nature human behaviour

SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41562-017-0234-y

In the format provided by the authors and unedited.

Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth

Marcel Adam Just[1*], Lisa Pan[2], Vladimir L. Cherkassky[1], Dana L. McMakin[3], Christine Cha[4], Matthew K. Nock[5] and David Brent[2]

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [3]Department of Psychology, Florida International University, Miami, FL, USA. [4]Clinical Psychology Department, Columbia University, New York, NY, USA. [5]Department of Psychology, Harvard University, Cambridge, MA, USA. *e-mail: just@cmu.edu

Supplementary Methods

Participants

There were no reliable differences between the included and excluded suicidal ideator participants in ASIQ, number of attempts, and PHQ; CTQ was lower for included than excluded participants (included: 41.3, excluded: 57.1, $t(36) = 2.75$, $p<0.01$) and WASI IQ was higher for included than excluded participants (included: 124, excluded: 113, $t(36) = 2.52$, $p<0.02$). In the case of the control participants, there were no differences between those who were included versus excluded from the neurosemantic analysis, except for the gender ratio (included male proportion: 17.6%, excluded male proportion: 62.5%, $\chi^2(1) = 7.99$, $p < 0.00$).

The two full groups of participants (38 ideators and 41 controls) (Supplementary Table 1) differed in age (years): ideators 24.7 (SD=5.7), controls 22.0 (SD=2.9), $t(77) = 2.72$, $p < .009$ and in gender ratio (suicide ideators male proportion: 21.1%, controls male proportion 43.9%, $\chi^2(1)=4.58$, $p<0.04$), with a greater number of female participants in the suicidal ideator group. This is consistent with population studies that indicate higher rates of reported suicidal ideation in women than men [1]. Participants in the suicidal ideator group were significantly more symptomatic than the control group on all measures. Within the suicidal ideator group, participants who reported a prior suicide attempt had significantly higher scores on self-reported depression (PHQ-9) [2] (attempters: 14.0, non-attempters: 9.8, $t(36) = 2.21$, $p < 0.035$) and suicidal ideation (A-SIQ) [3] (attempters: 68.7, non-attempters: 43.5, $t(36) = 2.75$, $p < 0.01$) compared to those without a past suicide attempt.

fMRI data acquisition and preprocessing

The fMRI data were acquired on a Siemens Verio 3.0 Tesla scanner. Parameters for echo-planar pulse sequences were: TR (repetition time) = 1000 ms, TE (or echo time) = 30 ms, flip angle = 60 degrees, FOV (field of view) = 20 cm, matrix size = 64 x 64, and voxel size of 3.125 x 3.125 x 5 mm thick (skipping 1mm between slices) in 20 AC-PC aligned brain slices that cover the cerebrum. The images were slice-timing- and motion-corrected, and spatially normalized to the MNI template without changing voxel size (3.125 x 3.125 x 6 mm) using SPM8 (Wellcome Dept. of Cog. Neurology). The % signal change relative to fixation was computed at each gray matter voxel for each stimulus item at each presentation. The input for subsequent analyses consisted of the mean % signal change of each voxel averaged over the four images acquired within a 4s window, offset 4s from the stimulus onset, to account for hemodynamic delay. These mean images for each stimulus item were then normalized.

Selecting voxels with stable semantic tuning curves. The analyses focus on a subset of all the voxels (each ~50 mm^3) whose profile or semantic tuning curve of activation over the stimulus items is *stable* across the multiple presentations of the set of items. Voxel stability is measured by the mean correlation of the vector of activation levels (across the set of stimulus items) over the multiple pairs of presentations. High stability is thus an analytic for the replicability of the voxel's semantic tuning curve. The voxel selection is based on only the training data for the model in each cross validation fold and is then applied to the test data.

Group membership classification. Two main types of group classification were performed: 1. suicidal ideator vs control group, consisting of 17 participants in each group, and, 2. within the suicidal ideator group, attempters (n=9) vs. non-attempters (n=8). Both types of classification were based on fMRI data in the sets of group-level stable clusters identified for both groups.

