypothalamus, which is in general an integrative region that brings together a range of inputs relating to the internal environment, compares them to setpoints (ideal ranges) and activates autonomic, endocrine and behavioural responses in order to maintain the internal state within these ranges (Saper and Lowell, 2014)) 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 orbitofrontal cortex (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 medial orbitofrontal cortex (mOFC) within this framework related to findings from the neuroscience of compassion and the hypothesised effects of the Self-Attachment bonding protocols.

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 (MI) (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-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 (which is involved in a broad range of social-emotion processing and regulatory tasks, including the inhibition of socially inappropriate and impulsive behaviours (Beer et al., 2006) and the rapid response in the parent to a range of infant cues (Parsons et al., 2013)) 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)).

A Model of Personal Distress and Empathic Concern

As discussed previously, empathy plays an important role in psychotherapy and particularly in attachment-based therapies. In this section we propose (and simulate) a computational neural model to distinguish between states of personal distress and (weak and strong forms of) empathic concern. An understanding of how representations of self and other might mediate between such states is important, since therapists and other health workers engaging in empathic resonance 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, and self-focused emotional distress (Wagaman et al., 2015; Zenasni et al., 2012; Gleichgerrcht and Decety, 2013). The model developed here will also serve as a basis for examining the role of empathy in Self-Attachment therapy in the following section. The overall architecture is shown in Fig. 3.

Here, we have attempted to extract key findings from the studies by Zaki et al. (detailed in Section 2.2) with respect to self- and other-pain networks in order to model states of personal distress and empathic concern. In particular, we con-

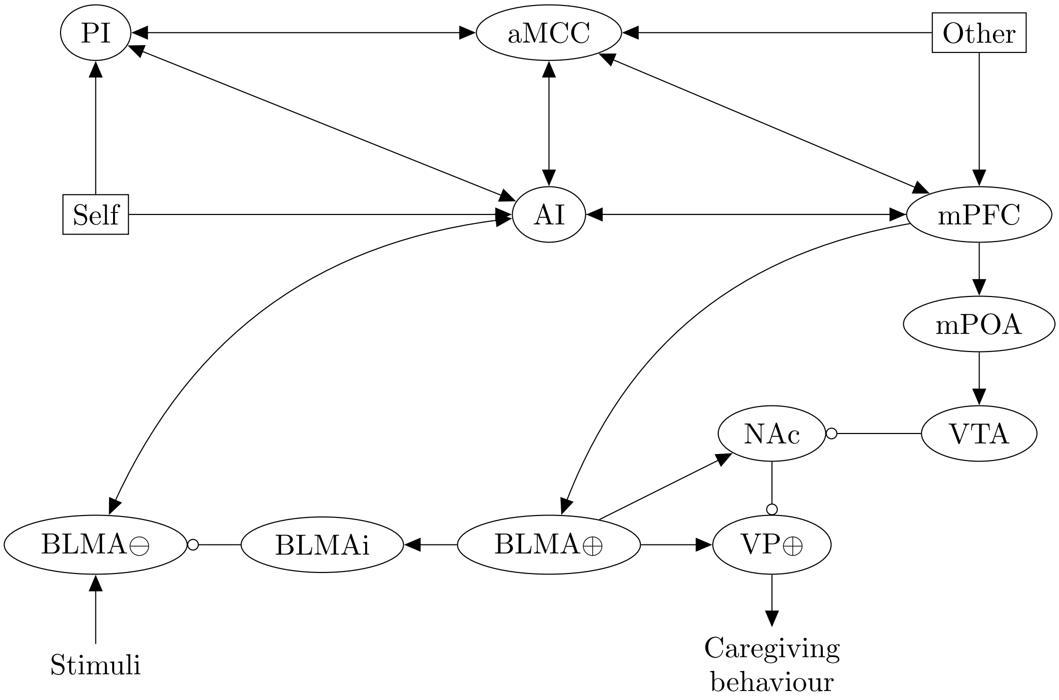


Figure 3: Model capturing how self-other distinction can mediate empathically-motivated caregiving behaviour in response to pain-inducing stimuli. The model is based on an expanded version of Numan's framework (Fig. 2) to incorporate data on the self-other distinction within the context of pain, in order to account for the distinct states of personal distress, and (weak and strong) empathic concern. During the self-pain condition (associated with personal distress), the posterior insular (PI) and negatively-valent neurons in the basolateral and basomedial amygdala (BLMA⊖), associated with egoistic withdrawal/avoidance pathways, activate the anterior insular (AI) and anterior midcingulate cortex (aMCC). During the other-pain conditions (associated with empathic concern), positively-valent BLMA⊕ neurons are activated to potentiate caregiving behaviour and inhibit BLMA⊖ avoidance circuitry. The strong empathic concern state (involving close-other representations) is associated with stronger medial prefrontal cortex (mPFC) activation of prosocial pathways compared to a weak empathic concern state (associated with relatively more distant-other representations). Ellipses represent neural populations, arrows are excitatory connections, 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 (see text for detail). ⊕ denotes positively-valent neurons (i.e. populations that subsequently activate prosocial/approach pathways) and ⊖ negatively-valent neurons (that subsequently activate withdrawal/avoidance pathways).

sider two points of current injection (Self and Other). 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 of these regions. 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 the precuneus, which is anatomically connected with the mPFC via the PCC.

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 negatively valenced neurons of the basolateral and basomedial nuclei of the amygdala (BLMA$\ominus$) (representing presence of a stimulus of an individual in distress); the AI and PI (representing input from the self-pain network); and the aMCC and mPFC (representing input from the other-pain network). Note that $\oplus$ 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. 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 $G \in \{\mathrm{BLMA}\ominus, \mathrm{AI}, \mathrm{PI}, \mathrm{aMCC}, \mathrm{mPFC}\}$, 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 (see Table A2). In particular, the mPFC has been connected to the mPOA and positively valenced neurons of the basolateral and basomedial nuclei of the amygdala (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. A1).

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.

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 for this state, and since personal distress is associated with withdrawal rather than prosocial behaviour, we should additionally have low activity in the VP (which drives caregiving behaviour). Relatively high activity in the BLMA$\ominus$-AI-PI network, 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 stimulating 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 $c_t(\text{PI}) \subseteq$

$g_t(\mathrm{PI})$ of PI neurons (with $|c_t(\mathrm{PI})| \sim \mathcal{U}\{|\mathrm{PI}| * M, |\mathrm{PI}| * H\}$. 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 stimulating random neuron subsets $c_t(\mathrm{aMCC}) \subseteq g_t(\mathrm{aMCC})$ (with $|c_t(\mathrm{aMCC})| \sim \mathcal{U}\{|\mathrm{aMCC}| * L, |\mathrm{aMCC}| * M\}$); and $c_t(\mathrm{mPFC}) \subseteq g_t(\mathrm{mPFC})$ (with $|c_t(\mathrm{mPFC})| \sim \mathcal{U}\{|\mathrm{mPFC}| * L, |\mathrm{mPFC}| * M\}$). Current is injected into mPFC neural populations encoding a close-other representation (Fig. A1).

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⊖ (as a result of input stimuli representing a distressed other), 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⊕-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 the strong empathic concern state considered next).

During the second phase of the simulation, in addition to current injections into the BLMA⊖, we thus inject current at a low level into both the AI and PI. This corresponds to stimulating a random subset $c_t(\mathrm{AI}) \subseteq g_t(\mathrm{AI})$ of AI neurons (with $|c_t(\mathrm{AI})| \sim \mathcal{U}\{|\mathrm{AI}| * L, |\mathrm{AI}| * M\}$), and a random subset $c_t(\mathrm{PI}) \subseteq g_t(\mathrm{PI})$ of PI neurons (with $|c_t(\mathrm{PI})| \sim \mathcal{U}\{|\mathrm{PI}| * L, |\mathrm{PI}| * M\}$. 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 stimulating random neuron subsets $c_t(\mathrm{aMCC}) \subseteq g_t(\mathrm{aMCC})$ (with $|c_t(\mathrm{aMCC})| \sim \mathcal{U}\{|\mathrm{aMCC}| * M, |\mathrm{aMCC}| * H\}$); and $c_t(\mathrm{mPFC}) \subseteq g_t(\mathrm{mPFC})$ (with $|c_t(\mathrm{mPFC})| \sim \mathcal{U}\{|\mathrm{mPFC}| * M, |\mathrm{mPFC}| * H\}$). Current is injected into mPFC neurons encoding more distant-other representations (Fig. A1).

