nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: Relationship between nuclei-specific amygdala connectivity and mental health dimensions in humans
Corresponding author name(s): Miriam C Klein-Flügge

Reviewer Comments & Decisions:

Decision Letter, initial version:

11th May 2020

Dear Dr Klein-Flugge,

Thank you once again for your manuscript, entitled "Anatomically precise relationship between specific amygdala connections and selective markers of mental well-being in humans", and for your patience during the peer review process.

Your Article has now been evaluated by 2 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions to address all the points raised. In particular, before sending this manuscript back to review, we would require to see a replication in an additional sample, as requested by Referee #2. A patient dataset would be ideal for this purpose, and we believe that this would increase the impact of the manuscript, but validation of the results in at least one additional healthy participant dataset would also be acceptable.

Regarding Referee #1's concern about the high number of exclusions of datasets prior to analysis using this method, we agree with Referee #1 that you should explore what results your method produces when applied to a broader range of data. However, we do not feel that requiring high-quality data is necessarily a limitation of your method.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its

consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. I have also attached a template manuscript file that exemplifies our policies and formatting requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript we would hope to receive it within 6 months. We understand that the COVID-19 pandemic is causing significant disruptions which may prevent you from carrying out the additional work required for resubmission of your manuscript within this timeframe. If you are unable to submit your revised manuscript within 6 months, please let us know. We will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

• Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #1:
Remarks to the Author:

This manuscript uses resting-state correlations to identify amygdala subnuclei and cross-validation predictive approaches to relate subnucleus connectivity to mental well-being. Finding brain correlates

of subclinical mental health variation is an important endeavor and has therefore been studied extensively. The novelty of this manuscript is in its methods and its use of a large, high-quality dataset. However, only a small subset of this dataset was used here. Further, there are major methodological issues that need to addressed before the conclusions are supported.

A major issue is that the data being used for hierarchical clustering and delineating the amygdala subnuclei is the same as what is being used to correlate with mental well-being. If certain connections (resting-state correlations) are variable such that they are predictive of mental well-being in a healthy population, they would likely also be variable in delineating the subnuclei. A less circular approach would be to use an atlas or perhaps diffusion data for identifying the subnuclei vs. predicting mental well-being.

Exclusionary criteria and generalizability of findings: Please further clarify the exclusionary criteria for removing 800 subjects from analysis. The HCP dataset is supposed to be state of the art. How useful is the information learned here if it cannot be applied to a majority of that state of the art data? If the models are applied to the remaining subjects, what predictors are revealed (apart from the respiratory-sensitive brainstem correlations)? Do the results largely replicate? This point seems especially important when the discussion clearly states that these results may be informative for therapeutic purposes. "Despite the importance of large network approaches, an advantage of the current approach is that it provides specific regions and connections as targets for therapeutic intervention involving a range of approaches such as pharmacological, neurostimulation, neurofeedback, or cognitive interventions."

Preprocessing & generalizability of findings: "refined data pre-processing pathway that focused on the removal of breathing related artefacts that allowed us to examine activity even in brainstem regions, several of which exhibit very specific interactions with particular amygdala subnuclei." How was this refined data pre-processing validated? Is it possible that this preprocessing introduced confounds instead of removing them?

Amygdala clusters: Why were 7 amygdala clusters chosen? Were the clusters consistent across participants? The assignment of labels to the amygdala clusters is unclear and would greatly benefit from a more detailed description.

The predictors are not anatomical connections but rather resting-state correlations. Please revise throughout accordingly, and especially in the abstract. Further, some statements such as " likely that the negative coupling found between them reflects an indirect interaction mediated by another brain region (p.8)" are not based on data but are rather conjectures so should be saved for the Discussion.

How were the ROIs defined and identified? Maximally adjusting thresholds to maximise anatomical plausibility (p. 29) does not seem well justified. Why were those particular references chosen as the ROIs that interconnects with the amygdala nuclei? How big are the ROIs? Functional correlations may be impacted by the size of the regions. Also, some of the regions are likely tiny and may have less reliable/more noisy fMRI signal due to having one or two voxels (e.g. BNST).

On p. 8, "Having established and validated the network" was confusing because I was not able to find a validation section in the Results. If the authors mean the connections they listed in the paragraph above, I do not think this counts as validation. The list of ROIs/connections was in no way exhaustive

of amygdala subnucleus connections. What connections would tell us that the network was invalid?

What were the model fits like? The comparison between models from the whole amygdala vs. subnuclus parcellation seem arbitrary but perhaps would benefit from clarification or comparison of model fits. Please clarify what is meant by "If the probability of the parcellated amygdala connection is lower than the threshold set by the whole amygdala, we can infer that the parcellation increased our sensitivity." p. 13. Is this how the models were compared? Why is it inferred that the parcellation increased sensitivity (and is this ROC sensitivity?) And are these tests corrected for multiple comparisons across the 7subnuclei X ROI connections?

What is meant by 'reliably predicted'? p.16. Perhaps 'significantly predicted' would be more appropriate.

Rather than perform the CV 10,000 regression model approach, why not do a factor analysis on the connectivity data as well, and then perform the predictive modeling approach between the connectome factors and the mental well-being factors? What does that analysis yield?

Using a factor approach may be more informative than single connections; it is possible but perhaps unlikely that social & life satisfaction depends on one connection, for example.

Further, it is still unclear to me how the connections are unaffected by existing correlations between predictors (given that 5 random connections are chosen for each regression model, and therefore could be highly correlated with one another). Are the beta coefficients normalized across models?

Reviewer #2:
Remarks to the Author:

Klein-Flügge and colleagues use data from 200 HCP subjects to explore the link between (i) functional connectivity (FC) measures between 7 nuclei of the amygdala and 28 (mainly cortical) ROIs with known synaptic connections to the amygdala, and (ii) 33 measures of mental well-being.

The focus on amygdala is motivated by its known role in emotion regulation and the parcellation into seven nuclei was defined from resting-state functional MRI data while also being consistent with previous post-mortem histological results. The 33 mental well-being measures were summarized using four interpretable latent factors reflecting sleep quality, life satisfaction, negative emotions, and anger. The weights of these factors in each subject were then related to the subjects' 196 FC metrics (7 * 28) involving the amygdala. Several of the 196 FC links were found to be predictive of the four behavioral latent scores, and this is supported using multiple statistical tests. These tests essentially consist in comparisons against predictions using FC connections that either do not involve the 7 amygdala nuclei or do not involve the other 28 ROIs used in the original analysis.

I found the paper remarkably well written and I enjoyed reading it. Below are a few questions regarding the reach and significance of the results, followed by a few smaller comments.

While the statistical analysis relating FC and behavioral metrics is well executed, I am wondering

whether the interpretation of these results could be developed. For example, the authors discuss the neurotransmitters known to be present in the FC connections that are predictive of the four latent behavioral factors. Beyond this, would it be possible to show a distribution map of neurotransmitters of interest (or their main pathways) and compare it to the identified significant connections? This would allow to both highlight the relevance of the nuclei clustering while also allowing to elaborate on the relevance of the proposed partitioning of well-being measures in terms of the underlying neural circuitry.

The statistical analysis consists in comparisons against other FC connections that essentially do not share one of the two ends of the 196 original FC links. I think it is also important to test for the absolute significance of the predictive power of the 196 FC connections using a classical non-parametric test (e.g., shuffling subject labels). This could be seen as a sanity check since comparison against other FC connections is a priori a stronger test, but it would clarify the significance level of the results. Then, the set of other FC connections considered to build the null-data is quite restrictive. I think it is important to explore the predictive power of any pair of ROIs: it might be that some cortico-cortical FC connections are more predictive than the ones being considered and this is not explored. Finally, since most results are of reasonable but not strong statistical significance (e.g., few connections with uncorrected p-value < 0.001; p-values not corrected considering all the tests performed across the paper), I believe replicating these results in an independent dataset would significantly increase impact of the paper. This could ideally be done in a separate dataset of patients with a disease related to the behavioral factors identified in the paper (e.g., sleep disorders), or using other HCP subjects. If the results are robust, I believe it should be possible to find another HCP cohort of similar size within the 1200 release with acceptable data quality.

Minor comments

• Why did the authors prefer using factor analysis over PCA to identify main behavioral subspaces of variation? Unlike the former (Fig. 3B), the latter produces uncorrelated components which might be easier to interpret.

• Two measures of the Penn Emotion Recognition Test were included but these measures (32&33 in Table 1) have a close-to-zero weight in all four factors identified (Figure 3A). Why did the authors include these measures?

• The four runs available for each HCP subject are used to compute FC metrics. Since correlation is usually robustly estimated from a limited number of time points, I would expect the results to be reproduced using only one run (1200 time points) per subject, have the authors tested this? If this is the case, it would allow to relax data quality constraints for a subject to be included in the analysis.

• It is mentioned throughout the paper that 200 HCP subjects were used but it seems from the methods that only 195 were used because 5 subjects were considered as outliers. If this is the case I would specify N=195 throughout the whole paper.

• The acquisition reconstruction software version is included as a confound and regressed from the data. Is this common practice, and what is the regression weight of this confound?

• I found the double grey-scale bars in Figure 5A&B not easy to interpret. For example, I was expecting them to be of equal height but this does not seem to be the case; is this a plotting inaccuracy or am I misunderstanding something?

• Line 107: 'ratio' is missing

Author Rebuttal to Initial comments

Response to reviewers

Reviewer's Comments

Reviewer #1:

Remarks to the Author:

This manuscript uses resting-state correlations to identify amygdala subnuclei and cross-validation predictive approaches to relate subnucleus connectivity to mental well-being. Finding brain correlates of subclinical mental health variation is an important endeavor and has therefore been studied extensively. The novelty of this manuscript is in its methods and its use of a large, high-quality dataset. However, only a small subset of this dataset was used here. Further, there are major methodological issues that need to addressed before the conclusions are supported.

1. A major issue is that the data being used for hierarchical clustering and delineating the amygdala subnuclei is the same as what is being used to correlate with mental well-being. If certain connections (resting-state correlations) are variable such that they are predictive of mental well-being in a healthy population, they would likely also be variable in delineating the subnuclei. A less circular approach would be to use an atlas or perhaps diffusion data for identifying the subnuclei vs. predicting mental well-being.

Thank you for this comment. We agree that it is important to ensure there is no circularity in our argument. We realise that we should have been clearer about one important distinction: the hierarchical clustering used to derive the amygdala parcellation was conducted on the **group average connectome, i.e., the average functional connectivity across all participants**. This ignores any variability present across subjects, but instead what matters is the **variability across voxels** within the amygdala that is shared across all subjects. In other words, this first step of our analysis pipeline generates a group atlas exactly as the reviewer is suggesting. By contrast, the average connectivity of amygdala voxels is completely ignored (due to z-scoring) in the latter part of the manuscript where we relate variation in subnuclei coupling to variation in mental well-being. We have made several changes to the manuscript

to make this distinction clearer and hope this has helped to clarify that there is no circularity between these two analyses. We thank the reviewer for raising this point.

==========

CHANGES IN MANUSCRIPT

Results p.7:

To identify subdivisions within the amygdala, hierarchical clustering was performed on the similarity matrix which summarized the similarities between the functional connectivity of each amygdala voxel and all other points across the whole brain. The similarity matrix was computed based on the group average connectome. It thus ignored variability present across subjects, but instead focused on variability across voxels within the amygdala in terms of their functional connectivity to the rest of the brain that was present across subjects. […] We replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**; see Methods for further details). Because this parcellation was not obtained for each individual, but using the group connectome, it did not introduce bias in subsequent analyses focusing on individual differences.

Results p.11:

In the next analysis step, we asked whether the functional connectivity between specific amygdala nuclei and ROIs carried information about mental well-being as captured by the four latent mental health dimensions. Unlike for the amygdala parcellation which used group mean functional connectivity values, here we were interested in interindividual differences in functional connectivity.

Methods p.31:

We used the parcellation generated from the first 3T dataset for further analyses. Importantly, since this parcellation was obtained from the group connectome, rather than for each subject individually, it did not introduce bias in subsequent analyses focusing on individual differences.

==========

2. Exclusionary criteria and generalizability of findings: Please further clarify the exclusionary criteria for removing 800 subjects from analysis. The HCP dataset is supposed to be state of the art. How useful is the information learned here if it cannot be applied to a majority of that state of the art data? If the models are applied to the remaining subjects, what predictors are revealed (apart from the respiratory-sensitive brainstem correlations)? Do the results largely replicate? This point seems especially important when the discussion clearly states that these results may be informative for therapeutic purposes. "Despite the importance of large network approaches, an advantage of the current approach is that it provides specific regions and connections as targets for therapeutic intervention involving a range of approaches such as pharmacological, neurostimulation, neurofeedback, or cognitive interventions."

Apologies for not presenting the rationale for participant selection in sufficient detail in the original version of the manuscript. We fully agree with the reviewer that the HCP data is a state-of-the-art

neuroimaging dataset, and this is the main reason for choosing to use it here. Despite its quality, however, there are ongoing discussions around the relatively weak signal in subcortical structures, compared to cortical structures, in the 3T-HCP data on its user list (hcp-users@humanconnectome.org) and the majority of publications that have come out of the HCP data release thus far focus on cortical regions. By contrast, our focus is on the amygdala. Not only is the amygdala itself a subcortical structure but its subnuclei are partly distinguished by the specific patterns of connections they have with other small subcortical nuclei, for example some in the brainstem.

When we started this project, the 7T-HCP data had not been released, and the consensus in the field was that for looking at subcortical and brainstem regions which are greatly impacted by physiological noise, it is necessary to clean-up the data as much as possible for these artefacts to obtain reliable data. However, unfortunately, only **n=764** out of n=1206 3T-HCP participants have physiological noise regressors for all four resting-state runs. Initially, we inspected these traces manually and selected the n=200 participants which had the best-quality physiological noise recordings, while also ensuring a good spread in their DSM scores to allow making meaningful predictions about mental well-being. N=200 felt like a large sample size given most studies in patients and healthy people had focused on n<50 to date. While for studies focusing on cortical regions, including all n=1206 participants should be beneficial, our study was specifically designed to focus on connections with the amygdala, and adding participants with little signal in subcortex would only increase the noise in the data.

Nevertheless, based on the reviewer's comment, we went back to the remaining 3T participants and inspected their data more thoroughly. Without our original n=200 participants, **n=436 candidate 3T participants** had a complete set of behavioural scores (required for factor analysis), complete resting-state data, and physiological traces for all runs. For these participants, we started inspecting each cardiac and respiratory trace (4 runs x 2 types of traces = 8 traces) and categorized them as 'good' (no deficiency), 'mild' (insignificant or transient deficiencies), 'moderate' (noticeable but transient deficiencies) and 'severe' (pronounced or prolonged deficiencies which render the trace meaningless). We first considered participants with DSM scores at the upper and lower ends of the distribution to retain a reasonable spread in mental health dimensions. After inspection of nearly 250 additional participants, we realised that for attempting to double our numbers of participants from n=200 to n=400, while retaining some behavioural variance, we would have to include some participants with severe problems in their physiological traces. The statistic is shown below: of the newly included 3T participants, about half (n=110) had no severe problems with their physiological traces (but could have mild or moderate problems), the other half had severe problems in 1, 2 or 3 traces (out of 8). This means the majority of their data could still be processed in the same way as the n=200 participants we had originally included, but for some runs, the physiological noise clean-up might be somewhat suboptimal.

	total_severe <int>	count <int>	cumaltive_sum <int>
1	0	110	110
2	1	27	137
3	2	54	191
4	3	11	202
5	4	35	237
6	5	2	239
7	6	1	240
8	8	1	241

We have therefore now doubled our numbers to n=400 3T participants in this new version of the manuscript, while maintaining our careful and rigorous physiological noise clean-up and pre-processing routines.

In addition, we now replicate our findings in the 7T-HCP data in the revised version of the manuscript. Out of the n=176 7T-HCP participants with complete resting-state data, we included all n=98 that are unique and non-overlapping with our n=400 3T-HCP participants. The 7T HCP data has the advantage of improved signal in subcortical regions even in the absence of additional preprocessing to correct for physiological noise (which was not possible because the required cardiac and respiratory traces have not been recorded in the 7T-HCP data).

For all analyses performed on the group connectome (Figs 1-2), we therefore included **n=400 3T and n=98 7T participants**, and thus **nearly 500 data sets**. We hope the reviewer will agree that this is a considerable volume of data and a good trade-off between size and quality. For all analyses conducted on individual participant's resting-state coupling values (from Fig4 onwards), we reject outliers as done previously (if more than 10% of their coupling values across all connections deviate more than 3.5 standard deviations from the mean across participants, see Methods), and retain n**=393 3T participants and n=97 7T participants**.

As a result of these additions, our manuscript has changed substantially. As we will show below, we were able to replicate (a) the factor analysis conducted on behavioural/questionnaire scores, and (b) the amygdala parcellation. For the last part of the manuscript, to relate coupling strength to dimensional variation in mental health, we now focus less on functional connectivity in individual edges, which were not sufficiently robust to replicate in each case, and more on the patterns generated based on multiple edges. For this section of the manuscript, we use the 3T dataset to generate hypotheses which we then replicate in the independent 7T data.

==========

CHANGES IN MANUSCRIPT

(1) We have now added sections describing our subject inclusion criteria in more detail:

Results p.6:

We did not include all 1206 HCP participants because these additional pre-processing steps required good quality physiological recordings of respiration and cardiac activity which were not available or not of sufficient quality in a considerable number of HCP participants (see Methods for further details on subject inclusion criteria).

Methods p. 27/28:

Participants

Data and ethics were provided by the Human Connectome Project (HCP), WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Several datasets were included for analysis. First, an initial dataset comprised a subset of n=200 out of the full set of n=1206 3T subjects from the HCP young adults data set (n=200; mean age 29 ± .26; age range 22-36; 108 females, 92 males). These were chosen from the full HCP data set (https://www.humanconnectome.org/) based on two criteria: the quality of the physiological variables acquired (both cardiac and respiratory; inspected visually and using summary measures such as their variance over time) and their total DSM/ASR (DSM_Depr, ASR_Totp) score to allow us to maximise subclinical variance across participants (resulting mean DSM score: 4.46, variance: 16.64; mean of all 1206 HCP participants: 4.25; variance: 12.24; mean/variance total ASR score n=200: 37.91, 669.08; n=1206: 37.43, 523.82). The DSM/ASR scores were not used in any of the key analyses. Working on a subset of all 1206 HCP participants was necessary because one key aspect of the pre-processing was to correct rs-fMRI data for physiological noise, which particularly affects the key regions of this study such as the amygdala and brainstem. However, the quality of physiological data that has been acquired varies substantially across HCP participants. For example, only n=764 out of 1206 3T-HCP participants have recordings of physiological noise regressors for all four resting-state runs, and only n=636 have complete resting-state data, physiological recordings and behavioural scores (required for factor analysis, see below). Thus, a second dataset of the same size as the first (n=200) was selected for replication from the remaining n=436 3T-participants with complete data. It was not possible to match the variance in DSM/ASR scores to the first dataset and participants did not have quite as high-quality physiological noise recordings, but we still prioritized inclusion of participants at the upper and lower ends of the distribution of DSM scores (resulting DSM mean/variance in n=200: 3.96, 10.71; ASR mean/variance: 35.72, 481.67) and those with the largest number of runs with high-quality physiological noise recordings (severe problems in a maximum of 3 out of 8 = 4 cardiac + 4 respiratory traces). The resulting demographics in this second dataset of n=200 3T participants were mean age 28 ± .28; age range 22-36; 99 females, 101 males. A third dataset contained all 7T-HCP young adult participants not already included in either of the 3T datasets and with full resting-state and behavioural data, which left us with n=98 7T-HCP participants (mean age 29 ± .33; age range 23-36; 59 females, 39 males; DSM mean/variance: 3.43, 5.73; ASR mean/variance: 31.79, 253.43; **Supplementary Table 3**). Physiological noise recordings are not available in the 7T data, but the higher field strength significantly improves the tSNR in subcortical regions.

(2) We have also included sections showing the replicability of all key findings in the manuscript:

Part 1 – Replication of amygdala parcellation and average functional connectivity pattern

Introduction p. 4:

This parcellation was replicated in two additional datasets acquired at 3T (n=200) and 7T (n=98).

Results p.6:

This average amygdala functional connectivity pattern was replicated in two additional HCP datasets (3T: n=200; 7T: n=98; **Supplementary Fig 3**; see Methods for further details).

Results p.7:

We replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**; see Methods for further details).

Results p.9:

These average functional connectivity patterns between amygdala nuclei and ROIs were replicated in two separate datasets (entire matrix: n=200 3T: Pearson's *r*=0.97, p=8.82e-119; n=98 7T: Pearson's *r*=0.88, p=3.92e-66; **Supplementary Fig 4B**).

Discussion, p.20:

This amygdala parcellation replicated across several data sets.

Methods p.31 + p.36:

Following the exact same procedure, we closely replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**). However, a parcellation using the data from all n=1206 3T-HCP participants, which had not been corrected for physiological noise, showed less symmetry and anatomical plausibility despite relying on more data.

[…]

This notion was corroborated by the mean pattern of functional resting-state connectivity between amygdala nuclei and ROIs shown in **Fig2B** which we robustly replicated in two independent data sets (**Supplementary Fig 4B**).

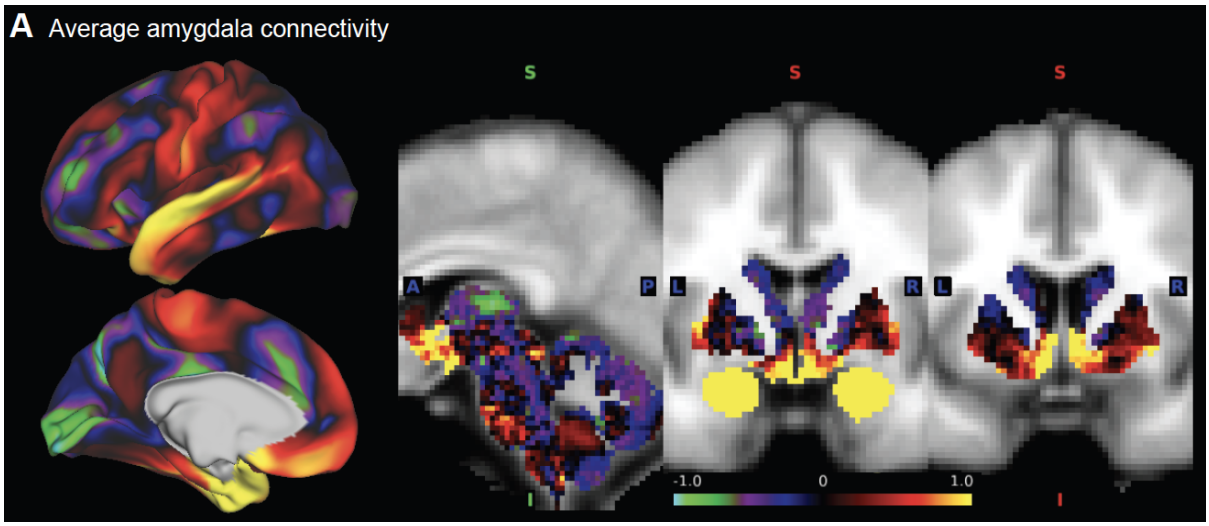
Figures/Figure legends (p.57 and Supplement)

Figure 1, Average amygdala functional connectivity and definition of amygdala clusters, **A,** A group connectome was generated from resting-state fMRI (rs-fMRI) data of n=200 3T young-adult HCP participants using an improved pre-processing pipeline to correct for physiological noise (**Supplementary Fig 1**). The average functional connectivity of all amygdala voxels to the rest of the brain, corrected for global absolute connectivity strength, shows patterns that would be expected from tracer studies, for example strong connectivity of the amygdalae with subgenual ACC, hypothalamus, and ventral striatum. This pattern was replicated in two independent datasets containing n=200 additional 3T-HCP participants and n=98 7T-HCP participants (**Supplementary Fig 3**). **B**, Hierarchical clustering was performed on the similarities between the whole-brain functional connectivity patterns of different amygdala voxels to identify amygdala subdivisions sharing connectivity profiles. Seven subdivisions were identified (left:

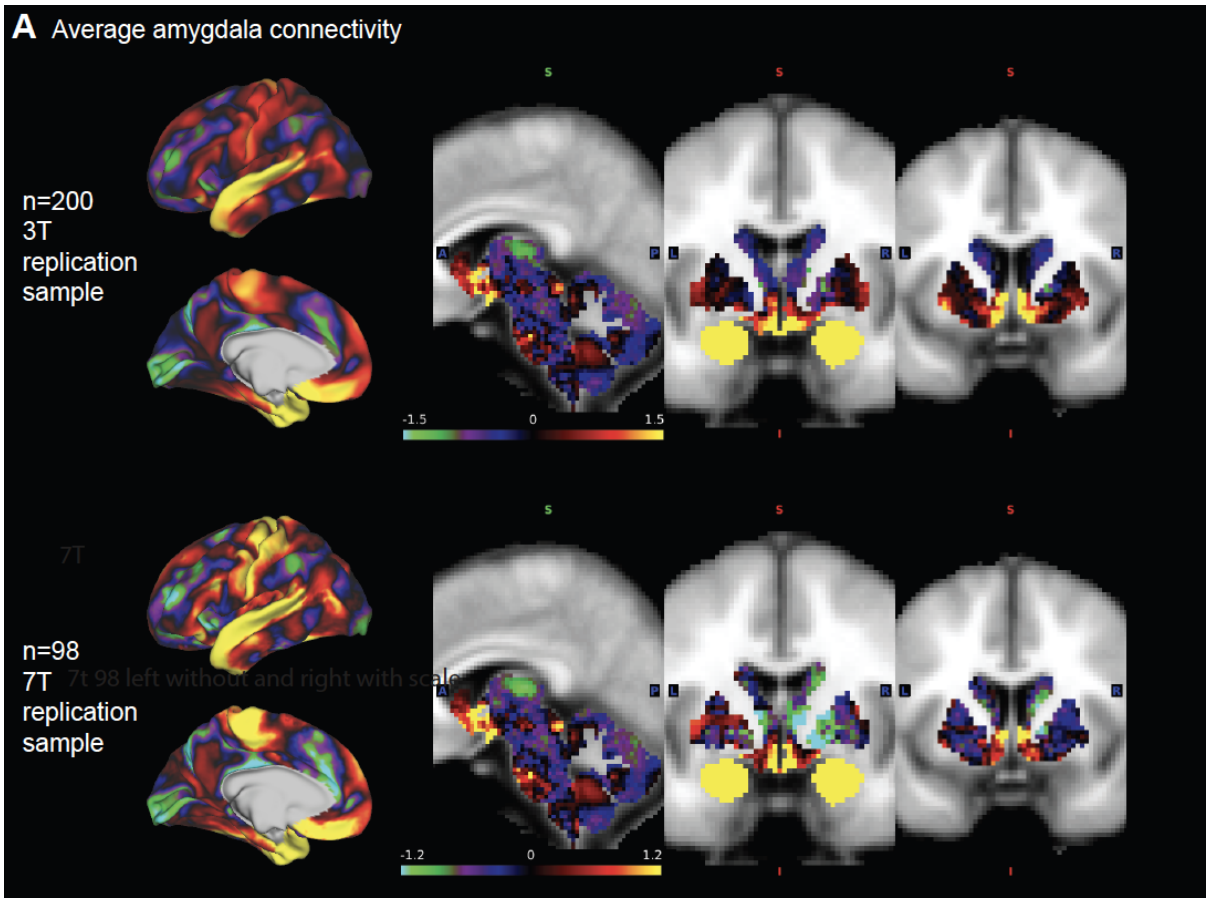
horizontal; middle: coronal; right: saggital view), showing strong symmetry across hemispheres and strong resemblance with subdivisions identified from histology and high-resolution *post-mortem* structural neuroimaging. The parcellations obtained in two independent datasets closely reproduced these nuclei subdivisions (**Supplementary Fig 4A**).

Figure 2, Amygdala nuclei and their profile of functional connectivity to regions of interest, **A,** Labels assigned to the seven amygdala subdivisions obtained from hierarchical clustering: Ce = central nucleus, CoN = cortical nuclei, B = basal, AB/BM = auxiliary basal/basomedial, LaV/BL = lateral (ventral part) containing aspects of basolateral, LaI = lateral (intermediate part), LaD = lateral (dorsal part). **B**, Average resting-state functional connectivity from the seven nuclei to 28 regions of interest (ROIs) defined *a priori* based on their known connectivity with the amygdala and potential role in regulating emotions and mental well-being. This highlights strong functional connectivity of subgenual cortex (area 25) to the entire amygdala, but particularly to basal subdivisions, in line with tracer work. Similar profiles are observed for posterior OFC (pOFC) and the subgenual portion of area 32 (s32). By contrast, subcortical and brainstem regions most strongly connect with the central nucleus as expected. The mean functional connectivity between ROIs and nuclei was replicated in two datasets (**Supplementary Fig 4B**). **C**, Masks of all ROIs used in this study. For details on their definition, please refer to the Methods. NAc=Nucleus Accumbens; BNST=bed nucleus of the stria terminalis; vl/dPAG=ventrolateral/dorsal periaqueductal grey; SN=substantia nigra; RN_DR/RN_MR=dorsal and median raphe nuclei; LC=locus coeruleus. Definitions of cortical regions were taken from Glasser et al., 2016.

Figure 1A (original n=200 3T HCP participants)



Supplementary Figure 3 (replication in n=200 additional 3T and n=98 7T participants)



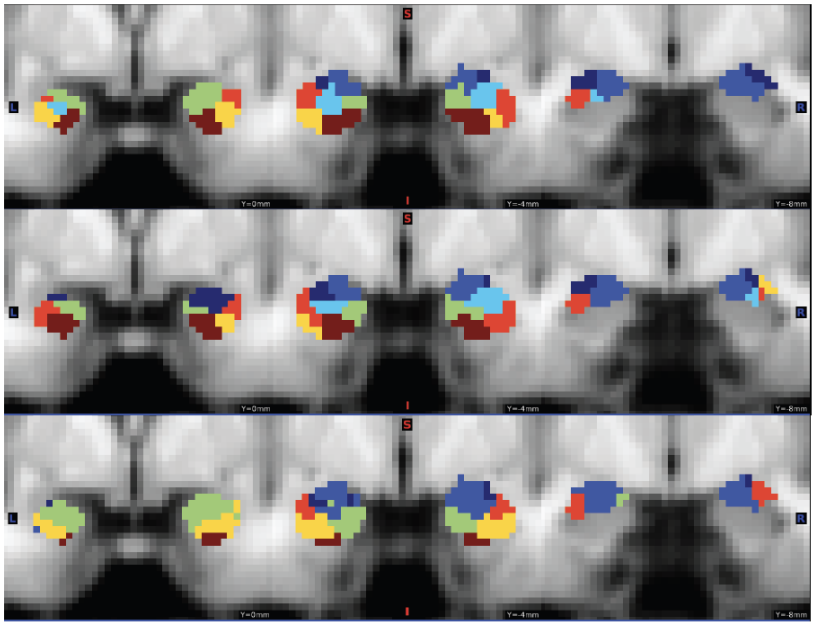
Supplementary Figure 3, Replication of average amygdala functional connectivity The group connectome shown for the original n=200 3T young-adult HCP participants presented in Fig1A was replicated in two other cohorts containing n=200 non-overlapping 3T-HCP participants and n=98 non-overlapping 7T-HCP participants. In the 3T data, we used an improved pre-processing pipeline to correct for physiological noise, as before. This was not possible in the 7T data where physiological noise regressors were not available. However, the 7T resting-state data has improved signal-to-noise in subcortical regions due to the higher field strength.

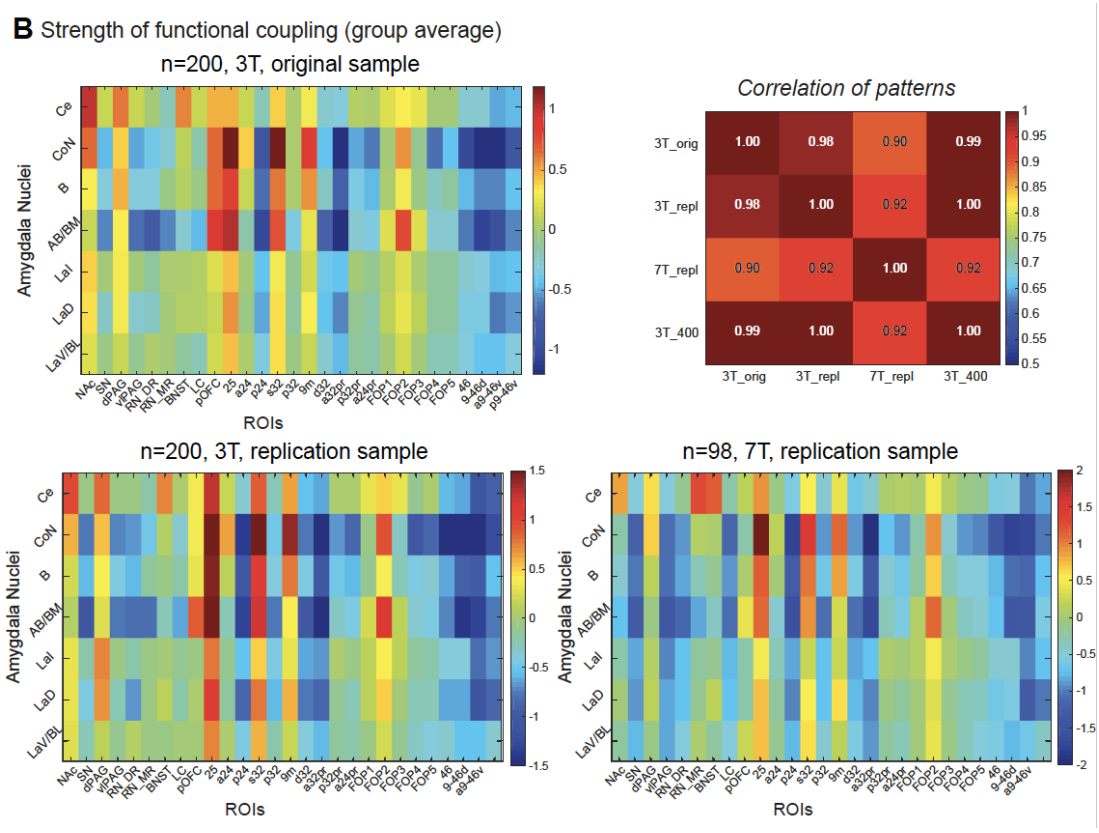
A Replication of amygdala parcellation

n=200
3T
original
sample

n=200
3T
replication
sample

n=98
7T
replication
sample





Supplementary Figure 4, Replication of amygdala parcellation and nuclei mean functional connectivity, A For comparison, the parcellation of the amygdala obtained in the original n=200 3T participants is shown for the n=200 3T replication sample and the n=98 (all non-overlapping) 7T participants (compare Fig 1B). This shows that the key subdivisions of the amygdala were replicated in these two additional parcellations. **B**, The average amygdala nuclei to ROI functional connectivity replicates across cohorts (top: original, bottom left: replication 3T, bottom right: replication 7T; compare **Fig 2B**), as confirmed in the strong correlation between these patterns (top right).