Obtaining clusters of stable voxels to use as feature locations. The various reported classifications all used a Gaussian Naïve Bayes classifier (described below), but the sets of features (based on the activation levels for some concepts in some brain locations) differed. To select a set of brain locations potentially useful for participant's characterization, particularly for the classification of the ideator and control groups, initially the 1000 most stable voxels were selected in each participant (excluding bilateral occipital lobes, to minimize inclusion of the encodings of the printed stimulus word). To obtain a map of the clusters of stable voxels that characterized each group, a hit map was computed for the ideator group and the control group, such that only the voxels with a contribution of at least 4 (of 17) participants were considered. The choice of starting with a large number of stable voxels (1000) and a hit-map threshold of 4 participants was motivated by the goal of obtaining 10-15 spatial clusters in the hit maps, on the assumption that a small number of activation regions (less than 25) characterized the relevant components of the neural representations of the stimulus concepts, as they have in several previous studies [4–6].

The voxels in the hit map were spatially clustered, and only the clusters containing at least 5 voxels were included in the stability map of each group. The clustering algorithm was spatial clustering (as implemented in the spm_cluster function of SPM8). The minimum cluster size is consistent with the cluster sizes used in the previous studies [4,5]. Large clusters (with radii > 11 mm) were subdivided into smaller clusters by finding the within-cluster local maxima of the number of hits and/or splitting the cluster into two at the midpoint of its longest axis (x, y, or z in voxel space), such that the resulting clusters contained all of the voxels of the original (large) cluster. The rationale for subdividing large clusters is that very large clusters have a higher probability of containing voxels with dissimilar functions. The resulting clusters of stable voxels for both groups are shown in Figure 1.

To determine which of these clusters would be useful in the group classification, the most discriminating clusters were identified using a stepwise procedure described below. Five such clusters were identified, as shown in Figure 2. These clusters were populated by stable voxels from one or both groups. During the group membership classification, the locations of the five discriminating clusters were recomputed to exclude the data of the left out participant (in order that the test participant's data be excluded from the training of the classifier). The number of starting voxels was increased in the cross-validation folds from 1000 to 1200 with the goal of obtaining 10-15 clusters, as above, when the number of participants was reduced by the leaving-out of the test participant's data.

Classifier features: Activation levels of stable voxels in the discriminating locations in the neural representations of discriminating concepts. For the suicidal ideator-control group membership classification, each participant was characterized by a vector of activation levels (assessed in a set of discriminating brain locations, described below) for each of the discriminating concepts (identified by the procedure described below). In the case of the main classification, in which there were 6 discriminating concepts and 5 discriminating locations, the resulting 30-element vector consisted of 30 activation levels

(6 concepts times 5 locations). Within each location (characterized as a cluster), the activation measure was computed as the mean of 5 of the most representative (most stable) voxels within the cluster.

The 34 vectors of activation levels (each describing the activation of one of the 34 participants) were used in cross-validated classification (leaving out the vector for one participant and training the classifier on the data from the other 33 participants). In each of the cross-validation folds, only 50% of the 30 features that were most discriminating for the group membership within the training set (i.e. excluding the participant under test) were used (evaluated by a between-group t-test within the training set). The accuracies obtained in these cross-validation folds were then averaged. Note that the data of the participant that was being classified in a fold was always excluded from all aspects of the classification, such as the determination of the clusters of stable voxels that were computed separately for each fold.

For the group membership classification of suicidal ideators with prior suicide attempt versus those without attempt, the vectors characterizing the participant's activation were obtained similarly, but using only 3 concepts and 3 locations identified as discriminating, resulting in 9-element vectors. For this classification, all 9 features were used in cross-validation folds. The cross-validated mean accuracy of membership classification was computed following a similar procedure.

Identifying the most discriminating concepts and locations. To identify the most discriminating concepts, a reiterative procedure analogous to stepwise regression was performed. In the first iteration, the group classification was performed using only one concept at a time, determining which single concept of the 30 resulted in the highest classification accuracy. In the second iteration, the classification was performed using pairs of concepts, namely the single concept that produced the highest accuracy in the first iteration as well as each of the 29 other concepts. All pairs that produced at least as high an accuracy as achieved on the previous iteration, were explored in the third iteration, where triplets of concepts were used, namely the pairs that produced the highest accuracy in the previous iteration, plus each of the remaining 28 concepts. Such stepwise addition of discriminating concepts continued until adding any one of the remaining concepts resulted in a decrease in accuracy. An analogous procedure identified the most discriminating locations. The search for discriminating concepts followed the search for discriminating locations.