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 empathic concern 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$\oplus$-VP and mPOA-VTA-NAc-VP pathways.

Current injections for the third phase of the simulation capturing strong empathic concern are the same as for the weak empathic concern state, except that current is now injected into mPFC neurons encoding close-other representations rather than distant-other representations (Fig. A1).

Simulation Results

Here we describe results of simulations of our network over the three phases described in Section 3.1. 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 (with inputs from the other-pain network targeting 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 compared to the weak empathy state. Our simulations are implemented using the CARLsim 3.1 framework (Beyeler et al., 2015) with Izhikevich neurons and a network topology as described in Appendix A, and results presented are representative of a typical simulation run.

Fig. 4 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 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. 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,

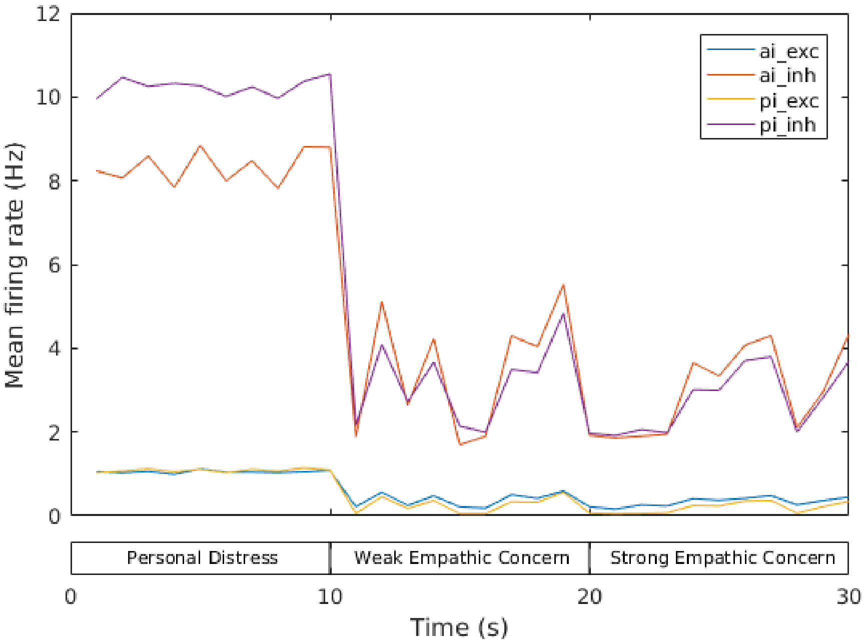


Figure 4: Mean firing rates for the AI and PI in the model of Personal Distress and Empathic Concern (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

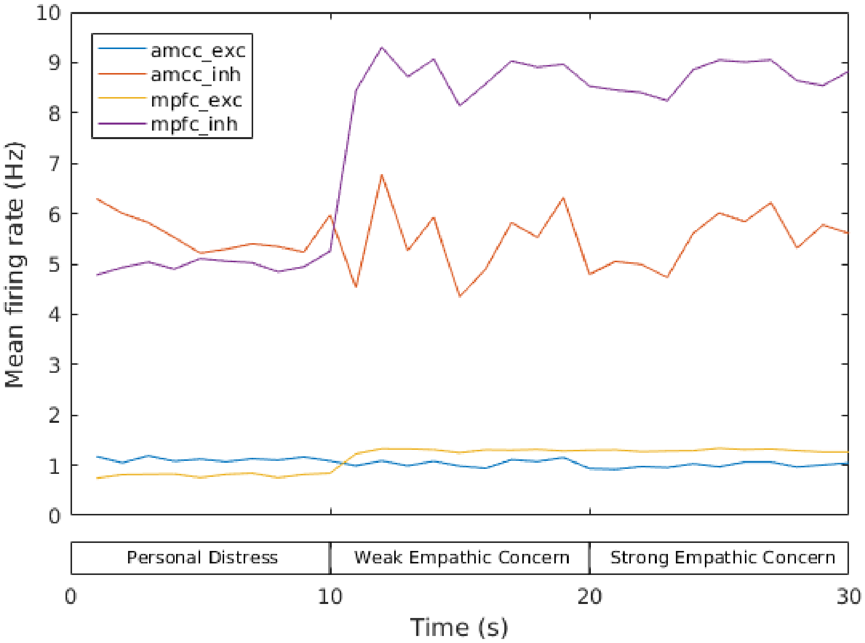


Figure 5: Mean firing rates for the aMCC and mPFC in the model of Personal Distress and Empathic Concern (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

and 1.05/5.61 Hz at 30s. This relative stability is in accordance with the findings of the experiments described in Section 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 external current during the self-pain condition).

Fig. 6 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 inhibitory interneurons of the basolateral and basomedial nuclei of the amygdala (BLMAi).

Finally, MFR for the mPOA, VTA, NAc and VP are shown in Fig. 7. Activity in the mPOA, VTA and NAc is relatively low during the first phase, and rises

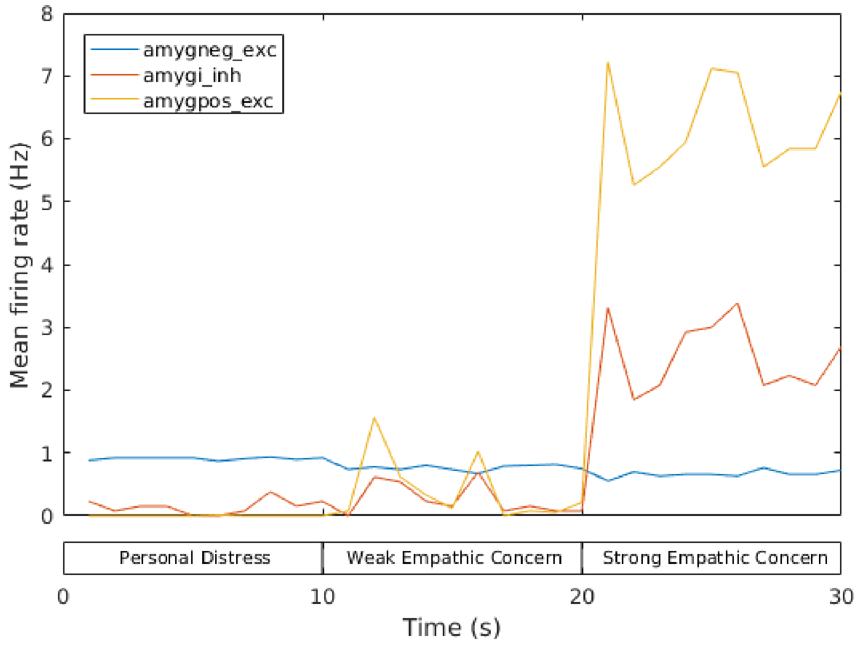


Figure 6: Mean firing rates for the three BLMA neural populations in the model of Personal Distress and Empathic Concern (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

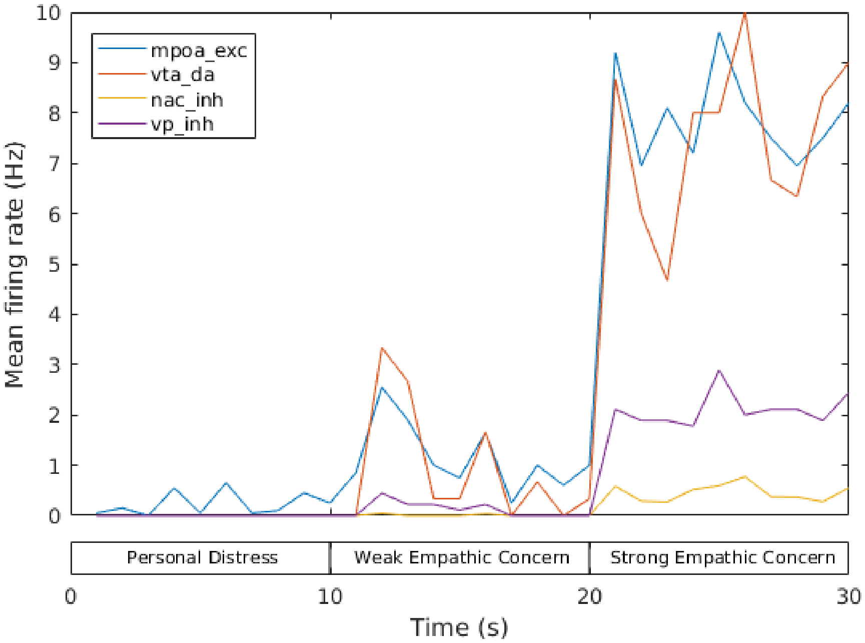


Figure 7: Mean firing rates for the mPOA, VTA, NAc and VP in the model of Personal Distress and Empathic Concern (exc = excitatory neurons, inh = inhibitory neurons, da = dopaminergic neurons). See text for details.