Part 2 – Replication of behavioural factor analysis

Introduction p.4

[…] we were able to define latent behaviours by applying a factor analysis to a set of questionnaire scores which generated four highly replicable dimensions of mental health.

Results p.10:

The factor analysis was replicated in several other datasets (**Supplementary Fig 6**). Notably, because the factor analysis focuses on behavioural data rather than neural data, it can be employed with all HCP

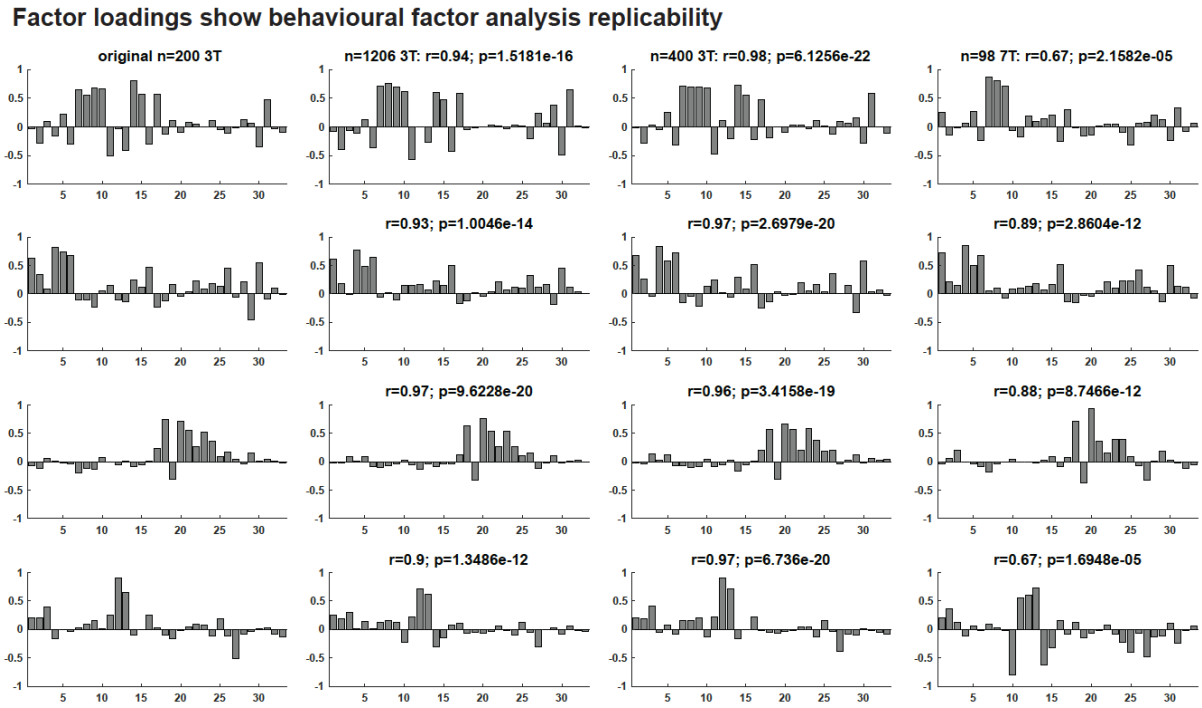
datasets and not just the subset of data with the highest quality physiological recordings of respiration and cardiac activity. When the analysis was repeated on the complete set of n=1206 3T HCP participants for maximal robustness, the resulting factors were highly similar (Pearson's correlation between factor loadings n=200 vs n=1206 3T participants: *r*=.94, *p*=1.5e-16, *r*=.93, *p*=1e-14, *r*=.97, *p*=9.6e-10, *r*=.9, *p*=1.3e-12; **Supplementary Fig 6**).

Methods p.38:

Moreover, the factor analysis replicated to several other datasets (**Supplementary Fig 6**), most importantly to the full set of all n=1206 3T HCP participants (Pearson's correlation between factor loadings n=200 vs n=1206 3T participants: *r*=.94, *p*=1.5e-16, *r*=.93, *p*=1e-14, *r*=.97, *p*=9.6e-10, *r*=.9, *p*=1.3e-12), but also the n=400 3T and the n=98 7T participants included for neural analyses (**Supplementary Fig 6**).

Figures + Figure legends (p. 57/58):

Figure 3, Latent behavioural dimensions capture distinct aspects of mental well-being, **A,** A factor analysis conducted based on 33 behavioural scores (**Table 1**) available as part of HCP revealed four factors. The loadings for each factor are shown in different colors, corresponding to the four rows. The highest five contributing behavioural scores are shown in order of their contribution (absolute loading) on the right. This shows that the four factors capture quite distinct dimensions of participants' mental well-being which we summarized as 'Social and life satisfaction', 'Negative emotions', 'Sleep' (problems), 'Anger and rejection'. The four factors replicated when the factor analysis was performed on all 1206 HCP participants (see Methods), or only the subset of 3T and 7T participants included here (**Supplementary Fig 6**). **B**, Correlations between factor loadings.



Supplementary Figure 6, Replication of factor analysis The factor analysis computed to generate mental health dimensions in our original n=200 3T participants (left) replicated in all n=1206 HCP participants (2nd column) and the full set of n=400 3T and n=98 7T participants used in this manuscript (3rd and 4th column). Correlation coefficients and p-values refer to the similarity with the original pattern shown on the left.

Part 3 – Replication of amygdala functional connectivity behaviour relationships
This section was changed most extensively due to the inclusion of several new datasets and is not fully reproduced here. The complete set of changes can be found in the Results p.11-19, Methods p.38-44, Figures 4-7, and Suppl Figs 7-10.

Abstract p. 2
Connectivity in circumscribed amygdala networks predicted behaviours in an independent dataset.

Introduction p.4
In our final step, we selected the best predictors in terms of the functional connectivity between amygdala nuclei and other brain regions for each of the four behavioural dimensions. We showed that functional connectivity in a select number of specific amygdala connections predicted each mental health dimension in an independent dataset.

Selected new Results sections:

Results p.11:

We first established whether relationships between nuclei-specific amygdala functional connectivity and mental health dimensions replicated between the two independent (3T and 7T) datasets. We fitted robust linear regression coefficients to capture the relationship between functional connectivity values in each 'edge' (e.g., Ce to NAc) and behavioural dimension (e.g., sleep problems), separately for the 3T and 7T dataset. This resulted in 196 regression coefficients (7 amygdala nuclei x 28 ROIs) for each of four behaviours and two datasets. If amygdala nuclei functional connectivity carries no information about mental health dimensions, then, by chance, the correlation between regression coefficients obtained across behaviours in the 3T versus 7T datasets should be zero. To formally test this, we generated a null distribution by shuffling the subject order of the behavioural scores n=10,000 times while keeping the functional connectivity values unchanged. Indeed, by chance, the across-dataset replication of the pattern of regression coefficients was centred on zero (**Fig4A**). The similarity between 3T and 7T regression coefficients in the actual data, however, was significantly greater than chance (Pearson's r=0.26; p=0.0313; **Fig 4A,B**), showing that relationships between nuclei-amygdala functional connectivity and mental health dimensions were similar across datasets.

Results p.12/13:

Having established that the overall relationship between nuclei-specific amygdala functional connectivity and mental health dimensions is similar between datasets, we next asked whether we could predict individual 7T participants' behavioural scores using regression coefficients estimated from the 3T data. In other words, we examined whether we could predict mental health dimensions in completely held-out data (7T) using a weighting of nuclei-specific amygdala functional connectivity values derived from an independent dataset (3T). For each behavioural dimension, the 196 robust regression coefficients estimated from the 3T data for all nuclei functional connectivity values were applied to the functional connectivity values of individual 7T participants to obtain their predicted behavioural scores. This out-of-sample prediction was significant for life satisfaction, negative emotions, and anger (correlation between predicted and true behaviour for the 7T data: lifeSat: r=0.187, p=0.0335; negEmot: r=0.219, p=0.0155; anger: r=0.226, p=0.0143), but did not reach significance for sleep (r=0.05, p=0.31; **Fig4E**).

Given that medial temporal lobe areas are considered areas of high drop-out and low signal-to-noise, we performed a second replication to further demonstrate the consistency with which nuclei-specific amygdala functional connectivity predicted dimensional variation in mental health. This time, we examined the consistency within-subject rather than across-dataset (pooling across 3T/7T datasets). We divided the full resting-state data of each participant in two halves (run 1+2 versus 3+4, acquired in separate sessions on different days) and again computed robust regression coefficients to predict the four behavioural dimensions, but this time separately using resting-state functional connectivity values extracted from only the first or second half of the full resting-state data. A null distribution obtained using shuffled behavioural values estimated the similarity of regression coefficients between the two halves (i.e., sessions) expected by chance. The true functional-connectivity-behaviour relationship between the first and second half was significantly greater than expected by chance

(Pearson's r=0.47; p=0.014; **Fig4A**). The anatomical pathways where functional connectivity most contributed to this within-subject replication were highly similar to those most contributing to our previous across-dataset replication (**Fig4D**). Together, these analyses show that despite substantial noise and difficulties in neuroimaging subcortical regions such as the amygdala[29], specific patterns of variation in functional connectivity both within- and between-datasets consistently related to participants' mental health dimensions.

Discussion, p. 23:

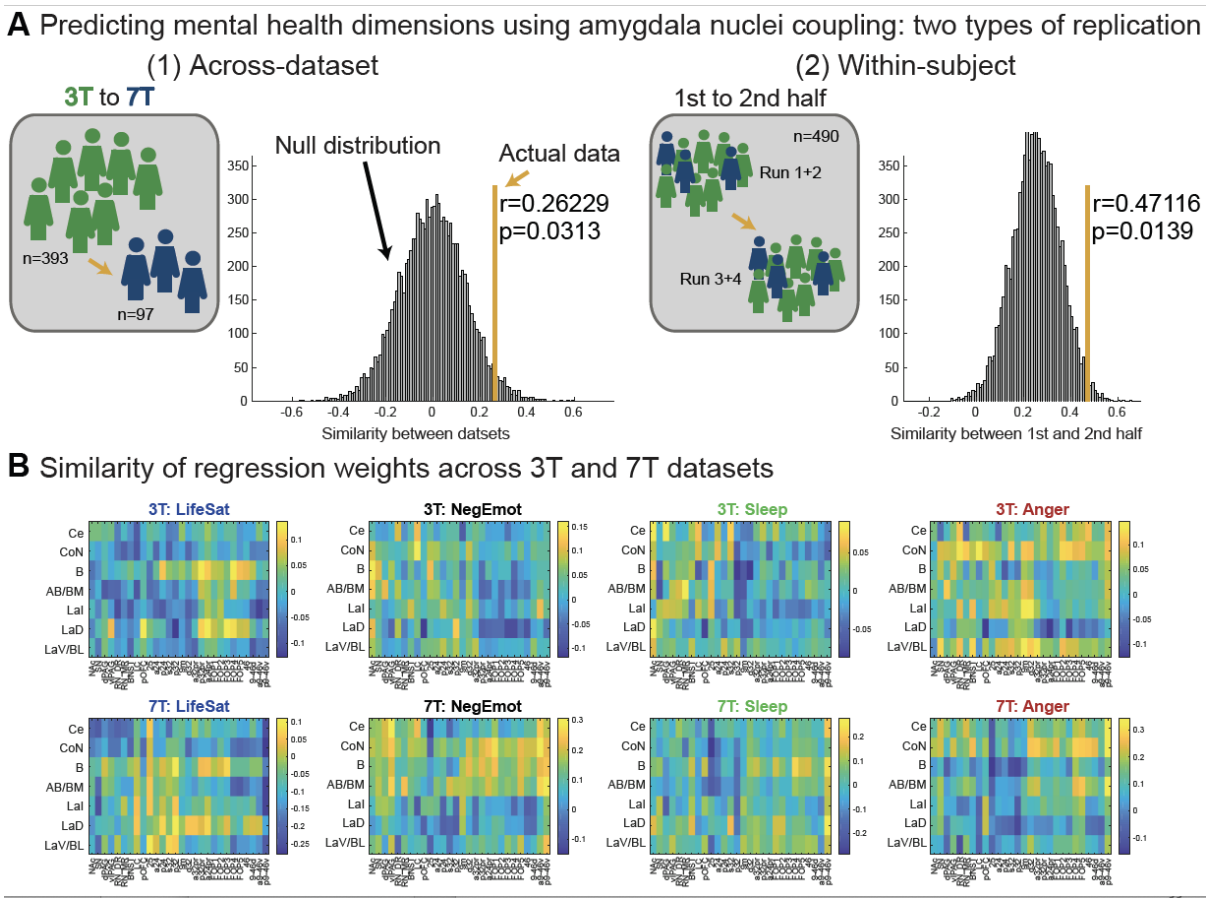
The anatomical features of the amygdala networks identified for the different latent behaviours seem plausible in the context of previous work, and consistent across two types of replications (across-dataset and within-subject; **Fig 4**). We note that, importantly, both feature selection – which determined the anatomical networks to focus on – and estimation of regression weights was performed in an initial dataset (n=393 3T participants) and all predictions were generated out-of-sample (n=97 7T participants).

Methods p.39/40:

To test whether the obtained robust regression weights captured meaningful relationships between the functional connectivity of individual amygdala nuclei and the four mental health dimensions, we tested whether regression weights were (a) similar between the 3T and 7T datasets ('across-dataset replication') and (b) similar between two halves (corresponding to MR sessions) of the experiment ('within-subject replication'; **Figure 4A**). First, for the across-dataset replication, we computed Pearson's correlation coefficient between the overall pattern of regression weights obtained for the 3T and 7T data (row 1 versus 2 in **Figure 4B**). To establish whether the obtained correlation was better than predicted by chance, given the level of noise present in brain connections with the amygdala and given our number of connections, we generated a null distribution (**Fig 4A**) by shuffling the vectors **y** containing the behavioural dimension n=10,000 times and recomputing the correlation coefficient between the overall pattern of regression coefficients.

Second, to attempt a within-subject replication, functional connectivity values were extracted from half of the resting-state data, separately from either run 1+2 or run 3+4 (which were acquired in two separate sessions on separate days). Robust regression weights capturing the relationship between FC (X) and behavioural scores (y) were then computed separately for the FC values obtained from runs 1+2 versus 3+4. Here we used the merged data from all n=490 3T+7T participants. The similarity between the overall pattern of robust regression weights obtained for the two experimental halves was computed using Pearson's correlation. Again, to test whether their similarity was greater than expected by chance, a null distribution was generated by repeating this procedure n=10,000 times using shuffled behavioural scores (**Figure 4A**).

Figures and Figure legends – in addition to the new figures pasted below, **Suppl Figs 7-10** have also newly been added:


A Predicting mental health dimensions using amygdala nuclei coupling: two types of replication
(1) Across-dataset
3T to 7T
n=393
n=97
Null distribution
Actual data
r=0.26229
p=0.0313
Similarity between datsets
(2) Within-subject
1st to 2nd half
n=490
Run 1+2
Run 3+4
r=0.47116
p=0.0139
Similarity between 1st and 2nd half
B Similarity of regression weights across 3T and 7T datasets
3T: LifeSat
3T: NegEmot
3T: Sleep
3T: Anger
7T: LifeSat
7T: NegEmot
7T: Sleep
7T: Anger
Ce
CoN
B
AB/BM
Lal
LaD
LaV/BL

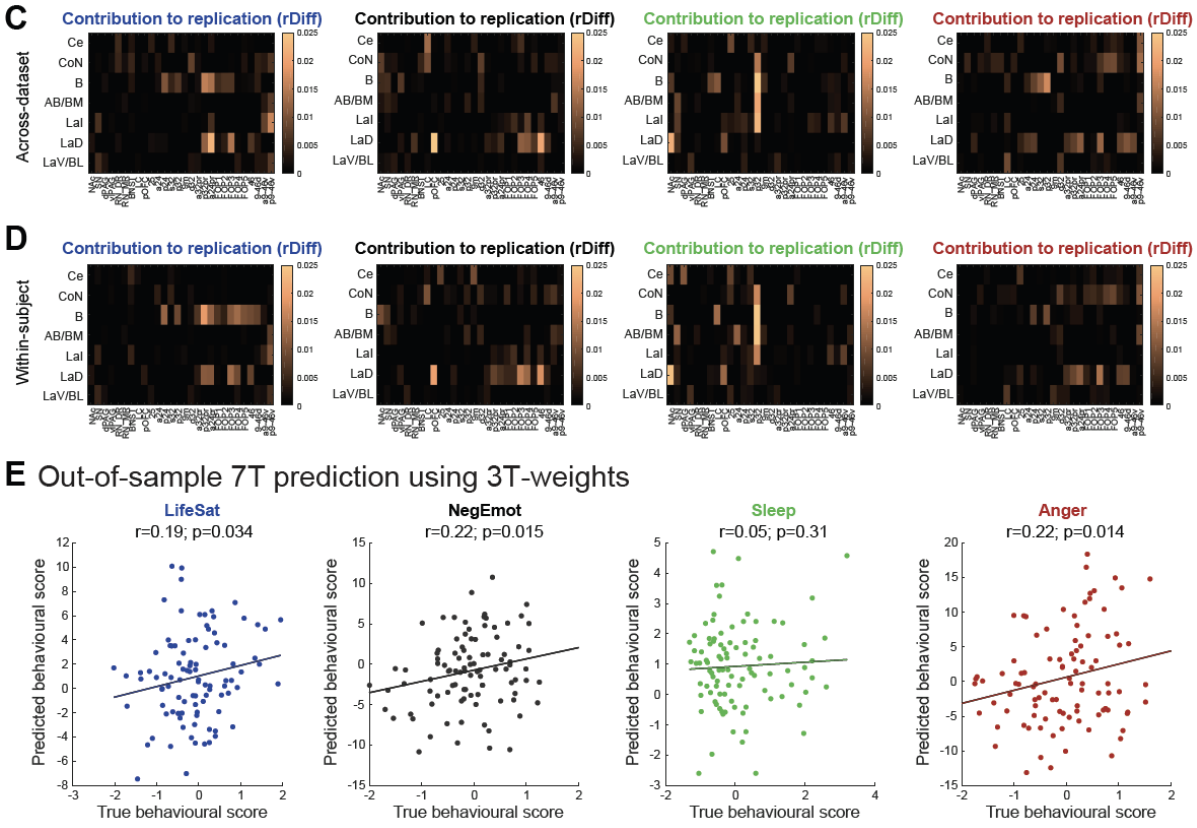


Figure 4, Nuclei-specific amygdala functional connectivity shows consistent relationships with interindividual variation in mental health dimensions, **A,** Relationships between interindividual variation in nuclei-specific amygdala functional connectivity and mental health dimensions were examined in two HCP datasets containing n=393 3T and n=97 non-overlapping 7T participants (following outlier rejection). Despite challenges with neuroimaging signals in subcortical regions, relationships were robust and replicable, as established in two ways: (1) Across-dataset replication: the similarity of robust regression coefficients capturing the relationship between resting-state functional connectivity for each 'edge' (e.g., Ce to NAc) and each of four mental health dimensions was greater than expected by chance across datasets (null distribution generated using shuffled behavioural scores; n=10,000 iterations). (2) Within-subject replication: robust regression coefficients estimated on half of the resting-state data (runs 1+2 versus 3+4, from separate sessions) also showed greater-than-chance similarity. **B**, Visualization of obtained robust regression coefficients for each edge, mental health dimension (columns) and dataset (rows) illustrates their similarity across cohorts. **C**, **D,** For each edge, its contribution *rDiff* to the across-dataset (**C**) and within-subject (**D**) similarity was computed as the difference between the correlation achieved when excluding this edge (195 values) and when including all 196 edges (28 ROIs x 7 nuclei). Visual inspection of *rDiff* values highlights strong similarities between *rDiff* values in the two replications (C vs D), clear differences between the four behavioural dimensions, and anatomical specificity – e.g., the importance of cortical connections with B and LaD nuclei for predicting life

satisfaction, for connections with NAc, other subcortical regions and medial frontal area p32 for predicting sleep, and functional connectivity with the cortical nuclei (CoN) for predicting anger. **E**, Regression coefficients estimated from the 3T-participants applied to 7T-functional connectivity values to predict 7T-mental health dimensions showed significant out-of-sample predictions for all mental health dimensions except sleep problems.

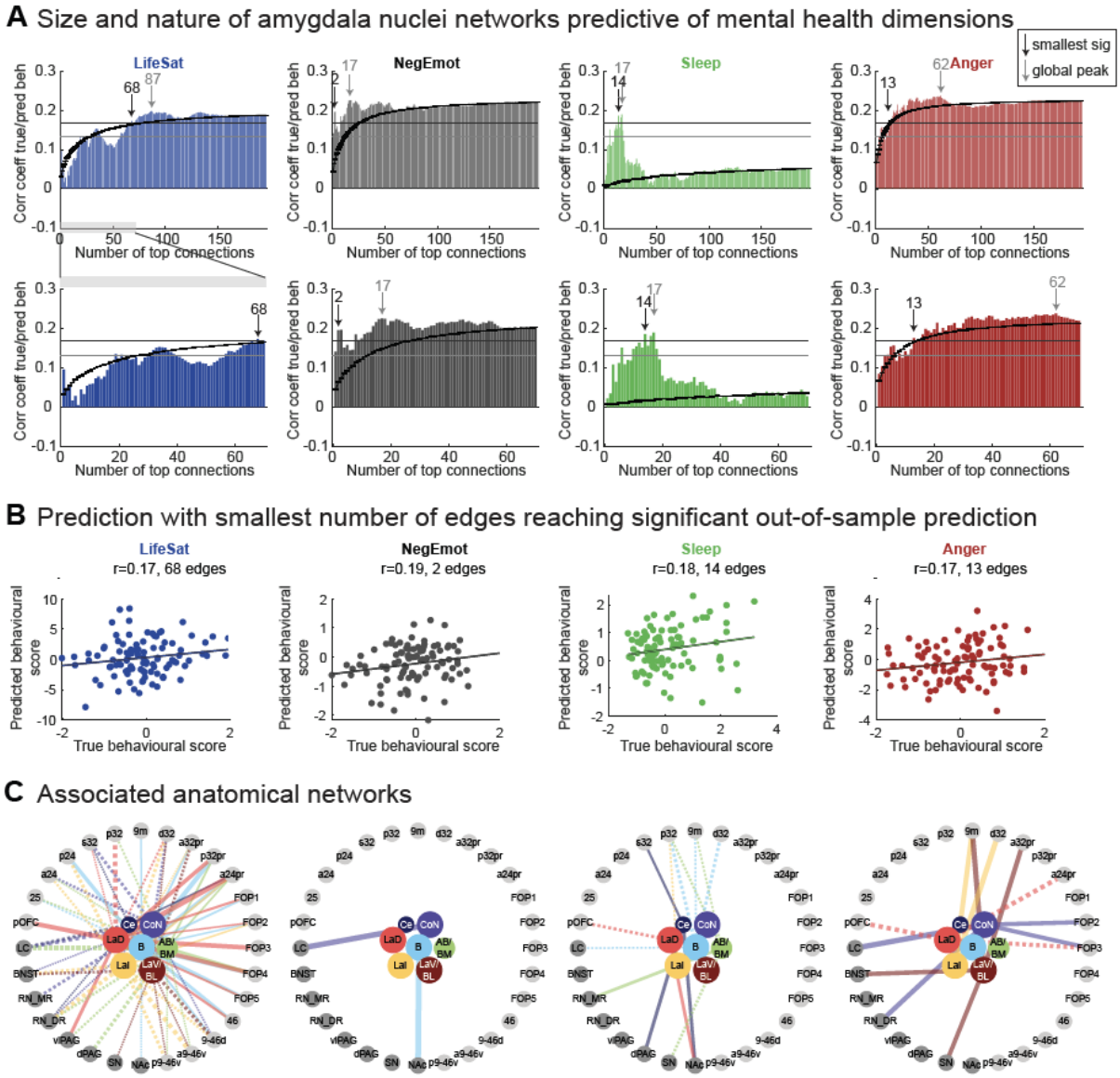
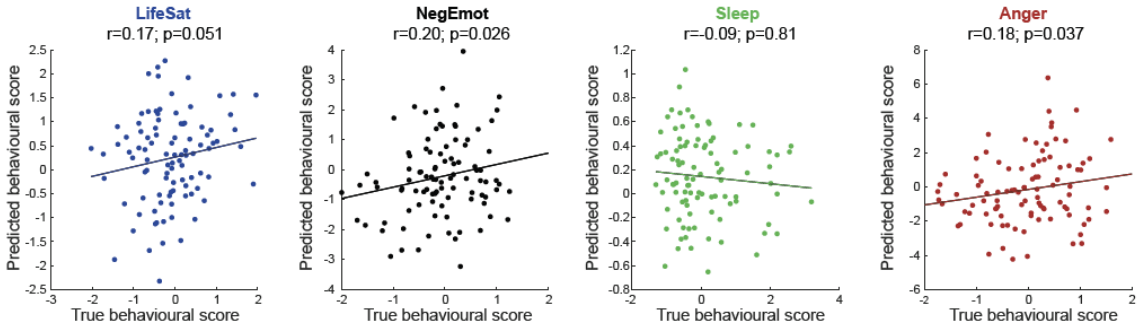


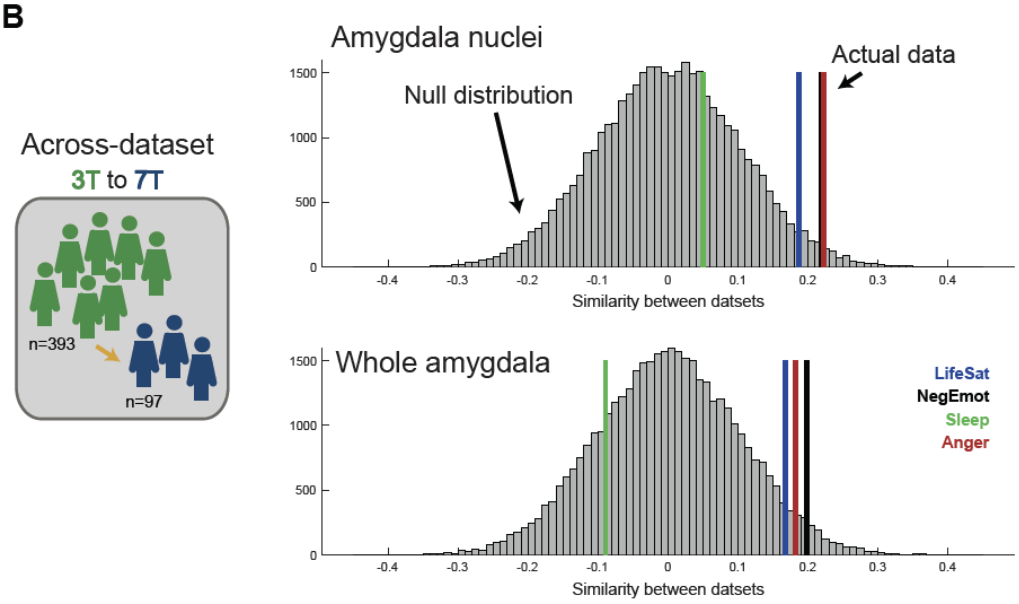
Figure 5, Functional connectivity in smaller sets of specific amygdala nuclei connections is predictive of interindividual variation in mental health dimensions. **A,** Predictions achieved with subsets of edges between 1 and 196: robust regression coefficients estimated from 3T-participants were applied to 7T-functional connectivity values to predict interindividual differences in mental health dimensions in the

held-out 7T data (as done in **Fig4E** using all 196 edges). Prediction accuracies are shown as the correlation between true and predicted mental health dimensional scores in the 7T participants, but were only statistically evaluated at the peak (grey arrow: 'global peak') and to derive the smallest number of edges that reached a significant out-of-sample prediction (black arrow: 'smallest sig'; see Methods); coloured bars show accuracy when including edges in order of their absolute regression coefficient in the 3T data; black curve indicates performance using the same number of edges but included in random order (n=10,000 shuffles; error bars denote SEM); black line at r=0.168 indicates threshold for significance at $p<0.05$ purely for visualization (grey line: $p<0.1$); second row shows the same but zoomed in on the first 70 edges. For all behavioural dimensions, smaller sets of amygdala nuclei functional connectivity values achieve a significant out-of-sample prediction. In general, except for life satisfaction, using the top 3T edges is better than a random selection of the same number of edges. **B**,**C** Illustration of the prediction (scatterplot, **B**) and contributing edges (fingerprint, **C**) for the smallest network that achieved a significant prediction (indicated using a black arrow in **A**). Fingerprints shows ROIs on the circumference (dark=subcortical), amygdala nuclei in the centre (colour-coded); line width denotes the size of the absolute 3T regression coefficient; line style denotes its sign (continuous=positive; dashed=negative).

A Sensitivity of whole-amygdala as opposed to nuclei-specific amygdala coupling



B



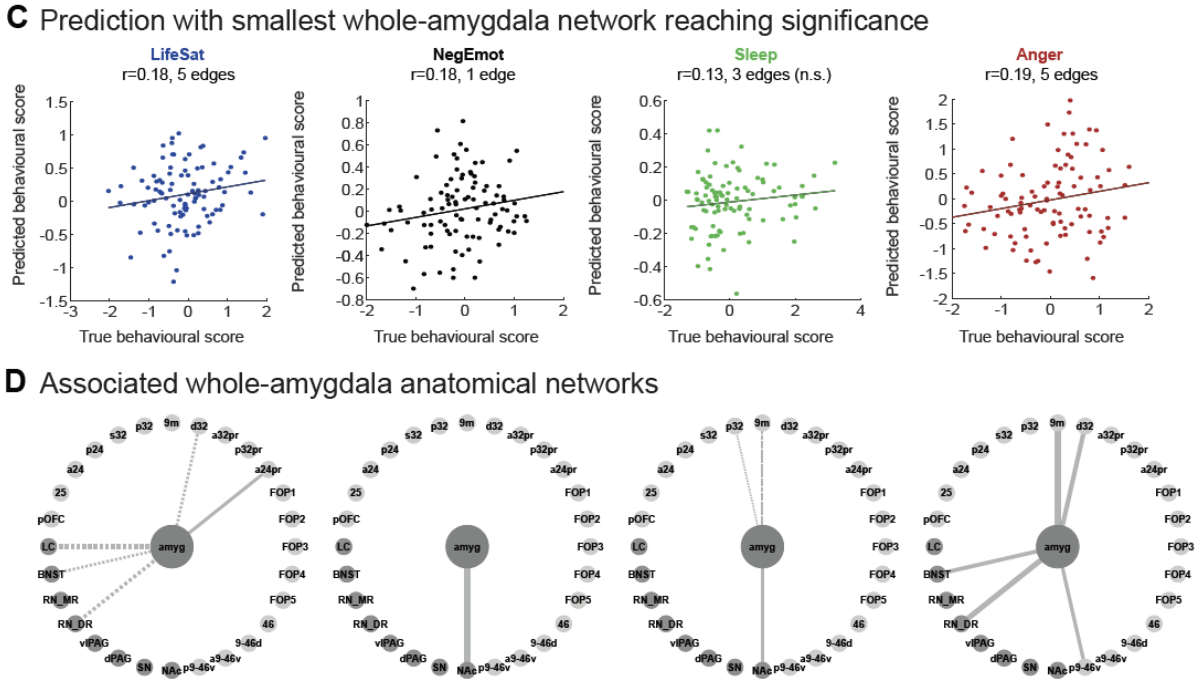


Figure 6, Parcellating the amygdala improves the accuracy for predicting interindividual differences in mental health dimensions, **A**, Out-of-sample predictions achieved when considering the functional connectivity of the whole amygdala with our 28 *a priori* ROIs, instead of nuclei-specific functional connectivity, are less precise, but still significant for two of the four mental health dimensions (negative emotions + anger). **B**, To control for differences in the number of predictors for nuclei-specific and whole-amygdala functional connectivity (28 vs 196), out-of-sample 7T predictions (shown in Fig 4A and 6A) were evaluated based on their own null distribution. Nuclei-specific predictions were still superior to whole-amygdala predictions for all four mental health dimensions (coloured bars indicate Pearson's r overlaid on the null distribution; for statistics see main text); this was also true when looking at peak predictions achieved using smaller sets of edges (**Supplementary Fig 9**). **C,D,** As in **Fig5**, the prediction (scatterplot, **C**) and contributing edges (fingerprint, **D**) are shown for the smallest set of edges that reached significance (sleep never reached significance; life satisfaction was significant when using fewer than 28 edges) which highlights clear differences between mental health dimension.

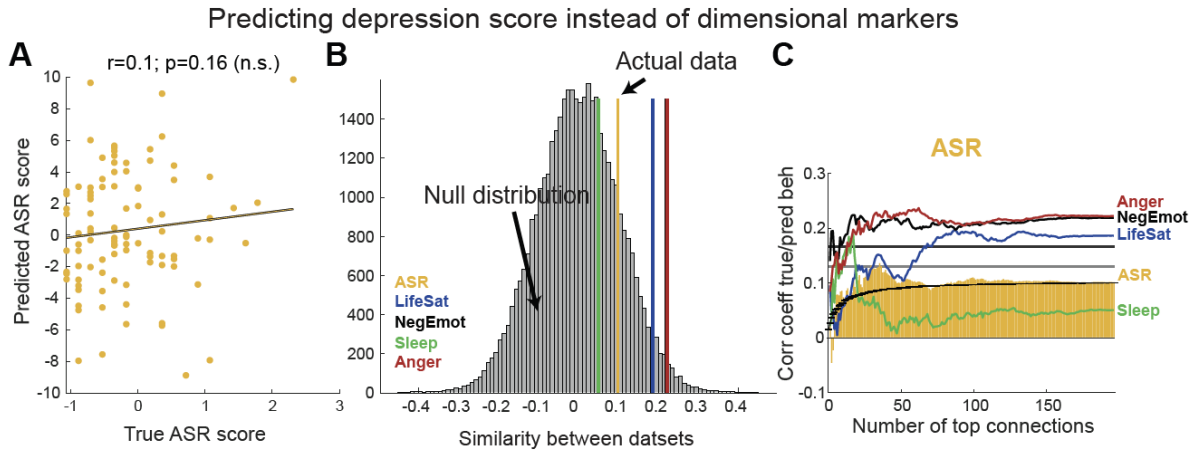


Figure 7, Amygdala functional connectivity relates better to dimensional behaviours than overall depression scores, **A, B** Out-of-sample prediction of 7T participant's ASR_AnxD scores, using nuclei-specific functional connectivity and regression weights estimated from 3T participants as in **Fig 4E**, is (**A**) not significant (for DSM, see **Supplementary Fig 10**) and (**B**) less accurate than three out of four of our dimensional behaviours. **C,** Overall predictions are worse for ASR compared to dimensional scores when using smaller sets of edges (bars shown in Fig5A for dimensional behaviours are overlaid as coloured lines for comparison); plotting conventions as in **Fig5**.

