Mechanisms of Change in Mentalization-Based Treatment of BPD

Peter Fonagy
University College London and The Anna Freud Centre

Anthony W. Bateman
St. Ann’s Hospital, North London, and University College London

There are very few less contentious issues than the role of attachment in psychotherapy. Concepts such as the therapeutic alliance speak directly to the importance of activating the attachment system, normally in relation to the therapist in individual therapy and in relation to other family members in family-based intervention, if therapeutic progress is to be made. In group therapy the attachment process may be activated by group membership. The past decade of neuroscientific research has helped us to understand some key processes that attachment entails at brain level. The article outlines this progress and links it to recent findings on the relationship between the neural systems underpinning attachment and other processes such as making of social judgments, theory of mind, and access to long-term memory. These findings allow intriguing speculations, which are currently undergoing empirical tests on the neural basis of individual differences in attachment as well as the nature of psychological disturbances associated with profound disturbances of the attachment system.

In this article, we explore the crucial paradoxical brain state created by psychotherapy with powerful clinical implications for the maximization of therapeutic benefit from the talking cure. © 2006 Wiley Periodicals, Inc.

Keywords: mentalization-based therapy; borderline personality disorder; attachment

Overview

Borderline personality disorder (BPD) is a complex and serious mental disorder that is characterized by a pervasive pattern of difficulties with emotion regulation, impulse control, and instability both in relationships and in self-image, with a mortality rate, associated with suicide, that is 50 times that of the general population (Skodol, Gunderson,
Pfohl, et al., 2002). The dysfunction of self-regulation is particularly apparent in the context of social relationships (e.g. Posner et al., 2002).

The regulation of emotion and the catastrophic reaction to the loss of intensely emotionally invested social ties together place borderline personality disorder in the domain of attachment. A number of theorists have drawn on N. Bowlby’s ideas in explanation of borderline pathology (Holmes, 2003). Gunderson (1996) carefully described typical patterns of borderline dysfunction in terms of exaggerated reactions of the insecurely attached infant, for example, clinging, fearfulness about dependency needs, terror of abandonment, and constant monitoring of the proximity of the caregiver. Lyons-Ruth and Jacobovitz (1999) focused on the disorganization of the attachment system in infancy as predisposing to later borderline pathology. Crittenden (1997) has incorporated in her representation of adult attachment disorganization the specific style of BPD individuals who are deeply ambivalent about and fearful of close relationships. We (Fonagy, Target, & Gergely, 2000) have also used the framework of attachment theory but emphasize the role of attachment in the development of symbolic function and the way in which insecure disorganized attachment may generate vulnerability in the face of further turmoil and challenges.

Further support for the central role of attachment in the disorder is found in the evidence that psychotherapy is the most effective treatment modality although pharmacotherapy may enhance its effects. However, the mechanism of change remains unknown. Given the strong suggestion of abnormal (disorganized) attachment processes and the consequent instability in emotions and relationships in BPD, we suggest that the mechanism mediating change indeed occurs via the improved regulation of neuropsychological systems underpinning the organization of interpersonal relationships (Bateman & Fonagy, 2004).

Mechanisms of Change: Causing Change in Causal Mechanisms

Discussing the therapeutic action of psychotherapy with BPD patients assumes that psychotherapy is indeed therapeutic. Although evidence for this claim is gradually accumulating (Roth & Fonagy, 2005), we do not know how psychotherapy effects change, and given the relatively rapid rate of spontaneous improvement (Zanarini, Frankenburg, Hennen, & Silk, 2003) in BPD, the observations of symptomatic improvements in uncontrolled studies should be considered with great caution. If we were to start from first principles, an understanding of the processes underpinning psychopathology would inform treatment innovation, which in turn would then be subject to empirical investigation. Kazdin (2004) has outlined a radical and rigorous program for psychotherapy research based on these principles. The first stage follows from the proposition that treatment should reflect what we know about the processes that directly bear on the onset and course of a clinical problem. In Kazdin’s model, demonstration that a specific process was present in a sizable proportion of individuals with a specific presentation would be the basis of treatment development. Further, rather than assuming that all individuals who have the same presentation would be equally responsive to treatment, further work would aim to detect subtypes of a dysfunction, multiple pathways to the same presentation, as well as risk and protective factors. The second step asks questions about the processes by which a treatment method achieves change, first by specifying the processes or factors responsible for change, then by developing measures of these processes, and

---

1A more fully referenced version of this article is available from the authors on request.
finally by showing that these processes change before therapeutic change occurs. On this
basis, manualization becomes feasible, on the presumption that the manual now includes
a high dosage of “effective” ingredients. Evaluation of outcome can then follow, along
with process-outcome studies that aim to examine moderator variables (helping us to
discover more about what actually does work for whom).

Following Kazdin’s proposal we shall first consider a comprehensive model of bio-
logical and psychosocial factors that are present together in BPD. Each of the compo-
nents and the evidence associated with dysfunction in the domains specified are outlined.
Second, we consider how the therapeutic technique we favor, mentalization-based treat-
ment (MBT), links to components of the proposed mechanisms underlying the pathology
of BPD. Finally we offer preliminary (inevitably somewhat speculative) ideas as to how
psychotherapeutic intervention might be seen to address these dysfunctions. In so doing
we consider psychotherapy more broadly than our own approach as it is inherently unlikely
that specific mechanisms exist for each of the therapeutic models apparently relatively
successfully used in this clinical context.