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 facilitates caregiving behaviour) is thus low during personal distress, increased for the weak empathic concern state, and highest for strong empathic concern.

In summary, our model replicates data on relative AI and PI activation across self- and other-pain states (with relatively higher activation in these areas during self-pain), and activation in the aMCC (for which no difference was found across these conditions). Relatively higher activation of mPFC during other-pain conditions compared to self-pain conditions is also broadly consistent with a noted role for this region in empathy but not personal distress or experience of pain in the self. The model additionally makes some predictions relating to the two mPFC-mediated caregiving pathways, that are a result of the weights we have chosen. In particular, our model predicts that both of the mPFC-mediated caregiving pathways (mPFC-BLMA$\oplus$-NAc/VP and mPFC-mPOA-VTA-NAc-VP) will be active during both weak and strong empathic concern states, and that they will both increase activation across these phases (whereas it might be that, for example, activation in the mPFC-mPOA-VTA-NAc pathway is relatively stable across these phases, with increases in caregiving output in the strong empathic concern state mediated mostly by increases in stimulation of the VP by the BLMA). According to the model, mPOA and VTA activation will be highest in the strong empathic concern state relative to weak empathic concern, with mPOA and VTA firing rates roughly equal and greater than VP firing rates, which are in turn greater than those for the NAc, across both states. During the personal distress state, mPOA firing rates are predicted to be too low to result in any significant firing in the VTA and NAc (as opposed to activation along this pathway, but insufficient BLMA$\oplus$ stimulation of NAc in order to release VP from inhibition, which could result in similar caregiving output). These predictions can potentially be tested empirically (for example, in empathic scenarios involving a behavioural response in which stimuli depicting self, close other, and distant other are used with the aim of differentially activating the vmPFC), with weights in the model tuned accordingly to account for any discrepancies.

A Model of the Self-Attachment Empathy Protocols

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-valenced emotional state of another (e.g. the client), and 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 consider Self-Attachment therapy, in which the individual takes the role of both psychotherapist and client. In particular, we propose a hypothetical model of how Self-Attachment therapy might gradually progress from a state of personal distress, to one of strong empathic concern (as the conceptualisation stage is completed), to one of compassion (as a result of empathically-motivated self-directed bonding) in which inner-child directed caregiving behaviour is mediated by a positive (compassionate) rather than negative (empathic) emotional state.

Bonding Protocols

Before detailing our proposal regarding the hypothesised effects of empathic resonance with the inner-child, we briefly review our previously presented neurobiological hypothesis with regards to the effects of the Self-Attachment bonding protocols (Cittern and Edalat, 2015; Cittern, 2017) which this empathic concern is aiming to motivate as behavioural output. The bonding protocols are concerned with the individual forming an imaginative but passionate affective bond with the inner-child (from the perspective of the adult-self), which is subjectively experienced as falling in love with them. The bond-making process can further be enhanced with the use of activities such as self-massage and (overt and/or imagined) song and dance directed towards the inner-child, which are proposed to stimulate reward circuitry.

At a neural level, the internal working model (attachment schema) has been theorised to be based in unconscious and implicit memories, rooted mainly in right hemisphere (RH) brain regions centred on the OFC, amygdala and hypothalamus (Schore, 2003; Cozolino, 2006, p.139); areas known to be central to, and crucially involved in, a broad range of social cognition and emotional processing functions. 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, which represents many types of primary reward (including positive tactile stimulation) Rolls (2013) and parts of which have been found to preferentially activate in mothers viewing images of own as opposed to other infants (with activation levels correlating with self-reported pleasant mood ratings) (Nitschke et al., 2004; Minagawa-Kawai et al., 2009), with medial regions crucially involved in the learning of stimulus-reward associations (Walton et al., 2010). The OFC is believed to mediate stress and facilitative reactivity to social stimuli via projections to the dorsomedial hypothalamus (dmH) and the paraventricular nucleus of the hypothalamus (PVN). The parvocellular part of the paraventricular nucleus of the hypothalamus (PVNp) releases corticotropin-releasing hormone (CRH) (the precursor to stress hormone cortisol), which stimulates stress circuitry focused on the central nucleus of the amygdala (CeA) (a major output nucleus of the amygdala involved in processing pain and fear responses (Zimmerman et al., 2007)) and the locus coeruleus (LC) (which is involved particularly in arousal and alertness aspects (Benarroch, 2009)), while the magnocellular part of the paraventricular nucleus of the hypothalamus (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 within circuitry involving the vmPFC, amygdala and hypothalamus (Atzil et al., 2017). 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, we have previously hypothesised (Cittern and Edalat, 2015; Cittern, 2017) that a main effect of the bonding protocols 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 self-massage or directed singing (Salimpoor et al., 2011; Jeffries et al., 2003; Kleber et al., 2007) with inner-child imagery (Strathearn et al., 2008)). At a neural level, we proposed that these new stimulus-reward associations in mOFC would result in a rebalancing in activation of stress and facilitative circuitry in response to such classes of social stimuli. As the bonding protocols progress, the mOFC should gradually come to learn new reward associations, increasingly facilitating natural endogenous OXT release and inhibiting CRH release; while dopaminergic reward-prediction errors should drive a vmPFC-mediated inhibition of stress circuitry via strengthening of an intercalated cells of the amygdala (ITC)-CeA pathway, inhibiting activity in the amygdala and stress circuitry.

Bonding Protocols and Compassionate States

The Self-Attachment bonding protocols are closely related to the concept of compassion. In a compassionate state, as for an empathic state, there is a strong self-other distinction along with knowledge of the emotional state of another (Gonzalez-Liencres et al., 2013). 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. With respect to Self-Attachment, our proposal is 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 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) (Engen and Singer, 2015), 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.

Self-Attachment Empathy Protocols: A Hypothesis

Based on our model of personal distress and (weak and strong) empathic concern states, we now attempt to form a hypothesis with respect to how the Self-Attachment empathy protocols might result in neural activation corresponding to a strong empathic concern state towards the inner-child, and how repetitive application of the self-directed caregiving and bonding (that is proposed to result as motivated behaviour from this strong empathic concern) might in turn facilitate a gradual shift towards a pattern of network activation corresponding more to a compassionate state within the adult-self.

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 might 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. As overviewed in Section 4.1, we previously hypothesised that one effect of the Self-Attachment bonding protocols would be new attachment-related stimulus-reward associations in the mOFC (an area that represents many forms of primary reward and is involved in learning stimulus-reward associations (Rolls, 2013); is known to increase firing in response to stimuli that predict rewards with activation levels that are correlated with reward value (Gottfried et al., 2003); and has been associated with positive emotional states and compassionate stances, with heightened activation found in this area following compassion training in human subjects (Klimecki et al., 2013)). Thus, within the context of our model of empathically-motivated caregiving behaviour here, we hypothesise that with repetitive application of the bonding protocols, mOFC activation should increase in response to the distressed inner-child stimulus. In this way, the mOFC should be expected to increasingly stimulate both inhibitory BLMAi (which inhibit the negatively valent BLMA$\ominus$ neurons) and positively valent BLMA$\oplus$, resulting in caregiving behaviour that gradually comes to be facilitated via a positive (OFC-mediated) rather than negative (BLMA$\ominus$-mediated) emotional state.