==========

Finally, we turn to the question raised by the reviewer concerning the general informativeness and wider utility of amygdala connectivity data. We believe careful preprocessing is essential to reveal the effects that we report. However, as we show in the 7T data, a higher field strength can make up for the low signal to some extent. It is true that care both in data acquisition and data analysis is needed and the approach that we outline may not be applicable to some existing data sets. However, it is equally important to realise that there is little to stop approaches of this sort being used more widely. Equipment for physiological noise monitoring is a comparatively small cost within the context of a large-scale neuroimaging study and it is relatively easy to use. In addition, the quality of data that can be acquired in a given amount of time is constantly improving. We think it is highly likely that more datasets with sufficient resolution and SNR will become available, which will make our method more broadly applicable to the neuroscientific community and to the study of other subcortical brain structures, including work in clinical populations. To facilitate such future work, upon publication, we will make our parcellation and all analysis scripts available as part of the Open Science Framework (OSF). We have now added a paragraph outlining this issue to the Discussion.

==========

CHANGES IN MANUSCRIPT

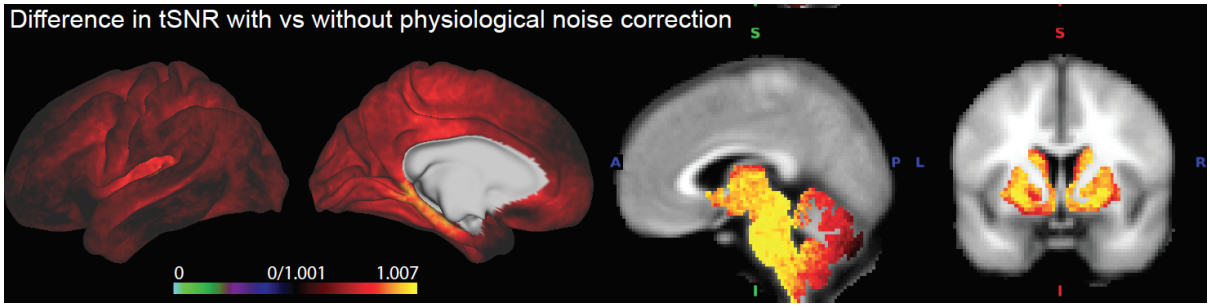
Patients and datasets – p.22 Discussion:

The finer grained parcellation we obtained reflected improved image quality and preprocessing pipelines that better controlled for physiological noise. While this could be seen as a limitation of our method because care both in data acquisition and data analysis is needed, we note that scan hardware, software and imaging sequences are developing fast and that acquiring physiological cardiac and respiratory traces is easy, cheap and uses standard equipment available in most MR facilities. It is true that our approach may not be applicable to some existing data sets, but we believe more data with sufficient resolution and SNR is likely to become available, which will make our method more broadly applicable to the neuroscientific community and to the study of other subcortical brain structures, including in clinical populations.

==========

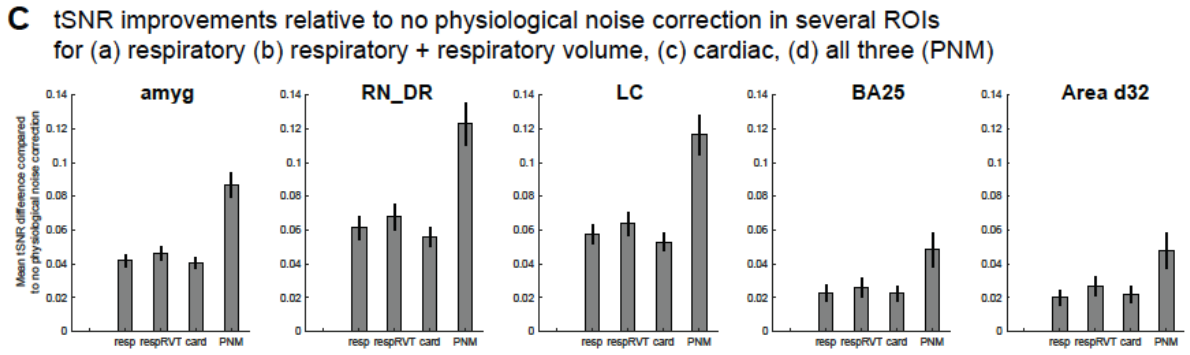
3. Preprocessing & generalizability of findings: "refined data pre-processing pathway that focused on the removal of breathing related artefacts that allowed us to examine activity even in brainstem regions, several of which exhibit very specific interactions with particular amygdala subnuclei." How was this refined data pre-processing validated? Is it possible that this preprocessing introduced confounds instead of removing them?

It is unfortunately not possible to verify this in resting-state data where the ground truth is unknown. However, there is prior evidence to show that physiological noise clean-up helps – particularly in regions closer to major vessels and pulsating fluid-filled spaces (as is the case for the brainstem and medial temporal lobe structures) (e.g., Harrison et al., 2021; Van den Brink et al., 2019). In addition, to illustrate where the additional data clean-up made a difference, we presented the temporal signal-to-noise (tSNR) maps in Figure S1 of the original manuscript, which shows a measurable change in tSNR that is prominent in brainstem and other subcortical as well as medial and orbital cortical areas of interest to us (pasted here again for ease of access). In the revised manuscript we now give this figure greater emphasis so that other readers will be aware of this result.



In response to the reviewer's comment, however, we have now examined the improvements gained from the physiological noise clean-up step further in several ways:

(1) First, we tried to understand which components of the physiological noise clean-up - respiratory volume over time (RVT), pulse (cardiac) or breathing (respiratory) - helped the most. To do this, in a subset of n=20 3T participants (all with good quality physiological noise traces), we produced tSNR maps with five versions of data clean-up: no physiological noise correction, correction for only respiratory measures, respiratory measures + RVT, only cardiac, or all three types of regressors (PNM: respiratory + RVT + cardiac). The below plot shows the difference in tSNR relative to the version without physiological noise correction in three representative subcortical regions and two representative medial cortical regions (from left to right in the below figure: amygdala, dorsal raphe, locus coeruleus, area 25, area d32). This shows that all regressors contributed significantly to improvements in tSNR. As expected, the effects are especially notable in the amygdala itself and in other subcortical areas. This figure has now been added to **Supplementary Figure 1**.



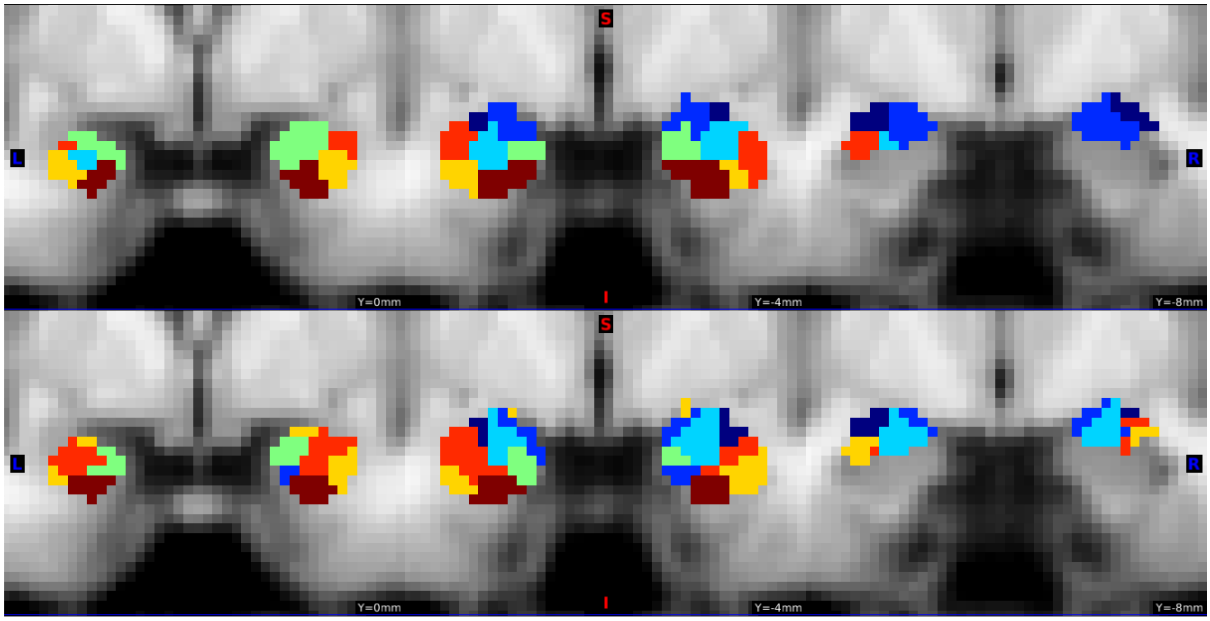
(2) Second, we have now repeated the amygdala parcellation using all n=1200 3T participants (using the group connectome made available by HCP – which is absent of any physiological noise correction). If the physiological data clean-up step made no difference or possibly even introduced more noise, this comparison should be biased **against** our original one which was based on only n=200 participants. We found that the parcellation obtained from the physio-corrected data in n=200 participants produced coherent, contiguous parcels, for which we have strong anatomical priors from animal work. However, this is not the case for the parcellation obtained from the complete set of n=1200 HCP participants. While the core features of the parcellation replicate, there seems to be no improvement gained from including more participants. On the contrary:

a. The parcellation on n=1200 participants produces some non-contiguous voxels in some clusters as can be seen below (for example, in yellow = AB and mid-blue = CoN).
b. An important indicator of the reliability of a parcellation is its approximate similarity across hemispheres. In essence a similar parcellation in two hemispheres amounts to a type of replication. We developed a symmetry index by flipping the left hemisphere onto the right hemisphere (using fslswapdim followed by a non-linear registration of the entire swapped left hemisphere onto the right hemisphere). This transformation was then applied to each nucleus, and the proportion of overlap was determined for the corresponding binary nuclei of the two hemispheres. This showed that our original parcellation was slightly more

symmetrical across hemispheres than one based on a larger amount of less suitable data (mean symmetry (% overlap): 0.58 in n=200 vs 0.56 in n=1200). This is even though it relied on only one sixth of the data, which speaks in favour of the physiological noise processing being beneficial.

c. And finally, our original parcellation shows closer similarity with post-mortem atlases in terms of the location and size of the nuclei we would expect to find.

Top: n=200 original 3T participants, with physio correction
Bottom: n=1200 complete set of 3T participants, no physio correction



We note that for 7T participants, unfortunately no physiological traces were recorded. Nevertheless, we can expect more signal in subcortical regions in general at higher field strength, so we decided to use these data for our replication, nevertheless. We are not aware of any other equally high-quality dataset available that would have been more suitable.

Taken together, without knowledge of the ground truth, we cannot be sure, but we believe there is good reason to believe that the additional preprocessing for physiological noise correction helps: it changes the tSNR in regions known to be affected by pulsation artefacts and it improves the parcellation in a subcortical structure despite relying on a smaller amount of data. The parcellation is solely evaluated based on known anatomical criteria (contiguous clusters; symmetry across hemispheres; similarity to post-mortem work).

Nevertheless, to strengthen our conclusions, we have now performed several replications of our data – including to a 7T dataset without physiological noise correction but with better signal-to-noise in subcortex. We have summarized these changes fully under question 2 above and hope that this addresses the reviewer's concern.

==========

CHANGES IN MANUSCRIPT

Results p.6

This pattern of tSNR changes is in line with prior work that shows physiological noise particularly affects brainstem and other subcortical regions because of their proximity to major vessels and pulsating fluid-filled spaces[29].

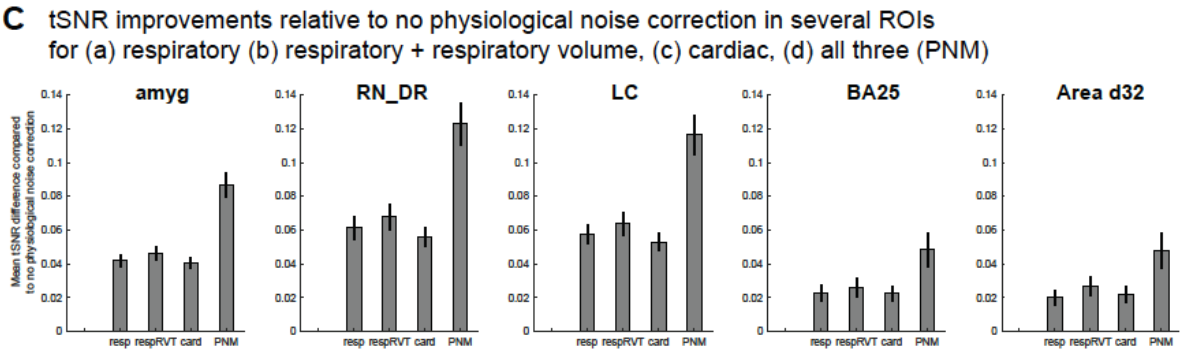
Methods p.29 + p.31

It is not possible to validate the additional pre-processing step for physiological noise correction without knowledge of the ground truth, which is not available in resting-state data. However, there is good reason to believe that our additional pre-processing helped recover meaningful signal in our data. First, because it resulted in tSNR changes in regions known to be most affected by pulsation artefacts (**Supplementary Fig 1A**), consistent with prior work. Second, because it improved a parcellation of the amygdala in two subsets of n=200 participants corrected for physiological noise, compared to the full dataset of all n=1200 3T-HCP participants not corrected for physiological noise artefacts (see below); the parcellation was solely evaluated based on known anatomical criteria (contiguous clusters; symmetry across hemispheres; similarity to post-mortem work).

[…]

Following the exact same procedure, we closely replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**). However, a parcellation using the data from all n=1200 3T-HCP participants, which had not been corrected for physiological noise, showed less symmetry and anatomical plausibility despite relying on more data. We used the parcellation generated from the first 3T dataset for further analyses. Importantly, since this parcellation was obtained from the group connectome, rather than for each subject individually, it did not introduce bias in subsequent analyses focusing on individual differences.

New Supplementary Figure 1C



Supplementary Figure 1, Additional preprocessing to account for physiological noise, A, […] **C**, The mean tSNR difference achieved with subsets of the physiological noise regressors is shown compared to the baseline of not performing any physiological noise correction. Improvements are illustrated for several regions of interest (ROIs) including amygdala, dorsal raphe (RN_DR), locus coeruleus (LC), and areas 25 and d32 in medial PFC. The regressors used were either just respiration, both respiration and respiratory volume over time (RVT), just cardiac activity, or all of the above (which is what was ultimately used in the main analysis). This shows that subcortical ROIs benefited more from physiological noise correction, with greatest improvements in brainstem nuclei, and that respiratory and cardiac regressors contributed about equally to the improvement, with the greatest improvements achieved when including all noise regressors.

==========

4. Amygdala clusters: Why were 7 amygdala clusters chosen? Were the clusters consistent across participants? The assignment of labels to the amygdala clusters is unclear and would greatly benefit from a more detailed description.

The parcellation was only performed once on the group average dense connectome, not for each participant separately. Hence, the same parcellation was used for all participants which is important because it ensures subsequent analyses on interindividual differences are completely independent. This relates to the comment raised above in point 1 (and the response that we provided).

Cluster labelling

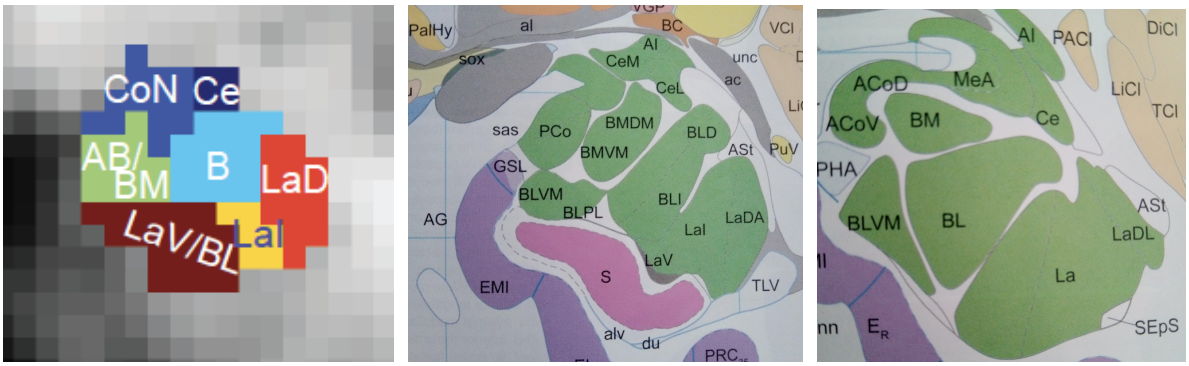
We had devoted a paragraph in our Methods to convey the rationale for our labelling of the clusters. We are sorry if this was not sufficiently clear or convincing. We realise that, while there is little ambiguity for some nuclei, others could be labelled in multiple ways. We mainly followed the Mai & Paxinos terminology, but also a high-resolution post-mortem study by Saygin and colleagues.

- Saygin, Z. M. et al. High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. NeuroImage 155, 370–382 (2017).
- Mai, J. K., Majtanik, M. & Paxinos, G. Atlas of the Human Brain. (Academic Press, 2015).

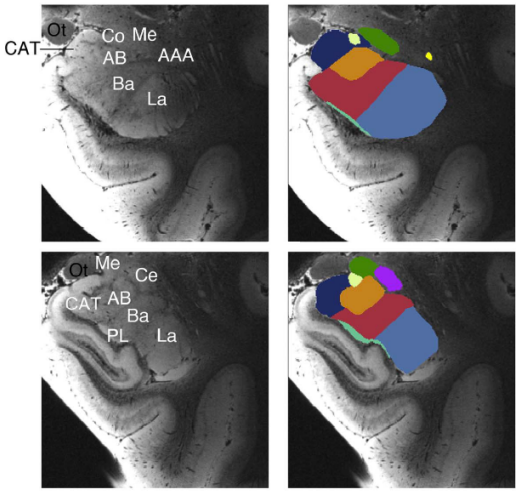
To make it easier to understand how we chose our labels, we have now tabulated the main nuclei labels in the literature: our own alongside the two human parcellations mentioned above, and the two most influential non-human primate amygdala nomenclatures. Each row shows the label we chose to give to our cluster followed by the most closely corresponding nucleus and its label in previous parcellations (* denotes labels changed in response to this reviewer's comment):

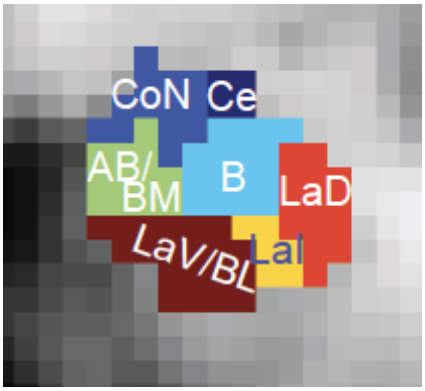
This study (human)	**Mai et al. (human)**	**Saygin et al (human)**	**Allen Brain Institute (34yo-human)**	**Fudge et al (macaque)**	**Barbas et al (macaque)**
CoN	PCo/ACoV/ACoD	CAT/Co	CoPv/CoPd	PAC/Me	Co (AAA/Me)
Ce	CeM/CeL	Ce/Me	CEm/CEl	CeN	Ce
B	BLI/BLD/BMVM/BMDM	Ba	BLD/BMD/BLI/BMV	Bi/Bmc	BL
AB/BM*	BM/BLVM/BLPL	AB	AHA/BMV	ABmc/ABpc	BM
LaD	LaD (L/A/M)	La	LaD	L	La
LaI	LaI	La	LaI	L	La
LaV/BL*	LaV/BLI	La/Ba/PL	LaV/BLVI	L/Bpc	La/BL

This table has now been included as Supplementary Table S1 and shows that there is close correspondence between many of our labels – including Ce, LaD, LaI and LaV, and the Mai et al terminology (1st and 2nd column) which can hopefully be appreciated when putting our parcellation next to two coronal slices from Mai et al.'s brain atlas:



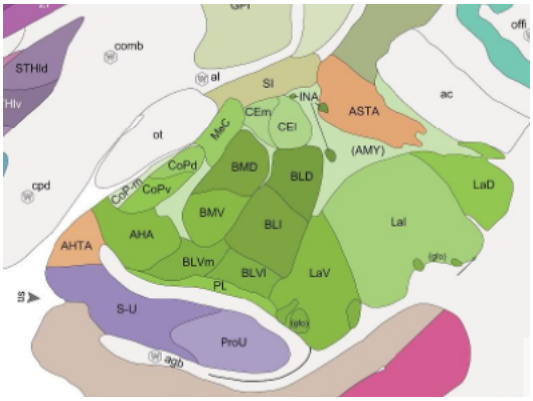
Similarly, there is close correspondence between our labels and those used by Saygin et al. in their post-mortem work – including Ce, B, AB and the lateral nucleus (which we subdivide into several subdivisions). Saygin's cortico-amygdaloid transition (CAT) area most closely corresponds to the cortical nuclei (CoN) in our parcellation. Again, this might be best seen when putting these two parcellations side-by-side:

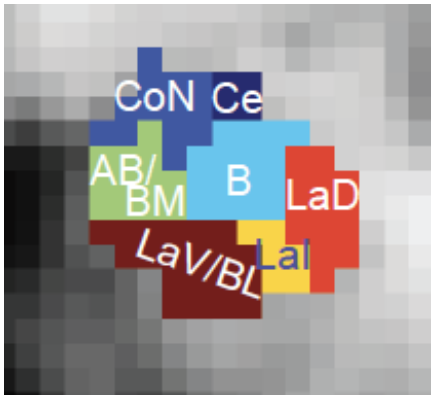




We note that what is referred to as BM in the Mai et al. terminology, is referred to as AB in the Saygin atlas. To reflect both of these traditions, we have therefore decided to re-label our AB nucleus as **AB/BM.**

Finally, our labels match the Allen Brain Insitute 34-year old human reference atlas well (https://atlas.brain-map.org/), as can be appreciated in the picture below and the table above, but in their parcellation, they further subdivide the basal nucleus.





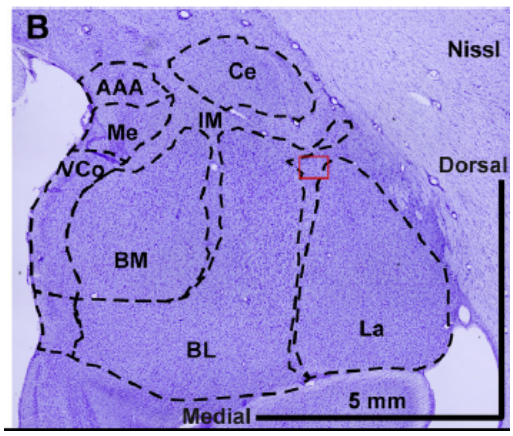
One small nucleus that we did not delineate in our parcellation is the medial (Me) nucleus. In our Methods, we explained how our central (Ce) nucleus likely contains the medial and lateral divisions of

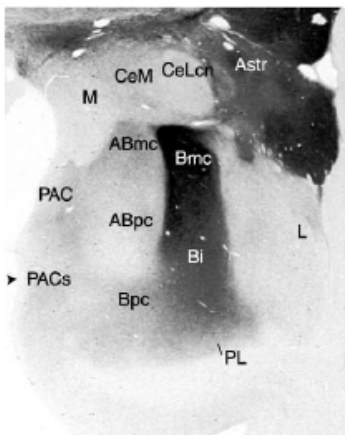
the central nucleus, but how it is less clear whether the medial amygdaloid nucleus (Me) is included as part of Ce or CoN in our parcellation.

Furthermore, we elaborate on the fact that our basal amygdala region (B) likely contained Mai et al.'s ventral and dorsal basomedial nuclei (BMVM and BMDM), and probably also its basolateral paralaminar and intermediate subdivision (BLPL and BLI), and thus the majority of basomedial and basolateral aspects of the basal nucleus.

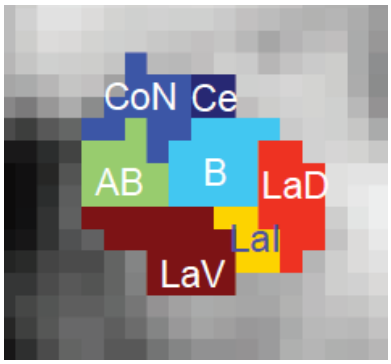
While those are the references that we reported in the manuscript, and which reflect work conducted in humans, we also looked at similar atlases in macaque monkeys when evaluating and labelling our parcellation. This is appropriate given subcortical regions such as the amygdala are highly preserved across primate species. The below figures show the standard nomenclature used by two of the leading neuroanatomists in the field – Helen Barbas and Julie Fudge. Once again, as noted above for the Mai vs Saygin terminology, what Julie Fudge tends to refer to as BM (basomedial) is often labelled AB (ABmc and ABpc) in work from Helen Barbas, which is now better reflected in our new label for this nucleus as **AB/BM.** Similarly, it seems that the most ventral portion of the lateral nucleus in our parcellation would – given its anatomical position – be considered as part of BL by Helen Barbas and labelled Bpc by Julie Fudge (which refers to the parvicellular subdivision of the basal nucleus), while the more dorsal part we labelled B would comprise what Julie Fudge labels Bi and Bmc – the basal nucleus' intermediate and magnocellular subdivisions). To reflect that it is mostly the human LaV (especially given the expanded size of the lateral nucleus in humans), but might contain a portion of BL, we have now labelled this nucleus **LaV/BL** and included additional references to this non-human literature in our manuscript. We have revised the Methods section on the labelling accordingly but would of course welcome any further suggestions on how to best label our nuclei from this reviewer.

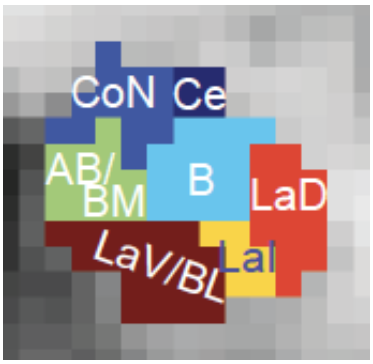
Left: Nissl stain from Zikopoulos … Barbas 2016
Right: Coronal section stained for AChE from Fudge Haber, 2002





To summarize, our original labels have been changed as follows:



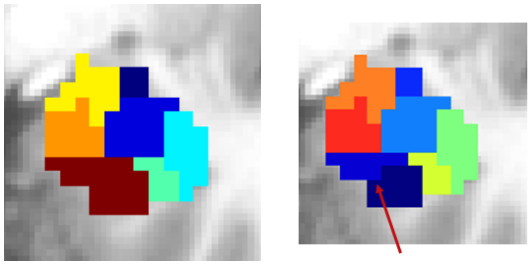


Depth of clustering

With respect to the depth of the hierarchical clustering, we carefully inspected parcellations into more clusters (for example, to see whether the medial nucleus might emerge), but we felt that our parcellation already provided a lot more detail than any previous *in vivo* atlas of the human amygdala (e.g., three subdivisions of the lateral nucleus), and that further parcellation steps might not aid the sensitivity of our planned analyses. However, we now describe the next clustering steps here and in **Supplementary Figure 2**, so that other researchers can choose their desired clustering depth.

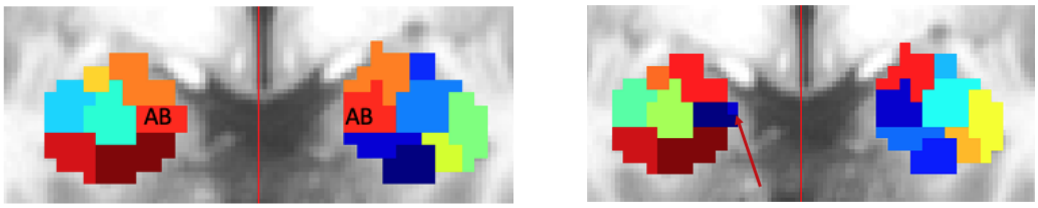
In the next clustering step, the LaV/BL nucleus shown in dark red on the left (our chosen parcellation), gets broken down into two clusters in the right hemisphere (middle, two shades of blue).

Hierarchical depth 12 Hierarchical depth 13



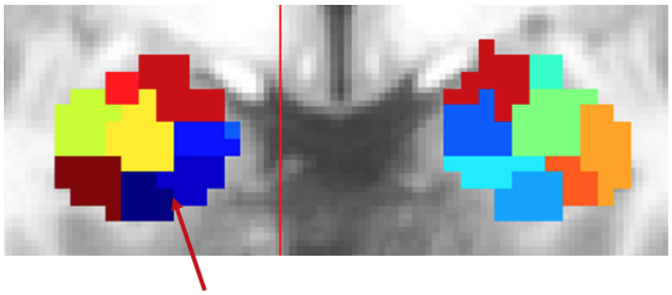
In the next step, the AB clusters in both hemispheres (which were so far part of one cluster spanning both hemispheres; shown in red) are split into two, but one voxel is included in the wrong hemisphere which is less anatomically plausible and reduces the symmetry of the parcellation (two shades of blue; see arrow).

Hierarchical depth 13 Hierarchical depth 14



The next step restores some symmetry across hemisphere because LaV/BL is broken down into two clusters in the left hemisphere as well, roughly matching the right hemisphere again (see below). We have now included these further steps in **Supplementary Figure 2** so that other researchers interested in those particular subdivisions can use this even more detailed parcellation if preferred.

Hierarchical depth 15



In summary, while future work could further improve the parcellation (e.g., based on higher resolution data), we have already received requests from several labs to share the parcellation with them (based on our BiorXiv preprint) and we hope that it would present a useful tool for other researchers interested in the more fine-grained anatomy of the amygdala.

==========

CHANGES IN MANUSCRIPT

Results p.7/8

Hierarchical clustering resulted in parcellations of the amygdalae into increasing numbers of clusters. We evaluated parcellations obtained at different clustering depths, based on several anatomical considerations. First, we carefully compared the location and size of the obtained clusters to known anatomical subdivisions of the amygdala [30–33]. Given post-mortem work in humans and a large body of work in animals, we expected an anatomically plausible cluster solution to show good symmetry across hemispheres (see Methods) and to contain spatially contiguous clusters. Using these criteria, we chose a parsimonious and anatomically plausible parcellation for further analyses. This parcellation contained seven subdivisions in each hemisphere (**Fig 1B**; see **Supplementary Fig 2A** for shallower and deeper clustering steps).

[…]

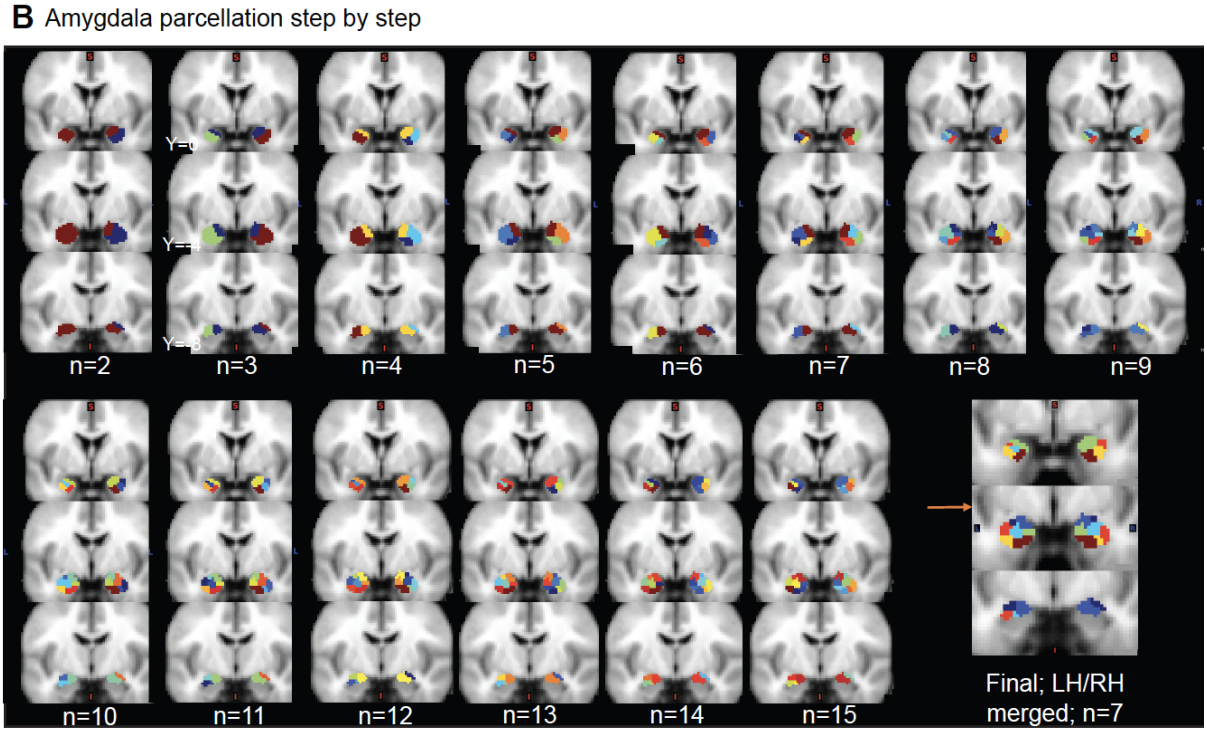
To facilitate links to other studies, we assigned each cluster a putative label, corresponding to nuclei that have previously been identified (see Methods). As a guide, we used the best match in size and position when comparing our clusters with several atlases of the human amygdala [32–34] (**Fig 2A**). The seven nuclei were labelled central nucleus (Ce), cortical nuclei (CoN), auxiliary basal or basomedial nucleus (AB/BM), basal nucleus (B), and lateral nuclei (dorsal portion: LaD; intermediate portion: LaI, ventral portion, containing portions of basolateral: LaV/BL). The rationale for our choice of these labels is further explained in the Methods, and its correspondence to nuclei labels in other macaque and human investigations is summarized in **Supplementary Table 1**.

