Neural correlates of fears of abandonment and rejection in borderline personality disorder [version 1; peer review: 2 not approved]

Liana Romaniuk, Merrick Pope, Katie Nicol, J. Douglas Steele, Jeremy Hall

1 Division of Psychiatry, University of Edinburgh, Edinburgh, UK
2 Medical Research Institute, University of Dundee, Dundee, UK
3 Neuroscience and Mental Health Research Institute, Cardiff University, Cardiff, UK
4 MRC Centre for Neuropsychiatric Genetics and Genomics, Cardiff University, Cardiff, UK

Abstract

Background: Borderline personality disorder (BPD) is a prevalent and disabling psychiatric condition commonly associated with early life adversity. Social difficulties are a prominent symptom of BPD, particularly a fear of abandonment and rejection. There has recently been a growing interest in the neural basis of these social symptoms and their relationship to early experience.

Methods: In the current study, we examined social brain function and learning in BPD using functional MRI. Participants with BPD (n=20) and healthy controls (n=16) completed a computerized parametric social exclusion task (the “Cyberball” task). Brain activation was compared between groups and related to social symptom status and experiences of childhood trauma. Additional analyses were conducted using a reinforcement learning model treating social inclusion as a rewarding event.

Results: Participants with BPD demonstrated a group effect of decreased right temporoparietal junction (TPJ) activation (p < 0.013, FWE-corrected). Increased fear of abandonment in BPD was associated with reduced inclusion-related activation of the inferior frontal gyrus (p = 0.003, FWE-corrected). Across all participants, TPJ inclusion-related activation was modified by prior experience of childhood physical neglect (p < 0.001, FWE-corrected). Reinforcement learning modelling revealed decreased midbrain responses to social inclusion in BPD participants (p = 0.028, FWE-corrected within midbrain mask), with decreased anticipatory midbrain activation in anticipation of social inclusion specifically associated with fears of abandonment (p = 0.019, FWE-corrected within a midbrain mask).

Conclusions: The findings demonstrate alterations in social brain function and social reinforcement learning in BPD, which are...
influenced by both early life experience and symptom status.

**Keywords**
Borderline personality disorder, functional neuroimaging, social rejection, childhood trauma, inferior frontal cortex, temporoparietal junction, midbrain, reinforcement learning
Introduction

Borderline personality disorder (BPD) is a common and serious disorder associated with significant impact on the individual, their families and a wide range of healthcare providers\(^1\)\(^-\)\(^4\). Social difficulties are a core feature of BPD\(^5\), and include marked fears of abandonment and social rejection\(^5\)\(^-\)\(^7\). Indeed these factors underlying their characteristically disturbed relationships have been proposed to be a phenotype of this personality disorder\(^8\). Both genetic and early environmental factors have been shown to contribute risk for developing BPD symptoms\(^9\)\(^-\)\(^11\), with evidence that early environmental factors, such as childhood adversity\(^12\), play a particularly prominent role in the development of BPD\(^12\) and interpersonal symptoms in particular\(^3\)\(^,\)\(^10\)\(^,\)\(^13\).

Fear of abandonment is one of the core symptoms of BPD recognized in both Diagnostic and Statistical Manual of Mental Disorder (DSM)-\(^5\) and International Statistical Classification of Diseases and Related Health Problems (ICD)-\(^10\). Fearing abandonment is one of the most stable features of the disorder, and is associated with negative core beliefs\(^13\)\(^-\)\(^15\) and an increased propensity for self-injurious behavior\(^16\)\(^-\)\(^18\). The central nature of this symptom is further confirmed by formal testing, in which people with BPD showed heightened sensitivity to rejection and abandonment in questionnaire measures\(^19\), and increased psychophysiological reactions to scripts relating to abandonment\(^20\).

The marked fear of abandonment seen in people with BPD is reflected in the disorganised attachment patterns observed in these patients\(^21\)\(^-\)\(^23\). Abnormalities in the development of secure attachment patterns are considered to underlie the development of persistent fear of rejection seen in people with BPD\(^1,\)\(^2,\)\(^3\). The development of stable attachment relationships requires the experience-dependent maturation of specific brain circuitry mediating key aspects of social function and learning, such as mentalizing\(^2\)\(^,\)\(^3\). Such regions include the superior temporal cortex, temporoparietal junction (TPJ), and insula and inferior frontal gyrus (IFG)\(^2\). Theoretical analyses of BPD have suggested that altered social learning during development, potentially secondary to childhood adversity, may lead to the emergence of key symptoms of BPD, such as fear of abandonment\(^2\)\(^-\)\(^4\).

Within BPD, several imaging studies have assessed the processing and regulation of emotion, self-injury, cognitive disturbance and impulsivity (see reviews 26,27). However, there has been an increasing interest in imaging the neurobiology mediating interpersonal difficulties. The behavioural and neural correlates of social rejection and social learning can be studied experimentally using computer games\(^28\)\(^,\)\(^29\). A commonly applied task is the “Cyberball” game, in which the participant takes part in a computerised ball-throwing game with two on-screen figures, where the degree to which the participant is included or excluded from the game is systematically varied\(^29\). Imaging studies using the Cyberball task have revealed a network of brain regions that are sensitive to the degree of inclusion and exclusion, including frontal and temporal brain regions implicated in social cognition\(^10\)\(^-\)\(^13\). It has also been shown that social inclusion events produce reward learning signals in a range of brain regions, including the superior frontal cortex and posterior temporal areas that are known to be involved in social behavior\(^31\)\(^,\)\(^34\).