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. 8) which are proposed to increase in strength as the bonding protocols (stage 3) 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 A3.

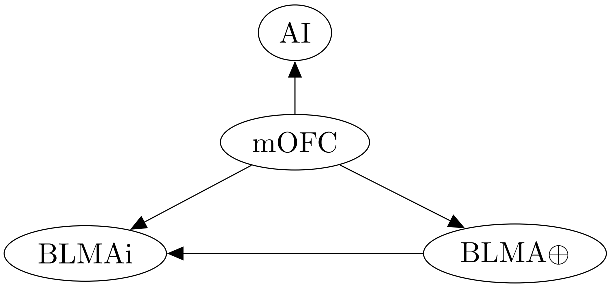


Figure 8: Additional connectivity emanating from the medial orbitofrontal cortex (mOFC) to the anterior insular (AI), and inhibitory (BLMAi) and positively-valent (BLMA$\oplus$) neurons of the basolateral and basomedial amygdala. The remaining architecture is as in Fig. 3

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 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 3.1), 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).

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 the lack of self-other distinction might be expected to result in personal distress: in accordance with the findings of Zaki et al. described previously, AI and PI activations should be relatively high for this state, which involves a form of self-pain. Since personal distress is associated with withdrawal rather than prosocial behaviour, we should additionally 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 hypothesised to correspond to personal distress during the early stages of Self-Attachment.

Current injections for the first simulation phase ($t \in (0, 10]$ seconds) are as previously defined for the personal distress state (Section 3.1.1), 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 in stage 3 of the therapy). This corresponds to stimulating random neuron subsets $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ (with $|c_t(\text{mOFC})| \sim \mathcal{U}\{|\text{mOFC}| * L, |\text{mOFC}| * M\}$).

Strong Empathic Concern with Strengthening Conceptualisation

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 self-directed bonding in stage 3 of the therapy) results as motivated behaviour. During this state of strong empathic concern, we should again have activation in the BLMA$\ominus$ (stimulated as a result of 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, during this phase of the simulation we want to capture a transition from personal distress (with weak self-other distinction) to strong empathic concern (with a well developed self-other distinction and a close-other representation), which is proposed to occur as conceptualisation of the inner-child (stage 2 of the therapy) progresses. This transition is assumed to correspond to a linear shift in input current into the insular, aMCC and mPFC, representing a gradual shift between the states of personal distress and strong empathic concern (that we previously defined) that might occur as the concept of the inner-child as a distinct entity is developed. 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 stimulating a random subset $c_t(\mathrm{AI}) \subseteq g_t(\mathrm{AI})$ of AI neurons (with $|c_t(\mathrm{AI})| = ((|\mathrm{AI}| * (L - H)) * (t/p)) + |\mathrm{AI}| * (2 * H - L)$, and a random subset $c_t(\mathrm{PI}) \subseteq g_t(\mathrm{PI})$ of PI neurons (with $|c_t(\mathrm{PI})| = ((|\mathrm{PI}| * (L - H)) * (t/p)) + |\mathrm{PI}| * (2 * H - L)$. 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 stimulating a random subset $c_t(\mathrm{aMCC}) \subseteq g_t(\mathrm{aMCC})$ of aMCC neurons (with $|c_t(\mathrm{aMCC})| = ((|\mathrm{aMCC}| * (H - L)) * (t/p)) + |\mathrm{aMCC}| * (2 * L - H)$; and a random subset $c_t(\mathrm{mPFC}) \subseteq g_t(\mathrm{mPFC})$ of mPFC neurons (with $|c_t(\mathrm{mPFC})| = ((|\mathrm{mPFC}| * (H - L)) * (t/p)) + |\mathrm{mPFC}| * (2 * L - H)$.

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. One possibility is 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 in circuits involving these regions, which might in turn account for a change in activation in self- and other-pain network activity (modelled here in terms of a simple linear shift in current injection). As we will discuss in Section 5, 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 (to induce levels of activation corresponding to network states before which application of the bonding protocols have begun to take effect) with stimulation of a random subset $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ of mOFC neurons (with $|c_t(\text{mOFC})| \sim \mathcal{U}\{|\text{mOFC}| * L, |\text{mOFC}| * M\}$).

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 (stage 3) and consequentially updated stimulus-reward associations in the mOFC. 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. In our model of the empathy protocols here, 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 stimulating 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 stimulating 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 $|c_t(\text{mPFC})| \sim \mathcal{U}\{|\text{mPFC}| * M, |\text{mPFC}| * H\}$. To capture successful application of the bonding protocols, we linearly increase the proportion of mOFC neurons receiving input across the phase, by stimulating a random subset $c_t(\text{mOFC}) \subseteq g_t(\text{mOFC})$ of mOFC neurons (with $|c_t(\text{mOFC})| = ((|\text{mOFC}| * (H - L)) * (t/p)) + |\text{mOFC}| * (3 * L - 2 * H)$. As we have discussed, increasing activation here is proposed to correspond to increasing levels of reward associated with the inner-child stimulus as a result of this self-directed bonding.

Simulation Results

Fig. 9 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. 10. 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 personal distress and strong empathic concern states, and is in accordance with the findings of the experiments described in Section 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

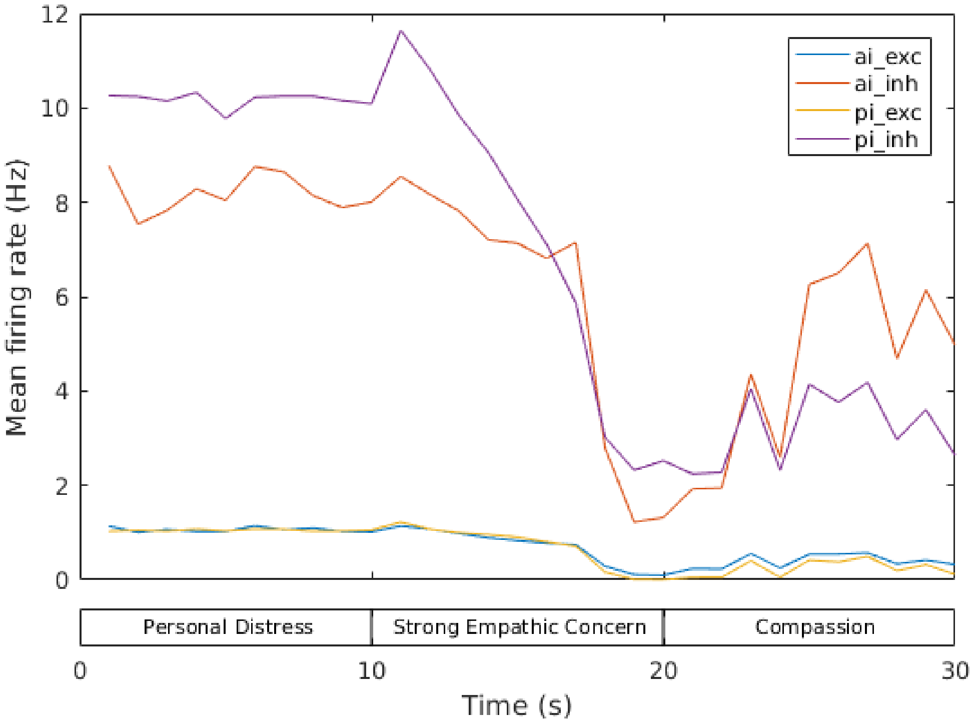


Figure 9: Mean firing rates for the AI and PI in the model of the Self-Attachment Empathy Protocols (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