BPD as a Deficit in Attachment-Related Mentalization

Our view of BPD and attachment has undergone a number of changes as a result of
accumulating data. In 1989 we suggested that a better understanding of the features of
borderline personality disorder could be achieved if we assumed that patients who had
this diagnosis had a limited capacity to mentalize, that is, to comprehend and use their
knowledge of their own and others’ states of mind (Fonagy, 1989). Our original theory of
mentalization dysfunction suggested that the apparent inability to process mental states
effectively and appropriately in an attachment context was a defensive reaction to phys-
ical or sexual abuse, which led to a decoupling of mental processes underpinning think-
ing about feelings and thoughts in self and others (Fonagy, 1991). Later on we added that
constitutional factors were also likely to be involved (Bateman & Fonagy, 2004).

Building on accumulating evidence from developmental psychopathology, the men-
talization theory of borderline personality disorder suggests that individuals either con-
titutionally vulnerable (Torgersen et al., 2000) and/or exposed to neglect in early
relationships (Battle et al., 2004) in which their emotional experience is not adequately
mirrored by the caregiver (Crandell, Patrick, & Hobson, 2003) develop with an enfeebled
or fragile capacity to represent affect and effortlessly control their attentional capacities
(Posner et al., 2002). When confronted with trauma (Battle et al., 2004), these individuals
are more likely to react by decoupling their capacity to deal with their own or others’
mental states comprehensively, particularly in an attachment context (Fonagy et al., 1996).
We may think of this decoupling as adaptive, a deliberate avoidance of the state of mind
of the perpetrator of maltreatment. The child cannot be expected to accept fully what
must be the frankly hostile intentional stance of the abuser. It may be easier not to think
about mental states as a whole. There may be more to this than an adaptation. Early
trauma may cause changes in the neural mechanisms of arousal that lead to a relatively
ready triggering of the arousal system underpinning posterior cortical activation while
taking the frontal mentalizing parts of the brain “offline” in response to relatively mild
emotional stimuli (Arnsten, 1998).

Whatever the immediate cause of the decoupling, its consequence is the reemergence
of modes of thinking about internal states that antedate the fully fledged mentalizing
capacity of the adult. In previous work we discussed three of these that we claim are
relatively readily observable in typical patterns of thinking of individuals who have BPD:
the psychic equivalent, the pretend, and the teleological mode of representing the internal
world (Fonagy, Gergely, Jurist, & Target, 2002). All these are modes of representing subjectivity that developmentally antedate the emergence of full mentalization (Flavell & Miller, 1998; Gopnik, 1993). Children of 2 or 3 years of age tend to assume a direct correspondence between what is in their mind and what is physically true. A small child’s unshakable belief that there is a tiger under the bed is an example of psychic equivalence. At other times children can contemplate an internal reality as long as no connection is made to the real world, as in early pretend play. While psychic equivalence makes subjective experience too real, the pretend mode severs its connection with reality and at an extreme it is tantamount to dissociation. Finally the teleological mode refers to a unique dependence on what is observable in thinking about intentions. Often individuals who have BPD seek proof in physical action as part of a full confirmation of a subjective state. For example, one is being loved only if one is physically touched.

Within our theoretical frame of reference, dysfunctional attachment relationships not only are the consequence of the difficulty in holding a stable and consistent representation of others’ and one’s own mind in mind (Liotti, 2002), but also cause developmental distortions in self-organization. We have suggested a model that might explain this striking sequence. Representations of emotions cannot emerge without interaction with another person, who mirrors the infant’s experience and whose reflections of this experience are internalized by the infant. When the caregiver’s mirroring is not congruent, the infant organizes internal experience by internalizing the caregiver, rather than the caregiver’s mirroring the child, as suggested by Winnicott (1956). Thus such second-order representations of internal states are by definition “alien.” They do not match the constitutional state of the self. Consequently the self-organization evolves in a somewhat flawed manner.

A further factor, namely, trauma, contributes to the foregrounding of self-fragmentation (identity diffusion) in BPD. We suggest that traumatized individuals can use discontinuities within the self, in other words, the alien parts of the self, to adapt to incomprehensible assault from someone connected to them by attachment bonds by “identifying with the aggressor” (Freud, 1936). The tragic consequence of this manner of self-protection is the modification of self-organization so that the self now incorporates the abusive intention. This modification generates momentary experiences of unbearable psychic pain when the self feels attacked literally from within and almost overwhelmed by an experience of “badness,” which is hardly mitigated by reassurance. Experienced in the mode of psychic equivalence, the feeling of badness translates directly into “actual badness” from which self-destruction might appear the only escape. In our view, this state is commonly the trigger for acts of self-harm and suicide.

In summary we are proposing that a constitutionally vulnerable individual who experiences developmental trauma in an attachment context becomes psychologically vulnerable in later attachment contexts as a result of instability of the self. In an attempt to cope, the individual decouples the mind from others’ minds and relies on earlier psychological mechanisms to organize the experience and in doing so reveals fragments of the self.

If our suggestions are correct, the enhancement of mentalization and the reduction of the predominance of nonmentalizing modes of experiencing internal reality represent the path to cure.

The Nature of the Therapeutic Intervention and the Process of Change

Although there have been several attempts to describe mentalization-focused therapy from a psychoanalytic standpoint for children (e.g., Bleiberg, Fonagy, & Target, 1997) and adults (e.g., Fonagy et al., 2002), the first comprehensive description of the therapeutic approach is to be found in Bateman and Fonagy’s description of the partial hospital-
based treatment of BPD (see Bateman & Fonagy, 2004). There is limited but good-quality evidence for the effectiveness of the approach (Bateman & Fonagy, 1999, 2001, 2003) with further studies under way. We believe that psychotherapy has the potential to recreate an interactional matrix of attachments in which mentalization develops and flourishes. The therapist’s mentalizing in a way that fosters the patient’s mentalizing is seen as a critical facet of the therapeutic relationship and the essence of the mechanism of change. The crux of the value of psychotherapy is the experience of another human being’s having the patient’s mind in mind.

The key features of this psychotherapeutic approach may be summarized as follows: (1) The therapist is asked exclusively to focus on patients’ current mental state (their thoughts, feelings, wishes, and desires) with the aim of building up representations of internal states. (2) The therapist is asked to avoid situations in which the patient talks of mental states that he or she cannot link to subjectively felt reality. Thus there is deviation from traditional psychodynamic technique in that (a) there is a deemphasis of “deep” unconscious concerns in favor of conscious or near-conscious content and less focus on the past as it is represented in the present; (b) the aim of therapy is not insight but the recovery of mentalization: achieving representational coherence and integration for intentional states; (c) the therapist avoids describing complex mental states (such as conflict, ambivalence, unconscious) and is asked to make “small interpretations” referring to ideation that is only slightly beyond the boundaries of the patient’s conscious thinking. (3) In this way the therapy creates a transitional area of relatedness in which thoughts and emotions can be “played with.” (4) The inevitable enactments over the course of the treatment are not interpreted or understood in terms of their unconscious meaning but in terms of the situation and affects immediately before the enactment. For a fuller discussion of the key features of MBT and its similarities with and differences from other therapeutic approaches, see Bateman and Fonagy (2004).

The Biological Basis of Therapeutic Change in MBT

The Role of Neuroscience in Psychological Therapies

Returning to our original contention that change results from improvement in regulation of neuropsychological systems, we must consider the biological basis of the change caused by the therapeutic approach outlined. Although BPD is one of the most investigated of the personality disorders (PDs), its neurobiological basis is relatively unknown even though research is rapidly identifying neural correlates of complex subjective states (Adolphs, 2003), for example, concern about the mental states of another person (Frith & Frith, 2003). Natural recovery from BPD (Zanarini et al., 2003), sometimes dramatic ( Gunderson et al., 2003), and substantial treatment-related improvements (e.g., Linehan et al., 2002) suggest that the psychoneurobiological processes of BPD should relate to a reversible brain state rather than permanent abnormalities or alterations of function.

On the basis of the key dysfunctions of BPD, namely, impulsivity and interpersonal problems, there are two regions of brain function that link psychological processes of particular relevance to psychotherapy for BPD and biological processes: (1) reward circuits and attachment and (2) dysfunctions of interpersonal relatedness and mentalization. We shall briefly consider knowledge in these areas in turn and consider the links between them.

The Biology of Reward and Attachment

Our suggestion is that the poor functioning of the reward system in individuals who have BPD is linked to the dysfunction of the attachment system by their shared neural basis.
Individuals who have BPD make “impulsive choices” that result in appetitive reward in the short term, but have the potential to be self-damaging in the long run. Examples of these choices include gambling, irresponsible spending, binge eating, substance abuse, unsafe sex, and reckless driving. These impulsive choices all require the individual to assign greater value to immediate, short-term reward than to long-term rewards such as safety and security. We may think of this kind of impulsivity as characterized by actions that are poorly conceived and prematurely expressed (Daruna & Barnes, 1993).

The simplest analogues of this type of impulsive behavior have been investigated within laboratory studies (Rogers et al., 1999). Individuals who have BPD make poor estimations of likely outcomes in estimating the likelihood of monetary reward (Bazanis et al., 2002). The pattern of maladaptive performance was similar to that of individuals who had focal lesions to the orbitofrontal cortex (Rogers et al., 1999), a region previously associated with reward anticipation and valuation. A further study showed BPD patients choose short-delay small rewards over longer-delay, larger rewards (Dougherty, Bjork, Huckabee, Moeller, & Swann, 1999). Recent animal work implicates the nucleus accumbens (Nac) in temporal bias toward short-term over long-term rewards (Cardinal, Pennicott, Sugathapala, Robbins, & Everitt, 2001). This, combined with the dopaminergic reward system of the Nac’s being implicated in an animal model of impulsivity, suggests that the Nac specifically and the mesencephalic dopaminergic reward system (MDRS) in general may be involved in bias toward short-term reward and may be characteristic of the maladaptive valuation and decision making of individuals who have BPD.

The mesencephalic dopaminergic reward system has also been implicated in the process of drug addiction and the neurobiological process of attachment. The pathway of interest includes the ventral tegmental area (VTA), that projects directly and indirectly via the amygdala/bed nucleus of the stria terminalis to the nucleus accumbens. From here projections are to the ventral pallidum and from there to the thalamus. The thalamic projections go to the prefrontal and cingulate cortex, which are thought to activate cells that ultimately feed back to the VTA (Everitt & Wolf, 2002). It appears that, broadly speaking, drugs that lead to dopamine release in this system, such as psychostimulants, are addictive (Koob & Le Moal, 1997), although the neurochemical basis of addiction and substance use is probably more complicated than dopamine release (Insel, 2003). Given that it is unlikely that nature created a brain system specifically to serve drug and alcohol abuse, it seems most likely that addiction is the accidental by-product of the activation of a biological system that plays a crucial evolutionary function. MacLean (1990) speculated that substance abuse and drug addiction were attempts to replace opiates or endogenous factors normally provided by social attachments. Similarly, Panksepp (1998) suggested a common neurobiological process of mother–infant, infant–mother, and male–female attachment relationships linked to the mesocorticolimbic dopaminergic reward circuit. More recently, Insel (2003) summarized relevant data that seem to answer in the affirmative the question, “Is social attachment an addictive disorder?”

There is good evidence, reviewed by Insel (2003), that mesocorticolimbic pathways also mediate mother–infant interactions in rats: (1) dopamine is released and c-Fos (a protooncogene that encodes a 55,000-molecular-weight phosphoprotein, Fos, which is thought to assist in the regulation of “target genes” containing an activating protein-1 (AP-1) binding site), is activated in the nucleus accumbens when maternal females are exposed to their pups (Stack, Balakrishnan, Numan, & Numan, 2002); (2) lesions of the VTA and the nucleus accumbens reduce the females’ approach and interaction with pups (Lee, Li, Watchus, & Fleming, 1999); (3) cocaine or c-flupenthixol (a nonspecific dopamine antagonist) injected directly into the nucleus accumbens causes a decrease in pup retrieval (Vernotica, Rosenblatt, & Morrell, 1999); (4) Insel and his group demonstrated
that the neuropeptides oxytocin (OT) and vasopressin (AVP) are released by sociosexual experience and may serve as an important link by which parturition, copulation, and lactation can activate this reward circuit, leading to the suggestion that OT and AVP activity marks the generic reward states as specifically attachment-related (Young, Lim, Gingrich, & Insel, 2001).

There is further good evidence that the mesocorticolimbic pathways also mediate pair bonding in rodents (Insel & Young, 2000). Prairie and pine voles form partner preferences and pair bonds after mating, whereas montane and meadow voles do not form selective attachments. Thus it is of substantial interest that in prairie but not montane voles dopamine release in Nac is associated with mating (Gingrich, Liu, Cascio, Wang, & Insel, 2000) and that D2 (dopamine 2) receptor agonists infused into the Nac induce and antagonists prevent partner preferences (Wang et al., 1999). These and numerous other studies support the notion that mesolimbic dopamine activation of D2 receptors is necessary and sufficient for the development of a partner preference (in prairie voles). Vole research also suggests the hypothesis that mating releases OT and AVP, which amplify the dopamine signal in the Nac shell (Insel, 2003). We can add more recent evidence that the ventral pallidum, located within the ventral forebrain and part of the mesolimbic dopamine reward pathway, shows high density of vasopressin V1a receptors (V1aRs) in the monogamous prairie and pine voles, but not in the promiscuous meadow or montane voles (Lim, Murphy, & Young, 2004). Site-specific infusion of a selective V1aR antagonist into the ventral pallidum blocks pair bond formation in prairie voles (Lim & Young, 2004). Thus, V1aRs in the ventral forebrain appear to be crucial for pair bond formation, and this V1aR pattern seems to be correlated with monogamous social organization. Experimental evidence supporting this hypothesis is found in a single gene manipulation study (Lim et al., 2004). Partner preference formation (measured as time spent huddling with a partner as opposed to a stranger vole) in the socially promiscuous meadow vole could be increased by using viral vector V1aR gene transfer into the ventral forebrain. The overexpression of V1aR in the ventral pallidum leads to the development of partner preference in this interpersonally undiscriminating rodent. What we see here is a potential molecular mechanism for the rapid evolution of complex social behavior and the possible recreation of a singular critical evolutionary event in the laboratory.

Knowledge from preclinical (animal) models is confirmed by neuroimaging studies that demonstrate an association between functional brain activity related to attachment and cortical and subcortical sites in the human brain that contain a high density of the neurohormones OT and AVP. Imaging shows that the mesocorticolimbic dopaminergic pathway is activated while processing attachment related stimuli. Strathern and McClure (2002) studied the brain activation in mothers while viewing pictures of their own infants and compared it to the pattern of activation when they looked at infants who were familiar but not their own in three affective states (crying, smiling, and neutral expressions). They reported significant differences in the activation of the right globus pallidus–ventral striatum, the left putamen, the bed nucleus of the stria terminals, the nucleus accumbens, the amygdala, the bilateral hippocampi, and the fusiform face area (FFA). A very similar study (Nitschke et al., 2004) also reported activation in the orbitofrontal areas (OFC) in the mothers watching their own infants, which correlated with their hedonic mood state, which was also measured in the study. OFC activation is of course also observed in other studies in response to pleasant experiences: taste, smell, money, winning, positive feedback, nicotine, cocaine (Kawabata & Zeki, 2004). The most compelling recent study is that of Bartels and Zeki (2004), who, using the contrast of own versus other child controlled for age and familiarity, were able to demonstrate activity in the substantia nigra, globus pallidus, subthalamic nuclei, bed nucleus of the
stria terminalis, and ventral tegmental area, constituting almost all of the regions critical for the attachment-mediating neuropeptides in the human brain. As these workers had already reported a functional magnetic resonance imaging (fMRI) study of romantic love using a similar contrast design (Bartels & Zeki, 2000), they were able to confirm that most of the regions activated by maternal love were the same as those that they had found to be associated with romantic love, that is, those in the striatum (the putamen, globus pallidus, caudate nucleus), the middle insula, and the dorsal part of the cingulate cortex.

In summary it has been established that mesocorticolimbic dopamine is an important candidate in the mediation of reward, the capacity for deferred gratification, and addiction, but is also critical for maternal behavior in rats and pair bonding in voles. A circuit linking a vasopressin-sensitive mechanism within the anterior hypothalamus (medial preoptic area [MPOA]) to the VTA and the nucleus accumbens shell may be especially important for mediating the rewarding properties of social interaction, which are dysfunctional in BPD patients in line with their difficulties in forming normal attachments. The neuropeptides OT and AVP are released by sociosexual experience in rodents and humans. When we activate this reward circuit, a change in attachment behavior follows (at least in voles). fMRI studies indicate activation of same pathways in response to stimuli relating to the participant’s own infant and partner. We may conclude that a major neural system underpinning attachment has been identified.

Dysfunctional Interpersonal Relatedness and a Deficit in Mentalization

Disturbed interpersonal relatedness has also been identified as a key aspect of BPD relative to other PDs (e.g. Skodol, Gunderson, McGlashan, et al., 2002). This aspect can refer to a range of difficulties, including dramatic shifts from idealization of to disillusionment with others, frantic efforts to prevent perceived abandonment, and inappropriate interpersonal aggression. However, an emerging literature suggests that all of these may share common mechanistic and etiologic features. Specifically, disorganized attachment in the relationships of individuals who have BPD may mediate the expression of interpersonal problems among individuals. Individuals who have BPD have a model of relationships characterized by insecurity of attachment relative to other groups of Axis I or Axis II patients (e.g., Barone, 2003). On numerous measures of adult attachment, patients who have BPD are identified as insecure, preoccupied, and fearful in their relationships (e.g. Patrick, Hobson, Castle, Howard, & Maughan, 1994). Patients who have BPD have been characterized as having a specific type of disorganized, anxious–preoccupied attachment focused around an approach–avoidance dilemma in which the attachment figure is simultaneously perceived as a source of threat and a secure base (Crandell et al., 2003). A variety of lines of research suggest that the disordered attachment that is characteristic of BPD results from psychosocial experiences of maltreatment (e.g., Trull, 2001) and premorbid temperamental attributes of negative affectivity and aggressive impulsivity (e.g., Silk, 2000). Surprisingly, although there is strong evidence for both the psychosocial mechanism (abnormal attachment) and psychosocial cause (maltreatment) of interpersonal difficulties in BPD (e.g., Skodol, Siever, et al., 2002), a neurobiological account of disrupted interpersonal interactions remains elusive. Explorations of the subscales of a self report measure of interpersonal relatedness identified two types of problems experienced by BPD patients: one group have difficulty in achieving closeness to others and another group feel extremely submissive, unable to state needs, and avoidant of conflict (Leihener et al., 2003). At least the problems of the first group may
be related to a deficit of social cognition that makes accurate perception of the respective mental states of self and other and self–other differentiation difficult (Fonagy et al., 2000). Deficits of this aspect of interpersonal perception have been demonstrated in analogue studies using film clips (e.g., Arntz & Veen, 2001), affect recognition and alexithymic symptoms (e.g., Sayar, Ebrinc, & Ak, 2001), and narratives of childhood experience (Fonagy et al., 1996).

A deficit of interpersonal awareness implies an underlying failure to distinguish clearly between one’s own and others’ mental states. Some of the brain abnormalities identified in borderline patients are consistent with the suggestion that a failure of representation of self-states is a key dysfunction. Some evidence suggests that the anterior cingulate cortex plays a key role in mentalizing the self, at least in the domain of emotional states (Frith & Frith, 2003). Lane has proposed more specifically that implicit self-representations (i.e., phenomenal self-awareness) can be localized to the dorsal anterior cingulate, whereas explicit self-representations (i.e., reflection) can be localized to the rostral anterior cingulate (Lane, 2000). Activation of the medial prefrontal cortex has been demonstrated in a series of neuroimaging studies in conjunction with a wide range of mentalization inferences, in both visual and verbal domains (Gallagher et al., 2000). It appears that the prefrontal cortex is involved when mentalizing interactively in a way that requires implicitly representing the mental states of others. The mesial prefrontal cortex, the parieto-temporal junction, and the temporal poles constitute a network of areas that are invariably active when mentalizing activity is taking place (Gallagher & Frith, 2003). The same area of the brain is involved in other tasks that have been clinically described as challenging to patients who have borderline personality problems, including assessing social trustworthiness (Winston, Strange, O’Doherty, & Dolan, 2002), interpreting the meaning of facial expressions (Critchley et al., 2000), making moral judgments (Greene & Haidt, 2002), and performing tasks that entail attending to one’s own emotions (Gusnard, Akbudak, Shulman, & Raichle, 2001). It has been argued that exposure to stress impairs prefrontal cortical function and the impairment may be catecholamine mediated (Arnsten, 1998). In line with this suggestion is the observation that the concentration of N-acetyl-aspartate (NAA), a marker of neural integrity, is lowered in the anterior cingulated region of the medial prefrontal cortex of maltreated children and adolescents (De Bellis, Keshavan, Spencer, & Hall, 2000).

The Loose Coupling of Attachment and Mentalization

So far we have described two systems in which borderline functional deficits are evident, certainly at the behavioral level and, on preliminary neurobiological evidence, at the biological level. First, the mesolimbic dopaminergic reward system underpinning attachment may be dysfunctional in BPD. This dysfunction is suggested both by the poor ability of individuals who have BPD to delay gratification and the evident disorganization of their attachment system. The overlapping neural underpinnings of these deficits might suggest that therapeutically addressing either set of issues may indirectly benefit the other. Thus behavioral approaches, such as dialectical behavior therapy (DBT), which begin by addressing the reward systems in general and impulsivity in particular, might through such interventions also have an impact on the quality of functioning within the attachment system (although it was not part of the original aim of DBT). Equally the possibility of a common deficit underpinning these two aspects of BPD dysfunction might explain the beneficial effects that improvements in attachment relationships might generate for problems of impulsivity or decision making.
Second, the mesial prefrontal cortex, the parietotemporal junction, and the temporal poles are related to the deficits in mentalization in interpersonal interactions. The suggestion here is that a focus on mentalization addresses some of the social cognitive dysfunctions that generate inappropriate behavior in BPD, particularly suspiciousness, aggression, insensitivity to social situations, and inadequate capacity to focus attention on self-states to a degree pertinent to the interaction.

Attachment and mentalization systems are probably not independent of each other but in subtle ways might be “loosely coupled.” This suggestion is based on the finding that securely attached children are relatively precocious in the development of mentalization (Fonagy, Steele, Steele, & Holder, 1997; Meins, Fernyhough, Russel, & Clark-Carter, 1998) and the quality of parental mentalization of a child facilitates the development of secure attachment (Fonagy, Steele, Moran, Steele, & Higgit, 1991). More recently, the intersection of mentalization capacity and social experience has been more generally noted (Carpendale & Lewis, 2004), although the specific nature of the interface remains controversial.

Increased understanding of the neural mechanisms underpinning attachment may provide a further important clue. As previously mentioned, Bartels and Zeki (2000, 2004), in two separate studies, reported that maternal and romantic attachment appeared systematically to suppress brain activity in regions associated with emotionally charged memories, negative emotions, and those associated with mentalizing and social judgments. This finding suggests that strong emotional ties to an other (infant or partner) not only inhibit negative feelings but also impede the functioning of neural networks that might assist in generating social judgments about the attachment figure. These ideas are extremely important and deserve consideration in some detail.

**The Function of Networks Deactivated by the Attachment System**

Although maternal and romantic love clearly serve different functions, they share a number of subjective and objective qualities such as preoccupation, deep concern, and high level of commitment. Underpinning this may be a common set of brain mechanisms that are activated when attachment feelings are powerfully triggered but also deactivation of a characteristic set of other neural functions. Bartels and Zeki (2004) suggest grouping these reciprocally active areas into two functional regions. The first (we refer to it as system A) includes the middle prefrontal, inferior parietal, and middle temporal cortices mainly in the right hemisphere, as well as the posterior cingulate cortex. These areas are specialized for attention and long-term memory (Cabeza & Nyberg, 2000) and have variable involvement in both positive (Maddock, 1999) and negative (Mayberg et al., 1999) emotions. Their role in both cognition and emotion suggests that these areas may be specifically responsible for integrating emotion and cognition (e.g., emotional encoding of episodic memories; Maddock, 1999). In addition, studies of individuals who have lesions in these areas suggest a role in judgments involving negative emotions (Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000). It is possible that as projections from the affect-oriented limbic/paralimbic regions modulate the activity of these areas, they could subserve the ways mood can inhibit or enhance cognitive processing (Mayberg et al., 1999). Further, these areas may play a role in recalling emotion-related material and generating emotion-related imagery (Maddock, 1999) that may be relevant in relation to understanding the typology of attachment.

The second set of areas deactivated by the activation of the attachment system includes the temporal poles, parietotemporal junction, amygdala, and mesial prefrontal cortex (we
call this set of areas system B). Activation of these areas is consistently linked to negative affect, judgments of social trustworthiness, moral judgments, “theory of mind” tasks, and attention to one’s own emotions, and, in particular, they constitute the primary neural network underlying our ability to identify mental states (both thoughts and feelings) in other people (Frith & Frith, 2003; Gallagher & Frith, 2003). Mentalization pertains not just to states of mind in others but also reflecting on one’s own emotional and belief states and, consequently such tasks appear to be associated with activation in the same neural system (Gusnard et al., 2001). Making judgments that involve mental states has been shown to be associated with activation of the same system. Thus intuitive judgments of moral appropriateness (rather than moral reasoning) are linked (Greene & Haidt, 2002), as is assessment of social trustworthiness based on facial expression (Winston et al., 2002).

If confirmed by further studies, the pattern of activation of these three systems (the attachment system, system A, and system B) has important implications for our understanding of the nature of individual differences in attachment, the relationship of attachment, and mentalization and consequently our understanding of dysfunctions associated with BPD and the mechanisms underpinning its psychological treatment. The activation of the attachment system, mediated by dopaminergic structures of the reward system in the presence of oxytocin and vasopressin, inhibits neural systems that underpin the generation of negative affect (system A). This result is to be expected because a key function of the attachment system is to moderate negative affects in the infant and presumably later in development (Sroufe, 1996). The overwhelming negative affect associated with the loss of attachment figures, the need for attachment figures at times of sadness, and the hedonic effect of “finding love” are obvious common observations in line with these findings. Not only is the loss of attachment likely to be aversive because of the loss of “reward” (addiction) but also the previous inhibition of systems associated with the generation of negative affect is removed.

Equally consistent with expectations is the suppression of social and moral judgments (system B) associated with the activation of the attachment system. Judgments of social trustworthiness and morality serve to distance us from others but become less relevant to and may indeed interfere with our relationships with those to whom we are strongly attached (Belsky, 1999).

Some Implications of the Reciprocal Activation of Mentalization and Attachment

The apparently reciprocal relationship of mentalization and attachment is puzzling as, at first sight, it appears to contradict some of the assumptions concerning the facilitative relationship between the two systems outlined earlier (viz. the assumption that mentalization and attachment are positively correlated). Further scrutiny suggests a more complex relationship. First, the neural association between attachment and mentalization confirms the link we have identified between the two systems at a behavioral level (Fonagy et al., 2002). Second, we have demonstrated the way the parent’s capacity to mentalize in the context of an attachment relationship facilitates the development of secure attachment in the infant (Fonagy et al., 1991). It is possible, taking a sociobiological perspective, that the parent’s capacity to mentalize the infant or child serves to reduce the child’s experienced need to monitor the parent for trustworthiness. The relaxation of the interpersonal barrier serves to facilitate the emergence of the attachment bond. Third, we have seen that mentalization emerges precociously in children who were securely attached in infancy (e.g., Fonagy et al., 1997). At first this finding may seem inconsistent with the
inverse relationship between attachment and mentalization at brain level; however, if we consider the association developmentally, it is to be expected that in individuals whose attachment is secure, there are likely to be fewer calls over time for the activation of the attachment system, and that in turn accounts for the precocious development of mentalization.

Fourth, we have consistently suggested that the capacity for mentalization in the context of attachment was in some respects independent of the capacity to mentalize about interpersonal experiences independently of the attachment context (Fonagy et al., 2002). We have found that our specific measure of mentalization in the attachment context, reflective function (Fonagy, Target, Steele, & Steele, 1998), is predictive of behavioral outcomes not correlated with other measures of mentalization. For example, in a quasi-longitudinal study based on interviews and chart reviews with young adults some of whom had suffered trauma, we found that the impact of trauma on mentalization in attachment contexts mediated outcome measured as the quality of adult romantic relationships, but mentalization measured independently of the attachment context using the Reading the Mind in the Eyes test did not (Fonagy, Stein, Allen, & Fultz, 2003). It seems that measuring mentalization in the context of attachment measures a unique aspect of social behavior.

Mentalization in the Context of Attachment and the Classification of Attachment Types

Although from an evolutionary perspective mentalization may be generally less relevant in an attachment context than in other social contexts, nevertheless, it is quite likely that the ability to mentalize in the context of attachment relationships points to a highly desirable capacity. Individuals who are able to mentalize while thinking about romantic partners or infants are likely to manage these relationships better and may, for example, have less turbulent interpersonal relationships or perhaps be particularly effective at times of inevitable conflict and argument. This could be why secure attachment is marked by a relatively good capacity to mentalize and generate coherent narratives of even turbulent interpersonal episodes (Main, 2000). The simple empirical prediction that follows from these speculations is that individuals who are able to retain a relatively high activation of the temporal lobes, the parietotemporal junction together with the mesial prefrontal cortex (system B), in the presence of the activation of the dopaminergic mesolimbic pathways (attachment and reward system) are those most likely to be classified as secure in their attachment.

A New Understanding of the Deficit in Mentalization in BPD

We have arrived at a new understanding of the deficit of mentalization entailed in borderline personality disorder. We have observed that mentalization in individuals who have BPD frequently represented a challenge described (inappropriately) in the psychoanalytic literature as a failure of symbolization or concreteness of thinking (e.g., Grotstein, 1983). Yet mentalization in general does not appear to be a problem for most individuals who have BPD; it is only in the context of intimate relationships that the patient’s capacity to depict mental states in others accurately appears to falter (Fonagy et al., 2002).

One simple model that may account for this somewhat paradoxical observation is that the attachment system is “hypersensitive,” triggered too readily, and consequently
the reciprocally deactivated systems are often inefficient (relatively deactivated) in their functioning. Two of the core symptoms of BPD, frantic efforts to avoid real or imagined abandonment and the characteristic pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation, are perhaps directly linked to such hyperactivity or hypersensitivity (Gunderson, 2001). A number of frequently noted observations made about individuals who have BPD are consistent with this simple assumption. (1) Relationships of individuals who have BPD have a rapidly escalating tempo moving from acquaintance to great intimacy far faster than one might expect. (2) Hyperactivity of the attachment system removes the system responsible for maintaining a normal emotional barrier between self and others and generates an impression of entangled and preoccupied relationships (Agrawal, Gunderson, Holmes, & Lyons-Ruth, 2004) and frequently, somewhat unwisely, removes the need to assess the social validity of the social partner. (3) The excessively positive character of the initial phase relationships that individuals who have BPD form (often labeled idealization) may reflect the suppression of negative relationship-specific affects and the inability to integrate emotion and cognition. (4) Affective instability, particularly the characteristic intense brief episodes of dysphoria, may be the result of some form of rebound phenomena related to the hypersensitive or hyperactive attachment system. Such rebound effects might also account for outbursts of violent anger and interpersonal suspiciousness (paranoia), which might reflect overactivity in system B (Pickup & Frith, 2001). (5) The reduction of the influence of affectively laden episodic memory may relate to the chronic feelings of emptiness often encountered by individuals who have this diagnosis.

Of central concern to our theoretical and clinical propositions is the mentalization deficit we have reported in these patients (Fonagy et al., 1996). While standing by our description of the mental processes that emerge as characteristic of the intentional stance of the individual who has BPD, we would now like to specify the cause of this deficit. Previously, we have argued that the deficit was a self-induced one—a defensive reaction in vulnerable individuals when confronted with hostile states of mind in the context of interpersonal trauma. Although many of the extant data still appear to us to fit with this model (e.g., the association of early neglect with BPD, the undermining of the development of symbolization in the families of maltreated youngsters, the high prevalence of attachment trauma, the high prevalence of individuals who show no BPD symptoms after trauma), it seems highly likely that mentalization deficit can be secondary to the abnormal functioning of the attachment system (i.e., its hyperactivity). The latter of course is likely to be the consequence of developmentally early dysfunctions of the attachment system in combination with later traumatic experiences in an attachment context. We speculate that deficits of mentalization might sometimes occur because the capacity to mentalize is taken offline by the activation of the attachment system. Of course, this is not an “all or nothing affair.” We assume that in these individuals, who are insecurely attached and therefore would not be able to maintain mentalization normally in any case in the context of attachment relationships, the hyperresponsiveness of the attachment system, perhaps related to traumatic or other early experiences or genetic predisposition, has an unusually negative impact upon mentalizing with the expected effects already described. It remains probable that activating the capacity to mentalize in the context of attachment relationships generates substantial anxiety for traumatized individuals, which in turn increases the activation of the attachment system in even those who have an avoidant-dismissing pattern of attachment. Of course here we have a potentially extremely vicious cycle of heightened attachment, increasingly decoupled mentalization, and increased vulnerability to further interpersonal trauma. A number of testable hypotheses follow from this model. Predictions from this model include the following: (1) Mentalization...
dysfunctions should be observable only when the attachment system is active; (2) mentalization dysfunction should be associated with negative affect; (3) problems of accurate social and moral judgments should correlate with mentalization capacity; (4) the degree of disorganization of attachment relationships should correlate with the likelihood of mentalization problems; (5) there are likely to be deficits associated with the retrieval of emotion-laden memories when the attachment system is active; (6) ambiguous stimuli (e.g., polysemous words with attachment and nonattachment meanings) are more likely to trigger the attachment system of BPD individuals.

Implications for Therapy and the Mechanisms of Change

How can this model of the nature of deficit in BPD serve to focus our work with BPD patients, or rather how do we understand the changes we observe given the current focus of our work, which is to assist in the recovery of the capacity to think accurately about thoughts or feelings? There is an important proviso to this aim. MBT in a range of contexts attempts to enhance mentalization but always in the context of an attachment relationship. Both in individual and in group therapy the therapist through a range of largely unconscious techniques activates the attachment system. This activation occurs through (1) the discussion of current attachment relationships, (2) the discussion of past attachment relationships, (3) the therapist’s encouragement and regulation of the patient’s attachment bond to him or her through the creation of an environment that assists with the patient’s regulation of affect, (4) the therapist’s attempt to engender attachment bonds between members of the group in the context of group therapy. At the same time, paradoxically, the therapist attempts to enhance mentalization not just in the techniques defined in the therapy manual but perhaps more importantly and generically simply by his or her interest in the mental world of the patient. This interest creates what we now understand as a somewhat paradoxical situation in terms of brain activity insofar as psychological therapy simultaneously activates what may be two mutually inhibitory sets of systems. There are two other ways in which this somewhat paradoxical pattern of activation is maintained: (1) the titrated activation of negative emotions as the therapist encourages the confrontation of adverse or traumatic experiences and (2) the encouragement to retrieve affect-laden episodic memories. In these ways the individual who has BPD is encouraged to counteract the normal pattern of attachment related deactivation of mentalizing of negative emotions and social and moral judgments.

We can see that overall MBT encourages the patient not to relinquish mentalization at the slightest suggestion of attachment-related brain activation. This is likely to have an impact on the attachment system as well because we have seen that mentalization strengthens the development of secure patterns of attachment. In other words, at the same time as strengthening mentalization we speculate that MBT moves the pattern of arousal within these systems closer to that characteristic of a secure attachment. Evidently, to achieve this aim the therapist must be careful to balance the intensity of attachment relationships and the complexity of mentalization required of the patient. The technical recommendation in MBT is to focus the patient’s mentalization on relationships that have relatively low levels of involvement and only gradually to focus the patient’s thinking on relationships closer to the patient’s core self. Similarly the tasks of mentalization vary in demand characteristics, with clarification at the most superficial end and exploring of repudiated intense emotions in relation to the attachment figure at the more complex end. The therapist’s aim is to reduce the likelihood of an anxious catastrophic response to the introduction of the need to think about the states of mind of attachment figures.
Thus using this model may help us to differentiate it from alternative therapeutic packages. Although such distinctions are never more than caricatures, they do point to important differences in technique between therapeutic approaches that are probably broadly equivalent in terms of achieving substantial therapeutic change. Thus aspects of DBT clearly also focus on the enhancing of mentalization, such as, the encouragement of mindfulness. However, this process rarely occurs in the context of requiring individuals to conceptualize mental states in the context of attachment relationships. By contrast, supportive psychotherapy is less specifically focused on mental states but probably attempts to reduce the hypersensitivity of the attachment system of an individual who suffered severe attachment trauma. Transference-focused psychotherapy is perhaps closest to MBT in its orientation to mental states in attachment contexts; the difference is a far more limited emphasis on the titration of interventions. The observations reported by Levy, Clarkin, and Kernberg (Levy & Clarkin, in press) are consistent with the propositions here: attachment-related mentalization (reflective function) improves only in transference focused psychotherapy (TFP) and not in DBT or supportive psychotherapy.

Implications for the General Process of Change in Therapy

The therapeutic process we have described in relation to MBT with BPD is probably not specific to either this approach or this patient group. A strong case can be made that all forms of psychotherapy take some advantage of the simultaneous activation of these normally mutually inhibitory systems. It is possible that psychotherapy in general is effective because it arouses the attachment system at the same time it applies interpersonal demands (psychotherapy technique), which require the patient to mentalize, to confront and experience negative affect, and to confront and review issues of morality (superego). Why might this be helpful? We speculate that thinking about feelings, thoughts, and beliefs in the context of attachment is helpful because in this “paradoxical” brain state there may be more access to modifying preset ways of conceptualizing the contents of one’s own and other’s minds, as well as issues of morality and social judgment. Activating the attachment system harnesses brain biological processes partially to remove the dominance of constraints on the present from the past (long-term memory) and creates the possibility of rethinking, reconfiguring intersubjective relationship networks. The specific advantage of MBT in this process may be its focus on the simultaneous activation of the attachment system and encouragement of development of psychological processes that are normally inhibited as a result. To this extent MBT represents a confluence of biology and psychology and goes some way toward meeting Kazdin’s edict that an understanding of the processes underpinning psychopathology should inform treatment innovation.

References


Bazanis, E., Rogers, R.D., Dowson, J.H., Taylor, P., Meux, C., Staley, C., et al. (2002). Neurocognitive deficits in decision-making and planning of patients with DSM-III-R borderline personality disorder. Psychological Medicine, 32(8), 1395–1405.