Controlling for group differences in self-report clinical measures. The ideator and control groups differed with respect to other measures besides suicidal ideation (Spielberger Anxiety/State, PHQ, CTQ, and ASR (total problems), as reported in Table 1. To control for these differences in the process of performing the group membership classification, the following method, developed by other researchers[7,8] was applied. This method estimates the effects in the control group of the (nuisance) variables on the variables to be used in the classification (an impact uninfluenced by suicidality) using multiple regression, and then removes these effects from both participant groups' classifier feature data[7]. The obtained regression coefficients were applied to the data of both groups, and the residuals were used as features to perform the group membership classification as described above. The group membership classification accuracy increased slightly (from .91 to .94) as a result of this correction.

Classification based on emotional content of neural signatures. The discriminating concepts' activation was represented as a weighted sum of activation vectors characterizing four emotions. The emotion activation signatures were obtained from a study of emotions in neurotypical participants[6]. Activation corresponding to *sadness, shame, anger* and *pride* in the most discriminating locations (defined in the data from the current study) was averaged across the 10 participants (in the emotions study), with each of the locations being characterized by the mean activation level of the 5 most stable voxels. This decomposition resulted in a set of emotion weights for each discriminating concept for each participant. The table showing the four emotion regression weights for the six discriminating concepts for each participant is available for ***download.*** In addition, that table contains the correlations between the signatures of all 9 emotions examined in Kassam et al.[6] and the six discriminating concepts.

Each participant was characterized by a vector consisting of emotion weights, and similarly to the machine learning procedure described above, the participants' group membership was identified by cross-validated classification. This approach was used both to distinguish participants with suicidal ideation from controls and, within the ideators group, to distinguish participants with prior suicide attempt from those without such attempts.

Classification of 30 concepts. The concept classification accuracy, used to assess data quality, was computed separately for each participant, similarly to previous studies [4]. The cross-validation procedure included training the GNB classifier on data from any 4 presentations (selecting the 120 most stable in the training presentations voxels) and identifying the neural signatures of the 30 concepts (using data averaged over the two left-out presentations). The mean normalized rank accuracy of this classification served as a measure of participant's data quality. This classification is independent of the group membership classifications that are the main focus of this study.

The Gaussian Naïve Bayes (GNB) classifier. In this study, GNB classifiers were used primarily to identify the group membership of participants. The classifier's features differed for the several different reported classifications but the general approach was the same.

Each participant was characterized by a set of activation levels for a number of concepts in several brain locations. For example, for classification of 17 participants with suicidal ideation and 17 control participants, performed in 34 "leave one participant out" cross-validation cycles (folds), on each fold a subset of the data (33 participants' data) was used to train a classifier to associate fMRI data patterns with the group label of each participant. A classifier is a mapping function *f* of the form:

$$f\text{: } \textit{cluster activation levels} \rightarrow Y_i,\ i=1,\ldots,\mathrm{m},$$

where Y_i were the two groups (suicidal ideators or controls), and where the *cluster activation levels* were the mean activation levels of the selected voxels in topographically-specified clusters (brain locations). The classifier used here was a Gaussian Naïve Bayes (GNB)-pooled variance classifier. (We make no claim of superiority for GNB-pooled over other possible classifiers.) GNB is a generative classifier that models the joint distribution of class Y and attributes and assumes the attributes $X_1,\ldots,X_n$ are conditionally independent given Y. The classification rule is:

$$Y \leftarrow \arg\max_{y_i} \mathrm{P}(Y = y_i) \prod_j P(X_j \mid Y = y_i)$$

where $P(X \mid Y = y_i)$ is modeled as a Gaussian distribution whose mean and variance are estimated from the training data. In GNB-pooled variance, the variance of attribute X_j is assumed to be the same for all classes. This single variance is estimated by the sample variance of the pooled data for X_j taken from all classes (with the class mean subtracted from each value).