In the current study, we built on previous studies of BPD using the Cyberball task\(^31\)\(^,\)\(^33\)\(^-\)\(^36\), Ruocco et al. demonstrated that people with BPD demonstrate increased medial prefrontal cortex (PFC) activation during social exclusion\(^36\). Using functional MRI (fMRI), Domsalla and colleagues found that people with BPD exhibited generalised hyperactivation of dorsal medial PFC, anterior cingulate, precuneus and superior/inferior parietal lobules, and that this activity was not modulated by social inclusion, as was seen in controls\(^37\). We extended this work by examining the relationship between Cyberball brain activation and (a) BPD symptoms of social difficulty, named fears of abandonment and unstable relationships; and (b) reported early life adversity. We also analysed the results in terms of social reward learning models and demonstrate altered reward learning in BPD consistent with previous models of the disorder\(^38\). We hypothesised that within people with BPD, social inclusion-evoked brain activation would demonstrate associations with previous childhood adversity, and their current interpersonal symptoms. These associations would likely be demonstrated in areas mediating mentalizing, namely the superior temporal cortex, insula and IFG.

Methods and materials

Participants

A total of 20 people with borderline personality disorder (BPD) were recruited from outpatient and support services from around Edinburgh, Scotland. Diagnoses were confirmed using the Structured Clinical Interview for DSM-IV (SCID-II; http://www.scid4.org/). Current symptoms were assessed using the Zanarini Rating Scale for Borderline Personality Disorder (ZAN-BPD)\(^1\). Adverse childhood events were assessed using the Childhood Trauma Questionnaire (CTQ)\(^1\). In total, 15 BPD participants were receiving antidepressant medication and 12 were taking antipsychotic medication. A total of 20 age- and sex-matched controls were recruited from the community; however, 4 were excluded due to technical issues during scanning, leaving 16 controls. Exclusion criteria for all participants included pregnancy, MRI contraindications, diagnosis of a psychotic disorder, previous head injury or current illicit substance dependence. Controls met the additional criteria of no personal or familial history of major mental illness. Ethical approval was obtained from the Lothian National Health Service Research Ethics Committee (09/S1101/49), and all participants provided written informed consent before taking part. Demographic details are given in Table 1.

Experimental task

Participants performed the Cyberball social exclusion task\(^31\) during fMRI, adapted from a previous implementation by Kumar et al. 2009\(^31\). The task involves playing “catch” with two computer-controlled players, during which the participant can be systematically included or excluded from the game. We used this task as it assesses neural responses to social exclusion, is known to activate a range of social brain regions\(^39\), and is amenable to reinforcement learning modelling\(^31\). The task was modified such that inclusion was varied parametrically over four levels: 0%, 33%, 66% and 100%, achieved by arranging the task into blocks of nine throws, respectively involving zero, one, two or three throws to the participant (Supplementary material; Supplementary Figure 1). Here, 100% inclusion means the degree to which the participant was included
| Table 1. Study participant demographics. CTQ, Childhood Trauma Questionnaire; ZAN-BPD, Zanarini Rating Scale for Borderline Personality Disorder; PTSD, post-traumatic stress disorder; OCD, obsessive compulsive disorder. |
|---|---|---|---|
| Demographics | Borderline personality disorder (n = 20) | Healthy volunteers (n = 16) | P value |
| Age (Mean [SD]) | 35.8 (8.6) | 34.8 (9.6) | 0.97 |
| IQ (Mean [SD]) | 114.8 (7.9) | 114.5 (6.0) | 0.58 |
| Gender (F:M) | 17:3 | 13:3 | 0.76 |
| CTQ (Mean [SD]) Range 1–5 | | | |
| Physical abuse | 2.51 (1.45) | 1.09 (0.24) | 0.002 |
| Physical neglect | 2.14 (1.11) | 1.28 (0.32) | 0.010 |
| Emotional abuse | 3.79 (1.10) | 1.25 (0.41) | <0.001 |
| Emotional neglect | 3.35 (1.36) | 1.54 (0.77) | <0.001 |
| Sexual abuse | 3.14 (1.74) | 1.28 (0.32) | 0.001 |
| ZAN-BPD | Total | 13.7 (6.7) | 0.0 (0.0) | <0.001 |
| Medication | | | |
| n | % | |
| Antipsychotics only | 2 | 10 |
| Antidepressants only | 5 | 25 |
| Both | 10 | 50 |
| Comorbid Diagnoses | | | |
| n | % | |
| Total | 17 | 85 |
| Depression | 10 | 50 |
| Anxiety | 6 | 30 |
| Eating disorder | 4 | 20 |
| PTSD | 3 | 15 |
| OCD | 1 | 5 |
| Bipolar type 2 | 1 | 5 |

was equal to that of the other two players, with each receiving three throws per nine-throw block. Mean block duration was 24s (in part dependent on participants’ reaction times), with onsets denoted by the appearance of the cartoon figures following rest, and offsets by the conclusion of the final throw animation. Blocks were randomized, and interleaved with 13s rest blocks. Within blocks, throwing events were jittered to permit event disambiguation for reinforcement learning analysis. Reaction times were analysed within a repeated measures ANOVA (within-subjects: inclusion level; between-subjects: group). Response bias, that is the propensity to favour one computer player over the other, was also assessed (Supplementary methodology; Supplementary Results).