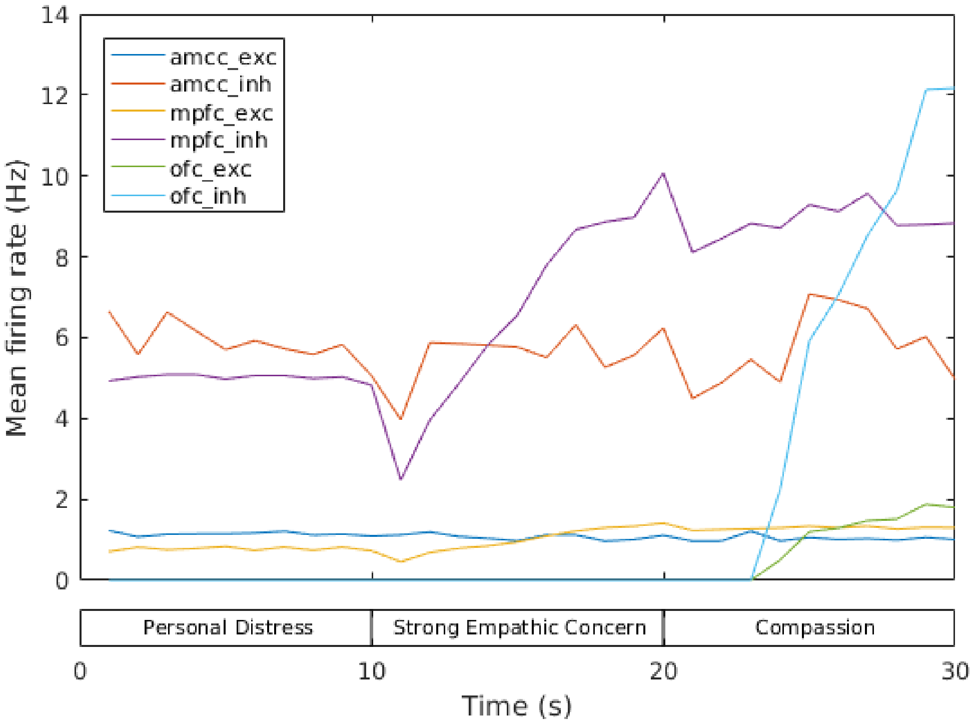


Figure 10: Mean firing rates for the aMCC, mPFC and mOFC in the model of the Self-Attachment Empathy Protocols (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

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. 11 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 to 13.62 Hz at 30s), whose neurons are stimulated by the mOFC. In contrast, the MFR of the BLMA$\oplus$ 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$\oplus$ and BLMAi, and relatively high activation in the BLMA$\ominus$ 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$\ominus$ falls significantly,

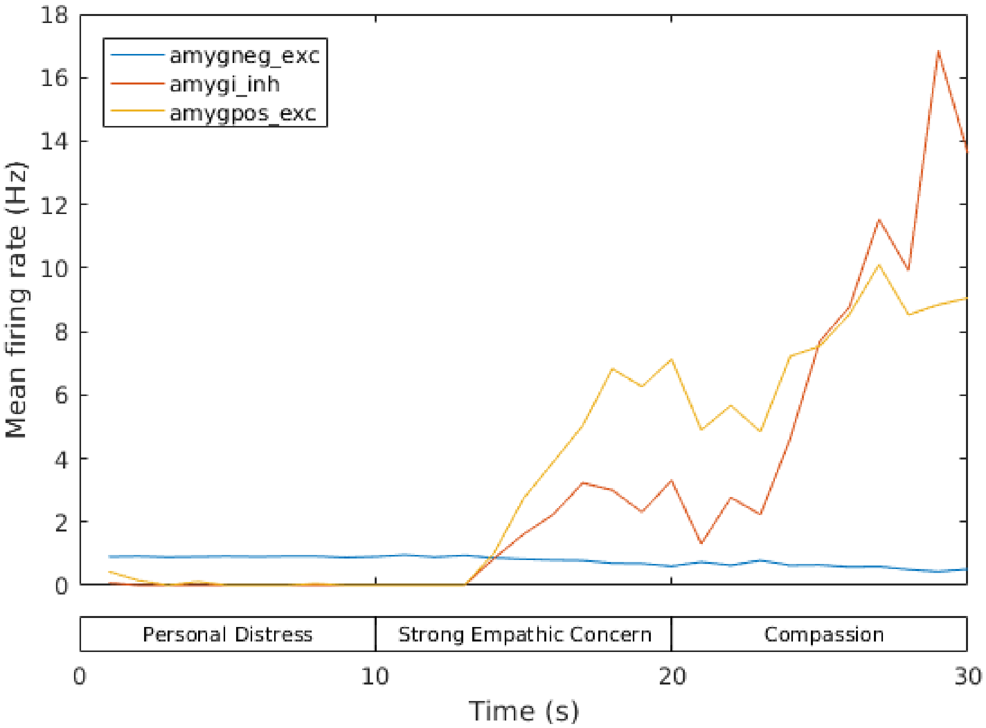


Figure 11: Mean firing rates for the three BLMA neural populations in the model of the Self-Attachment Empathy Protocols (exc = excitatory neurons, inh = inhibitory neurons). See text for details.

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. 12. 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. 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 increase in caregiving behaviour as the internal state of the adult-self transitions

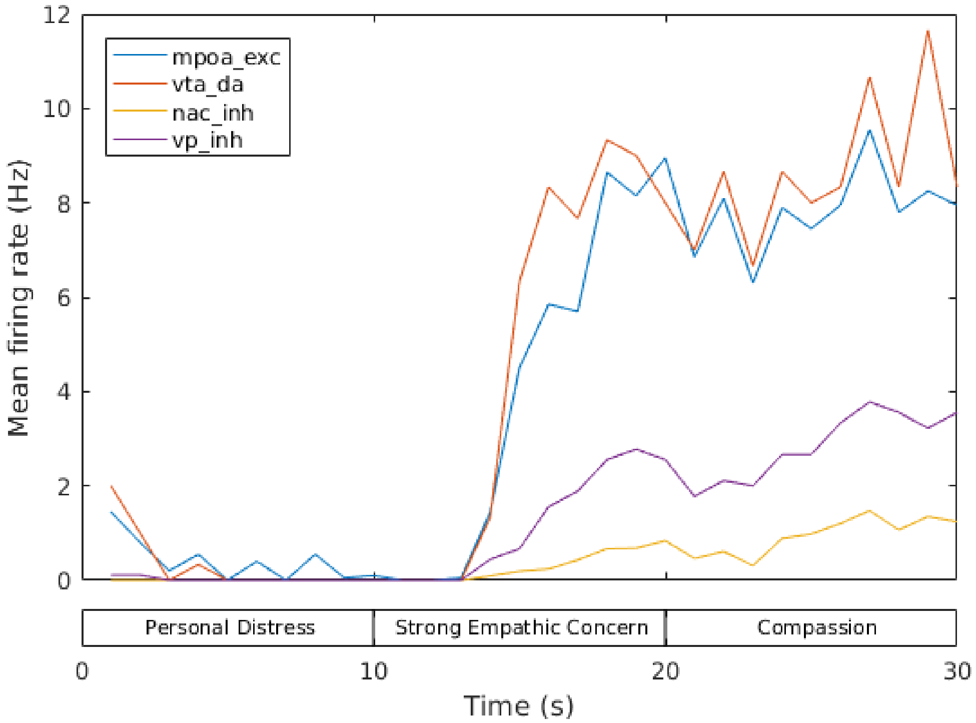


Figure 12: Mean firing rates for the mPOA, VTA, NAc and VP in the model of the Self-Attachment Empathy Protocols (exc = excitatory neurons, inh = inhibitory neurons, da = dopaminergic neurons). See text for details.

from an empathic state towards a more compassionate one.

In summary, our biologically-plausible model replicates existing data on relative AI and PI activation across self- and other-pain states (with relatively higher activation in these areas during self-pain), activation in the aMCC (for which no difference was found across self and other pain conditions), and relative activation in the mOFC in compassion compared to non-compassion (strong empathic concern and personal distress) states, resulting in appropriate output in caregiving pathways across these distinct states. Our model additionally makes a number of new predictions, which are partly a result of the particular weights that we have chosen in our simulations. These include that activation in mPFC neural populations will be relatively stable across strong empathic concern and compassion stages; that both mPFC-mediated caregiving pathways will be active during strong empathic concern and compassion, with rising activation across these states; and that inhibition on negatively valent BLMA$\ominus$ neurons mediating avoidance will be higher in compassion as compared to strong empathic concern states. As with those in Section 3.2, these predictions can potentially be tested (with suitably designed experimental scenarios) and the model refined accordingly; as can our hypotheses related to

Self-Attachment therapy (for example that inner-child representations will activate those vmPFC neurons encoding close/innocent others, and that mOFC-mediated inhibition of BLMA$\ominus$ will occur as a result of self-directed bonding).