Methods p.30-32

To evaluate the number of clusters, or in other words, the appropriate depth of the hierarchical clustering, we aimed for a balance between simplicity and detail, as well as anatomical plausibility. We carefully compared the location and size of the clusters obtained at increasing levels of depths to known anatomical subdivisions of the amygdala [30–33]. Given post-mortem work in humans and a large body of work in animals, we expected an anatomically plausible cluster solution to show good symmetry across hemispheres, i.e., to contain corresponding clusters across left and right hemispheres, and to involve spatially contiguous clusters. Another focus was on detail: […] Throughout the results, we therefore focussed on the depth 12 cluster solution, which when merging corresponding clusters in both hemispheres yielded seven final clusters (**Figs 1B and 2A**). Other shallower and deeper clustering depths are shown in **Supplementary Fig 2**. […] Similarities between our labelling and the labels of those two atlases as well as two common nomenclatures used in non-human primates are shown in **Supplementary Table S1**. **Supplementary Table S2** also summarizes the size of all clusters. The most dorsal, posterior and lateral nucleus (dark blue in **Figs 1B and 2A**) which was also the smallest in size (62 voxels across both hemispheres) perfectly matched in its size and position the central amygdaloid nucleus and was therefore labelled **Ce**. Judging from its position and size, it contained both medial and lateral divisions of the central nucleus (CeM and CeL) [32]. However, it is less clear whether it also contained the medial amygdaloid nucleus. The medial amygdaloid nucleus might have been part of this 'Ce' cluster or the adjacent cluster (middle blue in **Figs 1B and 2A**) which was positioned in a dorsal, posterior and medial location where the cortical amygdaloid nuclei are located (e.g. PCo=posterior cortical; ACoV and ACoD = anterior cortical, ventral & dorsal parts [32]; also referred to as CAT = cortico-amygdaloid transition area e.g. in [33]). We therefore labelled this adjacent cluster **CoN**, as an agglomeration of the cortical nuclei of the amygdala. It contained altogether 133 voxels across left and right hemispheres, and possibly comprised cortical nuclei as well as the medial nucleus. Ventral and anterior to the Ce and CoN nuclei, in a medial position within the amygdala (light blue in **Figs 1B and 2A**), we identified a portion of the basal amygdala which very likely contained Mai et al.'s ventral and dorsal basomedial (BMVM and BMDM), and probably also its basolateral intermediate and dorsal subdivision (BLI, BLD), and thus the majority of basomedial and basolateral aspects of the basal nucleus. We therefore refer to it simply as the basal nucleus **B**. It contained 74 voxels across hemispheres and was adjacent to a slightly more medial subdivision of the basal nucleus which we refer to as auxiliary basal/basomedial (**AB/BM**, green in **Figs 1B and 2A**) and which contained 104 voxels across hemispheres. This cluster, AB/BM, based on its size and location, would have contained the ventromedial part of the basolateral nucleus (BLVM) as well as paralaminar subdivision of the basal nucleus (BLPL) in [32]

and closely corresponded to what Saygin and colleagues [33] label AB as well. The remaining three clusters made up the lateral nucleus of the amygdala, namely its dorsal, intermediate and ventral portion (**LaD**, **LaI**, **LaV**, respectively, in red, yellow and dark red in **Figs 1B and 2A**). These clusters contained 84, 86 and 104 voxels, respectively. Because the ventral part LaV extended more medially, it likely contained parts or all of the basolateral intermediate (BLI) nucleus (or Ba/PL in [33]) and is therefore referred to as **LaV/BL**. Overall, there was good correspondence between the position and labelling of our nuclei, and those reported in [30–33] (**Supplementary Table 1**).

Supplementary Figure 2 now contains hierarchical steps n=13 to n=15:



Supplementary Figure 2, Amygdala parcellation at different clustering depths, A, […] **B**, Individual steps of the hierarchical clustering algorithm led to increasing subdivisions of the amygdala. All steps leading up to our final parcellation (depth 12), and a few additional clustering steps beyond it (up to depth 15), are shown. […]

==========

5. The predictors are not anatomical connections but rather resting-state correlations. Please revise throughout accordingly, and especially in the abstract. Further, some statements such as " likely that the negative coupling found between them reflects an indirect interaction mediated by another brain region (p.8)" are not based on data but are rather conjectures so should be saved for the Discussion.

Thank you – we have revised the manuscript throughout to reflect both changes. Apologies for our sloppy use of the word anatomical connection. We agree it is not precise. There is evidence that resting-state correlations reflect anatomical connectivity to some extent (see e.g. O'Reilly … Baxter, PNAS, 2013), but of course co-fluctuations are not restricted to monosynaptic connections. We have now made this point more explicitly and changed the wording as requested.

==========

CHANGES IN MANUSCRIPT (selected, but all highlighted in red in the revised manuscript)

Abstract p.2:

Finally, for each behavioural dimension, we identified the most predictive resting-state functional connectivity between individual amygdala nuclei and highly specific regions of interest such as the dorsal raphe nucleus in the brainstem or medial frontal cortical regions.

Discussion p.24:

There is evidence that resting-state correlations reflect anatomical connectivity[38]. However, it is worth noting that, while rs-fMRI was used as a proxy for anatomical connectivity here, the patterns in resting-state functional connectivity we identify do not necessarily correspond to monosynaptic connections. Monosynaptic connections might dominate the positive functional connectivity values in **Figure 2B** which illustrates the mean functional connectivity of the amygdala nuclei with other ROIs. However, the negative functional connectivity observed between dlPFC regions and the amygdala likely reflects an indirect interaction mediated by another brain region given limited direct connections between the homologue of this region and the amygdala in macaques. Similarly, it is possible that the relations we identified between functional connectivity and mental health indices (**Figures 4-6**) may rely on a multi-component connection pathway or may involve connections between two amygdala nuclei.

==========

6. How were the ROIs defined and identified? Maximally adjusting thresholds to maximise anatomical plausibility (p. 29) does not seem well justified. Why were those particular references chosen as the ROIs that interconnects with the amygdala nuclei? How big are the ROIs? Functional correlations may be impacted by the size of the regions. Also, some of the regions are likely tiny and may have less reliable/more noisy fMRI signal due to having one or two voxels (e.g. BNST).

ROI size

We agree, these are important points. We had already summarized the size of all ROIs in the Methods (in terms of the numbers of voxels or vertices) but have now included this information in a table to make it more easily accessible (pasted below). In doing so, we noticed that the size of BNST was missing from the manuscript – apologies for this oversight, it has 45 voxels which has now been added. All masks were generated at the beginning of the study before we started performing any of our key analyses. Note that our analyses used the average time course from each ROI. Thus, while a smaller ROI might have a less reliable average time course compared to a larger ROI, all ROIs were given equal weight in the analyses. In general, it is not the case that we find stronger effects for larger ROIs, and with the exception of the

median raphe (RN_MR), all ROIs had at least 20 voxels/vertices. We have now made this clearer in the revised manuscript.

ROI definition

***Subcortex*:** Given the size of some of the brainstem ROIs of interest, masks produced at a 2mm resolution will not be perfect, so we fully understand the concern of this reviewer. However, if our masks are suboptimal or too small or we are simply capturing noise, this would bias us against finding any meaningful effects. Instead, what we can show in **Figure 2B** is that the average connectivity between many of our subcortical regions is strongest with the central nucleus, as we would expect from tracer work in macaques, and we have been able to replicate this result in two independent datasets now (new Suppl Fig 4), even though the central nucleus is the smallest nucleus of the amygdala. From the literature and available higher-resolution atlases, we also have information about the shape, volume and position of each ROI. So while a 2mm voxel size will not allow us to track them perfectly, we carefully and visually inspected all subcortical ROIs individually at 2mm to try and achieve the best tracing of their shape, volume and position when compared with higher-resolution versions from published atlases. However, as the reviewer noted, this meant that we did not apply the exact same threshold for each ROI; we deviated from our standardized procedure in three cases to improve the masks: nucleus accumbens, dorsal and median raphe. In these three cases, slightly different thresholds provided more anatomically plausible ROIs. For full transparency, we are in each case reporting the atlas we used, the thresholding procedure we applied, we show the ROIs visually (Figure 2C) and we would of course be happy to publish all masks along with this manuscript should it be accepted. We have now added more detailed explanations to the revised Methods section.

Cortex: In the cortex, we used what is currently considered to be the best parcellation published in surface space (Glasser et al., Nature, 2016).

ROI selection

Our ROI selection was entirely hypothesis-driven and followed two key criteria: (a) knowledge about a region's mono- or disynaptic connectivity with the amygdala and (b) knowledge about a region's importance in mental health/mood disorders. Most of our ROIs fulfilled both criteria. For (a), there is a large literature using axonal tracing methods in macaque monkeys that has characterized the anatomical connections of the amygdala. The majority of the ROIs we included have monosynaptic connections with the amygdala (or some brainstem and dlPFC regions, di-synaptic via the hypothalamus or medial PFC, respectively), and most regions with strong amygdala connectivity were included. We left out some regions with monosynaptic connections to the amygdala – for example anterior temporal lobe regions – because we did not have strong reasons to believe they were relevant for the behaviours of interest to us. Other notable omissions are the hypothalamus and hippocampus which we would have needed to parcellate and which it would not have made sense to include as whole structures. We have now mentioned these regions in the Discussion. We chose regions with di-synaptic amygdala connections based on strong evidence for their role in mental health disorders. For example, we included dorsolateral PFC because manipulating dlPFC using non-invasive neurostimulation can directly alter activity in the amygdala and is used as a treatment for depression (Ironside et al., 2018; Furtado et al., 2013; Dichter et

al., 2015; Salomons et al., 2014). We have now tried to clarify our selection criteria in the corresponding section in the Methods. If the reviewer has particular regions in mind that we have overlooked, we would of course be interested to know what they are.

==========

CHANGES IN MANUSCRIPT

Supplementary Table 2: Vertex/voxel size of *a priori* ROIs and amygdala nuclei

ROI	Size
Subcortex	Voxels
SN	134
NAc	188
BNST	45
vlPAG	43
dPAG	45
RN_DR	23
RN_MR	8
LC	20
Cortex	Vertices
25	54
a24	89
p24	66
a24pr	75
s32	55
p32	122
d32	147
a32pr	163
p32pr	190
9m	408
pOFC	83
FOP1	61
FOP2	101
FOP3	83
FOP4	240
FOP5	193

46	316
9-46d	379
a9-47v	147
p9-46v	214
Amygdala nuclei	**Voxels**
Ce	62
CoN	133
BaL	74
AB/BM	104
LaI	84
LaD	86
LaV/BL	104

Results p.9

Supplementary Table 2 summarizes all included ROIs with their respective sizes.

Discussion p.20

These ROIs were selected *a priori* based on strong, often monosynaptic, connectivity with the amygdala in animal tracer work and their relevance for mental health related functional processes.

Methods p.33 and p.35-36

ROI selection

We had several *a priori* regions of interest which were informed by prior work, including anatomical work using tracers in macaque monkeys as well as work in humans with mental health disorders. Our ROI selection was thus entirely hypothesis-driven and followed two key criteria: (1) knowledge about a region's mono- or disynaptic connectivity with the amygdala and (2) knowledge about a region's importance in mental health/mood disorders. Most of our ROIs fulfilled both criteria. The majority of the ROIs we included are known to have monosynaptic connections with the amygdala (or in the case of some brainstem and dlPFC regions, di-synaptic connections that are made via the hypothalamus or medial PFC, respectively), and most regions with strong amygdala connectivity were included. We left out some regions with monosynaptic connections to the amygdala – for example anterior temporal lobe regions – because we did not have strong reasons to believe they were relevant for the mental health-related behaviours of interest in this study. Other notable omissions are the hypothalamus and hippocampus which we would have needed to parcellate and which it would not have made sense to include as whole structures. All our ROIs are illustrated in Fig **2C** and will be motivated one by one below. For an overview, all ROIs with their respective sizes are listed in **Supplementary Table 2**. ROI selection was performed prior to and independent of subsequent analyses focusing on individual differences.

[…]

The BNST mask was obtained from [100] and contained 45 voxels.

[…]

Probabilistic masks were binarized first, including all voxels with probability >.25, in other words, voxels that had a larger than 25% chance of being within the given region (NAc, SN). Binary files and all masks we received in binary format (BNST, PAG, LC, RN) were subsampled to 2mm. While most regions were then simply binarized again using any voxels >.25 in subsampled space, we carefully and visually inspected all subcortical ROIs individually at 2mm to try and achieve the best tracing of their shape, volume and position when compared with a higher-resolution versions from published atlases. However, this meant that we did not apply the exact same threshold for each ROI; we deviated from our standardized procedure in three cases to improve the masks: nucleus accumbens, dorsal and median raphe. In these three cases, slightly different thresholds provided more anatomically plausible ROIs. NAc thresholded at .25 would have yielded an unusually large ROI, so a threshold of >.75 was applied in the second step; for the raphe nuclei, thresholds were adjusted manually to maximise anatomical plausibility (>.6 and >.72 for dorsal and median, respectively). Importantly, this was done based on anatomical criteria prior to any subsequent analyses. All ROIs are shown in Figure 2C and published as part of the OSF repository ([reference to be inserted upon publication]).

In summary, we included a total of eight subcortical and brainstem regions (**Fig 2C**). A full list of all ROIs detailing their respective sizes in voxels/vertices is shown in **Supplementary Table 2**; apart from the median raphe (RN_MR), all ROIs had at least 20 voxels/vertices.

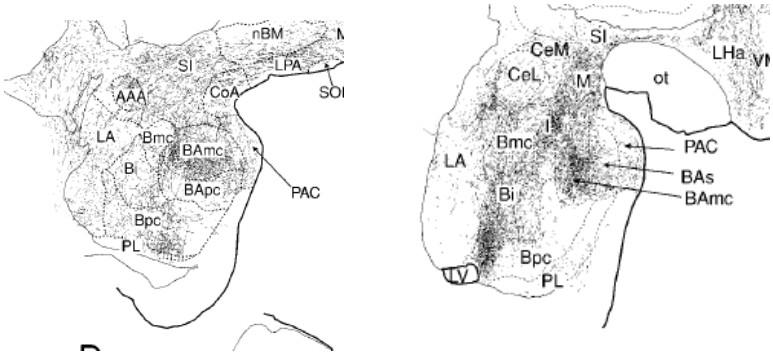
For analyses of group and individual data, we computed the average time course from each ROI. Thus, while a smaller ROI might have a less reliable average time course compared to a larger ROI, all ROIs were given equal weight in subsequent analyses. In general, we did not find that most effects pertained to larger ROIs, suggesting that even smaller ROIs were of sufficient size to capture meaningful signal. This notion was corroborated by the mean pattern of functional resting-state connectivity between amygdala nuclei and ROIs shown in **Fig2B** which we robustly replicated in two independent data sets (**Supplementary Fig 4B**).

==========

7. On p. 8, "Having established and validated the network" was confusing because I was not able to find a validation section in the Results. If the authors mean the connections they listed in the paragraph above, I do not think this counts as validation. The list of ROIs/connections was in no way exhaustive of amygdala subnucleus connections. What connections would tell us that the network was invalid?

We were referring to the average connectivity pattern shown in Figure 2B, but we agree calling it a validation was not appropriate and we have removed this formulation from the manuscript. Instead, what we did is to compare the average connectivity pattern between amygdala nuclei and ROIs with the pattern we would have excepted from tracer work in macaques. As an example, following retrograde tracer injections into area 25, Freedman et al (J of Comp Neurol, 2000) obtained the heaviest labelling of

cells in the parvocellular basal nucleus and magnocellular accessory basal nucleus, and a moderate amount of labelling in the medial/cortical nuclei (pasted below). In our resting-state data, we find that area 25 has the strongest connectivity with B, AB and CoN (Figure 2B), which seems to match Freedman et al.'s finding well (and also matches findings from a more recent study by Sharma and Fudge in CerCortex, 2020). This increased our confidence in our data both in terms of anatomical plausibility and the presence of meaningful signal.



==========

CHANGES IN MANUSCRIPT

We have reworded this section as follows to be clearer about what we meant:

Results p.9

Having established the anatomical plausibility of mean functional connectivity values between amygdala nuclei and our ROIs, given published tracer work, we next sought a characterization of dimensions related to participants' mental well-being.

==========

In addition, we have now replicated this average pattern of connectivity in two additional datasets as described above (n=200 additional 3T participants and n=98 7T HCP participants), which have both been included in Suppl Figs 3 and 4 pasted above in response to point 2.

8. What were the model fits like? The comparison between models from the whole amygdala vs. subnuclus parcellation seem arbitrary but perhaps would benefit from clarification or comparison of model fits. Please clarify what is meant by "If the probability of the parcellated amygdala connection is lower than the threshold set by the whole amygdala, we can infer that the parcellation increased our sensitivity." p. 13. Is this how the models were compared? Why is it inferred that the parcellation increased sensitivity (and is this ROC sensitivity?) And are these tests corrected for multiple comparisons across the 7subnuclei X ROI connections?

As described above, the last part of the manuscript that this comment refers to has been rewritten with the inclusion of the 7T data as a replication dataset and the use of a large sample of n=400 3T-HCP participants for feature selection.

Nevertheless, as before, to establish whether we achieve improvements in prediction accuracy from parcellating the amygdala into nuclei, we include a control analysis in which we use the functional connectivity of the whole amygdala, rather than its subnuclei. Our results are consistent with those reported in the original manuscript, showing that parcellating the amygdala leads to better out-of-sample predictions. However, the statistics associated with this result have been improved and we now correct for multiple comparisons (using bootstrapping to obtain comparable p-values across the version with 7 nuclei vs 1 amygdala, see below). We also use more careful wording in this section to be clear about what we have done and how we measure improvements in predictive power. Throughout the manuscript, prediction accuracies are evaluated as Pearson's correlation coefficient between the true behavioural scores and connectivity-based predictions of the behavioural scores. All our predictions use weights estimated in a separate sample (3T cohort) and are used to predict an independent dataset (7T cohort). Importantly, for each statistical test, we generate appropriate null distributions using bootstrapping (see also the related but rather extensive point #2 by Reviewer 2; the section summarizing the relevant changes can be found by searching "Paragraphs dealing specifically with permutation testing").

==========

CHANGES IN MANUSCRIPT

Results p.17/18:

Comparing the accuracy of nuclei-specific and whole-amygdala predictions

In the next step, we tested whether parcellating the amygdala into subnuclei increased the accuracy of predictions of mental health dimensions. We repeated the robust regressions for the amygdala as a whole, i.e., using functional connectivity between the entire amygdala and the same set of 28 ROIs. As before, the robust regression weights of all 28 edges derived from the 3T participants were applied to the 7T participants to obtain out-of-sample predictions. This produced significant predictions of negative emotions (Pearson's r=0.2, p=0.026) and anger (Pearson's r=0.18, p=0.037; **Fig6A**). However, predictions for all four behavioural dimensions were worse when considering whole-amygdala rather than nuclei-specific amygdala functional connectivity (LifeSat: nuclei: r=0.19, whole: r=0.17; NegEmot: nuclei: r=0.22, whole: r=0.20; Sleep: nuclei: r=0.05, whole: r=-0.09; Anger: nuclei: r=0.22, whole: r=0.18; compare **Figs 4E and 6A**). We note that – while the nuclei-specific predictions include seven times as many predictors as used in the whole amygdala control – the whole amygdala is made up of the exact same voxels as the seven nuclei together. If the subdivisions of our parcellation are not meaningful or if the fMRI signal in individual nuclei and therefore the derived functional connectivity values for individual nuclei are not robust, then separation into smaller subsets of voxels should merely increase the noise in our predictions. Alternatively, if there is value in considering amygdala nuclei separately, then using functional connectivity values from each nucleus might increase our accuracy for predicting mental health dimensions. Nevertheless, to account for the number of predictors, we generated separate null

distributions for the nuclei- and whole-amygdala versions to obtain directly comparable p-values (Methods). This slightly changed the precise p-values, but led to the same conclusions, suggesting negative emotions and anger could be predicted significantly (in relation to this this null distribution: p=0.026 and p=0.038, respectively). However, most importantly, the predictions of all four mental health dimensions still significantly benefited from considering the functional connectivity of separate amygdala nuclei as opposed to the amygdala as a whole (comparable p-values relative to appropriate null distribution: LifeSat: nuclei: p=0.034, whole: p=0.051; NegEmot: nuclei: p=0.016, whole: p=0.026; Sleep: nuclei: p=0.310, whole: p=0.807; Anger: nuclei: p=0.015, whole: p=0.038; **Fig 6B**). The same conclusion was reached when comparing the peak accuracy instead of the prediction achieved with the full set of edges (**Supplementary Fig 9A**). Peak accuracy using whole-amygdala edges occurred at n=16, 12, 3 and 9 for the four behavioural dimensions and fingerprints and scatterplots are shown for the earliest significant peak (**Fig 6C,D**) and global peak (**Supplementary Fig 9B**) for comparison with the nuclei version. However, even at the peak, prediction accuracies from whole-amygdala functional connectivity were worse than those achieved using nuclei-specific predictions reported above (correlation coefficients for whole-amygdala functional connectivity at the peak: r=0.185, 0.202, 0.127, 0.198; **Supplementary Fig 9**). Thus, despite their small size, considering the functional connectivity of individual subcortical nuclei benefitted predictions of mental health dimensions.

Methods p.43

Controlling for amygdala parcellation and dimensionality of behaviour

To show that parcellating the amygdala yielded improvements in prediction accuracy, we also repeated the regression procedure with only the connections from our ROIs to the entire amygdala instead of all individual nuclei (a total of 28 possible predictors). All figure panels related to connections with the whole amygdala, instead of its seven distinct nuclei, were generated using identical methods (**Fig 6 and Supplementary Fig 9**). Again, robust regression coefficients obtained from fitting the 3T participants were applied to the 7T participants' functional connectivity values to predict the 7T participants' behavioural scores (**Fig 6A**). However, instead of using 196 FC predictors, predictions were derived from only 28 FC values. Unlike in other situations, however, where a larger number of predictors is expected to perform better, here, the predictors captured the same information in both cases, namely functional connectivity with the total of all amygdala voxels. Subdividing the amygdala into smaller groups of voxels could have two potential effects: if subdivisions are meaningless or BOLD measurements in smaller sets of voxels too noisy, predictions generated from individual nuclei functional connectivity might be worse than those obtained from more robust whole-amygdala functional connectivity, despite including more predictors; alternatively, if amygdala subdivisions into nuclei are meaningful, this might increase the prediction accuracy of the nuclei-version over and above the whole-amygdala version for predicting mental health dimensions.

Nevertheless, to make predictions obtained from whole vs nuclei-specific amygdala functional connectivity more comparable, we generated null distributions in both cases. We shuffled the order of participants' behavioural scores in both 3T and 7T datasets and repeated the above procedure of predicting 7T-behavioural scores from 3T-weights 10,000 times. The p-values extracted from the respective null distributions account for the number of predictors, allowing a direct comparison (**Fig 6B**).

Thus, this test established if parcellating the amygdala into nuclei helped us improve the accuracy of our predictions.

==========

The new **Figure 5** associated with this section of the manuscript was pasted above in response to point 2.

9. What is meant by 'reliably predicted'? p.16. Perhaps 'significantly predicted' would be more appropriate.

This section has been rewritten and this comment no longer applies, but we have tried to use precise wording throughout.

10. Rather than perform the CV 10,000 regression model approach, why not do a factor analysis on the connectivity data as well, and then perform the predictive modeling approach between the connectome factors and the mental well-being factors? What does that analysis yield?
Using a factor approach may be more informative than single connections; it is possible but perhaps unlikely that social & life satisfaction depends on one connection, for example.

We note that the precise analysis approach that this comment is referring to has changed as part of the major revisions we have performed to address the comments raised by the two reviewers. We now simply use robust regressions to relate functional connectivity in each edge to each of the four mental health dimensions.

Nevertheless, we believe this is an interesting suggestion. However, the key goal of this manuscript was to identify anatomically interpretable networks that robustly relate to variation in mental health dimensions. There are already other papers showing that we can achieve high prediction accuracies (e.g., for predicting depression) if predictions are based on functional connectivity in a large network involving many parts of the brain. We believe the novelty here is that we can show that functional connectivity in a small but interpretable network with precise amygdala nuclei explains a significant portion of the variance related to mental health dimensions. As outlined in the introduction and Discussion, we believe this is important and valuable, for example because precise interventions in psychiatric disease (e.g., with deep brain stimulation) cannot usually target large complex brain networks, but rather specific nodes in a network (such as the amygdala, or subgenual ACC as done in treatment-resistant depression).

We fully agree with the reviewer that functional connectivity in single edges is unlikely to fully explain any variability related to dimensions of mental well-being such as social & life satisfaction. We note that this is part of the reason why in this new version of the manuscript, we focus on sets of edges that we found were robust and replicable in their prediction of mental health variation across datasets.

We are reluctant to try the suggestion of the reviewer as we believe it would change the focus of the manuscript and detract from our key message. Doing a factor analysis or PCA on the functional connectivity matrix would produce weighted sums of functional connectivity values which anatomically could no longer be 'labelled' or interpreted. Of course, if the reviewer thinks this would be a critical addition, even in light of the extensive changes that we have implemented in this round of revisions to improve the robustness of our findings, we would be happy to implement it.

11. Further, it is still unclear to me how the connections are unaffected by existing correlations between predictors (given that 5 random connections are chosen for each regression model, and therefore could be highly correlated with one another). Are the beta coefficients normalized across models?

As explained in response to the previous point, given the major changes implemented in the third section of the manuscript, this analysis approach has changed and been replaced with robust regressions performed separately for each functional connectivity value.

==========

CHANGES IN MANUSCRIPT

We will not paste all sections that have changed here (many have been reproduced above already). The complete set of changes can be found in Results p.11-19, Methods p.38-44, Figures 4-7, Suppl Figs 7-10.

The section explaining the robust regressions is as follows:

Results p.11

We first established whether relationships between nuclei-specific amygdala functional connectivity and mental health dimensions replicated between the two independent (3T and 7T) datasets. We fitted robust linear regression coefficients to capture the relationship between functional connectivity values in each 'edge' (e.g., Ce to NAc) and behavioural dimension (e.g., sleep problems), separately for the 3T and 7T dataset. This resulted in 196 regression coefficients (7 amygdala nuclei x 28 ROIs) for each of four behaviours and two datasets. If amygdala nuclei functional connectivity carries no information about mental health dimensions, then, by chance, the correlation between regression coefficients obtained across behaviours in the 3T versus 7T datasets should be zero. To formally test this, we generated a null distribution by shuffling the subject order of the behavioural scores n=10,000 times while keeping the functional connectivity values unchanged. Indeed, by chance, the across-dataset replication of the pattern of regression coefficients was centred on zero (**Fig4A**). The similarity between 3T and 7T regression coefficients in the actual data, however, was significantly greater than chance (Pearson's r=0.26; p=0.0313; **Fig 4A,B**), showing that relationships between nuclei-amygdala functional connectivity and mental health dimensions were similar across datasets.

Methods p. 39

A regression approach with two types of out-of-sample replication was used to identify whether amygdala nuclei functional connectivity was predictive of the four mental health dimensions. To generate regression weights, the data to be predicted, **y**, was a 393x1 or a 97x1 vector describing the true behavioural score for each 3T or 7T participant. The matrix of potential predictors **X** was a matrix with 393 x 196 (or 97 x 196) resting-state functional connectivity (FC) values for each participant and the 196 measures of functional connectivity between the areas described above (7 amygdala nuclei x 28 ROIs) which we refer to as "edges". Such functional connectivity measures are known to be indices of anatomical connectivity, although the connections are not always monosynaptic)[38]. Outlier participants from the original pool of 400 3T and 98 (non-overlapping) 7T participants were conservatively rejected based on their individual FC values if more than 10% of their FC values across all edges deviated more than 3.5 standard deviations from the mean across participants. This identified seven 3T and one 7T participants as outliers and all analyses were performed on the remaining 393 and 97 participants. Next, confounds were regressed out of the data in a similar way as described previously [112], and this was done separately in both the 3T and 7T data. Confounds included (1) a summary statistic quantifying average head motion; (2) weight; (3) height; (4) blood pressure – systolic; (5) blood pressure – diastolic; (6) haemoglobin A1C in blood; (7) cube-root of total brain volume; (8) cube-root of total intracranial volume. A total of 8 confounds were thus regressed out of the matrix X. Both y and X were z-scored.

For generating the plots in **Fig 4B**, we estimated robust linear regression models (Matlab's function robustfit) for functional connectivity in each of the 196 edges, four behavioural dimensions, and in the 3T and 7T cohorts. In each case, the resulting robust regression weight captured the relationship between FC and behaviour. Including all edges in one large regression model was not feasible due to the large number of regressors and existing correlations between them.

To test whether the obtained robust regression weights captured meaningful relationships between the functional connectivity of individual amygdala nuclei and the four mental health dimensions, we tested whether regression weights were (a) similar between the 3T and 7T datasets ('across-dataset replication') and (b) similar between two halves (corresponding to MR sessions) of the experiment ('within-subject replication'; **Figure 4A**).

==========

We would like to thank you again for your thoughtful and very helpful suggestions. We hope that you will find that the manuscript has improved as a result of the changes we have implemented.

Reviewer #2:

Remarks to the Author:

Klein-Flügge and colleagues use data from 200 HCP subjects to explore the link between (i) functional connectivity (FC) measures between 7 nuclei of the amygdala and 28 (mainly cortical) ROIs with known synaptic connections to the amygdala, and (ii) 33 measures of mental well-being.

The focus on amygdala is motivated by its known role in emotion regulation and the parcellation into seven nuclei was defined from resting-state functional MRI data while also being consistent with previous post-mortem histological results. The 33 mental well-being measures were summarized using four interpretable latent factors reflecting sleep quality, life satisfaction, negative emotions, and anger. The weights of these factors in each subject were then related to the subjects' 196 FC metrics (7 * 28) involving the amygdala. Several of the 196 FC links were found to be predictive of the four behavioral latent scores, and this is supported using multiple statistical tests. These tests essentially consist in comparisons against predictions using FC connections that either do not involve the 7 amygdala nuclei or do not involve the other 28 ROIs used in the original analysis.

I found the paper remarkably well written and I enjoyed reading it. Below are a few questions regarding the reach and significance of the results, followed by a few smaller comments.

Thank you very much for your positive feedback and careful reading of our manuscript. We have now addressed all comments below and hope that you will find that our changes have further improved the manuscript.

1. While the statistical analysis relating FC and behavioral metrics is well executed, I am wondering whether the interpretation of these results could be developed. For example, the authors discuss the neurotransmitters known to be present in the FC connections that are predictive of the four latent behavioral factors. Beyond this, would it be possible to show a distribution map of neurotransmitters of interest (or their main pathways) and compare it to the identified significant connections? This would allow to both highlight the relevance of the nuclei clustering while also allowing to elaborate on the relevance of the proposed partitioning of well-being measures in terms of the underlying neural circuitry.

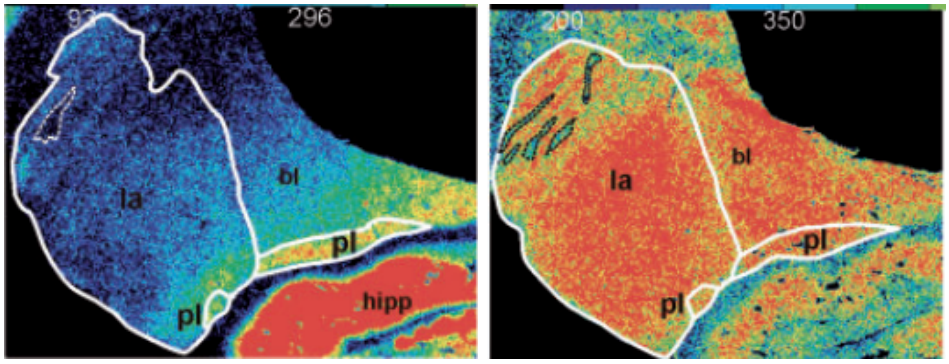
Thank you for this comment. We agree that we can improve the interpretation of our results. First, we note that based on comments from both reviewers, the way the predictive networks are presented has changed and become more nuanced in this new version of the manuscript. We provide more detail on this below; in brief, to improve the robustness and replicability of our results, the relevant networks and weights are now selected and trained based on one much larger dataset (n=393 3T participants) and then applied to an independent pool of participants (n=97 7T HCP participants). While our main conclusions hold and highly similar networks are highlighted, less emphasis is now placed on functional

connectivity in individual edges ("connections") and more emphasis on the smallest and most predictive networks (see new **Figures 5 and S7**). Fingerprints that visualise the networks now show not just which edges have meaningful resting-state functional connectivity, but also highlight the strength (width of the bar), direction (continuous vs dotted line) and relevant amygdala nucleus (colour) for which they predict behavioural scores (Fig5). While this does not address the comment on neurotransmitter systems, we believe this has improved the general interpretability of these anatomical fingerprints which contained less detail in the original version of the manuscript.

Neurotransmitters distributions

Looking at neurotransmitter distributions within the amygdala, or in relation to our other ROIs across the brain would be very interesting, but we feel is beyond the focus of our current investigation. Also, unfortunately, most available maps on neurotransmitter distributions in the human brain are at a relatively coarse spatial scale compared to our resolution of 2mm^3. This is because they were often derived from PET or other coarser methods (e.g., post-mortem multiarray assays). For example, Palomero-Ghallager & Zilles (2015) conclude on the amygdala "To the best of our knowledge, a comprehensive study of the subregional receptor expression is presently not available." Although one study by Graebenitz, Zilles & Pape (Brain, 2011) provides some details on neurotransmitter distribution within the amygdala and reports differences between nuclei (see screenshots below), conclusions of neurotransmitter distributions within the amygdala do not yet match up across investigations. Looking at serotonin receptors, for example, Graebenitz et al. suggest low mean 5-HT1a receptor densities across the amygdala on average, while Beliveau et al. (JNeurosci, 2017) find that 5-HT1a receptors are the most prominent receptor in the amygdala (see screenshot below).

Graebenitz, Zilles & Paper (Brain, 2011):
Serotonin 5-HT 1A (left) and 2 (right)



For comparison Beliveau et al. (JNeurosci, 2017), part of their Fig 4:

Given the complexity of neurotransmitter systems, inconsistencies in the currently available data in humans, and the fact that the HCP data used here does not include any information speaking to this question, we have decided to include a paragraph in the Discussion detailing how our work will need to be linked up to this literature for better interpretability in the future. While this is not quite the improvement the reviewer was asking for, we feel that it is difficult to provide any more conclusive statements given the currently available knowledge. However, we would be very curious to hear any further insights this reviewer might have on this point.

==========

CHANGES IN MANUSCRIPT

Discussion p.24:

Social and life satisfaction was most strongly predicted by functional connectivity between the amygdala and regions primarily located in medial and lateral frontal cortex, such as areas a24pr, p32pr, a32pr and d32 as well as several frontal opercular (FOP3, FOP4) and lateral prefrontal areas [37], with less pronounced negative connectivity between these areas and the basal or dorsal lateral amygdala nucleus predicting improved life satisfaction (Fig 4C,D; Fig 5C; **Supplementary Fig 7**). […] Medial prefrontal, frontal opercular and lateral prefrontal neurons express receptors for several neurotransmitters, including serotonin (e.g., 5HT-1A, 5-HT-2), noradrenaline (alpha1) and dopamine (D5)[68–72]. Receptors for these neurotransmitter systems are also expressed in the amygdala, with varying densities across nuclei[69,73]. Given the modality of the HCP data used here, it is difficult to establish which neurotransmitter systems are most likely to be mediating the observed effects. However, linking functional connectivity changes with changes in specific neurotransmitter pathways would be an interesting avenue for future research.