Neuroimaging

Full neuroimaging methods are provided in the Supplementary methodology. In summary, fMRI data was acquired at 3T, with TR 1560ms, 347 volumes, resolution 3.4 × 3.4 × 5mm. A T1 structural image was acquired with resolution 1mm isotropic. Data were preprocessed and analysed using SPM8 (Wellcome Department of Imaging Neuroscience, London, England; http://www.fil.ion.ucl.ac.uk/spm/software/spm8/) within MathWorks Matlab R2014a (http://www.mathworks.com), including its normalisation to MNI space, and smoothing using an 8mm FWHM Gaussian kernel.

At the first-level, each level of inclusion was modelled using separate regressors. Motion parameters were included as nuisance regressors. T-contrasts were generated for each inclusion level > rest, and taken into a random-effects second-level full factorial design matrix, modelling inclusion and diagnostic group. Both the main effects and the interaction were evaluated.

The relationships between current social symptoms and fMRI response to increasing inclusion were examined within a second level multiple regression of the first-level parametric contrast of increasing inclusion (100% > 66% > 33% > 0%), by the ZAN-BPD Total Disturbed Relationships sub-score items. Focusing on social interaction, this comprises items six (frantic efforts to avoid real or imagined abandonment; Z6) and nine (unstable and intense interpersonal relationships; Z9). It has been theorized that several characteristics of borderline psychopathology, especially those concerning interpersonal relationships, are secondary to changes
in self-other representations. Any assessment of one’s relationships with others implicitly involves the consideration of self, and could therefore be modulated by its disturbance, outside the boundaries of any one particular personality disorder. Therefore, we also included ZAN-BPD item four (identity disturbance: markedly and persistently unstable self-image or sense of self; Z4) in this single symptom regression analysis. The impact of past childhood adversity on fMRI responses to increasing inclusion were also assessed within a separate second level multiple regression incorporating the subscales of the CTQ (physical/emotional abuse/neglect, and sexual abuse).

We also adopted a reinforcement learning approach, examining how those with BPD learn to anticipate the rewarding outcome of social inclusion. Reinforcement learning models describe mathematically the learning processes that occur during conditioning and reward learning and closely relates to the observed changes in dopaminergic neuron firing seen in animals. Previous studies have demonstrated that reinforcement learning approaches are applicable to social learning tasks, including the Cyberball task and social exclusion.

A temporal difference reinforcement learning model represented the task as an instance of classical conditioning, where those moments when the ball was received by the participant being modelled as a rewarding unconditioned stimulus (outcome), and moments when the ball was held by the computer-controlled characters as a conditioned stimulus (anticipation). fMRI regressors describing the anticipatory and outcome phases of each ball throw were modulated by a parametric regressor modelling the prediction error at these moments, using MathWorks Matlab R2014a. The outcome phase represents trial-by-trial responses to inclusion (reward=1) and exclusion (reward=0), whereas anticipation captures learning regarding potential inclusion (see Supplementary methodology for more information). As well as group comparisons, patients’ contrasts were also regressed by the three social symptoms (Z4, Z6, Z9).

For all fMRI analyses, a voxel-wise threshold of $p < 0.001$ was applied, and only those regions additionally achieving a cluster-wise FWE-corrected significance of $p < 0.05$ were reported. As temporal difference models have been shown to provide a good account for midbrain activation during reward learning, the reinforcement learning analyses were corrected within an anatomically-defined mask comprising the substantia nigra and ventral tegmental area (Supplementary methodology).

**Results**

**Behaviour and psychopathology**

There were no performance differences between the groups as assessed by reaction times across the task ($F(1,34) = 0.21; p = 0.65$), indicating matched engagement with the task (Supplementary Figure 2). Participants with BPD had significantly greater Zanarini and CTQ scores than controls, as expected (Table 1).

**Main effect of inclusion on brain activation**

Examining the main effect of increasing inclusion on brain activation across the groups revealed robust activation of several brain regions known to be engaged during social cognition, including the IFG, inferior parietal lobule/TPJ, anterior cingulate cortex, caudate and thalamus (Supplementary Figure 3).

**Main effect of group on brain activation**

Across the task as a whole, participants with BPD showed lower activation of the right TPJ and left middle occipital cortex bilaterally (Figure 1).