On each fold, the trained classifier was tested on the data of the left-out participant. This procedure was reiterated for all 34 possible ways of leaving out one participant, yielding 34 classifications whose averaged accuracies are reported.

In the more general case, the ***rank accuracy*** (hereafter, simply *accuracy*) of the classification is the normalized rank of the correct label in the classifier's posterior-probability-ordered list of classes. If the classifier were operating at chance, the correct label would on average appear in the middle of the ranked list, producing a chance level accuracy of .50. Accuracies are calculated for each item in each fold and then averaged across folds, and then across items. Significance levels are obtained using random permutation testing (for the 30-class classification). In the case of classifying membership in two groups, simple accuracy is used and a binomial distribution is used to assess significance levels.

Supplementary Table

Supplementary Table 1. Demographic information and clinical variables for the full participant's groups

	Participants			
Measure	Suicidal Ideators (n = 38)	Controls (n = 41)	Test Statistic (df)	p-value
Gender (Male:Female)	8:30	18:23	χ^2 (1) = 4.58	0.032
Mean Age	24.74 (5.66)	22.02 (2.88)	t(77) = 2.72	0.008
WASI IQ	118 (14.28)	120.54 (9.50)	t(77) = 0.94	0.352
ASIQ	57.42 (30.54)	2.15 (4.68)	t(77) = 11.45	0.000
PHQ	12.16 (6.17)	0.49 (1.16)	t(77) = 11.9	0.000
Spielberger/Anxiety State	41.18 (6.16)	47.10 (4.41)	t(77) = 4.94	0.000
Spielberger/Anxiety Trait	48.92 (5.52)	46.68 (3.54)	t(77) = 2.16	0.034
CTQ	50.03 (19.11)	31.20 (7.43)	t(77) = 5.85	0.000
ASR internalizing problems	35.63 (10.98)	6.93 (6.17)	t(77) = 14.47	0.000
ASR externalizing problems	16.21 (8.66)	5.73 (4.69)	t(77) = 6.76	0.000
ASR total problems	87.76 (25.26)	24.27 (16.84)	t(77) = 13.23	0.000
Number of Attempts	1.24 (1.63)			
SIS	8.97 (9.11)			

Abbreviations:

WASI IQ: Wechsler Abbreviated Scale of Intelligence;

ASIQ: Adult Suicide Ideation Questionnaire;

PHQ: Patient Health Questionnaire;

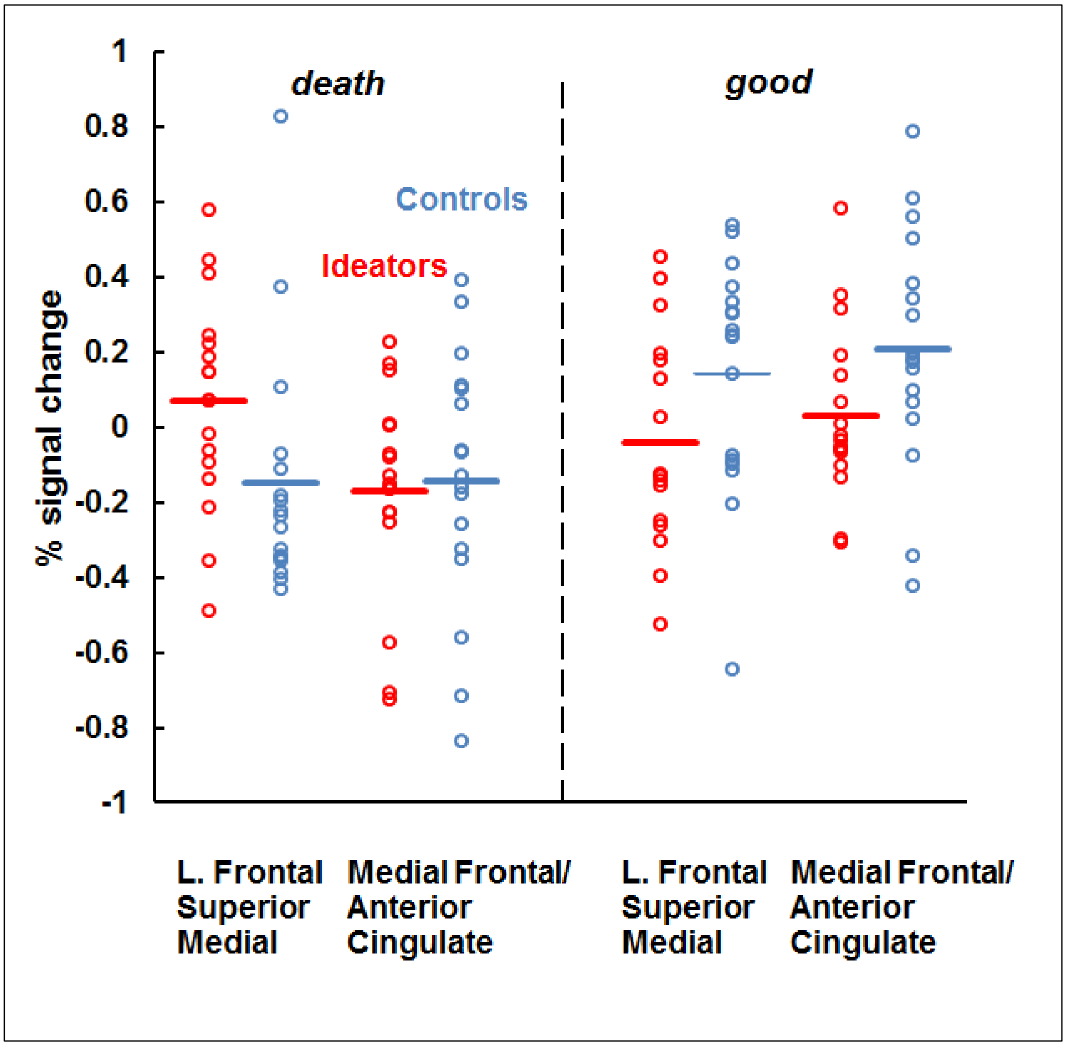
CTQ: Child Trauma Questionnaire;

ASR: Adult Self Report;

SIS: Suicidal Ideation Scale

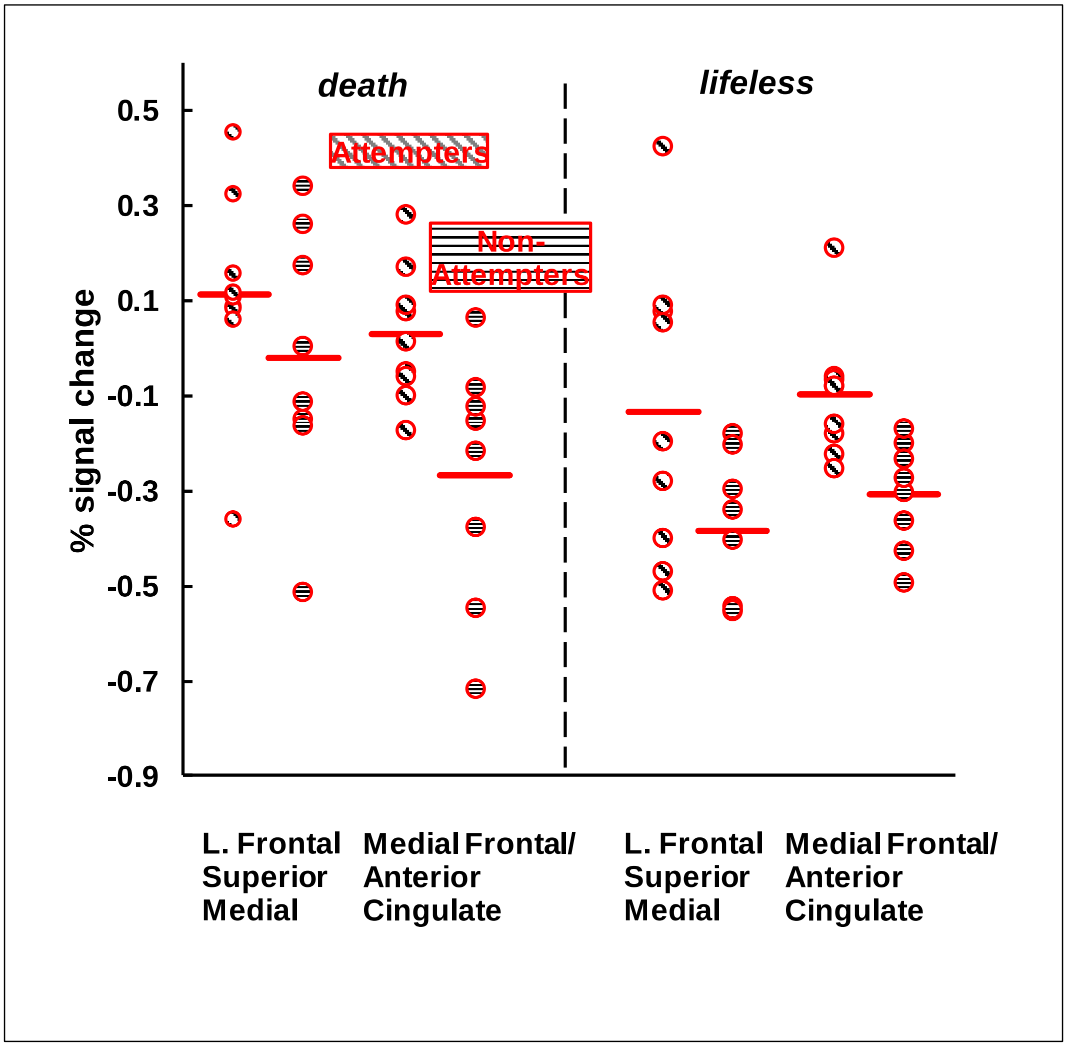
Standard deviations are shown in parenthesis.

Supplementary Figures



Supplementary Figure 1. Distributions of activation levels for 17 ideators and 17 controls for two concepts in two locations.

The individual participants' activation levels for the two discriminating concepts, *death* and *good*, in two discriminating brain locations (L. Frontal Superior Medial and Medial Frontal/Anterior Cingulate). Horizontal lines indicate group means. The figure illustrates some of the features on which the classification is based. The features collectively enable the discrimination between groups. The figure also illustrates that the group difference in activation at a given location may be in different directions for different concepts, showing that the group difference is not due to a consistent hypoactivation nor hyperactivation at a given brain location, but to a difference in how particular concepts are represented at that location.



Supplementary Figure 2. Distributions of activation levels for 9 ideators with a suicide attempt and 8 ideators without such an attempt for two concepts in two locations.

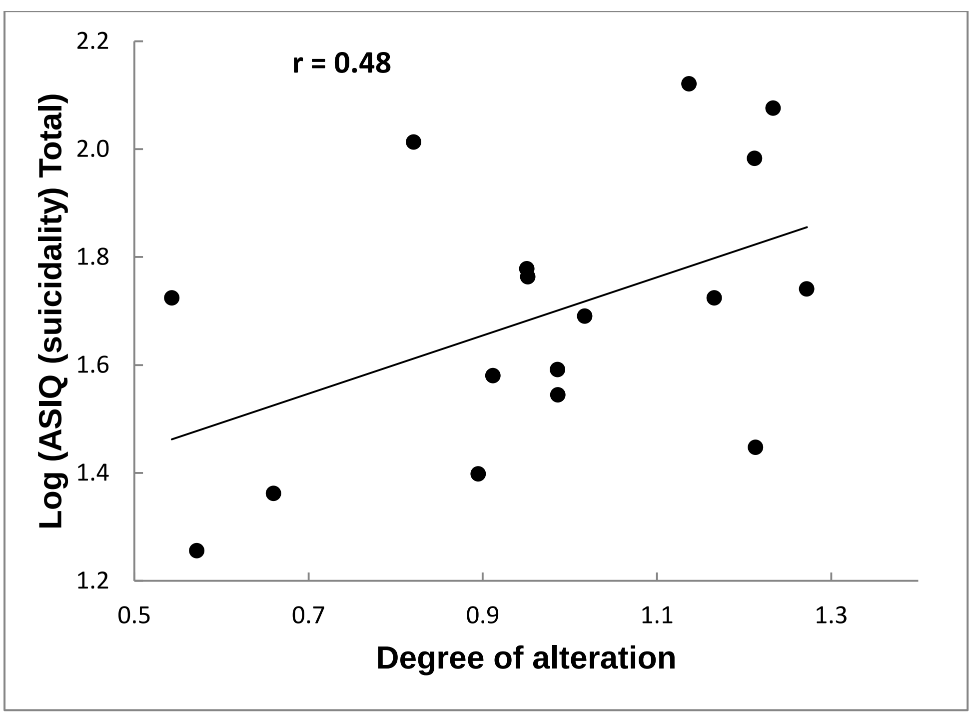
The individual participants' activation levels for the two discriminating concepts, *death* and *lifeless*, in two discriminating brain locations (L. Frontal Superior Medial and Medial Frontal/Anterior Cingulate). Horizontal lines indicate group means.

Supplementary Figure 3. Distributions of emotion regression weights for 17 ideators and 17 controls for two emotions, *sadness* and *shame*, for the two discriminating concepts (*death* and *good*).

The discriminating concepts' activation patterns were modeled with regression as a linear combination of four emotion signatures. Shown here are the resulting regression weights for the emotions *sadness* and *shame* in the modeling of the concepts *death* and *good*. Horizontal lines indicate group means.

Supplementary Notes

Correlations between neural alterations of concept representations and self-report measures of suicidal ideation. The degree of neural alteration of concepts in individual suicidal ideators can be quantitatively assessed and related to the self-reported measure of suicidal ideation. The neural representation here for each suicidal ideator participant was the vector of activation levels for the six most distinguishing concepts in the three most distinguishing brain regions (namely the control group locations shown in Table 3). The neurotypical norm to which this measure was compared was the mean of the corresponding vectors averaged across the control participants. The measure of alteration for each suicidal ideator was the *distance from this norm* (computed as one minus the correlation between the control group mean vector and the suicidal ideator participant's vector). There was a marginally reliable correlation ($r = 0.48$, $p < 0.051$) between the degree of concept alteration and the log-transformed self-reported ASIQ measure of suicidality, as shown in Supplementary Figure 4.



Supplementary Figure 4. Correlation between degree of alteration of discriminating concepts and log (ASIQ) self-report of suicidal ideation in 17 ideators

Supplementary References

1. Borges, G., Nock, M. K., Haro Abad, J. M., Hwang, I., Sampson, N. A., Alonso, J., Andrade, L. H., Angermeyer, M. C., Beautrais, A., Bromet, E., Bruffaerts, R., de Girolamo, G., Florescu, S., Gureje, O., Hu, C., Karam, E. G., Kovess-Masfety, V., Lee, S., Levinson, D., Medina-Mora, M. E., Ormel, J., Posada-Villa, J., Sagar, R., Tomov, T., Uda, H., Williams, D. R. & Kessler, R. C. Twelve-month prevalence of and risk factors for suicide attempts in the world health organization world mental health surveys. *J. Clin. Psychiatry* **71,** 1617–1628 (2010).
2. Kroenke, K., Spitzer, R. L., Williams, J. B. W. & Löwe, B. The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *Gen. Hosp. Psychiatry* **32,** 345–359 (2010).
3. Reynolds, W. M. *Suicidal Ideation Questionnaire (SIQ)*. (Psychological Assessment Resources, 1987).
4. Just, M. A., Cherkassky, V. L., Aryal, S. & Mitchell, T. M. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS One* **5,** e8622 (2010).
5. Just, M. A., Cherkassky, V. L., Buchweitz, A., Keller, T. A. & Mitchell, T. M. Identifying autism from neural representations of social interactions: Neurocognitive markers of autism. *PLoS One* **9,** e113879 (2014).
6. Kassam, K. S., Markey, A. R., Cherkassky, V. L., Loewenstein, G. & Just, M. A. Identifying emotions on the basis of neural activation. *PLoS One* **8,** e66032 (2013).
7. Dukart, J., Schroeter, M. L. & Mueller, K. Age correction in dementia – matching to a healthy brain. *PLoS One* **6,** e22193 (2011).
8. Koikkalainen, J., Pölönen, H., Mattila, J., van Gils, M., Soininen, H. & Lötjönen, J. Improved classification of Alzheimer's Disease data via removal of nuisance variability. *PLoS One* **7,** e31112 (2012).