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 as the therapy progresses.

One of the main considerations for any modelling effort is the selection of an appropriate level of abstraction. Since our hypotheses on the effects of Self-Attachment therapy have to date been formulated based on existing related data on firing and plasticity in particular brain regions, we chose to use a relatively detailed and biologically plausible spiking neural model. Such an approach clearly has limitations owing to the resulting complexity of the model, for example in terms of the large parameter space for connectivity weights, and the necessity to compartmentalise (and ignore regional interactions that might in future be determined to be significant for explaining the emergent properties of interest). Nonetheless, this method has certain advantages with respect to following an empirical hypothesis-testing approach to therapy development, in that predictions for changes in regional plasticity and firing rates made by such models can be relatively easily and directly tested (e.g. using imaging). Indeed, our aim is for future versions of the model to be able to capture the initial conditions (connectivity) characterising different groups of individuals (e.g. according to attachment type, or prevalence of borderline symptoms). In this way, given a target network state corresponding to a secure attachment schema, testable predictions made by the model might directly inform development of Self-Attachment therapy (for example by determining which pathways, and thus intervention methods, might be most effectively targeted in each case). The work presented here is a small step in that direction, and we discuss now some additional ways in which the model's biological accuracy might be improved in future iterations.

The first point to note is that we didn't 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 regular spiking (RS) (for mPOA), fast spiking (FS) (VP) and intrinsically bursting neuron (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) (Numan, 2014). 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.

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. 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. These studies suggest a strong interdependence between the bonding and empathy protocols in Self-Attachment, and point to a 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).

Acknowledgements

We would like to thank Michael Numan for his valuable comments and suggestions, and also the reviewers for their helpful feedback.

References

Abramowitz, J. S., Deacon, B. J., and Whiteside, S. P. (2012). *Exposure therapy for anxiety: Principles and practice.* Guilford Press.

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

Atzil, S., Touroutoglou, A., Rudy, T., Salcedo, S., Feldman, R., Hooker, J. M., Dickerson, B. C., Catana, C., and Barrett, L. F. (2017). Dopamine in the medial amygdala network mediates human bonding. *Proceedings of the National Academy of Sciences*, 114(9):2361–2366.

Baird, A., Dewar, B.-K., Critchley, H., Dolan, R., Shallice, T., and Cipolotti, L. (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 Baker, M. N. (1987). Heinz Kohut's self psychology: An overview. *American Journal of Psychiatry*, 144(1):1–9.

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

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

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

Bartz, J. A., Zaki, J., Ochsner, K. N., Bolger, N., Kolevzon, A., Ludwig, N., and Lydon, J. E. (2010). 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 Fonagy, P. (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.

Beer, J. S., John, O. P., Scabini, D., and Knight, R. T. (2006). Orbitofrontal cortex and social behavior: integrating self-monitoring and emotion-cognition interactions. *Journal of cognitive neuroscience*, 18(6):871–879.

Benarroch, E. E. (2009). The locus ceruleus norepinephrine system functional organization and potential clinical significance. *Neurology*, 73(20):1699–1704.

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

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

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

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).

Brudzynski, S. M., Wu, M., and Mogenson, G. J. (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.

Byne, W., Lasco, M. S., Kemether, E., Shinwari, A., Edgar, M. A., Morgello, S., Jones, L. B., and Tobet, S. (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., Egeland, B., and Sroufe, L. A. (2009). A prospective investigation of the development of borderline personality symptoms. *Development and Psychopathology*, 21(04):1311–1334.

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

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

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

Cittern, D. (2017). *Computational Models of Attachment and Self-Attachment*. PhD thesis, Imperial College London.

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

Cittern, D., Edalat, A., and Ghaznavi, I. (2017). An immersive virtual reality mobile platform for self-attachment. In *Proceedings of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB) Annual Convention*.

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

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

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).

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

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

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

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

Denny, B. T., Kober, H., Wager, T. D., and Ochsner, K. N. (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.

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

Edalat, A. (2015). Introduction to self-attachment and its neural basis. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*.

Edalat, A. (2017a). Self-attachment: A holistic approach to computational psychiatry. In Erdi, P., Bhattacharya, B. S., and Cochran, A., editors, *Computational Neurology and Psychiatry*, Springer Series in Bio/Neuroinformatics. Springer.

Edalat, A. (2017b). Self-attachment: A self-administrable intervention for chronic anxiety and depression. Technical Report 2017/3, Department of Computing, Imperial College London.

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

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

Falconer, C. J., Rovira, A., King, J. A., Gilbert, P., Antley, A., Fearon, P., Ralph, N., Slater, M., and Brewin, C. R. (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., Slater, M., Rovira, A., King, J. A., Gilbert, P., Antley, A., and Brewin, C. R. (2014). Embodying compassion: A virtual reality paradigm for overcoming excessive self-criticism. *PloS One*, 9(11):e111933.

Fehse, K., Silveira, S., Elvers, K., and Blautzik, J. (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., Adolphs, R., Damasio, A., and Tranel, D. (2011). The human amygdala and the induction and experience of fear. *Current Biology*, 21(1):34–38.

Feldman, R., Weller, A., Zagoory-Sharon, O., and Levine, A. (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., Hernandez-Reif, M., Diego, M., Schanberg, S., and Kuhn, C. (2005). Cortisol decreases and serotonin and dopamine increase following massage therapy. *International Journal of Neuroscience*, 115(10):1397–1413.

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

Freud, S. (2011). *The future of an illusion*. Martino Fine Books. First published in 1928.

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

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

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

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

Gleichgerrcht, E. and Decety, J. (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):824–827.

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

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

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

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

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

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

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

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

Hofmann, S. G. (2011). *An introduction to modern CBT: Psychological solutions to mental health problems*. John Wiley & Sons.

Höistad, M., Heinsen, H., Wicinski, B., Schmitz, C., and Hof, P. R. (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.

Hurlemann, R., Patin, A., Onur, O. A., Cohen, M. X., Baumgartner, T., Metzler, S., Dziobek, I., Gallinat, J., Wagner, M., Maier, W., 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. and Edelman, G. M. (2008). Large-scale model of mammalian thalamocortical systems. *Proceedings of the National Academy of Sciences*, 105(9):3593–3598.

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

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

Jeffries, K., Fritz, J., and Braun, A. (2003). Words in melody: an h215o pet study of brain activation during singing and speaking. *Neuroreport*, 14(5):749–754.

Jenison, R. L., Rangel, A., Oya, H., Kawasaki, H., and Howard, M. A. (2011). Value encoding in single neurons in the human amygdala during decision making. *Journal of Neuroscience*, 31(1):331–338.

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

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

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

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

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

Klimecki, O. M., Leiberg, S., Ricard, M., and Singer, T. (2013). Differential pattern of functional brain plasticity after compassion and empathy training. *Social Cognitive and Affective Neuroscience.*

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

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

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

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

Liotti, G. (1995). Disorganized/disoriented attachment in the psychotherapy of the dissociative disorders. In Goldberg, S., Muir, R., and Kerr, J., editors, *Attachment theory: Social, developmental, and clinical perspectives*, page 343–363. Analytic Press, Inc.

Lonstein, J. and De Vries, G. (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.

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

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

Main, M. and Hesse, E. (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.

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

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

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

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

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

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

McClure, W. O., Ishtoyan, A., and Lyon, M. (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.

Melchitzky, D. S., González-Burgos, G., Barrionuevo, G., and Lewis, D. A. (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., Masten, C. L., Ma, Y., Wang, C., Shi, Z., Eisenberger, N. I., and Han, S. (2012). Empathy for the social suffering of friends and strangers recruits distinct patterns of brain activation. *Social Cognitive and Affective Neuroscience*.

Mikulincer, M. and Shaver, P. R. (2005). Attachment security, compassion, and altruism. *Current Directions in Psychological Science*, 14(1):34–38.

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

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

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

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

Mogenson, G. J. (1987). Limbic-motor integration. *Progress in Psychobiology and Physiological Psychology*, 12:117–170.