Improved display of anatomical fingerprints – e.g. in **Figure 5C** (see also **Figure 6, Supplementary Figure 7+10**):

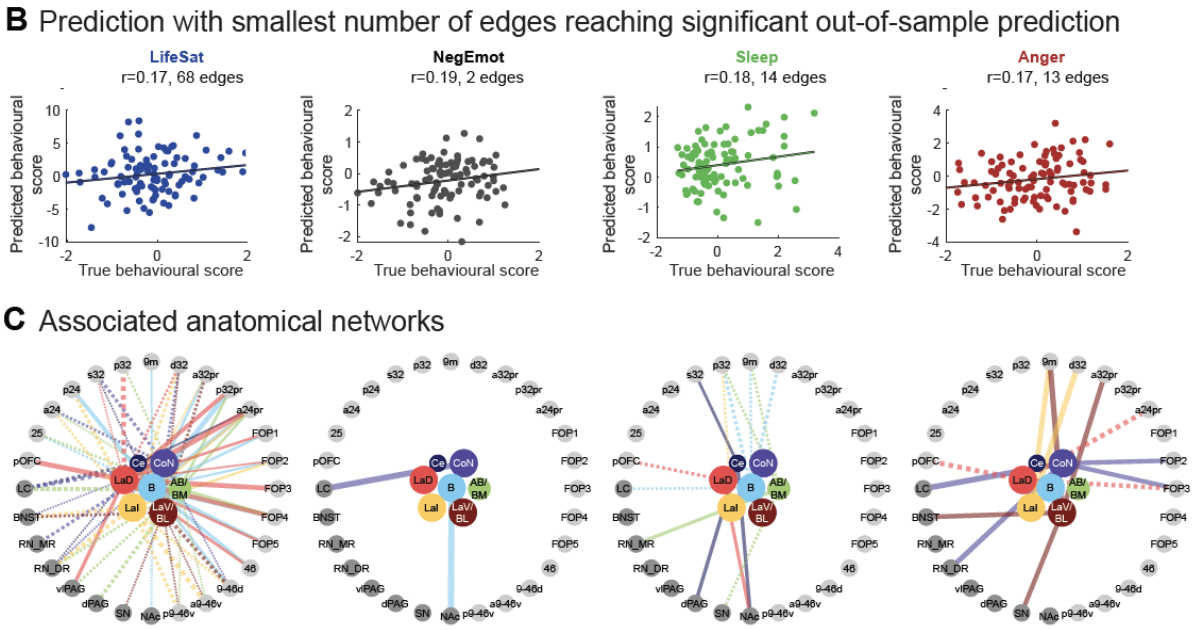


Figure 5, Functional connectivity in smaller sets of specific amygdala nuclei connections is predictive of interindividual variation in mental health dimensions. **A,** […] **B,C** Illustration of the prediction (scatterplot, **B**) and contributing edges (fingerprint, **C**) for the smallest network that achieved a significant prediction (indicated using a black arrow in **A**). Fingerprints shows ROIs on the circumference (dark=subcortical), amygdala nuclei in the centre (colour-coded); line width denotes the size of the absolute 3T regression coefficient; line style denotes its sign (continuous=positive; dashed=negative).

==========

2. The statistical analysis consists in comparisons against other FC connections that essentially do not share one of the two ends of the 196 original FC links. I think it is also important to test for the absolute significance of the predictive power of the 196 FC connections using a classical non-parametric test (e.g., shuffling subject labels). This could be seen as a sanity check since comparison against other FC connections is a priori a stronger test, but it would clarify the significance level of the results. Then, the set of other FC connections considered to build the null-data is quite restrictive. I think it is important to explore the predictive power of any pair of ROIs: it might be that some cortico-cortical FC connections are more predictive than the ones being considered and this is not explored. Finally, since most results are of reasonable but not strong statistical significance (e.g., few connections with uncorrected p-value < 0.001; p-values not corrected
considering all the tests performed across the paper), I believe replicating these results in an independent dataset would significantly increase impact of the paper. This could ideally be done in a separate dataset of patients with a disease related to the behavioral factors identified in the paper (e.g., sleep disorders), or using other HCP subjects. If the results are robust, I believe it should be

possible to find another HCP cohort of similar size within the 1200 release with acceptable data quality.

Thank you, we agree that these are all important points. We will respond to them individually under separate subheadings below.

Replication

In response to the editor's and both reviewers' comments, we have made substantial changes to show that our conclusions are robust and replicable:

- We have doubled the number of 3T-HCP participants that were included in the manuscript from 200 to 400. While there are more participants in some published databases such as the HCP database, we explain in detail below why the data from an even larger number of 3T subjects is unsuitable for our study
- We have replicated all key results in an independent cohort, the 7T HCP data. These data have improved signal in subcortical regions. We now include a replication of (a) the behavioural results, (b) the amygdala parcellation and (c) the relationships between behaviour and resting-state functional connectivity. Related to part (c), some results that were included in the first version of our manuscript, particularly those relying on the functional connectivity in individual connections, were not robust enough to warrant inclusion in the manuscript and have been removed. Instead, we now focus on the robust effects across small sets of edges which replicate across datasets.

Altogether, we therefore now report results from **almost 500 participants** using unprecedented anatomical detail in the amygdala, a subcortical structure of interest to a wide neuroscientific community. Thus, despite not including all available HCP participants, as far as we know, this is still one of the largest (possibly the largest) study of this kind conducted to date.

Patient dataset

We carefully looked at the possibility to include a patient dataset, which we agree would have provided the strongest test of our predictions. However, as far as we are aware, there is no available dataset which would match the HCP data in terms of image quality (volume of data, spatial resolution, optimized pulse sequences etc) and inclusion of physiological noise recordings or high field (7T) imaging to ensure that the signal is adequate in the amygdala and interconnected subcortical nuclei for the type of analyses we conduct. However, neuroimaging methods are improving fast, and we believe that more datasets with sufficient resolution and SNR will become available, which will ensure our method will be employable by the wider neuroscientific community and in future patient studies. We have now included a comment on this in the Discussion.

Participant selection

We reproduce here our rationale explaining why we could not include an even larger sample from the healthy young-adult HCP cohort (also included in response to R1's question 2 above):

The HCP data is a state-of-the-art neuroimaging dataset. Despite its quality, however, there are ongoing discussions around the relatively weak signal in subcortical structures, compared to cortical structures, in the 3T-HCP data on its user list (hcp-users@humanconnectome.org) and the majority of publications that have come out of the HCP data release thus far focus on cortical regions. By contrast, our focus is on the amygdala. Not only is the amygdala itself a subcortical structure but its subnuclei are partly distinguished by the specific patterns of connections they have with other small subcortical nuclei, for example some in the brainstem.

When we started this project, the 7T-HCP data had not been released, and the consensus in the field was that for looking at subcortical and brainstem regions which are greatly impacted by physiological noise, it is necessary to clean-up the data as much as possible for these artefacts to obtain reliable data. However, unfortunately, only **n=764** out of n=1206 3T-HCP participants have physiological noise regressors for all four resting-state runs. Initially, we inspected these traces manually and selected the n=200 participants which had the best-quality physiological noise recordings, while also ensuring a good spread in their DSM scores to allow making meaningful predictions about mental well-being. N=200 felt like a large sample size given most studies in patients and healthy people had focused on n<50 to date. While for studies focusing on cortical regions, including all n=1206 participants should be beneficial, our study was specifically designed to focus on connections with the amygdala, and adding participants with little signal in subcortex would only increase the noise in the data.

Nevertheless, based on the reviewer's comment, we went back to the remaining 3T participants and inspected their data more thoroughly. Without our original n=200 participants, **n=436 candidate 3T participants** had a complete set of behavioural scores (required for factor analysis), complete resting-state data, and physiological traces for all runs. For these participants, we started inspecting each cardiac and respiratory trace (4 runs x 2 types of traces = 8 traces) and categorized them as 'good' (no deficiency), 'mild' (insignificant or transient deficiencies), 'moderate' (noticeable but transient deficiencies) and 'severe' (pronounced or prolonged deficiencies which render the trace meaningless). We first considered participants with DSM scores at the upper and lower ends of the distribution to retain a reasonable spread in mental health dimensions. After inspection of nearly 250 additional participants, we realised that for attempting to double our numbers of participants from n=200 to n=400, while retaining some behavioural variance, we would have to include some participants with severe problems in their physiological traces. The statistic is shown below: of the newly included 3T participants, about half (n=110) had no severe problems with their physiological traces (but could have mild or moderate problems), the other half had severe problems in 1, 2 or 3 traces (out of 8). This means the majority of their data could still be processed in the same way as the n=200 participants we had originally included, but for some runs, the physiological noise clean-up might be somewhat suboptimal.

	total_severe <int>	count <int>	cumaltive_sum <int>
1	0	110	110
2	1	27	137
3	2	54	191
4	3	11	202
5	4	35	237
6	5	2	239
7	6	1	240
8	8	1	241

We have therefore now doubled our numbers to n=400 3T participants in this new version of the manuscript, while maintaining our careful and rigorous physiological noise clean-up and pre-processing routines.

In addition, we now replicate our findings in the 7T-HCP data in the revised version of the manuscript. Out of the n=176 7T-HCP participants with complete resting-state data, we included all n=98 that are unique and non-overlapping with our n=400 3T-HCP participants. The 7T HCP data has the advantage of improved signal in subcortical regions even in the absence of additional preprocessing to correct for physiological noise (which was not possible because the required cardiac and respiratory traces have not been recorded in the 7T-HCP data).

For all analyses performed on the group connectome (Figs 1-2), we therefore included **n=400 3T and n=98 7T participants**, and thus **nearly 500 data sets**. We hope the reviewer will agree that this is a considerable volume of data and a good trade-off between size and quality. For all analyses conducted on individual participant's resting-state coupling values (from Fig4 onwards), we reject outliers as done previously (if more than 10% of their coupling values across all connections deviate more than 3.5 standard deviations from the mean across participants, see Methods), and retain n**=393 3T participants and n=97 7T participants**.

As a result of these additions, our manuscript has changed substantially. As we will show below, we were able to replicate (a) the factor analysis conducted on behavioural/questionnaire scores, and (b) the amygdala parcellation. For the last part of the manuscript, to relate coupling strength to dimensional variation in mental health, we now focus less on functional connectivity in individual edges, which were not sufficiently robust to replicate in each case, and more on the patterns generated based on multiple edges. For this section of the manuscript, we use the 3T dataset to generate hypotheses which we then replicate in the independent 7T data.

Permutation testing

We now include permutation testing (by shuffling the subject labels) to generate all null distributions throughout the manuscript as suggested. We agree this is the gold standard and more transparent. Most analyses in the relevant part of the manuscript have changed due to the inclusion of additional data for replication (the key new sections and figures of the manuscript are pasted below). However, note that

we have also dramatically reduced the overall number of statistical tests we perform in the manuscript. For example, in Figure 5 which displays predictions achieved from 196 models with increasing numbers of predictors from 1 to 196 for each of the four behavioural dimensions, we do not evaluate each model's significance, but instead only evaluate the accuracy at the peak and the size of the earliest significant network across behavioural dimensions. For both of these statistics, we generate appropriate null distributions from bootstrapping as suggested, which carefully accounts for multiple comparisons.

Non-amygdala coupling

If we understand correctly, the reviewer is not suggesting to include functional coupling between any pair of two ROIs to address our main question, but rather to include coupling strength in non-amygdala edges for building the null distribution. One important reason for performing hypothesis-driven analyses focused on the amygdala and specific *a priori* ROIs in this manuscript is because the multiple comparisons problem explodes quickly if we consider the functional coupling between any pair of two ROIs (there are 180 cortical regions in the Glasser atlas, but this does not include subcortex, so just between cortical regions, there are already 32,400-180=32,330 edges). However, importantly, it is not our intention to conclude that amygdala nuclei functional connectivity outperforms any other similar size network one could construct anywhere else in the brain. Instead, our key message is that the functional coupling in circumscribed amygdala nuclei networks can predict dimensions of mental health in an independent cohort. We therefore think that it is correct to use the functional coupling between the same nodes for building the null distribution, as we now do in all our key statistical tests, because this ensures that the data used to generate the null is matched in all its statistical properties to the real data that our key inference is performed on. Importantly, however, we have now included several additional replications in various parts of the manuscript (pasted below) which we think have strengthened our conclusions. We also included a paragraph in the Discussion that emphasizes that the main advance of our paper is not to show that amygdala functional connectivity is superior in its predictive power to any other network in the brain, but rather (a) to show the value of parcellating subcortical structures into their nuclei so that our resolution matches the scale of circuit organization; (b) to show the importance of using functionally interpretable behavioural dimensions, rather than aggregate depression scores; and (c) finally, to show that small subcortical networks carry relevance for predicting these mental health dimensions.

We now summarize our key changes made in response to all these points below under (1) – (5) :

==========

CHANGES IN MANUSCRIPT

(1) Replications for each critical step of the manuscript:

Part 1 – Replication of amygdala parcellation and average functional connectivity pattern

Introduction p. 4:

This parcellation was replicated in two additional datasets acquired at 3T (n=200) and 7T (n=98).

Results p.6:

This average amygdala functional connectivity pattern was replicated in two additional HCP datasets (3T: n=200; 7T: n=98; **Supplementary Fig 3**; see Methods for further details).

Results p.7:
We replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**; see Methods for further details).

Results p.9:
These average functional connectivity patterns between amygdala nuclei and ROIs were replicated in two separate datasets (entire matrix: n=200 3T: Pearson's *r*=0.97, p=8.82e-119; n=98 7T: Pearson's *r*=0.88, p=3.92e-66; **Supplementary Fig 4B**).

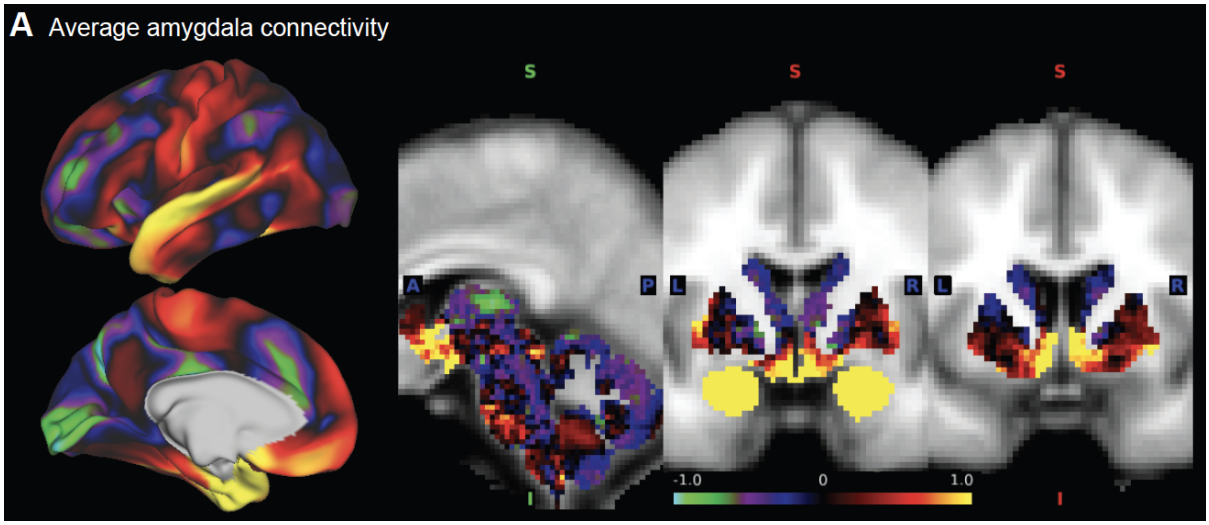
Discussion, p.20:
This amygdala parcellation replicated across several data sets.

Methods p.31 + p.36:
Following the exact same procedure, we closely replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4A**). However, a parcellation using the data from all n=1206 3T-HCP participants, which had not been corrected for physiological noise, showed less symmetry and anatomical plausibility despite relying on more data.
[…]
This notion was corroborated by the mean pattern of functional resting-state connectivity between amygdala nuclei and ROIs shown in **Fig2B** which we robustly replicated in two independent data sets (**Supplementary Fig 4B**).

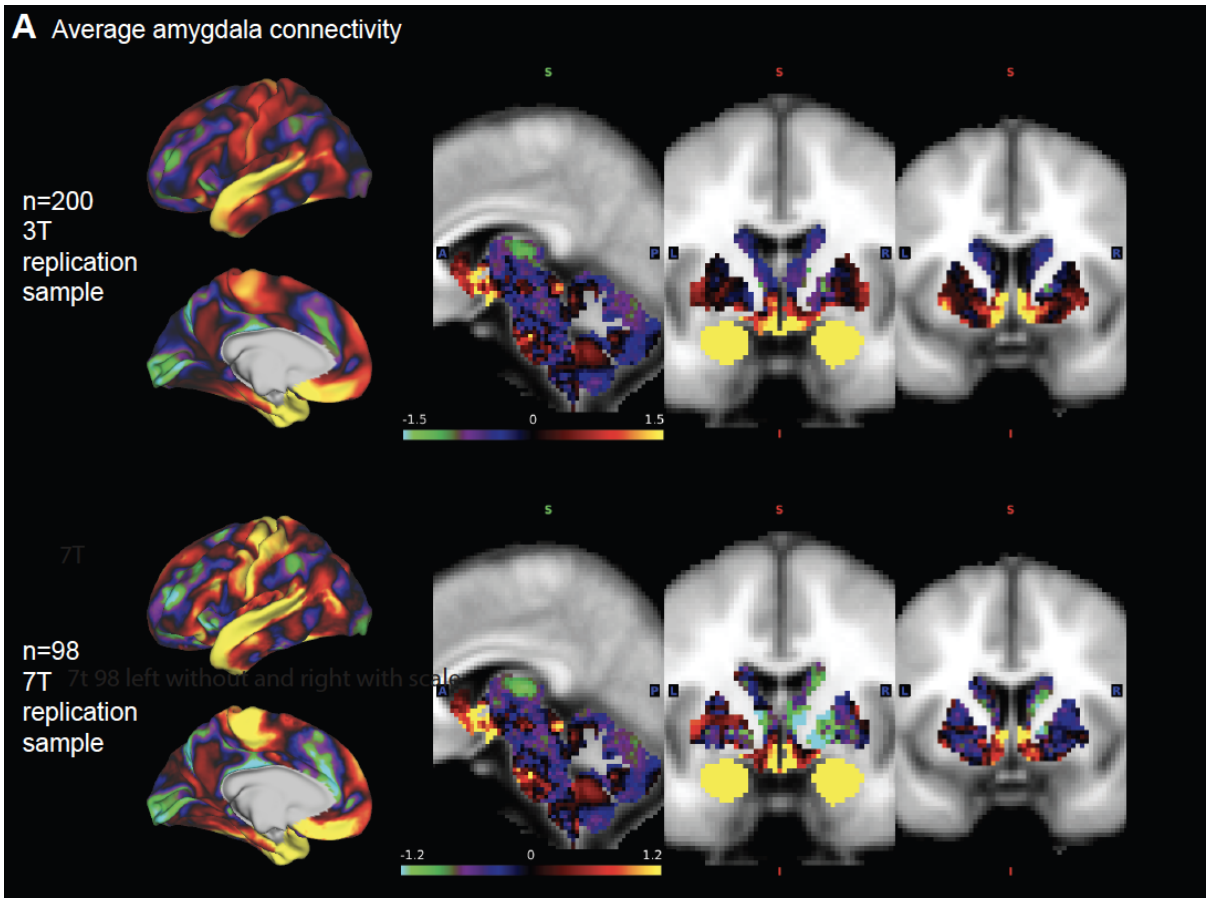
Figures/Figure legends (p.57 and Supplement)
Figure 1, Average amygdala functional connectivity and definition of amygdala clusters, **A,** A group connectome was generated from resting-state fMRI (rs-fMRI) data of n=200 3T young-adult HCP participants using an improved pre-processing pipeline to correct for physiological noise (**Supplementary Fig 1**). The average functional connectivity of all amygdala voxels to the rest of the brain, corrected for global absolute connectivity strength, shows patterns that would be expected from tracer studies, for example strong connectivity of the amygdalae with subgenual ACC, hypothalamus, and ventral striatum. This pattern was replicated in two independent datasets containing n=200 additional 3T-HCP participants and n=98 7T-HCP participants (**Supplementary Fig 3**). **B**, Hierarchical clustering was performed on the similarities between the whole-brain functional connectivity patterns of different amygdala voxels to identify amygdala subdivisions sharing connectivity profiles. Seven subdivisions were identified (left: horizontal; middle: coronal; right: saggital view), showing strong symmetry across hemispheres and strong resemblance with subdivisions identified from histology and high-resolution *post-mortem* structural neuroimaging. The parcellations obtained in two independent datasets closely reproduced these nuclei subdivisions (**Supplementary Fig 4A**).

Figure 2, Amygdala nuclei and their profile of functional connectivity to regions of interest, **A,** Labels assigned to the seven amygdala subdivisions obtained from hierarchical clustering: Ce = central nucleus, CoN = cortical nuclei, B = basal, AB/BM = auxiliary basal/basomedial, LaV/BL = lateral (ventral part) containing aspects of basolateral, LaI = lateral (intermediate part), LaD = lateral (dorsal part). **B**, Average resting-state functional connectivity from the seven nuclei to 28 regions of interest (ROIs) defined *a priori* based on their known connectivity with the amygdala and potential role in regulating emotions and mental well-being. This highlights strong functional connectivity of subgenual cortex (area 25) to the entire amygdala, but particularly to basal subdivisions, in line with tracer work. Similar profiles are observed for posterior OFC (pOFC) and the subgenual portion of area 32 (s32). By contrast, subcortical and brainstem regions most strongly connect with the central nucleus as expected. The mean functional connectivity between ROIs and nuclei was replicated in two datasets (**Supplementary Fig 4B**). **C**, Masks of all ROIs used in this study. For details on their definition, please refer to the Methods. NAc=Nucleus Accumbens; BNST=bed nucleus of the stria terminalis; vl/dPAG=ventrolateral/dorsal periaqueductal grey; SN=substantia nigra; RN_DR/RN_MR=dorsal and median raphe nuclei; LC=locus coeruleus. Definitions of cortical regions were taken from Glasser et al., 2016.

Figure 1A (original n=200 3T HCP participants)



Supplementary Figure 3 (replication in n=200 additional 3T and n=98 7T participants)



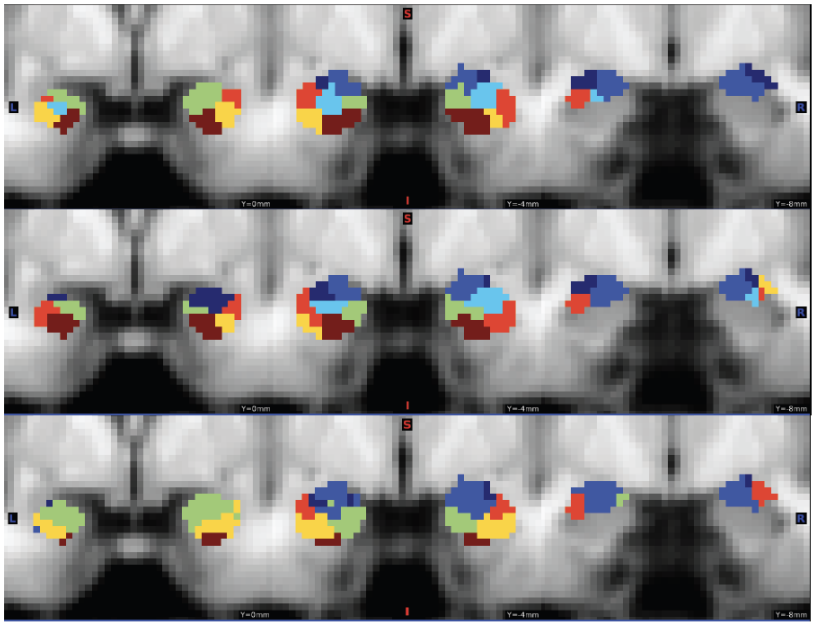
Supplementary Figure 3, Replication of average amygdala functional connectivity The group connectome shown for the original n=200 3T young-adult HCP participants presented in Fig1A was replicated in two other cohorts containing n=200 non-overlapping 3T-HCP participants and n=98 non-overlapping 7T-HCP participants. In the 3T data, we used an improved pre-processing pipeline to correct for physiological noise, as before. This was not possible in the 7T data where physiological noise regressors were not available. However, the 7T resting-state data has improved signal-to-noise in subcortical regions due to the higher field strength.

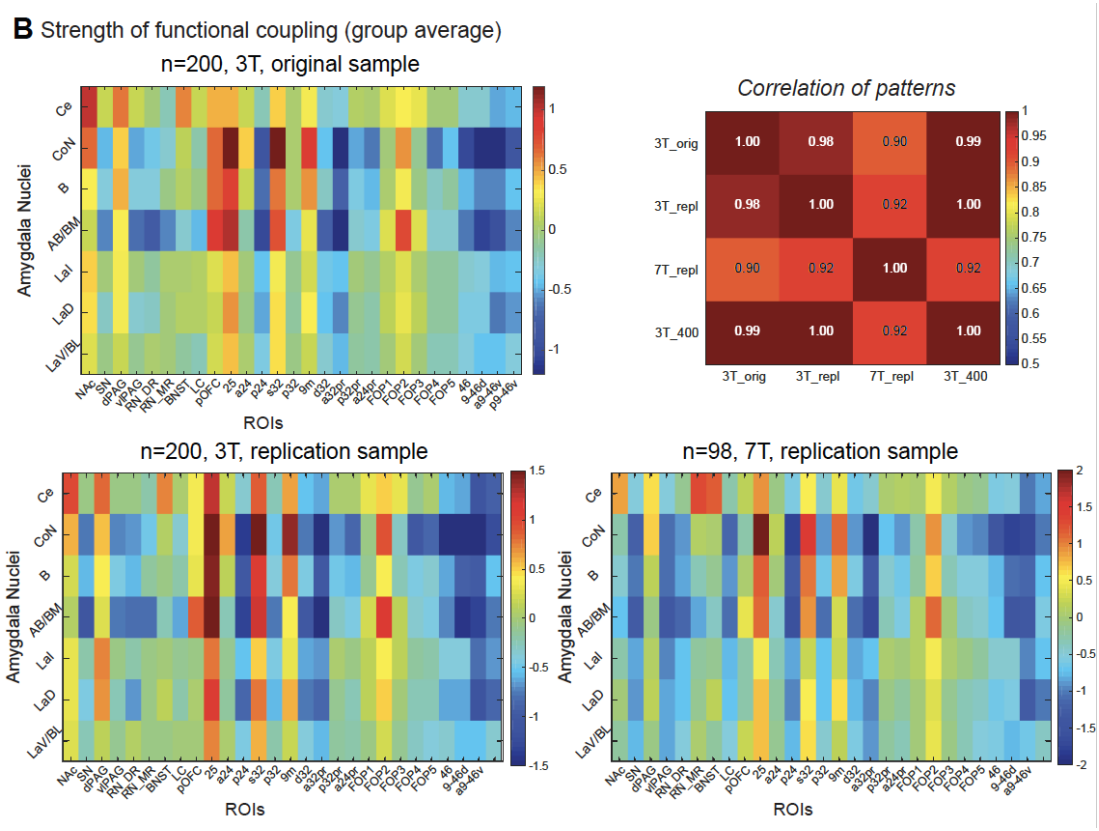
A Replication of amygdala parcellation

n=200
3T
original
sample

n=200
3T
replication
sample

n=98
7T
replication
sample





Supplementary Figure 4, Replication of amygdala parcellation and nuclei mean functional connectivity, A For comparison, the parcellation of the amygdala obtained in the original n=200 3T participants is shown for the n=200 3T replication sample and the n=98 (all non-overlapping) 7T participants (compare Fig 1B). This shows that the key subdivisions of the amygdala were replicated in these two additional parcellations. **B**, The average amygdala nuclei to ROI functional connectivity replicates across cohorts (top: original, bottom left: replication 3T, bottom right: replication 7T; compare **Fig 2B**), as confirmed in the strong correlation between these patterns (top right).

Part 2 – Replication of behavioural factor analysis

Introduction p.4

[…] we were able to define latent behaviours by applying a factor analysis to a set of questionnaire scores which generated four highly replicable dimensions of mental health.

Results p.10:

The factor analysis was replicated in several other datasets (**Supplementary Fig 6**). Notably, because the factor analysis focuses on behavioural data rather than neural data, it can be employed with all HCP

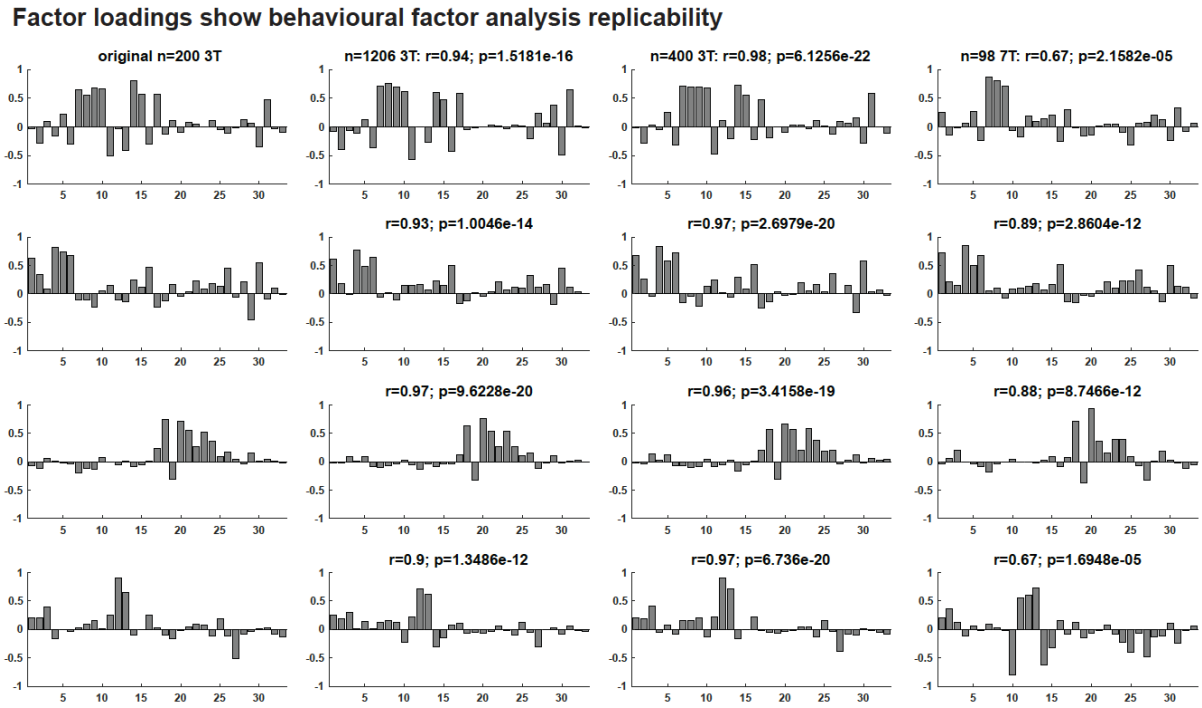
datasets and not just the subset of data with the highest quality physiological recordings of respiration and cardiac activity. When the analysis was repeated on the complete set of n=1206 3T HCP participants for maximal robustness, the resulting factors were highly similar (Pearson's correlation between factor loadings n=200 vs n=1206 3T participants: *r*=.94, *p*=1.5e-16, *r*=.93, *p*=1e-14, *r*=.97, *p*=9.6e-10, *r*=.9, *p*=1.3e-12; **Supplementary Fig 6**).

Methods p.38:

Moreover, the factor analysis replicated to several other datasets (**Supplementary Fig 6**), most importantly to the full set of all n=1206 3T HCP participants (Pearson's correlation between factor loadings n=200 vs n=1206 3T participants: *r*=.94, *p*=1.5e-16, *r*=.93, *p*=1e-14, *r*=.97, *p*=9.6e-10, *r*=.9, *p*=1.3e-12), but also the n=400 3T and the n=98 7T participants included for neural analyses (**Supplementary Fig 6**).

Figures + Figure legends (p. 57/58):

Figure 3, Latent behavioural dimensions capture distinct aspects of mental well-being, **A,** A factor analysis conducted based on 33 behavioural scores (**Table 1**) available as part of HCP revealed four factors. The loadings for each factor are shown in different colors, corresponding to the four rows. The highest five contributing behavioural scores are shown in order of their contribution (absolute loading) on the right. This shows that the four factors capture quite distinct dimensions of participants' mental well-being which we summarized as 'Social and life satisfaction', 'Negative emotions', 'Sleep' (problems), 'Anger and rejection'. The four factors replicated when the factor analysis was performed on all 1206 HCP participants (see Methods), or only the subset of 3T and 7T participants included here (**Supplementary Fig 6**). **B**, Correlations between factor loadings.



Supplementary Figure 6, Replication of factor analysis The factor analysis computed to generate mental health dimensions in our original n=200 3T participants (left) replicated in all n=1206 HCP participants (2nd column) and the full set of n=400 3T and n=98 7T participants used in this manuscript (3rd and 4th column). Correlation coefficients and p-values refer to the similarity with the original pattern shown on the left.

Part 3 – Replication of amygdala functional connectivity behaviour relationships
This section was changed most extensively due to the inclusion of several new datasets and is not fully reproduced here. The complete set of changes can be found in the Results p.11-19, Methods p.38-44, Figures 4-7, and Suppl Figs 7-10.

Abstract p. 2
Connectivity in circumscribed amygdala networks predicted behaviours in an independent dataset.

Introduction p.4
In our final step, we selected the best predictors in terms of the functional connectivity between amygdala nuclei and other brain regions for each of the four behavioural dimensions. We showed that functional connectivity in a select number of specific amygdala connections predicted each mental health dimension in an independent dataset.

Selected new Results sections:

Results p.11:

We first established whether relationships between nuclei-specific amygdala functional connectivity and mental health dimensions replicated between the two independent (3T and 7T) datasets. We fitted robust linear regression coefficients to capture the relationship between functional connectivity values in each 'edge' (e.g., Ce to NAc) and behavioural dimension (e.g., sleep problems), separately for the 3T and 7T dataset. This resulted in 196 regression coefficients (7 amygdala nuclei x 28 ROIs) for each of four behaviours and two datasets. If amygdala nuclei functional connectivity carries no information about mental health dimensions, then, by chance, the correlation between regression coefficients obtained across behaviours in the 3T versus 7T datasets should be zero. To formally test this, we generated a null distribution by shuffling the subject order of the behavioural scores n=10,000 times while keeping the functional connectivity values unchanged. Indeed, by chance, the across-dataset replication of the pattern of regression coefficients was centred on zero (**Fig4A**). The similarity between 3T and 7T regression coefficients in the actual data, however, was significantly greater than chance (Pearson's r=0.26; p=0.0313; **Fig 4A,B**), showing that relationships between nuclei-amygdala functional connectivity and mental health dimensions were similar across datasets.

Results p.12/13:

Having established that the overall relationship between nuclei-specific amygdala functional connectivity and mental health dimensions is similar between datasets, we next asked whether we could predict individual 7T participants' behavioural scores using regression coefficients estimated from the 3T data. In other words, we examined whether we could predict mental health dimensions in completely held-out data (7T) using a weighting of nuclei-specific amygdala functional connectivity values derived from an independent dataset (3T). For each behavioural dimension, the 196 robust regression coefficients estimated from the 3T data for all nuclei functional connectivity values were applied to the functional connectivity values of individual 7T participants to obtain their predicted behavioural scores. This out-of-sample prediction was significant for life satisfaction, negative emotions, and anger (correlation between predicted and true behaviour for the 7T data: lifeSat: r=0.187, p=0.0335; negEmot: r=0.219, p=0.0155; anger: r=0.226, p=0.0143), but did not reach significance for sleep (r=0.05, p=0.31; **Fig4E**).