**Figure 1. Controls > BPD across all levels of inclusion.** The behavioural personality disorder (BPD) group showed diminished temporoparietal junction activation across the Cyberball task as a whole (Table 2). Activation map displayed at a voxel-wise uncorrected threshold of $p < 0.001$; however, only clusters achieving a FWE-corrected significance of $p < 0.05$ are reported. Bar chart demonstrates R temporoparietal junction activation (TPJ) at each level of inclusion for both groups.
levels of reported childhood physical neglect were associated with a decreased inclusion response within bilateral superior temporal cortex, and middle anterior cingulate cortex (Table 2; Figure 2). These effects remained significant even when restricted to the BPD group only (p < 0.002). No effects were seen for the other CTQ subscales or total score.

**Group x inclusion interaction**
No regions showed a significant group by inclusion interaction.

**Relationship of childhood trauma to brain activation**
We next investigated whether experiences of childhood trauma correlated with neural responses to increasing inclusion. Higher levels of reported childhood physical neglect were associated with a decreased inclusion response within bilateral superior temporal cortex, and middle anterior cingulate cortex (Table 2; Figure 2). These effects remained significant even when restricted to the BPD group only (p < 0.002). No effects were seen for the other CTQ subscales or total score.

**Table 2. Functional imaging activation statistics.** Regions showing significant relationships between the inclusion response, and either Zanarini Rating Scale for Borderline Personality Disorder (ZAN-BPD) or Childhood Trauma Questionnaire (CTQ) measures, reported in MNI space. The latter two analyses concerning fear of abandonment and physical neglect were performed within the behavioural personality disorder (BPD) group only.

<table>
<thead>
<tr>
<th>Region</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>kE</th>
<th>Peak Z score</th>
<th>P value (cluster extent FWE-corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Controls &gt; BPD across all inclusion levels</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L middle occipital cortex</td>
<td>-30</td>
<td>-91</td>
<td>10</td>
<td>201</td>
<td>4.52</td>
<td>0.008</td>
</tr>
<tr>
<td>R temporoparietal junction</td>
<td>51</td>
<td>-37</td>
<td>13</td>
<td>128</td>
<td>4.84</td>
<td>0.037</td>
</tr>
<tr>
<td><strong>Fear of abandonment (Zanarini 6) x decreased inclusion activation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R inferior frontal cortex</td>
<td>39</td>
<td>29</td>
<td>10</td>
<td>3729</td>
<td>5.32</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Sub-peaks within this cluster</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L postcentral cortex</td>
<td>-30</td>
<td>-34</td>
<td>52</td>
<td>3729</td>
<td>5.07</td>
<td></td>
</tr>
<tr>
<td>L superior temporal cortex</td>
<td>-48</td>
<td>-28</td>
<td>-2</td>
<td>128</td>
<td>4.68</td>
<td></td>
</tr>
<tr>
<td><strong>Physical neglect (CTQ) x decreased inclusion activation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L temporoparietal junction</td>
<td>-39</td>
<td>-22</td>
<td>19</td>
<td>450</td>
<td>4.31</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>R middle cingulate cortex</td>
<td>6</td>
<td>-10</td>
<td>46</td>
<td>195</td>
<td>4.31</td>
<td>0.002</td>
</tr>
<tr>
<td>R insula cortex</td>
<td>45</td>
<td>-4</td>
<td>16</td>
<td>346</td>
<td>4.27</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Figure 2.** Associations between brain responses to inclusion, fear of abandonment and past physical neglect. Red indicates areas having a negative association between Zanarini item 6 (fear of abandonment) and the response to increasing inclusion. This is restricted to the behavioural personality disorder (BPD) group, as controls had no variance for any Zanarini item. Blue indicates areas having a negative relationship between Childhood Trauma Questionnaire (CTQ) physical neglect, and increasing inclusion. This incorporates both groups. (a) Sagittal slice at MNI X = 39, highlighting R inferior frontal gyrus and insula. (b) Sagittal slice at MNI X = -48, highlighting L superior temporal and temporoparietal junction structures. Maps are displayed at an uncorrected voxel-wise threshold of p < 0.001; however, only clusters with a FWE-corrected significance p < 0.05 are reported.
Relationship of social symptoms in BPD to brain activation
Analysis of the effect of social symptoms of BPD on brain activation was restricted to the BPD group, due to a floor effect for these symptoms in the control group. This analysis revealed that increasing fear of abandonment was associated with a reduced response to inclusion within the right IFG, and left postcentral gyrus within the BPD group (Table 2; Figure 2).

Temporal difference modelling of inclusion events as rewarding
We next used temporal difference modelling to examine neural responses to individual events of inclusion and exclusion, with inclusion events treated as rewarding. Compared to controls, participants with BPD showed significantly less activation in the midbrain during trials on which they were included, suggesting that their experience of inclusion was not imbued with the same salient or rewarding character as that felt by controls (Figure 3a; MNI -9 -13 -14; Z = 3.26; p = 0.028, FWE-corrected within the midbrain mask). There was no association between midbrain outcome responses and social symptom status, suggesting that decreased midbrain response to inclusion may represent a stable feature of the condition. We further examined brain activation during the anticipation phase of each trial. In participants with BPD, increasing fear of abandonment was associated with reduced midbrain activation, suggesting a diminished ability to learn to anticipate inclusion (Figure 3b; MNI 6 -13 -14; Z = 3.55; p = 0.019, FWE-corrected within the midbrain mask). There were no associations between midbrain activation and childhood adversity.