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

Morrison, S. E. and Salzman, C. D. (2010). Re-valuing the amygdala. *Current Opinion in Neurobiology*, 20(2):221–230.

Murcia, C. Q., Bongard, S., and Kreutz, G. (2009). Emotional and neurohumoral responses to dancing tango argentino the effects of music and partner. *Music and Medicine*, 1(1):14–21.

Murray, R. J., Schaer, M., and Debbané, M. (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.

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

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

Noriuchi, M., Kikuchi, Y., and Senoo, A. (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., Bress, J. A., Ranker, L. R., Gary, A. J., DeNicola, A. L., Bettis, J. K., and Knapp, S. E. (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., Numan, M. J., Schwarz, J. M., Neuner, C. M., Flood, T. F., and Smith, C. D. (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 Young, L. J. (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., Zaki, J., Hanelin, J., Ludlow, D. H., Knierim, K., Ramachandran, T., Glover, G. H., and Mackey, S. C. (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.

Olff, M., Frijling, J. L., Kubzansky, L. D., Bradley, B., Ellenbogen, M. A., Cardoso, C., Bartz, J. A., Yee, J. R., and van Zuiden, M. (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.

Packer, A. M. and Yuste, R. (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.

Parsons, C. E., Stark, E. A., Young, K. S., Stein, A., and Kringelbach, M. L. (2013). Understanding the human parental brain: a critical role of the orbitofrontal cortex. *Social Neuroscience*, 8(6):525–543.

Pavuluri, M. and May, A. (2015). I feel, therefore, I am: the insula and its role in human emotion, cognition and the sensory-motor system. *AIMS Neuroscience*, 2(1):18–27.

Phan, K. L., Wager, T., Taylor, S. F., and Liberzon, I. (2002). Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in PET and fMRI. *Neuroimage*, 16(2):331–348.

Rabinowicz, T., Dean, D. E., Petetot, J. M.-C., and de Courten-Myers, G. M. (1999). Gender differences in the human cerebral cortex: more neurons in males; more processes in females. *Journal of Child Neurology*, 14(2):98–107.

Rajkowska, G., Miguel-Hidalgo, J. J., Wei, J., Dilley, G., Pittman, S. D., Meltzer, H. Y., Overholser, J. C., Roth, B. L., and Stockmeier, C. A. (1999). Morphometric evidence for neuronal and glial prefrontal cell pathology in major depression. *Biological Psychiatry*, 45(9):1085–1098.

Riem, M. M., Bakermans-Kranenburg, M. J., van IJzendoorn, M. H., Out, D., and Rombouts, S. A. (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., Snyder, R., Steele, H., and Siever, L. J. (2013). The neurobiology of empathy in borderline personality disorder. *Current Psychiatry Reports*, 15(3):1–11.

Rockliff, H., Gilbert, P., McEwan, K., Lightman, S., and Glover, D. (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.

Rolls, E. T. (2013). *Emotion and decision making explained.* Oxford University Press.

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.

Rudy, B., Fishell, G., Lee, S., and Hjerling-Leffler, J. (2011). Three groups of interneurons account for nearly 100% of neocortical gabaergic neurons. *Developmental Neurobiology*, 71(1):45–61.

Sah, P., Faber, E. L., De Armentia, M. L., and Power, J. (2003). The amygdaloid complex: anatomy and physiology. *Physiological Reviews*, 83(3):803–834.

Salimpoor, V. N., Benovoy, M., Larcher, K., Dagher, A., and Zatorre, R. J. (2011). Anatomically distinct dopamine release during anticipation and experience of peak emotion to music. *Nature Neuroscience*, 14(2):257–262.

Saper, C. B. and Lowell, B. B. (2014). The hypothalamus. *Current Biology*, 24(23):1111–1116.

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

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K., and Van Hoesen, G. W. (1998). Limbic frontal cortex in hominoids: a comparative study of area 13. *American Journal of Physical Anthropology*, 106(2):129–155.

Sherman, S. M. and Guillery, R. (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 Rayport, S. (1994). Gaba synapses formed in vitro by local axon collaterals of nucleus accumbens neurons. *The Journal of Neuroscience*, 14(7):4548–4560.

Singer, T., Seymour, B., O'doherty, J., Kaube, H., Dolan, R. J., and Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, 303(5661):1157–1162.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439(7075):466–469.

Stewart, I. and Joines, V. (1987). *TA today: A new introduction to transactional analysis.* Vann Joines.

Stippich, C., Ochmann, H., and Sartor, K. (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.

Storr, A. (1988). *Solitude.* Harper Collins.

Strathearn, L., Fonagy, P., Amico, J., and Montague, P. R. (2009). Adult attachment predicts maternal brain and oxytocin response to infant cues. *Neuropsychopharmacology*, 34(13):2655–2666.

Strathearn, L., Li, J., Fonagy, P., and Montague, P. R. (2008). What's in a smile? Maternal brain responses to infant facial cues. *Pediatrics*, 122(1):40–51.

Tecuapetla, F., Carrillo-Reid, L., Bargas, J., and Galarraga, E. (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., Maruyama, T., Yoshida, S., Nishimori, K., Kato, T., Numan, M., and Kuroda, K. O. (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., Setzer, M., and Bohl, J. (2003). Ontogeny of the human amygdala. *Annals of the New York Academy of Sciences*, 985(1):22–33.

Voges, N., Schüz, A., Aertsen, A., and Rotter, S. (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., Andersson, F., Grandjean, D., Sander, D., and Vuilleumier, P. (2008). Individual attachment style modulates human amygdala and striatum activation during social appraisal. *PLoS One*, 3(8):e2868.

Wagaman, M. A., Geiger, J. M., Shockley, C., and Segal, E. A. (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.

Walter, H. (2012). Social cognitive neuroscience of empathy: concepts, circuits, and genes. *Emotion Review*, 4(1):9–17.

Walton, M. E., Behrens, T. E., Buckley, M. J., Rudebeck, P. H., and Rushworth, M. F. (2010). Separable learning systems in the macaque brain and the role of orbitofrontal cortex in contingent learning. *Neuron*, 65(6):927–939.

Wegiel, J., Flory, M., Kuchna, I., Nowicki, K., Ma, S., Imaki, H., Wegiel, J., Cohen, I. L., London, E., Wisniewski, T., 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.

Woodruff, A. R. and Sah, P. (2007). Networks of parvalbumin-positive interneurons in the basolateral amygdala. *The Journal of Neuroscience*, 27(3):553–563.

Young, J. E., Klosko, J. S., and Weishaar, M. E. (2003). *Schema therapy: A practitioner's guide.* Guilford Press.

Zaki, J. (2014). Empathy: a motivated account. *Psychological Bulletin*, 140(6):1608.

Zaki, J., Davis, J. I., and Ochsner, K. N. (2012). Overlapping activity in anterior insula during interoception and emotional experience. *Neuroimage*, 62(1):493–499.

Zaki, J. and Ochsner, K. (2011). You, me, and my brain: Self and other representation in social cognitive neuroscience. *Social neuroscience: Toward understanding the underpinnings of the social mind*, pages 14–39.

Zaki, J., Ochsner, K. N., Hanelin, J., Wager, T. D., and Mackey, S. C. (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.

Zenasni, F., Boujut, E., Woerner, A., and Sultan, S. (2012). Burnout and empathy in primary care: three hypotheses. *British Journal of General Practice*, 62(600):346–347.

Zhang, J., Muller, J., and McDonald, A. (2013). Noradrenergic innervation of pyramidal cells in the rat basolateral amygdala. *Neuroscience*, 228:395–408.

Zimmerman, J. M., Rabinak, C. A., McLachlan, I. G., and Maren, S. (2007). The central nucleus of the amygdala is essential for acquiring and expressing conditional fear after overtraining. *Learning & Memory*, 14(9):634–644.

Appendix

Implementation Details: Izhikevich Neurons

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:

$$\frac{dv}{dt} = (1/C)(k(v - vr)(v - vt) - u + I) \tag{A1}$$

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:

$$I_j^{syn} = \sum_{i=1}^{N} s_i w_{ij} \tag{A2}$$

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. A2, 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:

$$\frac{du}{dt} = a(b(v - vr) - u) \tag{A3}$$

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

$$v(v > vpeak) = c \tag{A4}$$

$$u(v > vpeak) = u + d \tag{A5}$$

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 (RS, FS, medium spiny neuron (MSN) and IB) are given in Table A1. Parameters for RS, MSN and IB come from the models in Izhikevich and

	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

Table A1: Izhikevich parameter values for the four types of neuron (RS, FS, MSN and IB) used in our model.