Given that medial temporal lobe areas are considered areas of high drop-out and low signal-to-noise, we performed a second replication to further demonstrate the consistency with which nuclei-specific amygdala functional connectivity predicted dimensional variation in mental health. This time, we examined the consistency within-subject rather than across-dataset (pooling across 3T/7T datasets). We divided the full resting-state data of each participant in two halves (run 1+2 versus 3+4, acquired in separate sessions on different days) and again computed robust regression coefficients to predict the four behavioural dimensions, but this time separately using resting-state functional connectivity values extracted from only the first or second half of the full resting-state data. A null distribution obtained using shuffled behavioural values estimated the similarity of regression coefficients between the two halves (i.e., sessions) expected by chance. The true functional-connectivity-behaviour relationship between the first and second half was significantly greater than expected by chance

(Pearson's r=0.47; p=0.014; **Fig4A**). The anatomical pathways where functional connectivity most contributed to this within-subject replication were highly similar to those most contributing to our previous across-dataset replication (**Fig4D**). Together, these analyses show that despite substantial noise and difficulties in neuroimaging subcortical regions such as the amygdala[29], specific patterns of variation in functional connectivity both within- and between-datasets consistently related to participants' mental health dimensions.

Discussion, p. 23:

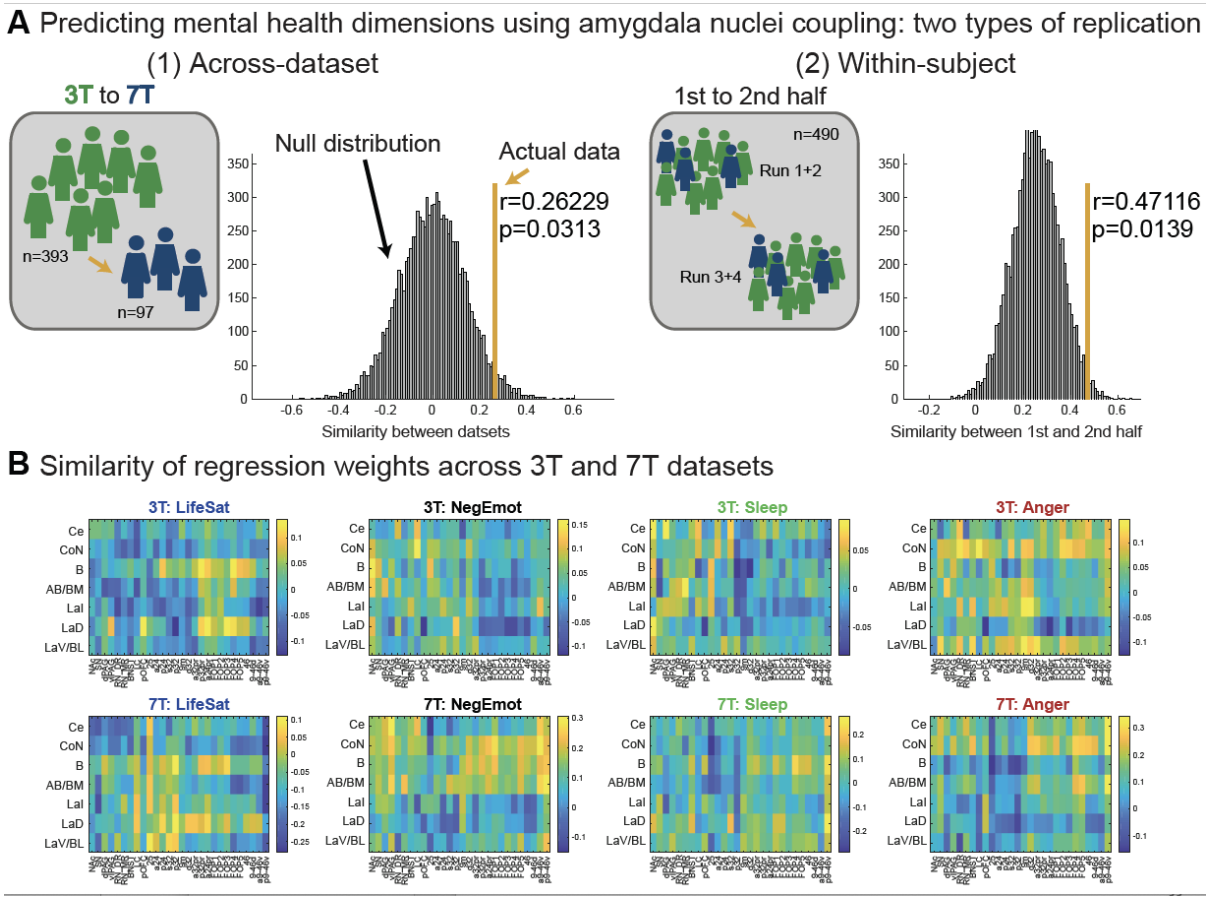
The anatomical features of the amygdala networks identified for the different latent behaviours seem plausible in the context of previous work, and consistent across two types of replications (across-dataset and within-subject; **Fig 4**). We note that, importantly, both feature selection – which determined the anatomical networks to focus on – and estimation of regression weights was performed in an initial dataset (n=393 3T participants) and all predictions were generated out-of-sample (n=97 7T participants).

Methods p.39/40:

To test whether the obtained robust regression weights captured meaningful relationships between the functional connectivity of individual amygdala nuclei and the four mental health dimensions, we tested whether regression weights were (a) similar between the 3T and 7T datasets ('across-dataset replication') and (b) similar between two halves (corresponding to MR sessions) of the experiment ('within-subject replication'; **Figure 4A**). First, for the across-dataset replication, we computed Pearson's correlation coefficient between the overall pattern of regression weights obtained for the 3T and 7T data (row 1 versus 2 in **Figure 4B**). To establish whether the obtained correlation was better than predicted by chance, given the level of noise present in brain connections with the amygdala and given our number of connections, we generated a null distribution (**Fig 4A**) by shuffling the vectors **y** containing the behavioural dimension n=10,000 times and recomputing the correlation coefficient between the overall pattern of regression coefficients.

Second, to attempt a within-subject replication, functional connectivity values were extracted from half of the resting-state data, separately from either run 1+2 or run 3+4 (which were acquired in two separate sessions on separate days). Robust regression weights capturing the relationship between FC (X) and behavioural scores (y) were then computed separately for the FC values obtained from runs 1+2 versus 3+4. Here we used the merged data from all n=490 3T+7T participants. The similarity between the overall pattern of robust regression weights obtained for the two experimental halves was computed using Pearson's correlation. Again, to test whether their similarity was greater than expected by chance, a null distribution was generated by repeating this procedure n=10,000 times using shuffled behavioural scores (**Figure 4A**).

Figures and Figure legends – in addition to the new figures pasted below, **Suppl Figs 7-10** have also newly been added:


A Predicting mental health dimensions using amygdala nuclei coupling: two types of replication
(1) Across-dataset
3T to 7T
n=393
n=97
Null distribution
Actual data
r=0.26229
p=0.0313
Similarity between datsets
(2) Within-subject
1st to 2nd half
n=490
Run 1+2
Run 3+4
r=0.47116
p=0.0139
Similarity between 1st and 2nd half
B Similarity of regression weights across 3T and 7T datasets
3T: LifeSat
3T: NegEmot
3T: Sleep
3T: Anger
7T: LifeSat
7T: NegEmot
7T: Sleep
7T: Anger
Ce
CoN
B
AB/BM
Lal
LaD
LaV/BL

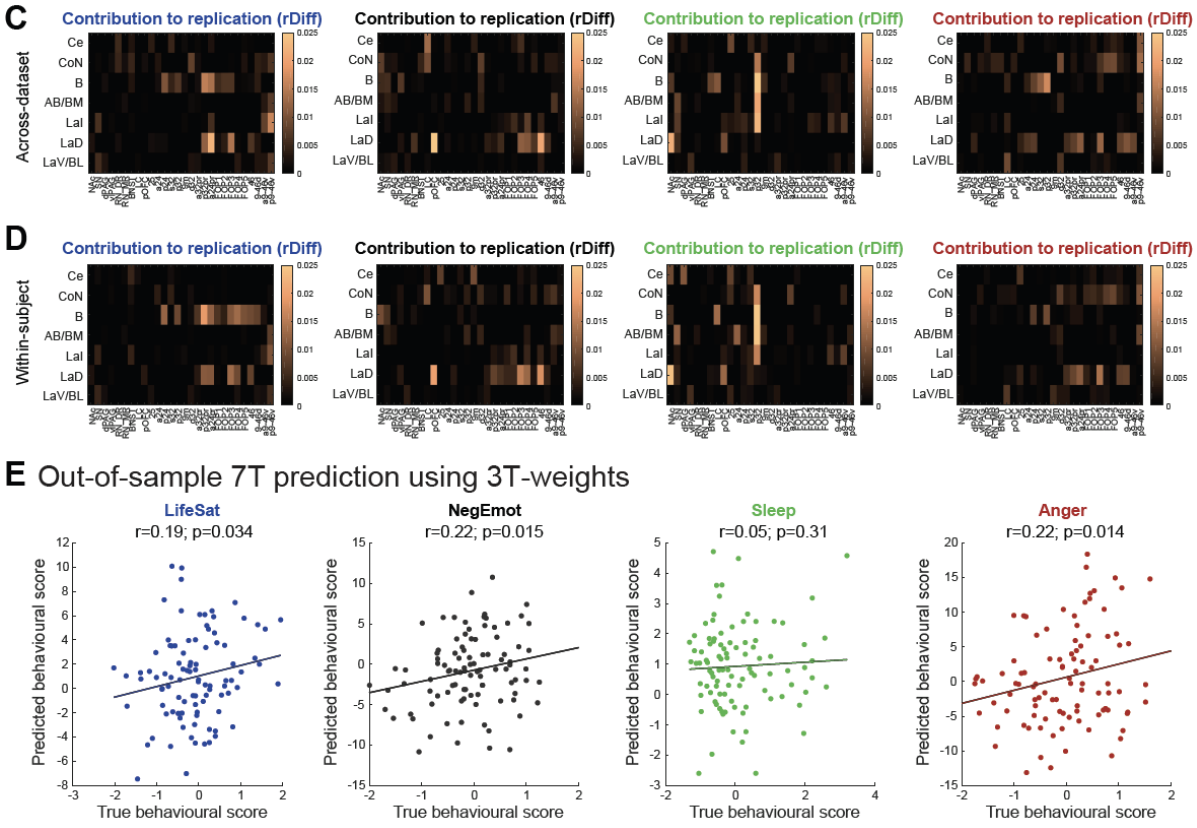


Figure 4, Nuclei-specific amygdala functional connectivity shows consistent relationships with interindividual variation in mental health dimensions, **A,** Relationships between interindividual variation in nuclei-specific amygdala functional connectivity and mental health dimensions were examined in two HCP datasets containing n=393 3T and n=97 non-overlapping 7T participants (following outlier rejection). Despite challenges with neuroimaging signals in subcortical regions, relationships were robust and replicable, as established in two ways: (1) Across-dataset replication: the similarity of robust regression coefficients capturing the relationship between resting-state functional connectivity for each 'edge' (e.g., Ce to NAc) and each of four mental health dimensions was greater than expected by chance across datasets (null distribution generated using shuffled behavioural scores; n=10,000 iterations). (2) Within-subject replication: robust regression coefficients estimated on half of the resting-state data (runs 1+2 versus 3+4, from separate sessions) also showed greater-than-chance similarity. **B**, Visualization of obtained robust regression coefficients for each edge, mental health dimension (columns) and dataset (rows) illustrates their similarity across cohorts. **C**, **D,** For each edge, its contribution *rDiff* to the across-dataset (**C**) and within-subject (**D**) similarity was computed as the difference between the correlation achieved when excluding this edge (195 values) and when including all 196 edges (28 ROIs x 7 nuclei). Visual inspection of *rDiff* values highlights strong similarities between *rDiff* values in the two replications (C vs D), clear differences between the four behavioural dimensions, and anatomical specificity – e.g., the importance of cortical connections with B and LaD nuclei for predicting life

satisfaction, for connections with NAc, other subcortical regions and medial frontal area p32 for predicting sleep, and functional connectivity with the cortical nuclei (CoN) for predicting anger. **E**, Regression coefficients estimated from the 3T-participants applied to 7T-functional connectivity values to predict 7T-mental health dimensions showed significant out-of-sample predictions for all mental health dimensions except sleep problems.

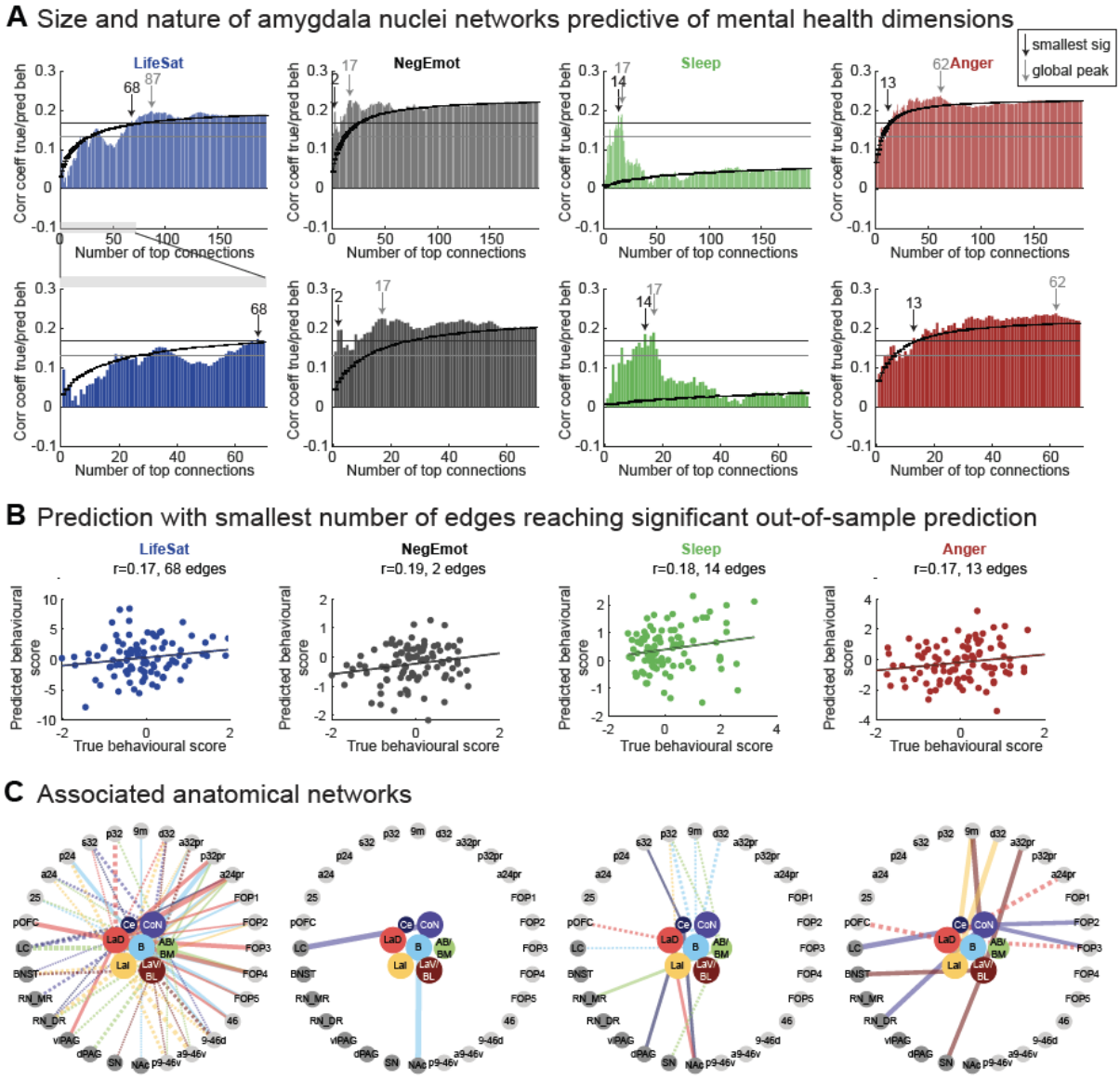
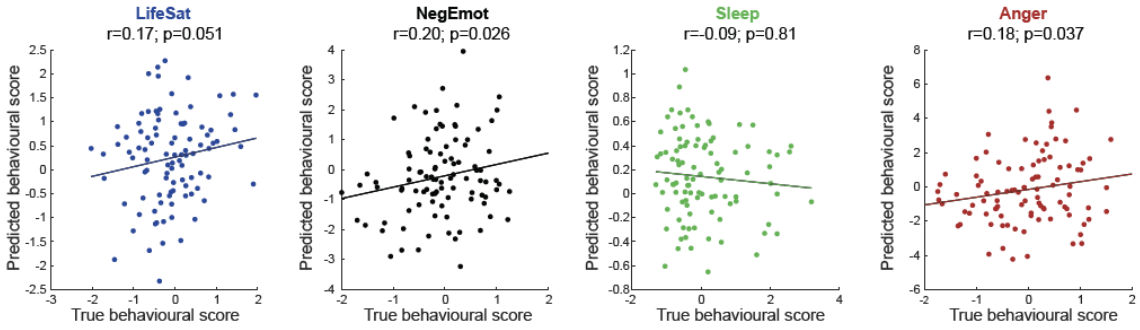


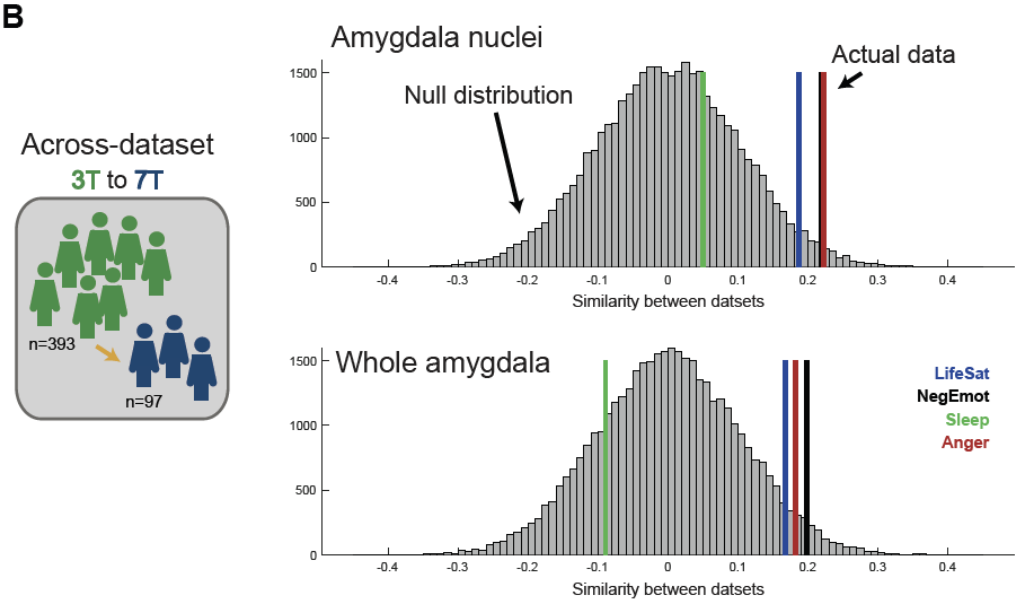
Figure 5, Functional connectivity in smaller sets of specific amygdala nuclei connections is predictive of interindividual variation in mental health dimensions. A, Predictions achieved with subsets of edges between 1 and 196: robust regression coefficients estimated from 3T-participants were applied to 7T-functional connectivity values to predict interindividual differences in mental health dimensions in the

held-out 7T data (as done in **Fig4E** using all 196 edges). Prediction accuracies are shown as the correlation between true and predicted mental health dimensional scores in the 7T participants, but were only statistically evaluated at the peak (grey arrow: 'global peak') and to derive the smallest number of edges that reached a significant out-of-sample prediction (black arrow: 'smallest sig'; see Methods); coloured bars show accuracy when including edges in order of their absolute regression coefficient in the 3T data; black curve indicates performance using the same number of edges but included in random order (n=10,000 shuffles; error bars denote SEM); black line at r=0.168 indicates threshold for significance at $p<0.05$ purely for visualization (grey line: $p<0.1$); second row shows the same but zoomed in on the first 70 edges. For all behavioural dimensions, smaller sets of amygdala nuclei functional connectivity values achieve a significant out-of-sample prediction. In general, except for life satisfaction, using the top 3T edges is better than a random selection of the same number of edges. **B**,**C** Illustration of the prediction (scatterplot, **B**) and contributing edges (fingerprint, **C**) for the smallest network that achieved a significant prediction (indicated using a black arrow in **A**). Fingerprints shows ROIs on the circumference (dark=subcortical), amygdala nuclei in the centre (colour-coded); line width denotes the size of the absolute 3T regression coefficient; line style denotes its sign (continuous=positive; dashed=negative).

A Sensitivity of whole-amygdala as opposed to nuclei-specific amygdala coupling



B



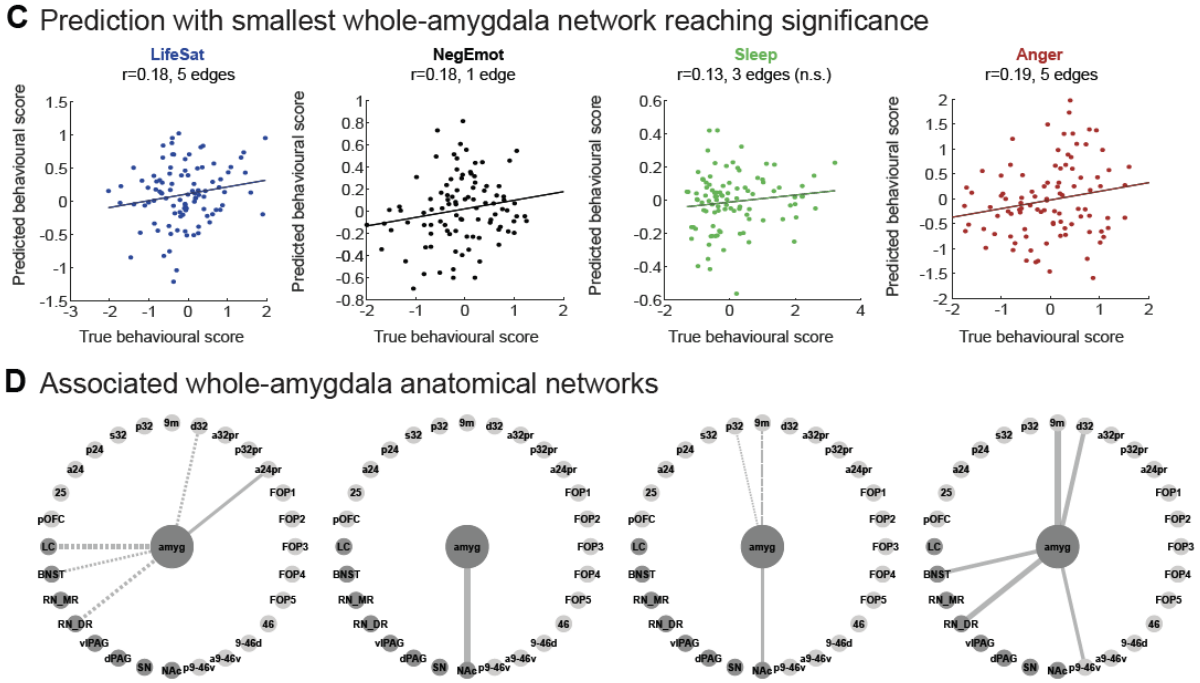


Figure 6, Parcellating the amygdala improves the accuracy for predicting interindividual differences in mental health dimensions, **A**, Out-of-sample predictions achieved when considering the functional connectivity of the whole amygdala with our 28 *a priori* ROIs, instead of nuclei-specific functional connectivity, are less precise, but still significant for two of the four mental health dimensions (negative emotions + anger). **B**, To control for differences in the number of predictors for nuclei-specific and whole-amygdala functional connectivity (28 vs 196), out-of-sample 7T predictions (shown in Fig 4A and 6A) were evaluated based on their own null distribution. Nuclei-specific predictions were still superior to whole-amygdala predictions for all four mental health dimensions (coloured bars indicate Pearson's r overlaid on the null distribution; for statistics see main text); this was also true when looking at peak predictions achieved using smaller sets of edges (**Supplementary Fig 9**). **C,D,** As in **Fig5**, the prediction (scatterplot, **C**) and contributing edges (fingerprint, **D**) are shown for the smallest set of edges that reached significance (sleep never reached significance; life satisfaction was significant when using fewer than 28 edges) which highlights clear differences between mental health dimension.

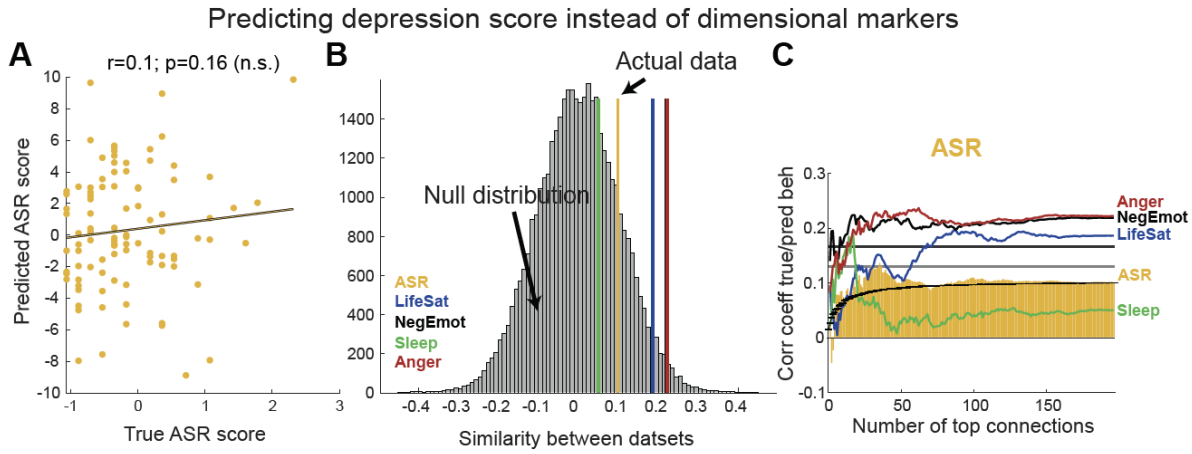


Figure 7, Amygdala functional connectivity relates better to dimensional behaviours than overall depression scores, **A, B** Out-of-sample prediction of 7T participant's ASR_AnxD scores, using nuclei-specific functional connectivity and regression weights estimated from 3T participants as in **Fig 4E**, is (**A**) not significant (for DSM, see **Supplementary Fig 10**) and (**B**) less accurate than three out of four of our dimensional behaviours. **C,** Overall predictions are worse for ASR compared to dimensional scores when using smaller sets of edges (bars shown in Fig5A for dimensional behaviours are overlaid as coloured lines for comparison); plotting conventions as in **Fig5**.

(2) Participant selection

Results p.6:

We did not include all 1206 HCP participants because these additional pre-processing steps required good quality physiological recordings of respiration and cardiac activity which were not available or not of sufficient quality in a considerable number of HCP participants (see Methods for further details on subject inclusion criteria).

Methods p. 27/28

Participants

Data and ethics were provided by the Human Connectome Project (HCP), WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. Several datasets were included for analysis. First, an initial dataset comprised a subset of n=200 out of the full set of n=1206 3T subjects from the HCP young adults data set (n=200; mean age 29 ± .26; age range 22-36; 108 females, 92 males). These were chosen from the full HCP data set (https://www.humanconnectome.org/) based on two criteria: the quality of the physiological variables acquired (both cardiac and respiratory; inspected visually and using summary measures such as their variance over time) and their total DSM/ASR (DSM_Depr, ASR_Totp) score to allow us to maximise subclinical variance across participants (resulting mean DSM

score: 4.46, variance: 16.64; mean of all 1206 HCP participants: 4.25; variance: 12.24; mean/variance total ASR score n=200: 37.91, 669.08; n=1206: 37.43, 523.82). The DSM/ASR scores were not used in any of the key analyses. Working on a subset of all 1206 HCP participants was necessary because one key aspect of the pre-processing was to correct rs-fMRI data for physiological noise, which particularly affects the key regions of this study such as the amygdala and brainstem. However, the quality of physiological data that has been acquired varies substantially across HCP participants. For example, only n=764 out of 1206 3T-HCP participants have recordings of physiological noise regressors for all four resting-state runs, and only n=636 have complete resting-state data, physiological recordings and behavioural scores (required for factor analysis, see below). Thus, a second dataset of the same size as the first (n=200) was selected for replication from the remaining n=436 3T-participants with complete data. It was not possible to match the variance in DSM/ASR scores to the first dataset and participants did not have quite as high-quality physiological noise recordings, but we still prioritized inclusion of participants at the upper and lower ends of the distribution of DSM scores (resulting DSM mean/variance in n=200: 3.96, 10.71; ASR mean/variance: 35.72, 481.67) and those with the largest number of runs with high-quality physiological noise recordings (severe problems in a maximum of 3 out of 8 = 4 cardiac + 4 respiratory traces). The resulting demographics in this second dataset of n=200 3T participants were mean age 28 ± .28; age range 22-36; 99 females, 101 males. A third dataset contained all 7T-HCP young adult participants not already included in either of the 3T datasets and with full resting-state and behavioural data, which left us with n=98 7T-HCP participants (mean age 29 ± .33; age range 23-36; 59 females, 39 males; DSM mean/variance: 3.43, 5.73; ASR mean/variance: 31.79, 253.43; **Supplementary Table 3**). Physiological noise recordings are not available in the 7T data, but the higher field strength significantly improves the tSNR in subcortical regions.

(3) Paragraphs dealing specifically with permutation testing (in addition to the below additions, see also Figure legends of Fig4 + Fig6):

Results p.11:

We first established whether relationships between nuclei-specific amygdala functional connectivity and mental health dimensions replicated between the two independent (3T and 7T) datasets. We fitted robust linear regression coefficients to capture the relationship between functional connectivity values in each 'edge' (e.g., Ce to NAc) and behavioural dimension (e.g., sleep problems), separately for the 3T and 7T dataset. This resulted in 196 regression coefficients (7 amygdala nuclei x 28 ROIs) for each of four behaviours and two datasets. If amygdala nuclei functional connectivity carries no information about mental health dimensions, then, by chance, the correlation between regression coefficients obtained across behaviours in the 3T versus 7T datasets should be zero. To formally test this, we generated a null distribution by shuffling the subject order of the behavioural scores n=10,000 times while keeping the functional connectivity values unchanged. Indeed, by chance, the across-dataset replication of the pattern of regression coefficients was centred on zero (**Fig4A**). The similarity between 3T and 7T regression coefficients in the actual data, however, was significantly greater than chance (Pearson's r=0.26; p=0.0313; **Fig 4A,B**), showing that relationships between nuclei-amygdala functional connectivity and mental health dimensions were similar across datasets.

Results p.13:

Given that medial temporal lobe areas are considered areas of high drop-out and low signal-to-noise, we performed a second replication to further demonstrate the consistency with which nuclei-specific amygdala functional connectivity predicted dimensional variation in mental health. This time, we examined the consistency within-subject rather than across-dataset (pooling across 3T/7T datasets). We divided the full resting-state data of each participant in two halves (run 1+2 versus 3+4, acquired in separate sessions on different days) and again computed robust regression coefficients to predict the four behavioural dimensions, but this time separately using resting-state functional connectivity values extracted from only the first or second half of the full resting-state data. A null distribution obtained using shuffled behavioural values estimated the similarity of regression coefficients between the two halves (i.e., sessions) expected by chance. The true functional-connectivity-behaviour relationship between the first and second half was significantly greater than expected by chance (Pearson's r=0.47; p=0.014; **Fig4A**). The anatomical pathways where functional connectivity most contributed to this within-subject replication were highly similar to those most contributing to our previous across-dataset replication (**Fig4D**). Together, these analyses show that despite substantial noise and difficulties in neuroimaging subcortical regions such as the amygdala[29], specific patterns of variation in functional connectivity both within- and between-datasets consistently related to participants' mental health dimensions.

Results p.13/14:

Characterizing the size and nature of amygdala nuclei networks predictive of mental health dimensions

Having established that the overall pattern of amygdala nuclei functional connectivity carries relevance for mental health, we next explored whether functional connectivity in smaller sets of nuclei-specific edges may be able to predict variation in our four mental health dimensions. So far, all predictions relied on the full set of 196 functional connectivity values from 7 amygdala nuclei to 28 ROIs, but our initial visualization of the pathways that most contributed to similarities both across and within datasets suggested that specific subsets of edges may be particularly important in each case (**Fig4C,D**); indeed, it is possible that inclusion of non-contributing predictors may only add noise to the prediction.

To test this in an unbiased way, we iteratively included an increasing number of functional connectivity values from 1 to 196 as predictors, based on their absolute robust linear regression coefficient in the 3T participants, and tested all predictions in the held-out 7T participants. In other words, as before, the 3T-coefficients were applied to 7T-functional connectivity values to generate out-of-sample 7T-predictions, but this time using subsets of increasing numbers of edges, determined based on the order of their contribution in the 3T data. Predictions were evaluated as the correlation between predicted and true behavioural scores in the 7T participants. By definition, adjacent predictions only differ from one another by the inclusion of one additional edge, which necessarily implies interrelationship between predictions. Thus, rather than evaluating each step between 1 and 196 edges, we used this analysis to identify (a) the smallest significant and (b) overall best prediction and examined these statistically in a way that corrected for the number of models (n=196) and behavioural dimensions (n=4). To examine significance in this way, we generated two null distributions with the same procedure but using shuffled behavioural scores (see Methods): (a) for the smallest number of edges to reach significance; and (b) for Pearson's r at the overall best prediction expected. This showed that our smallest

significant networks were smaller on average than expected by chance (lifeSat: n=68, negEmot: n=2, sleep: n=14; anger: n=13; resulting average: n=24.25, p=0.0032). In the same way, the average Pearson's r at the global peak was higher than expected by chance (lifeSat: r=0.197; negEmot: r=0.22; sleep: r=0.19; anger: r=0.24; average: r=0.2118, p=0.0068). Importantly, for all four mental health dimensions, the earliest significant and top prediction accuracy was reached before n=196 edges. **Fig5B/C** depict the smallest amygdala nuclei networks that reached significance for each behavioural dimension; **Supplementary Fig 7** additionally describes the edges associated with the top prediction accuracy.