Correlation with medication and depressive symptoms
Neither depressive symptoms nor medication status correlated with activation of the R IFG, TPJ or midbrain (Supplementary results; Supplementary Figure 4).

Discussion
In healthy individuals, social inclusion in the Cyberball task activated a network of brain regions including the inferior frontal gyrus (IFG) and temporoparietal junction (TPJ). Participants with BPD demonstrated a number of differences in brain activation during the Cyberball task. Firstly, participants with BPD showed reduced activation of the TPJ across all levels of inclusion. Secondly, individuals with BPD showed altered activation of the right IFG in relation to a key symptom of the disorder - fear of abandonment. Thirdly, TPJ, insula and anterior cingulate cortex (ACC) responses in both groups were modified by prior experiences of childhood physical neglect. Finally, participants with BPD showed a decreased reward response to social inclusion in the midbrain, and decreased anticipatory midbrain responses to social inclusion in relation to symptoms of fear of abandonment. Overall these results demonstrate sustained alterations in the TPJ and midbrain in BPD with additional modulation of RIFG function in relation

Figure 3. Altered midbrain responses in the BPD group. (a) Prediction error signal, controls > behavioural personality disorder (BPD), displayed at p < 0.001 uncorrected. (b) The relationship between the R inferior frontal gyrus (IFG) response to increasing inclusion, the anticipatory response of midbrain to inclusion, and fear of abandonment.
to the key symptom of fear of abandonment. Differences in TPJ function were in part related to adverse childhood experience.

There was a main effect of group within the right TPJ, with participants with BPD demonstrating reduced activation. The difference was not attributable to behavioural performance, which was matched across groups; however there was no significant task x group interaction in this region, meaning this could not be firmly attributed to altered responses to inclusion per se. Notably these findings are consistent with those reported in a previous study of BPD. However, the TPJ is known to play a key role in social behavior and has been implicated in social reward learning. The roles of posterior temporal regions in social function are complex, but are believed to sub-serve core components of reciprocal social interaction, imitation and relating self to others. The TPJ has in particular been implicated in considering the internal states of others, and in learning social rules about how much to trust another individual. Individuals with BPD can have difficulty in developing trust in relationships and show a tendency to rate others as untrustworthy on formal testing. They have also demonstrated reduced activation in nearby posterior superior temporal structures during a cognitive empathy task and attribution of emotional empathy tasks. Decreased activation of TPJ in individuals with BPD is thus likely to reflect an impaired ability to predict and trust other’s social intentions.

We also found that in BPD activation of the right IFG, a core component of the social brain network, was related to the symptom of fear of abandonment. The IFG is known in both primates and humans to contain neurons that are involved in social learning and show activation which “mirrors” responses seen in others, in particular encoding the intention of the other’s action. This role of the IFG in interpreting and responding to the actions of others extends beyond motor actions to recognizing and empathizing social and emotional responses. As such this region has been considered to form a key part of the brain mechanisms required for reciprocal social learning during development, developing an “internal working model” of social interaction, which governs expectations about the outcomes of social encounters. It is known that people with BPD demonstrate a bias in attributing malevolent intentions to others, and reduced IFG activation when attributing emotional intentions.

We additionally sought to determine whether experience of childhood adversity might contribute to the development of altered social brain responses across both groups. Regression of childhood trauma questionnaire scores against brain activation during the Cyberball task revealed that greater experiences of physical neglect during childhood were associated with reduced responses to social inclusion within left TPJ, right insula, and ACC. Physical neglect represents a definitive breakdown of normal reciprocal social interactions during childhood. These results therefore support the view that impaired nurturing from caregivers during childhood alters one’s understanding of the likely outcomes of social interactions. As the negative relationship was also seen in insula, this suggests that physical neglect also impacts on the normal development of one’s interoceptive sensations in the context of social interactions. Correlations with brain activation during the Cyberball task were not seen for the other four CTQ measures. This could however be due to a lack of power to detect an effect in this sample or because of resistance – conscious or otherwise - to the recall of physical or sexual abuse.

The IFG and TPJ have both been implicated in mediating the brain’s responses to the rewarding nature of social inclusion and interaction. Specifically these regions show temporal patterns of activation during social learning, which are consistent with those seen to a range of other rewarding stimuli. The underlying source of these reward-related brain signals is believed to be the firing of dopaminergic neurons in midbrain regions, which have been shown to be phasically responsive to reward learning in animal models and in human imaging studies. Here we examined the response of the midbrain to social inclusion events in BPD by modelling inclusion events as rewarding. This analysis showed that participants with BPD showed decreased midbrain activation to the reward of social inclusion than control participants, suggesting that people with BPD have a diminution of positive reward responses to social experience. In addition, participants with BPD with greater fear of abandonment also showed decreased activation of the midbrain in response to the anticipation of social inclusion. Phasic dopamine release has been argued to be more representative of the expectation-outcome mismatch i.e. reward prediction error, rather than simple hedonistic reward itself. It has been suggested that the pervasive rejection felt by people with BPD – even when being fully included – may be due to people with BPD biasing their perceptions to conform to their expectations, in which case, a dopaminergic “prediction error” signal is unlikely to be generated. This apparent selective blindness to inclusion in people with BPD is additionally supported by evidence from EEG studies. Indeed it seems that only during conditions of extreme over-inclusion do people with BPD report reduced negative emotions during social interaction.

A number of important limitations to the current study should be noted. Firstly, we did not directly measure the participants’ subjective experience of the different levels of social inclusion during the task. Previous behavioural studies have shown that individuals with BPD show greater feelings of exclusion during the Cyberball task. However we chose not to directly assess responses to the changing task contingencies in the present study, as doing so would necessarily interrupt the flow of the task and alert the participants to the covert changes in the degree of inclusion, whilst post-task analysis would be confounded by the differing degrees of inclusion/exclusion present across the task. Secondly, participants with BPD in the current study had a number of comorbid conditions and many were on some form of medication. However, we chose to include participants with typical outpatient presentations of BPD, increasing the generalisability of our findings. Notably, we did not identify any consistent relationship between the differences in social brain activation seen in the study and the most common comorbid conditions, such as depression, or treatment with either antidepressant or antipsychotic medications. Thirdly, our results appear to differ from those of Kumar et al., who also adopted a reinforcement learning approach to Cyberball data, but did not find TD-associated...
midbrain activation. This could be because their data was acquired at 1.5T or because the randomized nature of our inclusion blocks induced more abrupt violations of expectation. Our studies are qualitatively comparable, although we introduced a degree of jitter in considering the anticipation of inclusion, whereas their implementation was fixed, albeit on a similar time scale.

**Conclusions**
The current findings illuminate the neural basis of key social symptoms in borderline personality disorder and their relationship to developmental experience. We show that participants with BPD have differential activation of brain regions mediating social interaction, including the TPJ and midbrain. TPJ activation was also found to be related to childhood physical neglect. Furthermore, individuals with BPD experiencing a high fear of abandonment had altered brain responses during social learning in key brain regions mediating social interactions and reward, including the IFG. These results provide a framework for the further translational modelling of disrupted social reward learning in BPD. Future therapies for BPD may aim to reprogram this altered circuitry by training people with these symptoms to learn that social interaction can have sustained positive outcomes.

**Data availability**
Functional and structural MRI data, together with behavioral log files, demographics, medication and psychopathology data are available through OpenfMRI (accession number, ds000214; https://openfmri.org/dataset/ds000214/).

**Consent**
All participants provided written informed consent before taking part. Written informed consent for publication of the participants’ anonymised details and data were obtained from the participants.

**Author contributions**
JH conceived the study. LR, DS and JH designed the experiment. LR, MP, KN and JH carried out the research. LR and JH prepared the first draft of the manuscript. All authors were involved in the revision of the draft manuscript and have agreed to the final content.

**Competing interests**
No competing interests were disclosed.

**Grant information**
This work was supported by the Wellcome Trust [DEFINE, 100202/Z/12/Z]; a Scottish Senior Clinical Fellowship [SCD/10]; the MRC Centre [G0800509]; and the Waterloo Foundation [Changing Minds programme].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Acknowledgements**
We thank Stephen Giles for data curation and analysis support. We thank all the participants for their efforts in taking part.

**Supplementary material**
**Supplementary methodology:** Additional information is provided regarding the Cyberball task; behaviour, neuroimaging and reinforcement learning analyses.
Click here to access the data.

**Supplementary results:** Additional results regarding response bias, depressive symptoms and medication.
Click here to access the data.

**Supplementary Figure 1: The Cyberball task.** The two computer-controlled players are represented by figures on either side of the screen, and the participant by a hand near the lower centre. The trial starts with the ball being held by either one of the computer players, or the participant: here an inclusion throw is illustrated, where the participant receives the ball. After a jittered delay, the throwing of the ball is animated. When held by the participant, they select a computer player to throw to with a button press. The random jitter is adjusted in part to accommodate the participant’s reaction time from the previous trial. For the reinforcement learning analysis, the conditioned stimulus (CS) occurred halfway between the moment of the ball being caught and thrown again by another player; the unconditioned stimulus (US) was the time the ball was caught by another player (exclusion), or the participant (inclusion).
Click here to access the data.

**Supplementary Figure 2: Behavioural results.** Mean reaction times for BPD (blue) and healthy controls (green) across inclusion levels. Error bars indicate ±1 SE.
Click here to access the data.

**Supplementary Figure 3: BOLD responses to increasing social inclusion.** Activation maps are thresholded at p < 0.05 FWE-corrected. Table displayed activation peaks. Resel size: 16.4 16.0 16.4, count: 279.
Click here to access the data.
Supplementary Figure 4: Impact of medication status. Scatterplots displaying the relationships between medication status and three results described in the paper. Participants are divided according to whether they are unmedicated, taking antidepressants, taking antipsychotics, or taking both. (a) Right inferior frontal cortex response to increasing inclusion versus Zanarini fear of abandonment. Data is shown for those with BPD only. (b) CTQ physical abuse versus left superior temporal cortex response to increasing inclusion. Both those with BPD and controls are shown. (c) Midbrain anticipation of inclusion versus Zanarini fear of abandonment. This too incorporates only those with BPD.

Click here to access the data.

References


I agree with all of the comments made by Dr. Grinband. I have the following additional comments:

1. The sample size is very small and this is a major limitation of the study that should be acknowledged in the Discussion. The small sample size together with the medication confound make it difficult to know whether the findings from this study are valid and reliable.

2. More than half of the BPD patients were currently taking antipsychotic medications as per the text in the Participants section (p. 3). This description appears to contradict the sample description in Table 1. Clarity is needed given that psychoactive medication is a major potential confound on BOLD activity. Correlational analyses with “medication status” were conducted (Results, p. 7) but it is unclear how this was done and whether these results are interpretable given that 15 of the 16 BPD patients were taking psychoactive medications.

3. More details are needed about the comorbid diagnoses. Were the BPD patients psychotic at the time of the scan? In a major depressive episode? What were their Axis I diagnoses? The comorbid diagnoses need to be listed in a table format or listed in the text. Details are needed.

4. More details are also needed about the healthy control sample. What exactly were the technical issues that precluded four of their scans from being used? It's unclear what “no personal major mental illness” means exactly. Did they have any personality disorder(s)?

5. The midbrain results are not statistically significant (i.e. they are “uncorrected“) making Figure 3 misleading. The figure should be removed and the non-significant finding presented in the text. As pointed out by Dr. Grinband, the rationale for the midbrain mask is unclear.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
No

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Reviewer Report 15 March 2017

https://doi.org/10.21956/wellcomeopenres.11128.r20140

© 2017 Grinband J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jack Grinband
Department of Radiology, College of Physicians and Surgeons, Columbia University, New York, NY, USA

The study compares individuals with borderline personality disorder with healthy controls to assess the neural basis of the fear of abandonment and rejection. TPJ and inferior frontal gyrus showed group differences in BOLD activation during the Cyberball game. A reinforcement learning model suggested impairments in midbrain activity in BPD subjects. The approach is novel and the findings may have significant impact on the field. However, there are a number of technical issues that need to be resolved in order for this paper to be appropriate for publication. Specifically, (1) the methods are not adequately described, (2) some of the imaging results appears to be internally inconsistent, and (3) the theoretical basis of the behavioral data (RT, bias, and TD learning assumptions) do not appear to be valid. The current version of the paper cannot be approved until these issues are addressed.
1. The authors have defined 50% inclusion as "100%" inclusion. This is confusing and counterintuitive. Presumably, this would mean that the maximum probability of inclusion is 200%, which is non-nonsensical. The probability of inclusion does not need to be redefined - it has a clear definition:

\[
\text{(throws to subject)} / \text{(total throws made by other players)}
\]

Also, the methods state that each block contains nine throws where 100% inclusion represents each player receiving 3 throws. I'm not sure that this is a consistent definition. If player 1 throws the ball to player 2 once and then the subject chooses to throw the ball to player 2 on all three of his throws, then player 2 would have received the ball 4 times which appears to contradict what is stated in the methods.

Are the throws randomized, pseudo-randomized or non-random?

2. The definition of response bias is unclear. What is "r"? Why were the "66%" and "100%" levels chosen, rather than all the levels? If "unbiased" means equal probability of throwing the ball to each player, then why not simply use:

\[
(\text{throws from subject to player 1}) / (\text{throws from subject to players 1 or 2}) - 0.5
\]

This definition is intuitive and uses all the data. If you choose to use the existing formula for bias, please explain and justify the equation.

3. How was reaction time defined? In general, RT is not a meaningful measure in the Cyberball task. The subject's decision of who to throw to can be made at any point between the moment that the subject determines that he will be the recipient of the ball and the moment that the button is pressed, whereas the subject's report of his decision only comes after the end of the "player-to-subject" movie. Thus, in the Cyberball game there is not a consistent relationship between the decision time and the response time, which is why it is possible to have RTs on the Cyberball game that are less than 200ms (i.e. less than neural conduction delays). If you believe that your RT measure has a meaningful value in your paradigm, please explain how.

Showing that there are no significant differences in RT is fine. However, the interpretation of the lack of significance should be put into the context of whether RT is interpretable.

What are "backwards regressions"? Please specify the independent and dependent variables in the regression and the purpose of the regression.

4. Please describe the design matrix that was used in the first level fMRI GLM analysis. How many regressors? How were the regressors constructed? Canonical HRF? Was the first derivative of each regressor included? Which motion parameters? What were the contrasts at the first level?

The explanation of the second level analysis is unclear (i.e. "a second level multiple regression of the first-level parametric contrast of increasing inclusion"). Is the increasing inclusion a first level contrast or a second level contrast? Were all three social parameters
(Z4, Z6, Z9) included in the same design matrix? These social variables are likely to be correlated... was the design matrix well-conditioned? Which variables where used in the controls > bpd contrast? Was controls > bpd done at the second or third level?

5. The moment that the ball was caught is not the unconditioned stimulus (for the same reasons that the RT is not a meaningful representation of decision time). The unconditioned stimulus should be the moment the ball was thrown i.e. because immediately after the ball is thrown, the subject realizes who the recipient is and the remainder of the movie is uninformative. Moreover, if the design matrix more accurately represents the timing of the conditioned and unconditioned stimuli, I suspect that the model might be rank deficient i.e. after convolution with a HRF, the conditioned and unconditioned variables will be nearly identical.

There is no description of how the design matrix was constructed for the temporal difference model either at the first or second levels. Please list all the regressors in the model.

6. The midbrain mask is not justified. Please specify where in the O'Doherty papers it is stated that TD models provide a good account for midbrain activation.

Moreover, if the assumption of social reward learning is valid, then the ventral striatum should also be included and show strong activation to the TD model. The supplementary methods state that a ventral striatum mask was used, but no mention of it was made in the results.

In any case, the temporal difference signal should be spatially localized to the dopaminergic system; in other words, you need to show the spatial distribution of the TD signal to convince the reader that your activation is specific to TD and not simply a result of convenient masking.

7. It is not clear what is depicted in Fig 1a. Is this the contrast of the mean response across inclusion levels or is it a parametric response?

What is the y-axis for Fig 1b? Are these percent signal changes or parameter estimates or contrasts of parameter estimates?

There does not seem to be a significant parametric effect in Fig 1b. Please show the corresponding slices for the parametric effect for each group (rather than just the contrast) to demonstrate whether the voxels significant in the group contrast are also active for each group. If not, that should be mentioned in the discussion.

What is depicted in Fig S3? Is this the activity across all subjects?

Fig 1 shows BOLD activity vs inclusion. However, BPD patients are traditionally described as "rejection-sensitive", rather than "inclusion-sensitive". The authors should also show the negative contrast i.e. BOLD vs exclusion.

8. "Higher levels of reported childhood physical neglect were associated with a decreased
inclusion response within bilateral superior temporal cortex, and middle anterior cingulate
cortex". These data would be far more convincing on an inflated brain. At x=39, it appears
that the activity is in the frontal cortex, not temporal.

Also, it is not clear what is being represented (because the design matrices were not clearly
described). If this is a second level analysis where Z6 (fear of abandonment) is a modulator
of the inclusion activation and if higher Z6 results in lower inclusion activity, then the t
values should be negative, not positive. If I understand the figure correctly, it states that
fear of abandonment results in a bigger inclusion response.

More importantly, why would fear of abandonment and physical neglect (two variables that
should be strongly correlated), have inverse relationships within the regions of interest?

The regions that differ between controls and BPD subjects on inclusion are not the same as
the regions modulated by fear of abandonment/neglect. Doesn't this imply that the results
in Fig 1 are non-specific? Perhaps BPD subjects are more easily bored?

9. Why is bias being reported and what is the significance of a comparison with Zanarini
9? The bias does not appear to be related to the inclusion manipulation or the stimuli in any
way. So, it is not clear that a significant correlation between bias and Z9 is meaningful or
interpretable. If there is no meaningful interpretation of the correlation, then the
correlation is likely to be a false positive.

Statistical significance is a binary quality and a "trend" is not a meaningful statistical
concept. Please report all analyses as "significant" or "not significant".

10. The data show that increasing fear of abandonment is associated with reduced midbrain
activation, "suggesting a diminished ability to learn to anticipate inclusion". If the subject
believes that he is playing with two other human beings and if he believes that his choices
have consequences (i.e. that the throws of the other players are not random and that his
actions can influence other players), then the TD model should have independent learning
parameters for each player. However, it is hard to imagine that any probabilistic learning
can occur over the course of 3-5 trials (i.e. the number of trials each player throws the ball
within each block). If, on the other hand, both players are treated as a unit (as the model
appears to do), then that requires the assumption that the subject believes his choices do
not matter, which violates the assumptions of the game (i.e. that there is social interaction).

11. The rostral medial prefrontal cortex is one of the most reliably activated regions in
mentalization paradigms. Please comment on the fact that while it was activated in the task,
there was no difference between groups.

12. It is speculated that the reason the Kumar study did not find midbrain activation is because
it was acquired at 1.5T and characterized by lower SNR than the current study. This is
unlikely since they found lots of other expected activity (e.g. ventral striatum) which the
current study did not. That suggests that they had more sensitivity, not less.

13. There is no color scale in Fig 3. Please convert t-scores to z-scores so that the reader does
not have to look up the meaning of the t-values and make all the color bars equal in the two
images so that it is easier to make comparisons.

Is the work clearly and accurately presented and does it cite the current literature?
Partly

Is the study design appropriate and is the work technically sound?
Partly

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Partly

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Partly

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.