Moehlis (2008), and parameters for FS neurons come from the model in Izhikevich and Edelman (2008).

Details for Regions in Empathic States Model

Table A2 details the number of neurons, neuron types, and target connection probabilities and weights for all neural groups (regions) in our model of personal distress and (weak and strong) empathic concern states in Section 3. The number of neurons is derived from human non-clinical/control estimates (based on human bilateral volume and neuron density data, averaged across age and gender when applicable) which are scaled to give 25660 total neurons in the model. Neuron types and connection probabilities are based on animal (mainly rodent) data, due to a lack of human data, resulting in over 23 million synapses. 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 main 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 (weights can be tuned in future to account for additional data, and potentially also the results of model hypotheses). Pre- and post-synaptic groups are connected with the given weight randomly and uniformly according to the specified probabilities (except for mPFC connections to mPOA and BLMA$\oplus$, which are connected according to the negative-binomial distribution with $r = 7$ and $p = 0.0025$ (Fig. A1) intended to capture a simple model of the ventral-dorsal gradient for self-other representations in the mPFC [10].

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
AI						
	AI_exc	6212 [2]	Excitatory (RS) [9]			
				AI_exc	3.259 [3]	300
				AI_inh	13.036 [3]	50
				aMCC_exc	2.601 [3]	75
				aMCC_inh	1.157 [3]	25
				mPFC_exc	2.601 [3]	90
				mPFC_inh	1.156 [3]	100
				PI_exc	2.602 [3]	95
				PI_inh	1.155 [3]	70
				BLMA⊖	10 [15]	10
	AI_inh	1553 [2]	Inhibitory (FS) [9]			
				AI_exc	12.5 [4]	200
				AI_inh	2.632 [4]	750

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

[3] 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.

[4] 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 200μm) 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 450μm). Based on both this and data for the basolateral amygdala (see footnote 7) we set local connection probability inh→exc to 12.5%, from which other connection probabilities are calculated.

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
aMCC						
	aMCC_exc	1329 [5]	Excitatory (RS) [9]			
				aMCC_exc	3.289 [3]	100
				aMCC_inh	13.168 [3]	100
				AI_exc	0.417 [3]	750
				AI_inh	0.185 [3]	300
				mPFC_exc	0.417 [3]	50
				mPFC_inh	0.185 [3]	50
				PI_exc	0.417 [3]	750
				PI_inh	0.185 [3]	500
	aMCC_inh	332 [5]	Inhibitory (FS) [9]			
				aMCC_exc	12.5 [4]	200
				aMCC_inh	2.634 [4]	400

[5] 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) .

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
BLMA⊖		76 [6]	Excitatory (RS)			
				BLMA⊖	0.855 [7]	150
				BLMAi	5 [7]	350
				AI_exc	0.019 [7]	300
				AI_inh	0.008 [7]	25
BLMAi		13 [6]	Inhibitory (FS)			
				BLMAi	5 [7]	350
				BLMA⊖	10 [7]	1750
BLMA⊕		76 [6]	Excitatory (RS)			
				BLMA⊕	0.855 [7]	60
				BLMAi	5 [7]	225
				NAc	0.674 [7]	3000
				VP	0.674 [7]	550

[6] 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 (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 (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).

[7] As in the neocortex (see footnote 3), 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 intra-connectivity 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.

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
mPFC						
	mPFC_exc	9574 [8]	Excitatory (RS) [9]			
				mPFC_exc	3.315 [3]	150
				mPFC_inh	13.261 [3]	50
				aMCC_exc	7.459 [3]	25
				aMCC_inh	3.318 [3]	50
				AI_exc	7.460 [3]	35
				AI_inh	3.316 [3]	45
				BLMA⊕	10 [15] [10]	12.5
				mPOA	10 [15] [10]	20
	mPFC_inh	2393 [8]	Inhibitory (FS) [9]			
				mPFC_exc	12.5 [4]	25
				mPFC_inh	2.632 [4]	300

[8] 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 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 10). Density is from BA10 (vmPFC) (Rabinowicz et al., 1999).

[9] 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.

[10] 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.

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
mPOA		20 [11]	Excitatory [20] (RS)			
				mPOA	10 [15]	25
				VTA	50 [15]	1750
NAc		184 [12]	Inhibitory [13] (MSN)			
				NAc	2.5 [14]	25
				VP	10 [15]	25

[11] 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, 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.

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

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

[14] Tecuapetla et al. (2007) report a connection probability of 13% for MSN pairs with intersomatic distance within 100μm. Taking into account BLMA connection probability scaling (see footnote 7), and assuming more dense local connectivity, we set the connection probability to 2.5%.

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

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
PI						
	PI_exc	3106 [16]	Excitatory (RS) [9]			
				PI_exc	3.289 [3]	460
				PI_inh	13.149 [3]	100
				AI_exc	2.439 [3]	850
				AI_inh	1.084 [3]	25
				aMCC_exc	2.438 [3]	210
				aMCC_inh	1.085 [3]	80
	PI_inh	777 [16]	Inhibitory (FS) [9]			
				PI_exc	12.5 [4]	100
				PI_inh	2.630 [4]	200
VTA		3 [17]	Dopamine (IB)			
				NAc	1.85 [18]	10

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

[17] Total number of human dopamine (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.

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

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
VP		9 [19]	Inhibitory (FS) [20]			

Table A2: 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.

[19] 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 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).

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

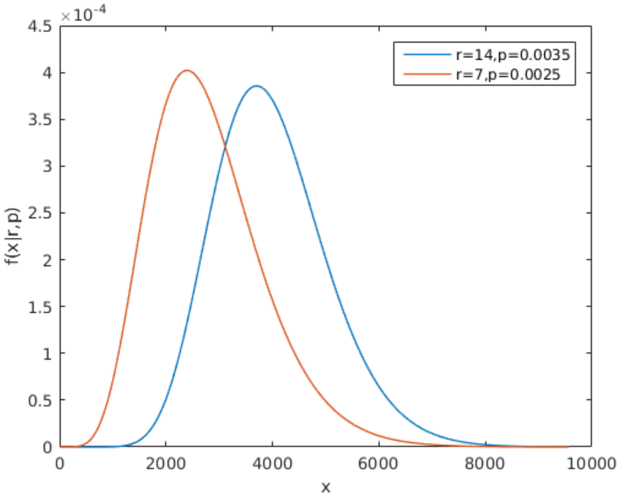


Figure A1: Negative-binomial distributions used to model close- and distant-other representations in the mPFC. The close-other distribution ($r = 7$, $p = 0.0025$) is used to connect the mPFC to the mPOA and BLMA, and to sample mPFC targets from the other-pain network encoding close-other representations. The distant-other distribution ($r = 14$, $p = 0.0035$) is used to sample mPFC targets from the other-pain network encoding distant-other representations.

Details for Additional Regions in Self-Attachment Model

Table A3 details the number of neurons, neuron types, and target connection probabilities and weights for the additional neural groups (regions) in our model of personal distress, strong empathic concern and compassion (as related to Self-Attachment) in Section 4. Details for the remaining neural groups in the model are as before (Table A2)

	Sub-Group	No. Neurons	Neuron Type	Target	Connection Probability (%)	Weight
mOFC						
	mOFC_exc	7472 [21]	Excitatory (RS) [22]			
				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 [15]	10
				BLMAi	10 [15]	35
	mOFC_inh	1868 [21]	Inhibitory (FS) [22]			
				mOFC_exc	12.5 [24]	100
				mOFC_inh	2.632 [24]	500

Table A3: 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 A2). The total number of neurons in the extended model is 35000, with over 32 million synapses.

[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 9) .

[23] We assume that the mOFC excitatory neurons follow the same neocortical connectivity profile as in footnote 3 .

[24] mOFC inhibitory neurons are assumed to follow the same neocortical connectivity profile as in footnote 4 .