Results p.17/18:

To test whether, for all four behavioural dimensions, the top predictors in the 3T data, i.e., the edges with highest absolute regression coefficients, indeed performed better at predicting 7T behavioural scores out-of-sample than the same number of randomly picked edges, we repeated the above procedure of including an increasing number of edges from 1 to 196 10,000 times but this time using randomly picked sets of 1 to 196 edges from the full set of 196 amygdala-to-ROI connections. The resulting null distribution is shown in black in **Fig4A**. For three out of four dimensions, all except life satisfaction, the top edges selected from the 3T data performed better in predicting the 7T behavioural scores than random subsets of edges of the same size (% of predictions above the null up to 70 edges (2nd row in **Fig5A**): lifeSat: 27.14%; negEmot: 98.57%; sleep: 74.29%; anger: 90.00%; % of predictions above the null up to 196 edges (1st row in **Fig5A**): lifeSat: 71.94%; negEmot: 69.39%; sleep: 60.71%; anger: 58.67%). This shows that there is consistency between datasets not only in terms of the overall pattern of regression weights, but also in terms of which smaller sets of connections carry particular relevance for predicting a given mental health dimension.

Comparing the accuracy of nuclei-specific and whole-amygdala predictions

In the next step, we tested whether parcellating the amygdala into subnuclei increased the specificity of predictions of mental health dimensions. We repeated the robust regressions for the amygdala as a whole, i.e., using functional connectivity between the entire amygdala and the same set of 28 ROIs. As before, the robust regression weights of all 28 edges derived from the 3T participants were applied to the 7T participants to obtain out-of-sample predictions. This produced significant predictions of negative emotions (Pearson's r=0.2, p=0.026) and anger (Pearson's r=0.18, p=0.037; **Fig6A**). However, predictions for all four behavioural dimensions were worse when considering whole-amygdala rather than nuclei-specific amygdala functional connectivity (LifeSat: nuclei: r=0.19, whole: r=0.17; NegEmot: nuclei: r=0.22, whole: r=0.20; Sleep: nuclei: r=0.05, whole: r=-0.09; Anger: nuclei: r=0.22, whole: r=0.18; compare **Figs 4E and 6A**). We note that – while the nuclei-specific predictions include seven times as many predictors as used in the whole amygdala control – the whole amygdala is made up of the exact same voxels as the seven nuclei together. If the subdivisions of our parcellation are not meaningful or if the fMRI signal in individual nuclei and therefore the derived functional connectivity values for individual nuclei are not robust, then separation into smaller subsets of voxels should merely increase the noise in our predictions. Alternatively, if there is value in considering amygdala nuclei separately, then using functional connectivity values from each nucleus might increase our prediction accuracy for predicting mental health dimensions. Nevertheless, to account for the number of predictors, we generated separate null distributions for the nuclei- and whole-amygdala versions to obtain directly comparable

p-values (Methods). This slightly changed the precise p-values, but led to the same conclusions, suggesting negative emotions and anger could be predicted significantly (in relation to this this null distribution: p=0.026 and p=0.038, respectively). However, most importantly, the predictions of all four mental health dimensions still significantly benefited from considering the functional connectivity of separate amygdala nuclei as opposed to the amygdala as a whole (comparable p-values relative to appropriate null distribution: LifeSat: nuclei: p=0.034, whole: p=0.051; NegEmot: nuclei: p=0.016, whole: p=0.026; Sleep: nuclei: p=0.310, whole: p=0.807; Anger: nuclei: p=0.015, whole: p=0.038; **Fig 6B**). The same conclusion was reached when comparing the peak accuracy instead of the prediction achieved with the full set of edges (**Supplementary Fig 9A**). Peak accuracy using whole-amygdala edges occurred at n=16, 12, 3 and 9 for the four behavioural dimensions and fingerprints and scatterplots are shown for the earliest significant peak (**Fig 6C,D**) and global peak (**Supplementary Fig 9B**) for comparison with the nuclei version. However, even at the peak, prediction accuracies from whole-amygdala functional connectivity were worse than those achieved using nuclei-specific predictions reported above (correlation coefficients for whole-amygdala functional connectivity at the peak: r=0.185, 0.202, 0.127, 0.198; **Supplementary Fig 9**). Thus, despite their small size, considering the functional connectivity of individual subcortical nuclei benefitted predictions of mental health dimensions.

Methods p.39/40:

First, for the across-dataset replication, we computed Pearson's correlation coefficient between the overall pattern of regression weights obtained for the 3T and 7T data (row 1 versus 2 in **Figure 4B**). To establish whether the obtained correlation was better than predicted by chance, given the level of noise present in brain connections with the amygdala and given our number of connections, we generated a null distribution (**Fig 4A**) by shuffling the vectors **y** containing the behavioural dimension n=10,000 times and recomputing the correlation coefficient between the overall pattern of regression coefficients.

Second, to attempt a within-subject replication, functional connectivity values were extracted from half of the resting-state data, separately from either run 1+2 or run 3+4 (which were acquired in two separate sessions on separate days). Robust regression weights capturing the relationship between FC (X) and behavioural scores (y) were then computed separately for the FC values obtained from runs 1+2 versus 3+4. Here we used the merged data from all n=490 3T+7T participants. The similarity between the overall pattern of robust regression weights obtained for the two experimental halves was computed using Pearson's correlation. Again, to test whether their similarity was greater than expected by chance, a null distribution was generated by repeating this procedure n=10,000 times using shuffled behavioural scores (**Figure 4A**).

Methods p. 41/42:

To establish whether (a) the size of the smallest significant network and (b) Pearson's r at the peak prediction across all 196 models were significant, we generated two null distributions using bootstrapping. As before, predictions of the 7T behavioural scores were based on increasing number of edges, using the order and weight of edges extracted from robust regressions on 3T participants, but this procedure was now repeated n=10,000 times using shuffled behavioural scores. For each iteration, we determined the size of the smallest network that reached significance (r>0.168) for a given behavioural dimension and averaged this number across the four behaviours to build the first null distribution (if no

significant predictions were achieved for a given iteration and behaviour, we used a score of 197, i.e. the maximum number of edges plus 1). In each iteration, we also determined the maximum Pearson's r achieved for each behaviour and averaged these four maximum r values across behaviours to build the null distribution. This ensured in both cases that the number of tests (n=196) and the number of behavioural dimensions (n=4) was accounted for. One p-value was then extracted from each of these two null distributions.

To examine if our predictions of 7T mental health dimensions benefited from using the top edges identified in the 3T data, i.e., to establish whether the order of inclusion of the edges mattered, we generated a further null distribution: we included the same number of edges (from 1 to 196), but this time they were randomly chosen from the full set of 196 rather than sorted in order of importance based on their absolute regression coefficient. This procedure was repeated 10,000 times at each step between 1 and 196. The resulting null distribution (black line in **Fig5A**) shows, at each step, how much adding any connection out of the original set of 196 is expected to help the prediction. We report the percentage of models out of the first 70 and all 196 where the 3T-informed order of edges produced better predictions than when edges were included in random order (i.e., the difference between the coloured bars and the black null distribution shown in Fig5A).

Methods p.43:

Nevertheless, to make predictions obtained from whole vs nuclei-specific amygdala functional connectivity more comparable, we generated null distributions in both cases. We shuffled the order of participants' behavioural scores in both 3T and 7T datasets and repeated the above procedure of predicting 7T-behavioural scores from 3T-weights 10,000 times. The p-values extracted from the respective null distributions account for the number of predictors, allowing a direct comparison (**Fig 6B**).

(4) Patients and datasets – p.22 Discussion:

The finer grained parcellation we obtained reflected improved image quality and preprocessing pipelines that better controlled for physiological noise. While this could be seen as a limitation of our method because care both in data acquisition and data analysis is needed, we note that scan hardware, software and imaging sequences are developing fast and that acquiring physiological cardiac and respiratory traces is easy, cheap and uses standard equipment available in most MR facilities. It is true that our approach may not be applicable to some existing data sets, but we believe more data with sufficient resolution and SNR is likely to become available, which will make our method more broadly applicable to the neuroscientific community and to the study of other subcortical brain structures, including in clinical populations.

(5) Main advance and comparison to other brain networks – Discussion, p. 23:

The amygdala networks we describe as carrying relevance for predicting mental health dimensions are not compared to other similarly sized networks elsewhere in the brain. It is therefore not possible to conclude that they are necessarily the best possible network of its size for predicting life satisfaction, negative emotion, sleep and anger. Instead, our data emphasize (a) the value of parcellating subcortical structures into their component nuclei when studying functional connectivity to match the scale at

which these circuits are organized; (b) the importance of using functionally interpretable behavioural dimensions, rather than complex aggregate scores; and (c) that small subcortical networks can carry meaningful relevance for predicting mental health dimensions.

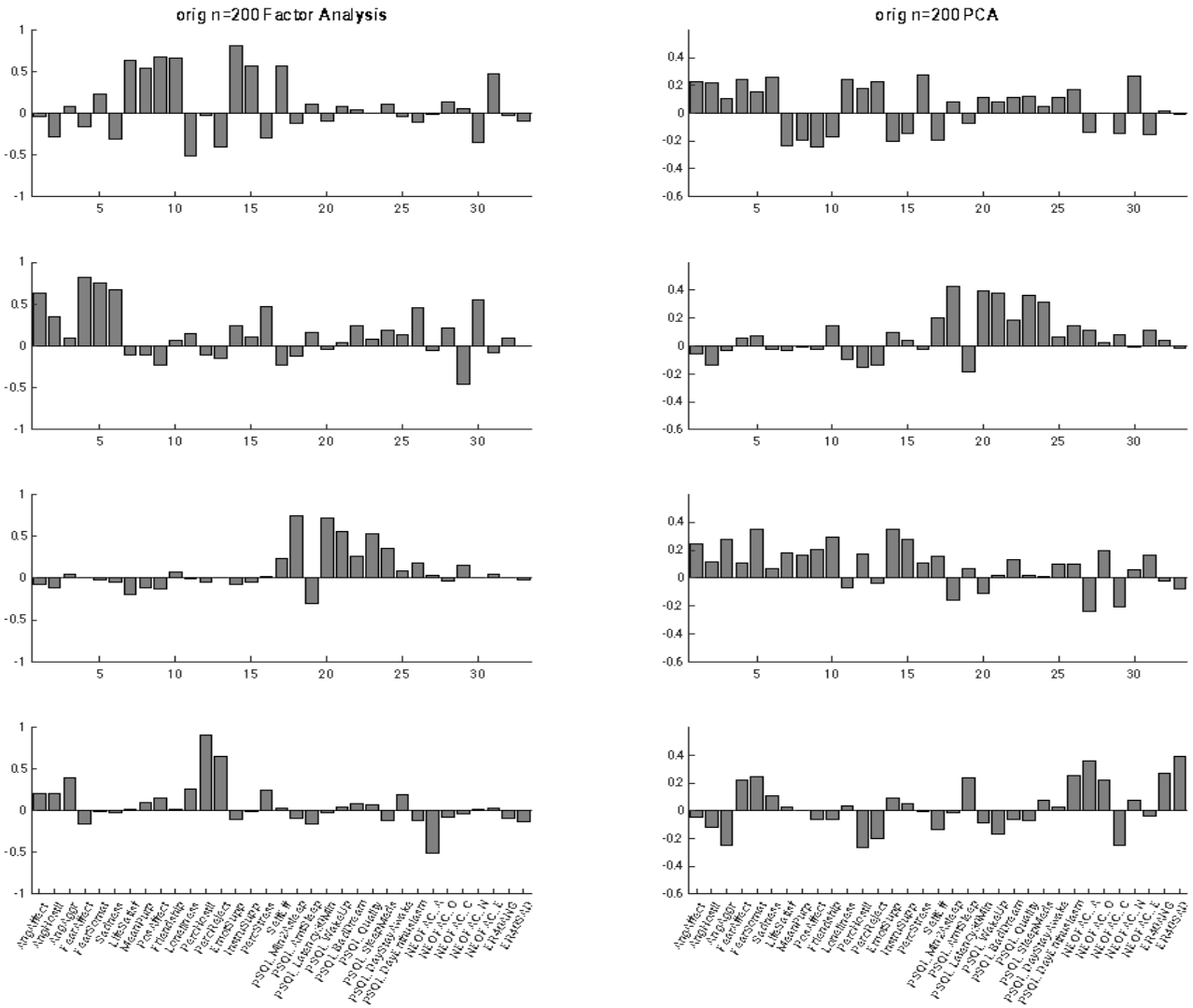
==========

Minor comments

• Why did the authors prefer using factor analysis over PCA to identify main behavioral subspaces of variation? Unlike the former (Fig. 3B), the latter produces uncorrelated components which might be easier to interpret.

The reviewer raises a very important question, and we realise that this is worth some further explanation in the revised manuscript. In summary, from a theoretical and statistical standpoint, a factor analysis generates a model of the latent variables that generated the scores and explains interrelationships among them (taking into account their unique and shared variance). By contrast, PCA is a data reduction technique that produces a weighted sum/linear combination of all scores (and assumes there is no unique variance), such that they explain the maximum amount of variance. Given the nature of our data (questionnaire scores that measure latent factors about participant's mental well-being), we were not interested in maximising variance but rather in latent variables that have maximal interpretability. Thus, a factor analysis seemed like the appropriate choice. Indeed, this is consistent with the field where factor analyses seem to be the dominant method used when dealing with questionnaire scores (e.g. Gillan et al., 2016, Rouault et al., 2018; Moutoussis et al., 2021; Scholl et al., accepted – all cited in our manuscript).

However, to respond to this comment, we ran a PCA to inspect what the first few principal components would look like and – in line with the above rationale – we find them less interpretable than our factors. The first PC has similarities with both the first and second factors (life satisfaction and negative emotions), and the 2nd PC is very similar to our third factor (sleep), but further PCs (3+4) seem not intuitive in their weighting of different variables, as shown below.



We have now added the rationale for this to the Results and a more detailed description for our choice of a factor analysis to the Methods section as follows:

==========

CHANGES IN MANUSCRIPT

Results p.10:

To capture such common 'latent' dimensions that produce these mental well-being scores, we performed a factor analysis which resulted in four main factors (see Methods; **Fig 3A and Supplementary**

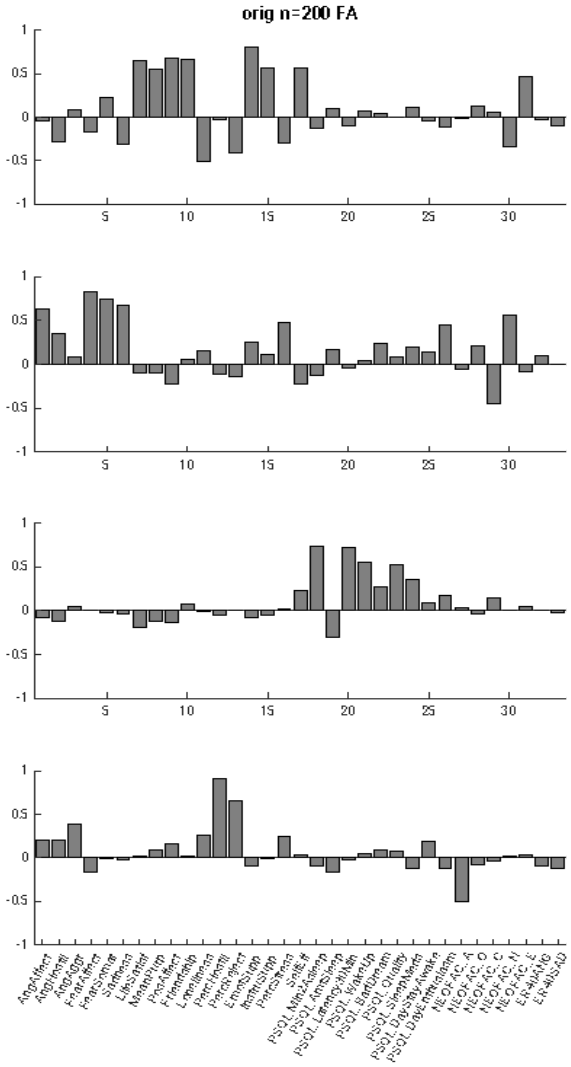
Fig 5). A factor analysis, rather than principal component analysis (PCA), is especially appropriate when the aim, as in the current study, is data reduction with maximal interpretability that is useful for establishing latent causes as opposed to data reduction into components explaining maximal variance[42–45].

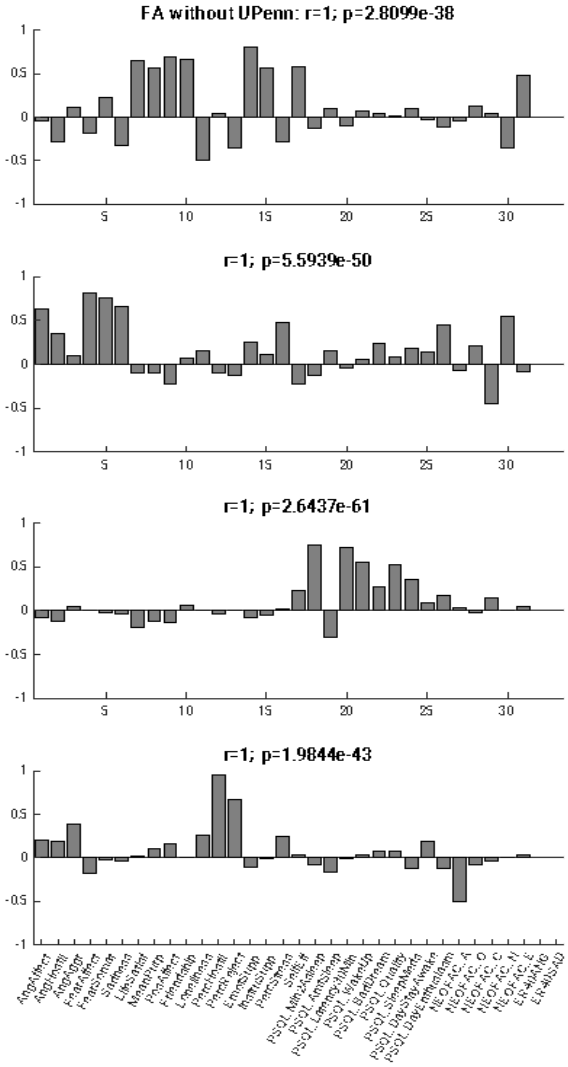
Methods p.38:
We conducted a factor analysis on these 33 behavioural markers (z-scored) using Matlab's function 'factoran', with a 'promax' rotation. A PCA would provide a data reduction of the behavioural scores that aims to obtain orthogonal principal components which explain maximal variance. By contrast, rather than merely reducing the data, here we were interested in a model of the latent variables that might have generated the behavioural scores and which explains interrelationships among them, taking into account their unique and shared variance and offering maximal interpretability [42–44]. A factor analysis therefore was the most appropriate choice of method.
==========

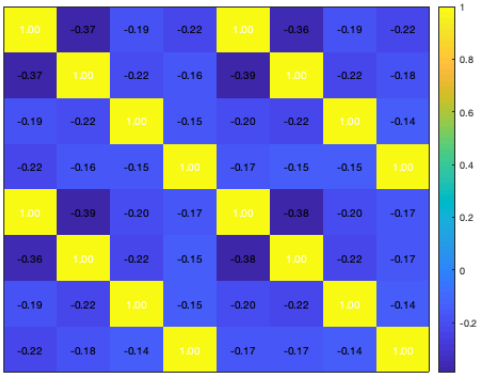
• Two measures of the Penn Emotion Recognition Test were included but these measures (32&33 in Table 1) have a close-to-zero weight in all four factors identified (Figure 3A). Why did the authors include these measures?

We did not iteratively exclude scores based on their weights, but we can see that it could have made sense given the low weights these two measures of the Penn Emotion Recognition Test are given. We basically included any behavioural measure that was part of the HCP data and that seemed relevant for mood disorders and emotional processing. The Penn Emotion Recognition Test fulfilled this criterion. We have now made this clearer in the revised version of the manuscript.

However, we have also now re-calculated the factor analysis without those two measures and as expected, the obtained factors correlate highly with the original factors. In fact, as shown in the plot below, the rounded Pearson's correlation coefficients all equal 1 when comparing the weights given to all other markers (n=31) when the two Penn Emotion Recognition scores were included vdersus excluded from the factor analysis. We would prefer to keep these two measures in, even if they effectively make no difference, because our inclusion of scores followed clear principles that we set out from the beginning. We hope the reviewer agrees with this choice.


orig n=200 FA


FA without UPenn: r=1; p=2.8099e-38
r=1; p=5.5939e-50
r=1; p=2.6437e-61
r=1; p=1.9844e-43


1.00 -0.37 -0.19 -0.22 1.00 -0.36 -0.19 -0.22
-0.37 1.00 -0.22 -0.16 -0.39 1.00 -0.22 -0.18
-0.19 -0.22 1.00 -0.15 -0.20 -0.22 1.00 -0.14
-0.22 -0.16 -0.15 1.00 -0.17 -0.15 -0.15 1.00
1.00 -0.39 -0.20 -0.17 1.00 -0.38 -0.20 -0.17
-0.36 1.00 -0.22 -0.15 -0.38 1.00 -0.22 -0.17
-0.19 -0.22 1.00 -0.15 -0.20 -0.22 1.00 -0.14
-0.22 -0.18 -0.14 1.00 -0.17 -0.17 -0.14 1.00

==========

CHANGES IN MANUSCRIPT

We have added a sentence to the Methods (p. 38):

Moreover, the factor analysis replicated to several other datasets (**Supplementary Fig 6**), most importantly to the full set of all n=1206 3T HCP participants (Pearson's correlation between factor loadings n=200 vs n=1206 3T participants: *r*=.94, *p*=1.5e-16, *r*=.93, *p*=1e-14, *r*=.97, *p*=9.6e-10, *r*=.9, *p*=1.3e-12), but also the n=400 3T and the n=98 7T participants included for neural analyses (**Supplementary Fig 6**). Note that a factor analysis performed after removing the two Penn Emotion Recognition scores (which did not contribute much to any of the four factors) resulted in identical factor loadings (all Pearson's r=1; p<1e=35).

==========

• The four runs available for each HCP subject are used to compute FC metrics. Since correlation is usually robustly estimated from a limited number of time points, I would expect the results to be reproduced using only one run (1200 time points) per subject, have the authors tested this? If this is the case, it would allow to relax data quality constraints for a subject to be included in the analysis.

This is an important observation. There are several ways to incorporate the four runs of each participant. For maximising robustness, we decided to concatenate the data from all four runs into one long time course with 4800 datapoints, and only compute the coupling between regions once using this single voxel/vertex-averaged concatenated time course for our initial analyses. One can imagine that if data from only a subset of runs are available in some participants, then functional connectivity measures will reflect the true underlying coupling less well.

Based on the comment made by this reviewer, however, we have now indeed considered runs separately to strengthen our conclusions. We extracted the functional connectivity separately based on the runs acquired in each session (runs 1 and 2 on day 1 versus runs 3 and 4 on day 2) to test whether relationships between functional connectivity and behavioural dimensions are replicable across the two halves of the data. This is shown in the new **Figure 4** (A & D, labelled "within-subject replication"; pasted above in response to point #2). The similarity between the connectivity-behaviour patterns measured across the two halves of the experiment was r=0.4711 which can be considered quite high given arguments among HCP users that subcortical brain regions contain no meaningful signal at all (despite HCP's improved resolution and neuroimaging sequences).

However, we decided not to include further participants who only have good quality data in a subset of runs because this would introduce changes in the reliability of estimates derived from different participants which would be difficult to incorporate when relating individual variation in functional connectivity to individual variation in behavioural dimensions.

However, as mentioned above, we have nevertheless now doubled our numbers to n=400 3T participants, while keeping the volume of data in each participant constant (i.e., all have four complete resting-state runs) and while maintaining our careful and rigorous physiological noise clean-up and

pre-processing routines. In addition, to replicate our findings in an independent dataset, we have now included all 7T-HCP participants that are non-overlapping with the n=400 3T-HCP participants (n=98). The changes made as a result of these additional data were summarized above under comment #2 made by this reviewer.

• It is mentioned throughout the paper that 200 HCP subjects were used but it seems from the methods that only 195 were used because 5 subjects were considered as outliers. If this is the case I would specify N=195 throughout the whole paper.

We have changed the number of included participants as explained above but now specify that we started with n=400 3T-HCP participants and n=98 (all non-overlapping) 7T-HCP participants. In other words, for all analyses performed on group connectomes (Figs 1-2), we included **n=400 3T and n=98 7T participants**, and thus **a total of 498 data sets**. For subsequent analyses conducted on individual participant's resting-state coupling values (from Fig4 onwards), we reject outlier participants from this pool of 498 participants as described before (if more than 10% of their coupling values across all connections deviate more than 3.5 standard deviations from the mean across participants, see Methods), which identified eight individuals as outliers (seven 3T and one 7T participant), and retain n**=393 3T participants and n=97 7T participants**, and so a total of n=490 participants, after outlier rejection. We have made this clearer throughout the manuscript now.

==========

CHANGES IN MANUSCRIPT
Here are several examples:

Results (p. 11):
Unlike for the amygdala parcellation which used group mean functional connectivity values, here we were interested in interindividual differences in functional connectivity. To improve the reliability of neural signals measured from individual participants, we rejected eight participants with outlier functional connectivity values, corrected for confounding variables, and used robust regressions throughout (for further details, see Methods). All subsequent analyses therefore relied on a total of n=393 3T and n=97 7T participants.

Methods (p. 39):
Outlier participants from the original pool of 400 3T and 98 (non-overlapping) 7T participants were conservatively rejected based on their individual FC values if more than 10% of their FC values across all edges deviated more than 3.5 standard deviations from the mean across participants. This identified seven 3T and one 7T participants as outliers and all analyses were performed on the remaining 393 and 97 participants.

Figure legend for Figure 4 (p. 58):

A, Relationships between interindividual variation in nuclei-specific amygdala functional connectivity and mental health dimensions were examined in two HCP datasets containing n=393 3T and n=97 non-overlapping 7T participants (following outlier rejection).

==========

• The acquisition reconstruction software version is included as a confound and regressed from the data. Is this common practice, and what is the regression weight of this confound?

We adopted the procedure used by one of the co-authors (Smith et al. Nat Neuro, 2015) for the 3T data and this was a binary indicator (an improved reconstruction method was implemented in the third quarter of the HCP data acquisition in year 1). However, this regressor is not available for the 7T data and based on the reviewer's comment and the fact that it is an unusual regressor to include, we have now removed it from all analyses and use the same confound regressors across both the 3T and 7T datasets.

==========

CHANGES IN MANUSCRIPT

Methods (p.39):

Next, confounds were regressed out of the data in a similar way as described previously [112], and this was done separately in both the 3T and 7T data. Confounds included …

==========

• I found the double grey-scale bars in Figure 5A&B not easy to interpret. For example, I was expecting them to be of equal height but this does not seem to be the case; is this a plotting inaccuracy or am I misunderstanding something?

The major revisions in this part of the manuscript detailed above have meant that we have generated new figures and this figure is no longer part of the manuscript. We hope that you will find the new figures easier to interpret but would be happy to make any further changes to aid clarity.

• Line 107: 'ratio' is missing

Thanks, this has been corrected.

We would like to thank you again for your thoughtful and very helpful suggestions. We hope that you will find that the manuscript has improved as a result of the changes we have implemented.

Decision Letter, first revision:

6th January 2022

Dear Miriam,

Thank you once again for your manuscript, entitled "Relationship between nuclei-specific amygdala connectivity and mental health dimensions in humans," and for your patience during the peer review process.

Your manuscript has now been seen by 2 reviewers, whose comments are included at the end of this letter. As I mentioned previously, Reviewer #1 was not able to re-review the manuscript. Reviewer #3 is a new referee who was asked to comment on your response to Reviewer #1.

While Reviewer #3 says that your response to Reviewer #1 is acceptable, you will see that they raise a concern of their own about the parcellation. I am afraid that as this is a technical concern of the type that would call the results into question if not addressed, we will require a new revision of the manuscript to respond to Reviewer #3's point. I understand that this is disappointing news especially given the extent of the revisions you have already carried out.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within four to eight weeks. We understand that the COVID-19 pandemic is causing significant disruption for many of our authors and reviewers. If you cannot send your revised manuscript within this time, please let us know - we will be happy to extend the submission date to enable you to complete your work on the revision.

With your revision, please:

• Include a “Response to the editors and reviewers” document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

• Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

[REDACTED]

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to

co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,
Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

REVIEWER COMMENTS:

Reviewer #2:
Remarks to the Author:

I thank the authors for the thorough revision of the manuscript, I have no further comment and I support publication. Congratulations for the nice work.

Reviewer #3:
Remarks to the Author:

The authors have done an excellent job addressing Reviewer #1's comments. I have only one remaining comment:

It is unclear what criterion was used to establish that the amygdala parcellation replicated across datasets. I could not find a quantitative assessment and supplementary figure 4A shows quite a few differences still. Quantitative support is provided for replication of the functional connectivity results, so is it possible that the parcellation differences reflect meaningful group differences in amygdala anatomy? If so, this could mean that there are substantial anatomical differences between subjects, which in turn could mean that the predictive effects are influenced by anatomical mislocation rather than brain circuitry differences per se. A similar concern applies to the ROI definitions in the brain stem and subcortex, which are unlikely to precisely locate and delineate the exact anatomical structures at the individual level. These caveats should be discussed in the light of the suggested implications for more targeted interventions.

Author Rebuttal, first revision:

Reviewer comments:

Reviewer #2:

I thank the authors for the thorough revision of the manuscript, I have no further comment and I support publication. Congratulations for the nice work.

Thank you very much for your helpful comments and positive feedback.

Reviewer #3:

The authors have done an excellent job addressing Reviewer #1's comments. I have only one remaining comment:

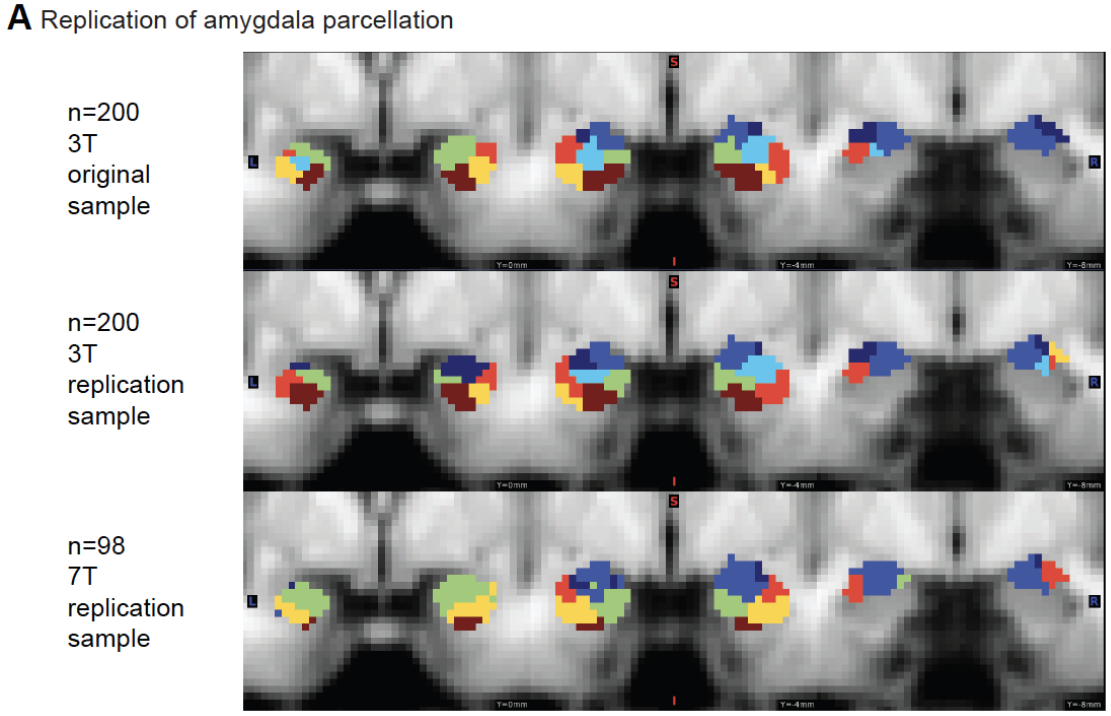
It is unclear what criterion was used to establish that the amygdala parcellation replicated across datasets. I could not find a quantitative assessment and supplementary figure 4A shows quite a few differences still. Quantitative support is provided for replication of the functional connectivity results, so is it possible that the parcellation differences reflect meaningful group differences in amygdala anatomy? If so, this could mean that there are substantial anatomical differences between subjects, which in turn could mean that the predictive effects are influenced by anatomical mislocation rather than brain circuitry differences per se. A similar concern applies to the ROI definitions in the brain stem and subcortex, which are unlikely to precisely locate and delineate the exact anatomical structures at the individual level. These caveats should be discussed in the light of the suggested implications for more targeted interventions.

Thank you for this thoughtful comment. We have now added additional analyses as well as further discussion to the text.

Clarification

*In the previous version of the manuscript, we thought that differences in the parcellation across the three study cohorts (n=200 3T original, n=200 3T replication; n=98 7T replication as shown in the original Suppl Fig 4A pasted below for convenience) were not a concern because **we used the same parcellation throughout the entire manuscript**. For all analyses reported in the manuscript, we consistently used the parcellation extracted from the first 200 participants (top row in the below figure) – even for analyses/replications performed on the two replication datasets. We will comment further below why we think it is unlikely that there are meaningful group differences in amygdala anatomy. We realise that the fact that we used the **same parcellation throughout** was maybe not sufficiently clear in our manuscript, so the first change we have implemented is to explicitly **state in multiple places that the same parcellation was used throughout**.*

As noted by the reviewer, we also reported that the mean functional connectivity results replicated across cohorts (original Suppl Fig 4B now moved to Suppl Fig5). Again, we believe we should have been clearer that this analysis was based on the same parcellation in all three cohorts (from the original n=200 3T participants). However, we believe this suggests that using this same amygdala parcellation throughout is justified and that the amygdala is organized similarly across cohorts.



We have made the following textual changes:

Results, p.7:

We replicated this parcellation in two additional datasets (3T: n=200; 7T: n=98; **Supplementary Fig 4**; see Methods and **Supplementary Figure 4** for further details and a quantification of their similarity), but we used the original parcellation throughout the manuscript.

Results, p.8:

This parcellation and the corresponding labels were used in all analyses reported from this point onwards.

Figure legends, p.58:

The parcellations obtained in two independent datasets closely reproduced these nuclei subdivisions (**Supplementary Fig 4**), but for consistency, the parcellation shown here was used for all analyses reported in this study.

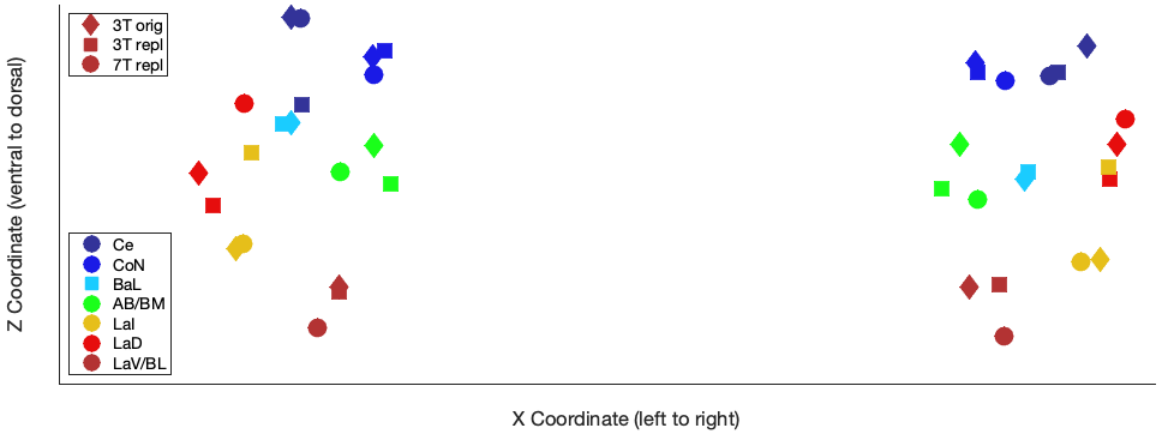
Supplementary Information, p.8

Supplementary Figure 5, Replication of amygdala nuclei mean functional connectivity, The average amygdala nuclei to ROI functional connectivity, in all cases extracted based on the amygdala nuclei from the original n=200 3T parcellation, replicates across cohorts (top: original, bottom left: replication 3T, bottom right: replication 7T; compare **Fig 2B**), as confirmed in the strong correlation between these patterns (top right).

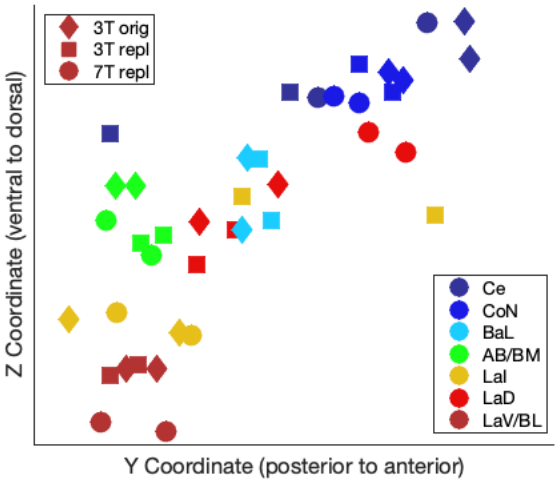
Analysis quantifying similarities

*The reviewer correctly noticed that we did not try and **quantify the similarities and differences between the three parcellations** in the previous version of the manuscript. We felt this was not crucial since we only used one parcellation throughout. Nevertheless, we have now tried to implement such an analysis which shows that our parcellations are **more similar than expected by chance**.*

First, to help compare the parcellations visually, we plotted the centroid of each of the seven nuclei for the original 3T parcellation (diamonds), the 3T replication (squares) and the 7T replication (circles) in a view corresponding to a coronal slice:



And also using a sagittal slice view:

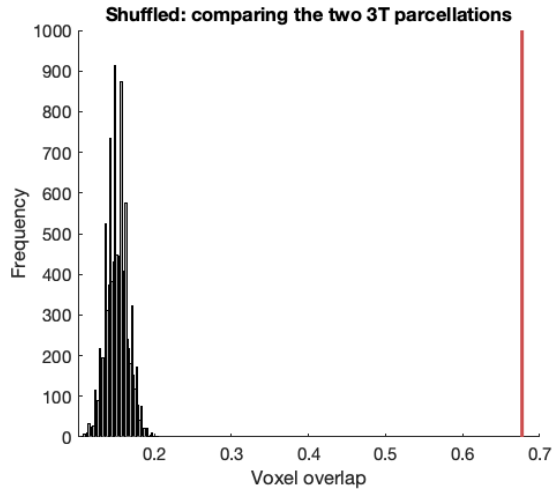


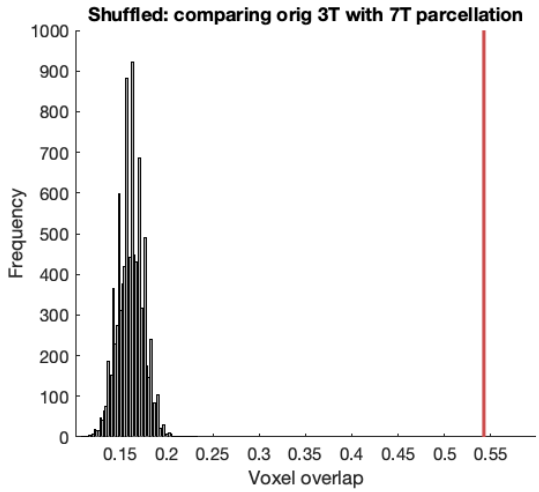
We believe both views show overall good correspondence between the centroids of the clusters. To quantify the similarity of the parcellations, we computed two metrics:

*First, the **percentage overlap** between voxels, where overlap is defined as being assigned the same nucleus label (chance corresponds to 14% or 1/7th because there are seven nuclei). The overlap between the original and the first replication was 68% and the overlap between the original and the 2nd 7T replication was 54%.*

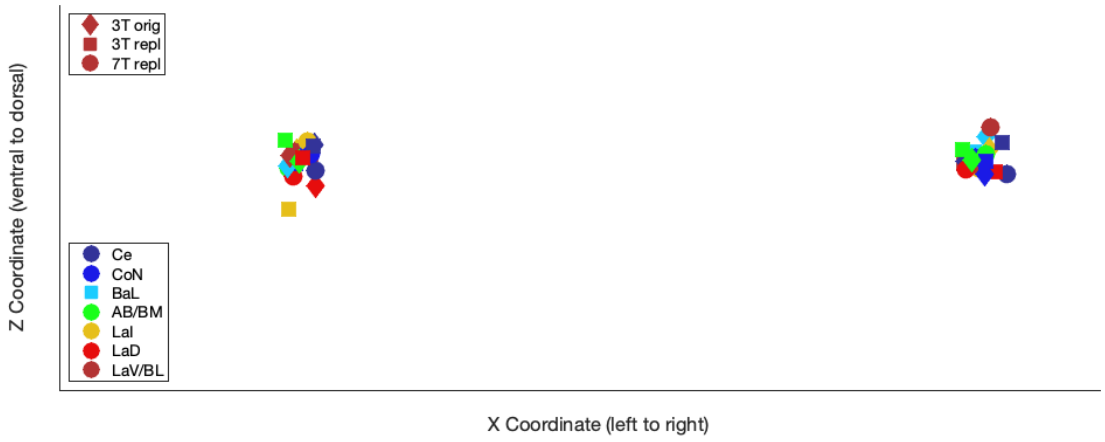
*Second, we computed the mean vector **distance between centroids** which was 1.42mm when comparing the two 3T parcellations and 1.30mm when comparing the original 3T with the 7T parcellation.*

However, it is not trivial to generate appropriate null distributions for deriving statistical p-values for these metrics. One basic null distribution could be based on simple shuffling of the nuclei labels. In that way, the nuclei sizes correspond to those in the original parcellation, but the extracted shuffled parcellations do not have spatially contiguous or symmetrical nuclei. Note that in the original parcellation, the clustering algorithm was not constrained to produce spatially contiguous or symmetrical nuclei of the sort that we found. Thus, it could be seen as appropriate that the shuffled parcellations should likewise not be constrained to have these features. According to this null distribution (shown below), the percentage overlap is associated with p=0 for both the similarity between the two 3T parcellations or the 3T and 7T parcellation (red line shows values in real data, grey histogram shows percentage overlap generated in n=10,000 shuffles).

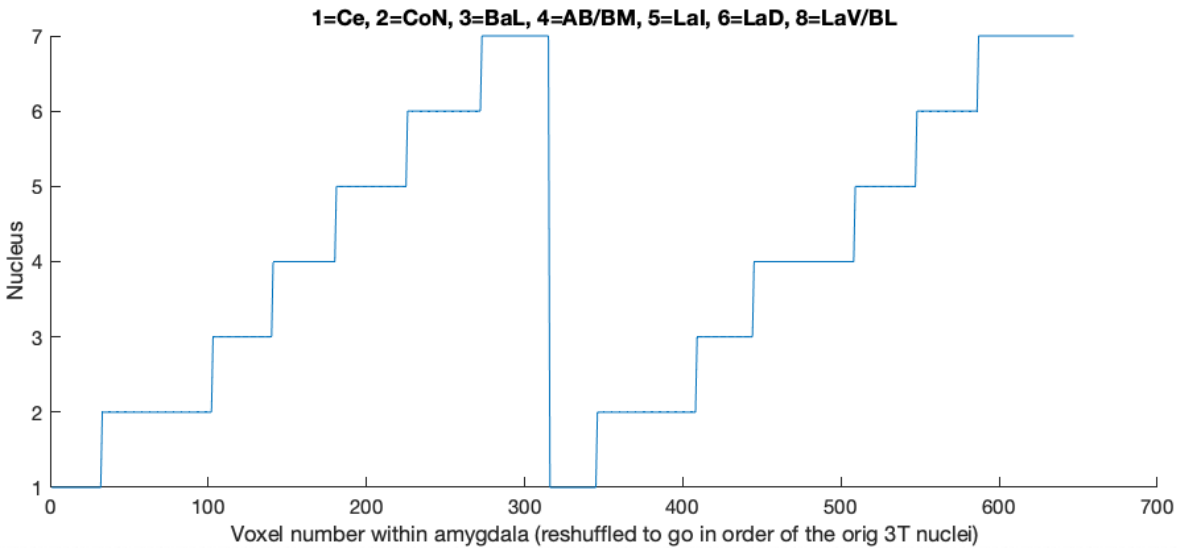




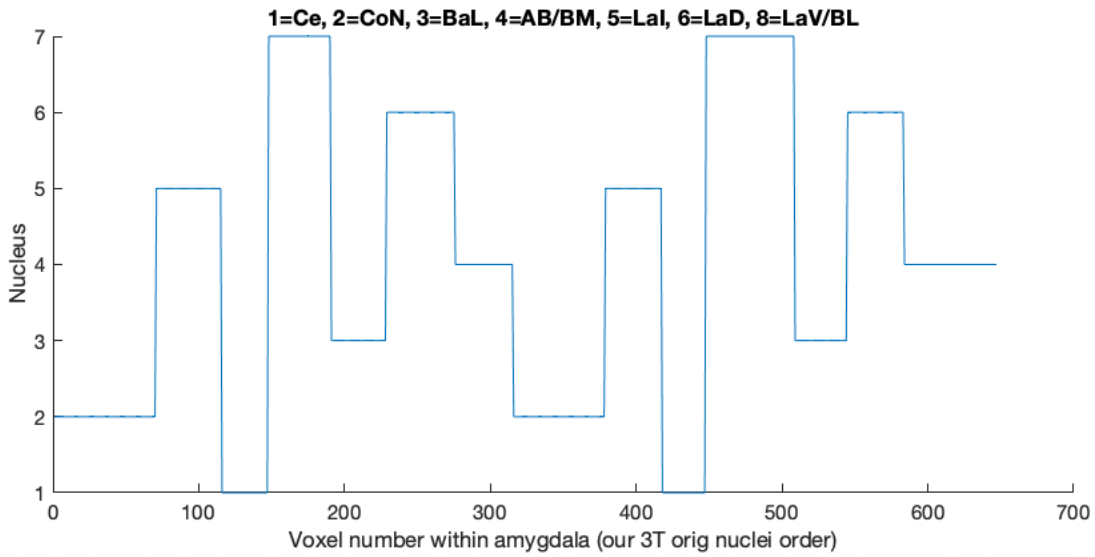
However, using the centroid distance as a metric does not make sense for this null distribution, as all clusters tend to have a centroid in the centre of the left and right amygdala, respectively (because voxels are dispersed randomly within each hemisphere), and are therefore close to each other. The figure below shows the centroids from one randomly chosen iteration in the same space as shown above for the real parcellations:



*We derived one **other form of null distribution**. Again, we kept the nuclei sizes the same as in the real parcellations, but we added some additional constraints to produce spatially contiguous and roughly symmetrical clusters, matching the features of our real parcellations. However, because it is not well-defined in a 3D space in which direction to expand a nucleus to ensure that it is contiguous, we used the order of the voxels and their corresponding MNI coordinates from our original 3T parcellation as reference. According to this order (shown on the x axis below), the voxels in our original parcellation would by definition be sorted by nucleus (1-7) and hemisphere (two repeats of 1-7):*

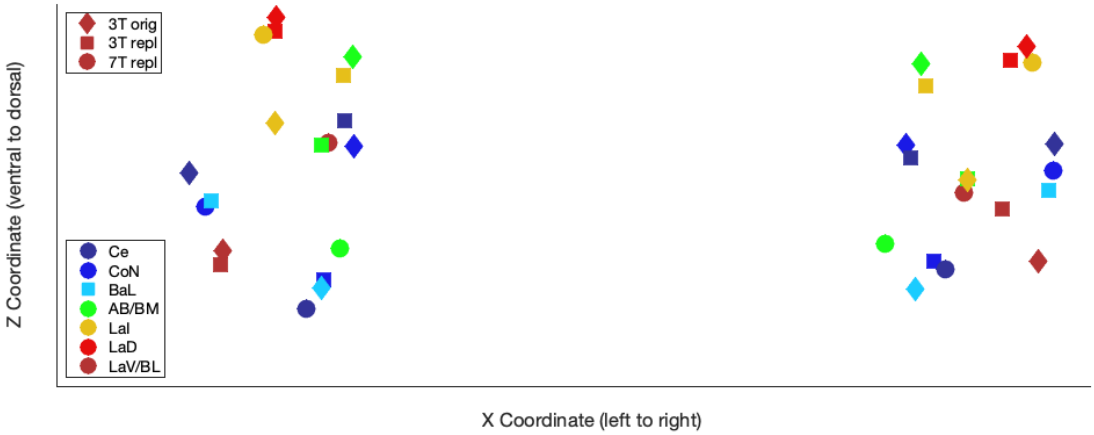


To generate shuffled versions of this parcellation, we kept the nuclei sizes and symmetry constraints identical, but placed the nuclei in a random order within the amygdala (the same order for right and left hemisphere), in non-overlapping voxels, following the order of our original 3T parcellation (plotted on the x axis). This means that a shuffled version might look like this (order of voxels plotted as above, nuclei sizes the same, labels symmetrical across hemispheres and continuous in space, but with a shuffled nuclei order):

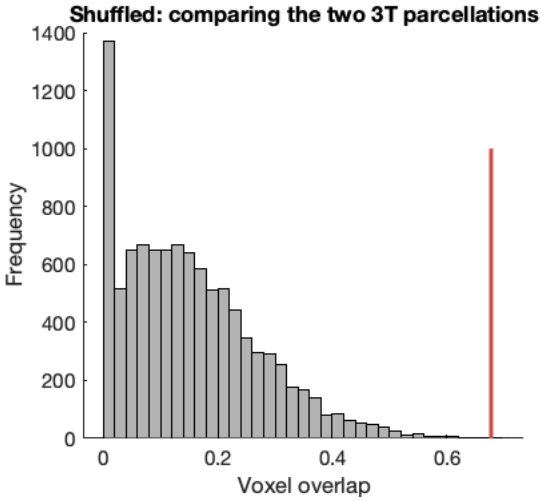


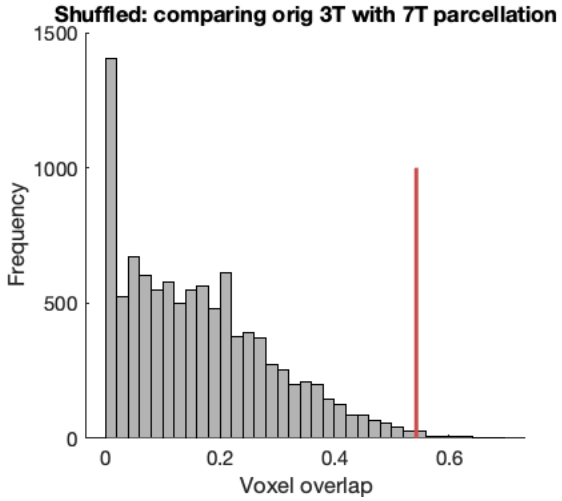
Thus, the obtained shuffled parcellations have similar spatial features and configurations as the real parcellations. This provides the basis for a more stringent test of the consistency of the parcellation results because, as noted, the parcellation algorithm was not actually constrained to produce spatially contiguous or symmetrical nuclei but these constraints are inherent to the shuffling process implemented here.

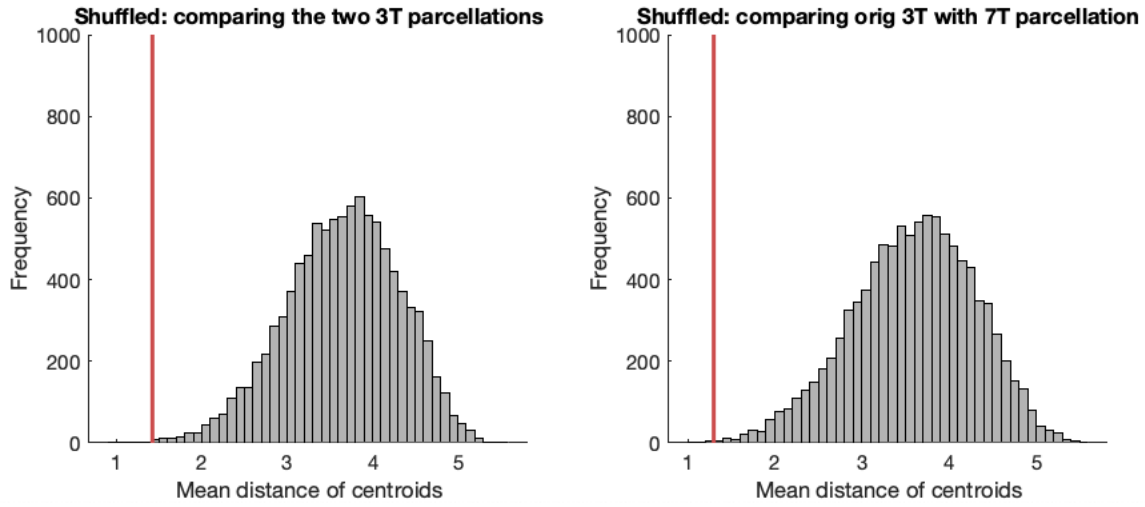
The centroids for the three shuffled parcellations might look like this (shown for a randomly chosen iteration, plotted in the same way as above for the real centroids):



Using this way of generating the null distribution and repeating the shuffling n=10,000 times generates the null distributions shown below. Importantly, all our measured values are significant indicating a better-than-chance similarity between our parcellations: overlap metric: comparing both 3T parcellations p=0.0003; comparing the original 3T and 7T parcellations: 0.0052; centroid distance metric: comparing both 3T parcellations p=0.0008; comparing the original 3T and 7T parcellations: p=0.0009.



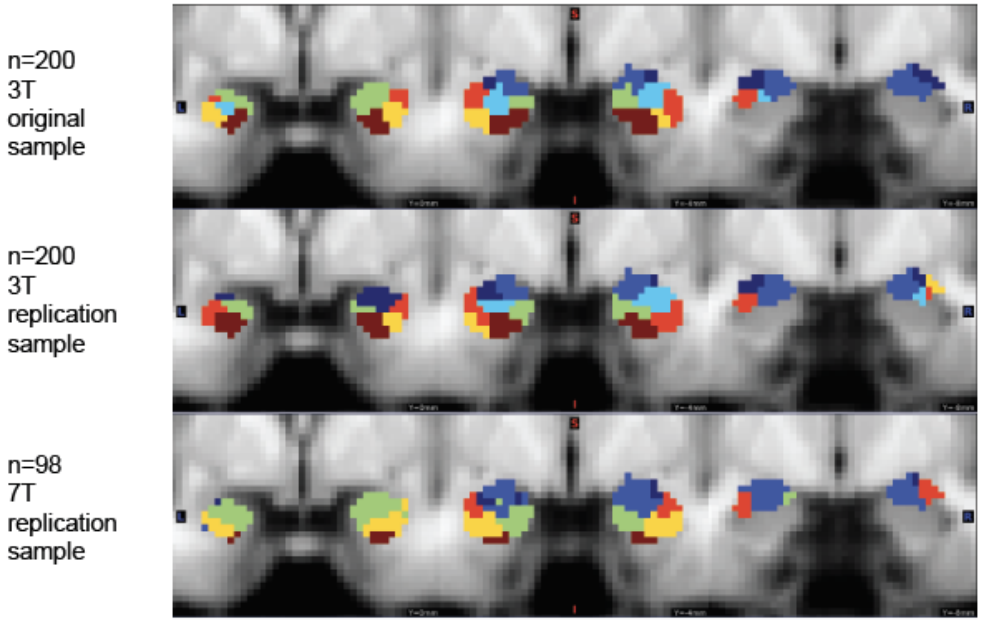




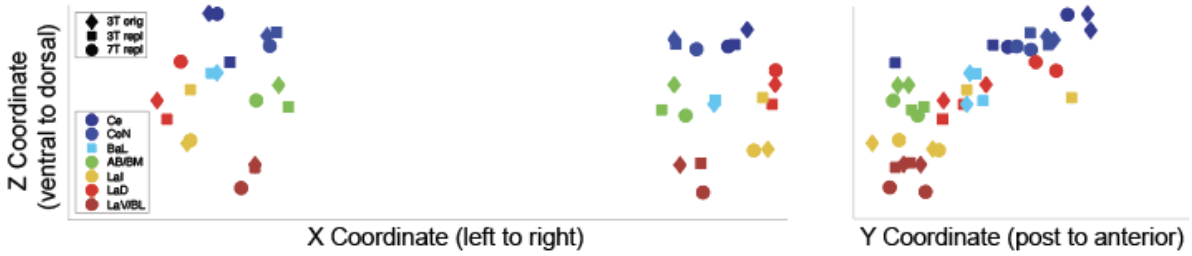
Here all random parcellations adhered to the size and symmetry constraints consistent with the original parcellation and contained contiguous clusters in space. Nevertheless, other basic anatomical constraints about where certain nuclei should be located based on prior knowledge are neglected. Nevertheless, based on this, we can conclude that our parcellations are more similar than expected by chance.

We have now added these new results to Supplementary Figure 4 and its legend (and put part of the original Suppl Fig 4 into a new Suppl Fig 5, shifting the numbering by 1).

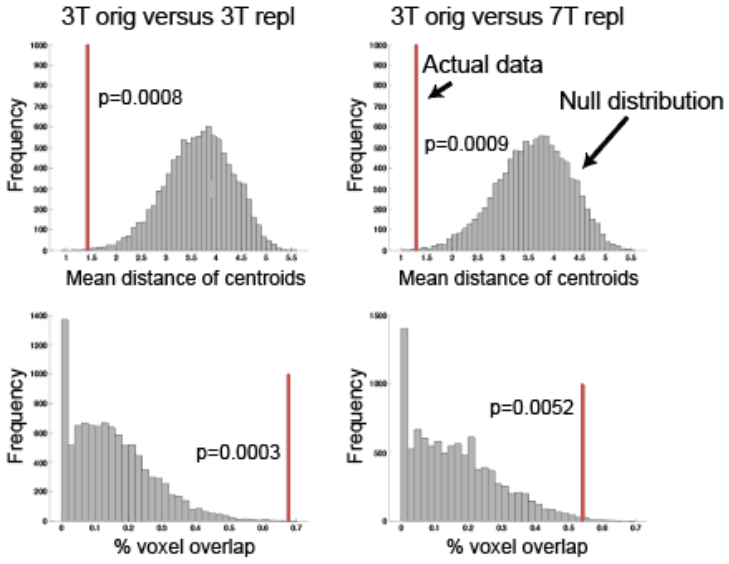
A Replication of amygdala parcellation



B Cluster centroids of the three parcellations



C Similarity of parcellations compared to null with contiguous and symmetrical clusters



Supplementary Figure 4 – related to Figure 2

Supplementary Figure 4, Replication of amygdala parcellation, A For comparison, the parcellation of the amygdala obtained in the original n=200 3T participants is shown for the n=200 3T replication sample and the n=98 (all non-overlapping) 7T participants (compare Fig 1B). This shows that the key subdivisions of the amygdala were replicated in these two additional parcellations. **B**, Visualization of the cluster centroids from a coronal (left) and sagittal (right) point of view illustrates the similarity of the three parcellations (diamond: original 3T parcellation; square: 3T replication; circle: 7T replication). **C**, To quantify the similarity between the parcellations, two metrics are reported: the mean distance of the centroids and the % of overlapping voxels (i.e., voxels with identical labels). Null distributions respect the size and symmetry of the original parcellations but shuffle the location of the nuclei in a way that yields contiguous but non-overlapping clusters. This shows that the parcellations (left: comparison with 3T replication; right: with 7T replication) are more similar than expected by chance. Importantly, however, throughout the manuscript, we use the original 3T parcellation across all analyses. In addition, the choice of parcellation (which is based on the mean group connectivity) is orthogonal to the key findings reported in the manuscript which relate to interindividual variation that is ignored when generating the parcellation.

We also added a short paragraph to the Methods, p.31:

We used the parcellation generated from the first 3T dataset for all further analyses. Importantly, since this parcellation was obtained from the group connectome, rather than for each subject individually, it did not introduce bias in subsequent analyses focusing on individual differences. We also quantified the similarity between the original and the two replication parcellations using two metrics: (a) the mean distance between cluster centroids and (b) the percentage overlap between parcellations (i.e., percentage of voxels with the same label). The overlap between the original parcellation and the 3T replication was 68% and the overlap between the original parcellation and the 7T replication was 54%. For further details and statistical analysis, see **Supplementary Fig 4B,C**).

Origin of group differences in the parcellation

*Without knowing the ground truth, of course we cannot rule out that there could be meaningful **group differences in amygdala anatomy**, as suggested by the reviewer. Nevertheless, we think this is unlikely to be driving the parcellation differences we see for at least two reasons:*

*(1) First, the three parcellations used the **group average** connectivity of all participants in each respective cohort. All three cohorts involved healthy young adults matched in their overall*

demographics (as can be seen in the below table). While the range of transdiagnostic scores was not matched perfectly (see previous response to the reviewers and manuscript), this would only matter for analyses considering interindividual variation. Thus, from a behavioural/demographic point of view, it is unclear what could have driven potential differences in amygdala anatomy.

Comparison of participant cohorts

	Original n=200 3T	Replication n=200 3T	Replication n=98 7T	All n=1206 HCP
Mean age	28.95	28.28	29.42	28.84
% females	54%	49.5%	60%	54%
DSM mean	4.46	3.96	3.43	4.25
ASR mean	37.90	35.72	31.79	37.43
Employment status *	1.53	1.50	1.57	1.52
Income	4.92	5.01	5.06	5.00
Education	14.98	15.09	14.89	14.86
Cognitive status (MMSE score)	29.1	29.03	28.98	28.98
Fluid intelligence PMAT ****	16.56	16.76	17.45	16.65

* not working = 0, part-time employment = 1; full-time employment = 2
** Total household income: <$10,000 = 1,10K-19,999 = 2, 20K-29,999 = 3,30K-39,999 = 4, 40K-49,999 = 5, 50K-74,999 = 6, 75K-99,999 = 7, >=100,000 = 8
*** Years of education completed: <11 = 11; 12; 13; 14; 15; 16; 17+ = 17
**** Penn Progressive Matrices: Number of Correct Responses

*(2) Second, from an anatomical standpoint, the amygdala may be a relatively more **preserved structure** than cortex. We have known for some time that there is individual variation in the presence, position, and organizational patterning of sulci and gyri in the cortex (Kubik and Ono, 1990; Petrides, 2018) but no individual variability in amygdala nuclei has been documented. In fact, there is a similar lack of documented variation in other subcortical nuclei in the brainstem and elsewhere. Therefore, differences in the location of amygdala nuclei in our three cohorts more likely simply reflect noise especially since subcortical BOLD data has a relatively low*

signal-to-noise ratio. In the Methods and our first reply to the reviewers (in round 1), we already commented on potential differences in image quality. Importantly, however, we believe all of this should, if anything, have made a replication more difficult and shows the robustness of the effects we report.

We have now added additional demographic information to Suppl Table 3 and commented on this in a new paragraph in the Discussion:

Discussion p.23:

We used an identical amygdala parcellation for all participants. This parcellation was based on our initial group dataset but could be replicated in additional 3T and 7T cohorts (**Supplementary Fig 4**). The amygdala is a relatively more preserved structure than cortex. There is individual variation in the presence, position, and organizational patterning of sulci and gyri in the cortex [62,63] but individual variability in amygdala nuclei has not been documented. Similarly, there is no evidence for major interindividual variation in other subcortical nuclei, for example in the brainstem. Nevertheless, examining whether the shape or location of amygdala nuclei or other brain regions might systematically relate to mental health markers is an interesting avenue for future work. There is evidence for atypical neurodevelopment linked to transdiagnostic markers [64] and resting-state connectivity at rest (taken as a proxy for long-range brain connections) can predict the shape of task-related activations [65]. If the location or shape of a given nucleus or region of interest varies with a persons' mental health, then using a group mask which captures a given participant's true region more or less well could contribute to the effects we report. Investigating this question will require future work to examine individual parcellations which was not attempted here.

Key effects are orthogonal to the choice of parcellation

Ultimately, we are not able to fully determine what is driving the minor parcellation differences. However, crucially, the main effects of interest reported in the manuscript are orthogonal to the choice of parcellation.

The main relationships of interest in our manuscript are ones between interindividual variation in mental health scores and connectivity. As mentioned above, we used the parcellation from the first cohort (i.e., a group parcellation) for any replication analysis. We believe this should have hindered rather than helped any attempts to replicate behaviour-connectivity relationships in the replication datasets.

We will give one example to illustrate this: let's assume that one nucleus (e.g., Ce) is not well captured in one of the replication datasets (e.g., the 7T dataset) given we are using the parcellation from the original n=200 3T participants. That means that likely we are 'smearing' the BOLD time course from the Ce nucleus with that from adjacent other nuclei when we extract it in the 7T cohort. Thus, relating Ce nucleus connectivity to mental health scores in the 7T data could reflect a mixture of other nuclei, or miss features specific the Ce nucleus. If there is truly a relationship with Ce-connectivity, this could mean we see a noisier replication or no replication at all, depending on the influence of these other adjacent nuclei over and above the Ce nucleus time series. It should not, however, lead to false positives, unless adjacent nuclei truly carry the relevant variance. However, this is unlikely because the influence of neighbouring nuclei should still be smaller than the influence of the Ce nucleus itself. So, for any successful replication, we can still conclude that the relevant nucleus (e.g., here Ce) is driving the observed effects.

Anatomical mislocation & future work on how shape relates to function

If the location or shape of a given nucleus varies with a persons' mental health, then using a group parcellation which captures the 'true' nucleus of a given participant better or worse could be driving some of our effects. We never tried generating a parcellation for each participant because we believe this would result in very noisy and difficult-to-compare parcellations. We think a systematic mislocation is unlikely (for example because we look at several mental health scores not just one, and at connectivity with other regions and the amygdala and our effects are behaviour and nucleus/connection specific). However, we cannot fully rule out this possibility and have therefore added a comment on this to the Discussion.

We think the general idea that function may relate to basic properties such as shape/location is interesting and deserves more attention in our discussion, and this equally applies to the amygdala parcellation and the other ROIs, as commented on by the reviewer. For example, Tavor, Jbabdi and colleagues found that resting-state connectivity at rest (taken as a proxy for long-range brain connections) can predict the shape/extent of task-related activation across a set of tasks, including theory of mind, language, relational reasoning and motor performance (Tavor et al., 2016). Ann Hermundstad, Danielle Bassett and colleagues have an interesting line of work showing structure-function relationships, e.g., relationships between white matter connectivity and both rest- and task-related functional connectivity (Hermundstad et al., 2014). Although their focus is not as much on individual differences in the localization of brain activation, it is

possible that differences in anatomy and function could lead to differences in the size, shape, or location of functional networks investigated here.

We believe this is an interesting avenue for future work. Examining whether the shape or location of a nucleus or a brain region might systematically relate to mental health markers is currently unknown as far as we are aware (although there is evidence for atypical neuro-development linked to transdiagnostic mental health markers: (Parkes et al., 2021)). We have now added a paragraph to the Discussion:

Discussion p.23 (also pasted above already):

We used an identical amygdala parcellation for all participants. This parcellation was based on our initial group dataset but could be replicated in additional 3T and 7T cohorts (**Supplementary Fig 4**). The amygdala is a relatively more preserved structure than cortex. There is individual variation in the presence, position, and organizational patterning of sulci and gyri in the cortex [62,63] but individual variability in amygdala nuclei has not been documented. Similarly, there is no evidence for major interindividual variation in other subcortical nuclei, for example in the brainstem. Nevertheless, examining whether the shape or location of amygdala nuclei or other brain regions might systematically relate to mental health markers is an interesting avenue for future work. There is evidence for atypical neurodevelopment linked to transdiagnostic markers [64] and resting-state connectivity at rest (taken as a proxy for long-range brain connections) can predict the shape of task-related activations [65]. If the location or shape of a given nucleus or region of interest varies with a persons' mental health, then using a group mask which captures a given participant's true region more or less well could contribute to the effects we report. Investigating this question will require future work to examine individual parcellations which was not attempted here.

In summary, we hope that our reply addresses the points you made in your comment. Thank you so much for your helpful suggestions which we believe have improved the clarity and rigour of our manuscript.

REFERENCES

Hermundstad, A.M., Brown, K.S., Bassett, D.S., Aminoff, E.M., Frithsen, A., Johnson, A., Tipper, C.M., Miller, M.B., Grafton, S.T., and Carlson, J.M. (2014). Structurally-Constrained Relationships between Cognitive States in the Human Brain. PLOS Comput. Biol. *10*, e1003591. https://doi.org/10.1371/journal.pcbi.1003591.

Kubik, S., and Ono, M. (1990). Atlas of the Cerebral Sulci by Stefan Kubik, M. Ono | Waterstones.

Parkes, L., Moore, T.M., Calkins, M.E., Cook, P.A., Cieslak, M., Roalf, D.R., Wolf, D.H., Gur, R.C., Gur, R.E., Satterthwaite, T.D., et al. (2021). Transdiagnostic dimensions of psychopathology explain individuals' unique deviations from normative neurodevelopment in brain structure. Transl. Psychiatry *11*, 232. https://doi.org/10.1038/s41398-021-01342-6.

Petrides, M. (2018). Atlas of the Morphology of the Human Cerebral Cortex on the Average MNI Brain - 1st Edition.

Tavor, I., Parker Jones, O., Mars, R.B., Smith, S.M., Behrens, T.E., and Jbabdi, S. (2016). Task-free MRI predicts individual differences in brain activity during task performance. Science *352*, 216–220. https://doi.org/10.1126/science.aad8127.

Decision Letter, second revision:

Our ref: NATHUMBEHAV-20029312B

28th April 2022

Dear Dr. Klein-Flugge,

Thank you for submitting your revised manuscript "Relationship between nuclei-specific amygdala connectivity and mental health dimensions in humans" (NATHUMBEHAV-20029312B). It has now been seen by the original referees and their comments are below. As you can see, the reviewers find that the paper has improved in revision.

We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

Jamie

Dr Jamie Horder
Senior Editor
Nature Human Behaviour

Reviewer #3 (Remarks to the Author):

The authors have thoroughly and adequately addressed my final concern and I have no further comments.

Final Decision Letter: