nature
human behaviour

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

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 McMakin[3], Christine Cha[4], Matthew K. Nock[5] and David Brent[2]

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[2]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA
[3]Department of Psychology, Florida International University, Miami, FL
[4]Clinical Psychology Department, Columbia University, New York, NY
[5]Department of Psychology, Harvard University, Cambridge, MA

The clinical assessment of suicidal risk would be significantly complemented by a biologically-based measure that assesses alterations in the neural representations of concepts related to death and life in people who engage in suicidal ideation. This study used machine-learning algorithms (Gaussian Naïve Bayes) to identify such individuals (17 suicidal ideators vs 17 controls) with high (91%) accuracy, based on their altered fMRI neural signatures of death and life-related concepts. The most discriminating concepts were ***death, cruelty, trouble, carefree, good,*** **and *praise*. A similar classification accurately (94%) discriminated 9 suicidal ideators who had made a suicide attempt from 8 who had not. Moreover, a major facet of the concept alterations was the evoked emotion, whose neural signature served as an alternative basis for accurate (85%) group classification. The study establishes a biological, neurocognitive basis for altered concept representations in participants with suicidal ideation, which enables highly accurate group membership classification.**

The assessment of suicide risk is among the most challenging problems facing mental health clinicians. The challenge is enormous because suicide is the second-leading cause of death among young adults [1] and at the same time, clinicians' predictions and patients' own predictions of their future suicide risk have been shown to be relatively poor predictors of future suicide attempt [2,3]. In addition, suicidal patients may disguise their suicidal intent as part of their suicidal planning or to avoid more restrictive care. Nearly 80% of patients who die by suicide deny suicidal ideation in their last contact with a mental healthcare professional [4]. This status identifies a compelling need to develop markers of suicide risk that do not rely on self-report. Biologically-based markers of altered conceptual representations have the potential to complement and improve the accuracy of clinical risk assessment [5,6].

In this study, we offer a new approach to the assessment of suicide risk that uses machine-learning detection of neural signatures of concepts that have been altered in suicidal individuals. This approach capitalizes on recent advances in cognitive neuroscience that use machine learning techniques to identify individual concepts from their functional magnetic resonance imaging (fMRI) signatures [7–9]. These fMRI signatures are common and reproducible across neurotypical individuals. Moreover, the signatures can be decomposed into meaningful components. For example, the concept of *spoon* includes a neural representation of the way it is manipulated (located in motor-related regions), as well as its role in eating (represented in gustatory areas such as insula and inferior frontal gyrus) [7]. By contrast, *house* is represented in regions related to shelter and physical setting or location (parahippocampal and parietal areas) [7]. This approach has previously been used to detect altered representations in a special population, enabling the discrimination between 17 participants with high functioning autism and 17 matched neurotypical individuals with 97% accuracy, based on their neural representations of 16 social interactions (such as to *hate* or *hug*) [10].

The current study applies this approach to determine whether the neural representations of positive, negative, and suicide-related concepts are altered in a group of participants with suicidal ideation, relative to a control group. If so, are the alterations sufficiently systematic to enable an individual participant to be accurately classified as a suicidal ideator versus a neurotypical control participant? The study also investigates whether *among* participants with suicidal ideation there is a classifiable difference between those who have attempted suicide and those who have not. Furthermore, the neural signature of the test concepts was treated as decomposable biomarker of thought processes that can be used to pinpoint particular components of the alteration. This decomposition attempts to specify a particular component of the neural signature that is altered, namely the emotion component, as described in more detail below.

Two lines of evidence within the suicide literature motivate the application of this approach to suicidal individuals. First, suicidal patients have demonstrated sensitivity to distinct concept alterations through their performance on behavioral measures. One of these measures is an adapted Emotional

Stroop Task that assesses reaction times in response to suicide-related words relative to neutral words [11]; another measure is an adapted Implicit Association Test that assesses reaction times in response to pairing suicide-related words and self-related words [3]. These studies indicate that people with a history of suicide attempts may represent certain concepts or concept pairs differently than non-attempters. Neural markers of these behavioral patterns have never been tested.

Building on these previous studies, the current investigation utilizes machine-learning multivoxel analysis that seeks a *pattern* of activation values (in a set of voxels distributed across a set of brain locations) that is associated with individual stimulus concepts, and which can identify an individual as suicidal or not.

Beyond detecting altered neural signatures of concepts, in the present study we also aimed to detect the emotion component of the neural signatures. To detect these emotion components, we drew on an archive of previously acquired identifiable neural signatures from neurotypical participants [8]. The archive contains nine different types of emotion such as sadness or shame. In the analysis of the current study, we searched for the presence of four of the archived emotion signatures that have previously been detected among suicidal individuals [12–18]: *sadness, shame, anger* and *pride*. We hypothesized that the groups would differ in the degree of presence of these emotion signatures in the neural representations of concepts such as *death*. (We assume that the *quality* of the emotions is similar between neurotypical and suicidal participants (e.g. anger, when it occurs, is similar). The ability to classify individual participants with respect to suicidal risk and to relate their altered activation patterns to altered emotional content associated with specific concepts would provide an interpretable, personalized profile for diagnosis and therapy.

In summary, we test three main hypotheses.

1. Participants with suicidal ideation will differ from non-suicidal control participants with regard to their neural representations of death- and suicide-related concepts, to a degree that a machine-learning classifier can accurately determine whether a participant is a member of the suicidal ideation group or the control group.

2. A similar machine-learning approach will accurately discriminate those members of the suicidal ideator group who have made an attempt at suicide from those who have not.

3. The neural signatures of discriminating concepts in suicidal ideators will contain different emotion component signatures (i.e. have different regression weights in a linear model) than the control group, and these group differences will allow a machine-learning classifier to accurately determine whether a participant is a member of the suicidal ideation group or the control group.

Results

The main neurosemantic analyses were performed on two groups of participants: 17 suicidal ideators and 17 healthy controls. The groups were balanced on gender ratio, age, and WASI IQ (Table 1). The stimuli were 30 concepts (as shown in Table 2) each presented for 3 sec, related to either suicide, positive affect, or negative affect. The brain locations that contain the main components of the neural representations of the 30 concepts, identified by the presence of stable voxels (those whose responses to the set of stimuli were similar over multiple presentations), are shown in Figure 1 (see Methods). Six of the concepts and five of the brain locations (shown in Figure 2) provided the most accurate discrimination between the two groups.

There were interpretable, clinically meaningful differences between the individuals in the suicidal ideator and control groups, and within the suicidal ideator group, there were differences between the attempters and non-attempters. The classification procedures identified the *concepts and brain locations* that were most predictive of the group membership for these two sets of contrasts (i.e. suicidal ideator vs. control and attempter ideator vs. non-attempter ideator).

Neurosemantic classification of suicidal ideator versus control group. A Gaussian Naïve Bayes (GNB) classifier trained on the data of 33 participants (leaving one out) predicted the group membership of the left-out participant with a high accuracy of 0.91, ($p < 0.000001$), correctly identifying 15 of the 17 suicidal participants and 16 of the 17 controls (Sensitivity = 0.88, Specificity = 0.94, PPV = 0.94, NPV = 0.89).

The classifier's features were the neural representations of the 6 most discriminating concepts (as described in more detail in Methods). The neural representation of each concept, as used by the classifier, consisted of the mean activation level of the 5 most stable voxels in each of the five most discriminating locations.

The concepts that most strongly discriminated the groups were *death, cruelty, trouble, carefree, good,* and *praise*. The most discriminating brain regions included the L. superior medial frontal area, medial frontal/anterior cingulate, R. middle temporal, L. inferior parietal, and L. inferior frontal (as shown in Figure 2 and Table 3). All of these regions, especially the first two, have repeatedly been strongly associated with *self*-referential thought (consistent with the behavioral findings in suicidal patients reported by [3]). The separation between the ideator and control groups in the multidimensional scaling of the activation features used by the classifier is shown in Figure 3. The distributions of the activation levels in two locations for the 17 ideator participants and 17 controls for the concepts *death* and *good* are shown in Supplementary Figure 1.

To determine how many and which concepts were most discriminating between ideators and controls, a reiterative procedure analogous to stepwise regression was used, finding the next most discriminating concept at each step. The procedure is further described in Supplementary Information.

This procedure identified *death* as the most discriminating single concept, and the concepts that followed in descending order of discriminating ability were *carefree, good,* and *cruelty*, and these were followed by *praise* and *trouble*. To determine how many and which brain locations were most discriminating between the ideators and controls, a similar stepwise procedure was used.

Because the ideator and control groups differed with respect to other measures besides suicidal ideation, it is useful to demonstrate that the high classification accuracy remains intact after statistically controlling for such differences (namely differences in Spielberger Anxiety/State, PHQ, CTQ, and ASR). When these differences were statistically controlled for (using methods described by [19,20] – see Supplementary Information for details), the classification accuracy slightly increased (from .91 to .94) (Sensitivity = 0.88, Specificity = 1, PPV = 1, NPV = 0.94), indicating the applicability of the model to groups that differ with respect to these clinical variables beyond suicidal ideation.

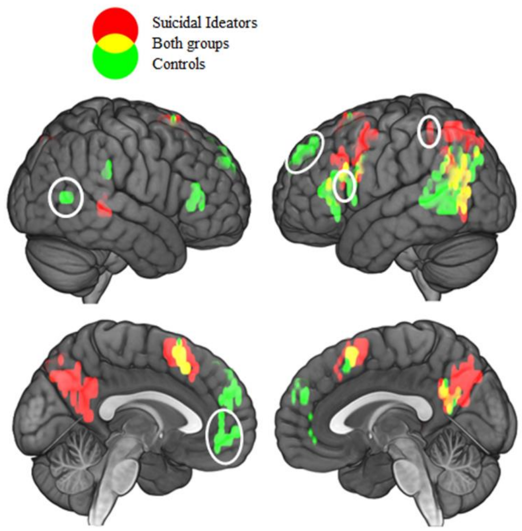


Figure 1. Clusters of stable voxels of the suicidal ideator group and the control group. White ellipses indicate the 5 discriminating locations.

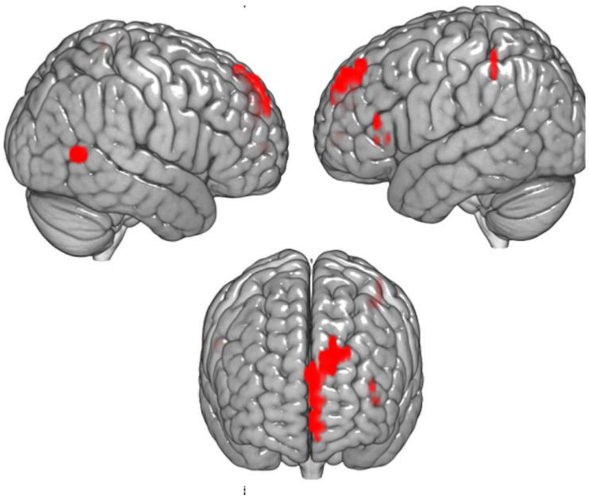


Figure 2. Discriminating brain locations for distinguishing suicidal ideator versus control group membership.

Table 1. Demographic information and clinical variables

Participants

Measure	Suicidal Ideators (n =17)	Controls (n = 17)	Test Statistic (df)	p-value
Gender (Male:Female)	5:12	3:14	χ^2 (1) = 0.63	0.42
Mean Age	22.88	22.06	t(32) = 0.74	0.46
	(3.57)	(2.84)		
WASI[1] IQ	124.1 (10.86)	121.12	t(32) = 0.82	0.420
		(9.70)		
ASIQ[2]	57.88 (34.38)	2.76	t(32) = 6.5	0.000
		(6.35)		
PHQ-9[3]	12.24	0.47	t(32) = 7.14	0.000
	(6.7)	(1.1)		
Spielberger/Anxiety State	40.12	46.88	t(32) = 3.59	0.001
	(6.14)	(4.77)		
Spielberger/Anxiety Trait	47.59	45.88	t(32) = 1.34	0.19
	(4.14)	(3.22)		
CTQ[4]	41.3	30.24	t(32) = 3.62	0.001
	(9.65)	(8.11)		
ASR[5] internalizing problems	35.6	5.9	t(32) = 9.46	0.000
	(11.9)	(5.0)		
ASR externalizing problems	13.9	4.8	t(32) = 3.60	0.001
	(9.8)	(3.5)		
ASR total problems	83.1	19.65	t(32) = 8.74	0.000
	(27.09)	(12.65)		
Number of Attempts	1.41 (2.0)			
SIS[6]	8.19 (9.06)			

Standard deviations are shown in parentheses.
[1] Wechsler Abbreviated Scale of Intelligence;
[2] Adult Suicide Ideation Questionnaire;
[3] Patient Health Questionnaire;
[4] Child Trauma Questionnaire;
[5] Adult Self Report;
[6] Suicidal Ideation Scale

Table 2. Stimulus concepts

Suicide	Positive	Negative
apathy	bliss	boredom
death	carefree	criticism
desperate	comfort	cruelty
distressed	excellent	evil
fatal	good	gloom
funeral	innocent	guilty
hopeless	kindness	inferior
lifeless	praise	terrible
overdose	superior	trouble
suicide	vitality	worried

Table 3. Cluster locations predictive for suicidal ideator-control group membership classification

Brain region	MNI centroid coordinates			Radius (mm)
	x	y	z	
Suicidal ideator group				
L. inferior parietal (LIPS)	-42	-43	50	5.0
LIFG triangularis	-42	29	8	5.1
Control group				
L. superior medial frontal	-11	52	33	10.5
Medial frontal/Anterior cingulate	-6	50	-3	8.3
R. middle temporal	56	-62	10	2.5

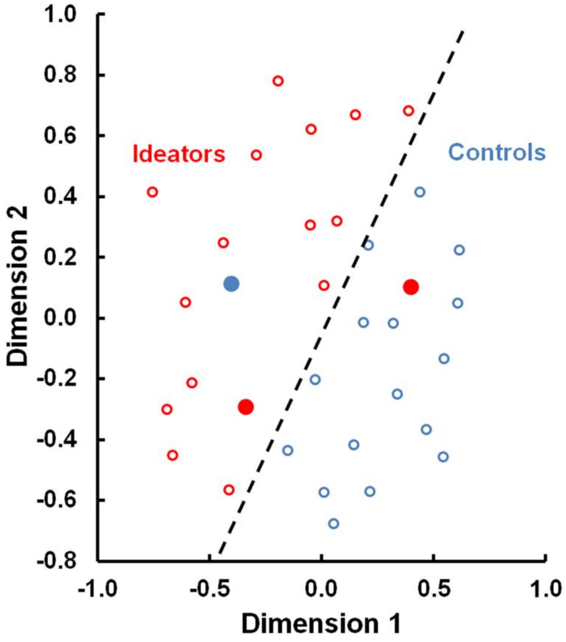


Figure 3. Group separation in the multidimensional scaling of the activation features of the 34 participants' (17 ideators and 17 controls) used by the classifier.

Ideators are indicated by red circles, controls by blue circles. Filled circles indicate misclassifications. The scaled features (activation levels in 5 brain locations for 6 discriminating words) were computed in 32 cross-validation folds, averaged across the folds. The dashed line shows the separability of the two groups in this 2D space.

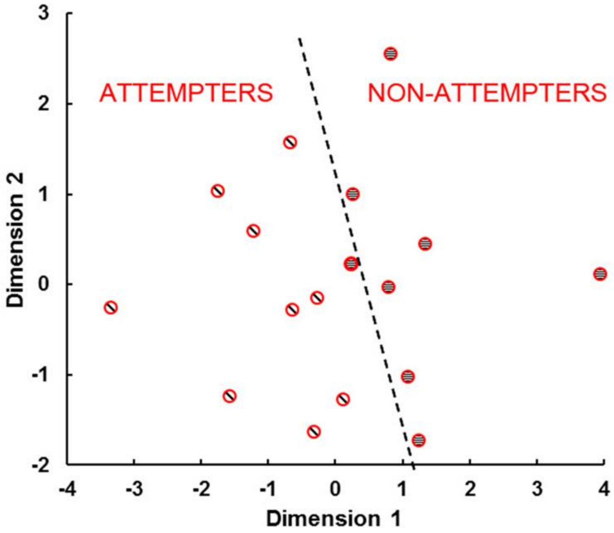


Figure 4. Group separation in the multidimensional scaling of the activation features of the 9 ideators with suicide attempts (diagonally filled circles) and the 8 ideators without attempts (horizontally filled circles) used by the classifier.

The features (activation levels in 3 brain locations for 3 discriminating words) were scaled in 2 dimensions. The dashed line shows the separability of the two groups in this 2D space.

An additional quantitative assessment of the generalizability of the model applied a more conservative cross-validation technique. Instead of training the model on data from all but one participant, this additional assessment left out the data of half of the participants (8 of 17) from each group for testing, and the model was trained on the remaining 9 participants' data. (Because there are a huge number of ways to leave out half of the participants from each group, 1000 random selections of such partitionings were performed and the outcomes were averaged). The result was that the classification accuracy remained at a highly reliable level of .76, showing that a model based on a much smaller sample of the participants generalizes to the other half. This constitutes an added test of model's generalizability.

Neurosemantic classification of suicidal ideators who have made an attempt vs. ideators who have not. Another classifier was able to distinguish, within the group of 17 suicidal ideator participants, those who had previously made an attempt (9 participants) from those who had not (8 participants). This classification resulted in a high accuracy of .94 (16 out of 17 correct, one non-attempter misclassified, p < 0.0002, Sensitivity = 1.0, Specificity = 0.88, PPV = 0.90, NPV = 1.0). The concepts that best discriminated between attempters and non-attempters were *death, lifeless,* and *carefree*. The most discriminating brain regions for this classification were a subset of the ones above that discriminated ideators from controls, namely L. superior medial frontal, medial frontal/anterior cingulate, and R. middle temporal. The most discriminating concepts and locations were obtained using the same stepwise reiterative procedure (described in Supplementary Information) that was used in the ideator-control classification. The separation between the attempter and non-attempter groups in the multidimensional scaling of the activation features used by the classifier is shown in Figure 4. The distributions of the activation levels in two locations for the 9 ideators with a suicide attempt and 8 ideators without such an attempt for the concepts *death* and *lifeless* are shown in Supplementary Figure 2.

Alterations in the emotional content of the neural representations of the discriminating concepts. Neurosemantic signature measures are interpretable activation patterns that contain information about the thought processes to which they correspond. This makes it possible to analyze the *psychological nature* of an alteration of a given concept in a clinical population. In the case of suicidal ideation, we postulated that the emotional content of the neural representations of the discriminating words would differentiate between the suicidal and control groups, consistent with previous behavioral findings [11].

In the analysis of the current results, we searched for the presence of four previously-acquired emotion signatures (*sadness, shame, anger,* and *pride)*[8] within the neural representations of the six concepts that best discriminated ideator and control groups. The reason for using only 4 of the 9 emotions for which signatures existed was that a model with all 9 emotions (*sadness, shame, anger, pride, disgust, envy, fear, lust, and happiness*) would overfit the data (activation levels in 5 most discriminating locations). The main rationale for choosing this particular set of four emotions (i.e. *sadness shame anger* and *pride*) is that it resulted in the highest classification accuracy of the two groups. Furthermore, most of these four emotions have been implicated as precursors and motives for suicidal behavior. Interpersonal discord (i.e., *anger*) and embarrassment are two prominent motivations for adolescent suicide attempts [21]. *Shame* is prominent in studies of male suicide attempters [22]. In a content analysis of over 1200 suicide notes, sadness (e.g., hopelessness, sorrow), anger (e.g., anger, blame), and guilt were particularly prominent, although positive emotions that expressed relief either on the part of the suicide victim or the intended recipient of the note were common [23]. However, note that our neurosemantic tests here probe for the emotional content in the representation of *particular* concepts (such as *death*), not for an enduring emotional trait.

The neurosemantic signature of each of the six discriminating concepts was modeled as a linear combination of sadness, shame, anger and pride, with the expectation that there would be group differences in the regression weights of the emotions. Consistent with this expectation, in the suicidal group, the concept of *death* reliably ($t(32)$=2.67, p < .012) evoked more (had a higher regression weight for) shame whereas the concept of *trouble* evoked reliably more sadness in this group ($t(32)$ = 2.24, p < .032). (These t tests are uncorrected for multiple comparisons, to provide an initial overview). *Trouble* also evoked reliably *less* anger ($t(32)$ = 2.78, p < .01) and *carefree* evoked less pride ($t(32)$ = 2.96, p < .006) in the suicide ideation group. In general, the negatively-valenced discriminating concepts evoked more sadness and shame but less anger in the suicidal ideation group.

In ideators who had made an attempt, the suicide-related concept *death*, evoked reliably less sadness ($t(15)$ = 2.91, p < .01) than in those who had not made an attempt, and the other suicide-related concept *lifeless* evoked reliably more anger ($t(15)$ = 3.58, p < .003) than in those ideators who had not made an attempt. Furthermore, in the ideators who had made an attempt, the positive concept *carefree* evoked reliably less anger ($t(15)$ = 2.34, p < .03).

These results are generally consistent with previous fMRI findings of altered emotion processing at the neural level (in response to face stimuli) in suicidal participants [24]. To more systematically assess the emotion signature group differences, the emotion signature weights were used as features of a classifier that attempted to identify group membership.

Identifying group membership on the basis of emotion signature differences in the distinguishing concepts. We investigated whether the emotional content of the neural signature of a concept could indicate whether a given participant was an ideator or a control participant, or, within ideators, whether they had made an attempt or not. The features that were used in this classification were the regression coefficients in the model above, indicating the degree of presence of each of the emotion signatures in their neural representation of each discriminating concept (e.g. how much shame was present in a participant's neural representation of *death*).

The GNB classifier correctly identified the group membership (ideator or control) of the 34 participants with 0.85 accuracy (14 ideators and 15 controls correctly identified, Sensitivity = 0.82, Specificity = 0.88, PPV = 0.88, NPV = 0.83). (Using the regression weights of only two of the emotions (pride and shame) resulted in the same classification accuracy (0.85) as using all four emotions). The distributions of emotion regression weights of *sadness* and *shame* in the representations of *death* and *good* for the 17 ideator participants and 17 controls are shown in Supplementary Figure 3.

The same approach of using emotion regression coefficients as features was applied to distinguishing the nine ideators who had made an attempts versus the eight who had not in the set of 17 ideators. Using the regression coefficients of the emotions of the three concepts that best discriminated attempters from non-attempters (*death*, *lifeless*, and *carefree*) as classifier features, it was possible to identify the group membership of the 17 participants as attempters or non-attempters with 0.88 accuracy (eight attempters and seven non-attempters were identified correctly; Sensitivity = 0.89, Specificity = 0.88, PPV = 0.89, NPV = 0.88). (As in the classification above, it was possible to achieve comparable accuracy using only a subset of the predictor variables.)

Thus the alterations of the neural signatures of the discriminating concepts in the ideator group and within that group, in the attempter sub-group, can be meaningfully attributed in large part to their evoking a different profile of specific emotions than in the comparison group. (These two classification accuracies based on the emotion signature weights, .85 and .88, were only slightly lower than the classification accuracies based directly on the activation data, .91 and .94.). This result indicates the emotional content is a significant way in which concepts are altered in suicidality and in suicidality after attempt, and thus provides potential targets for therapy.

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 Adult Suicidal Ideation Questionnaire (ASIQ) measure of suicidality, as shown in Supplementary Figure 4.

Locations of the neural representations (clusters of stable voxels) for the two groups. There was a substantial similarity in neural representation of 30 concepts between the two groups in terms of the involved brain locations, with one large exception. Only the control group had clusters of stable voxels in anterior frontal regions, namely the superior medial frontal and anterior cingulate areas, whereas the ideator group showed negligible stable activation in these frontal regions, as shown in Figure 1. By contrast, the ideator group had more clusters of stable voxels in the L. inferior parietal region. (Recall that stable voxels are those that have a similar semantic tuning curve across the 30 stimulus concepts in each of the multiple presentations of the stimulus set). These distinguishing brain locations play a substantial role in discriminating between the ideator and control participants on the basis of the neural activation evoked by the discriminating concepts. Notably, the accuracy of identifying which of the 30 stimulus items that the participant was thinking about based on its fMRI signature was similar for the two groups: .71 and .75 for the suicidal ideator and control groups respectively.

GLM univariate analyses of the same groups of participants (17 ideators and 17 controls) as in the main classification failed to show FDR- or family-wise-corrected significant between groups in the activation patterns for all 30 concepts considered together, nor for various subsets of the concepts, such as the six discriminating concepts, nor for any of the three categories of concepts. By contrast, the multivoxel analyses of the patterns corresponding to individual concepts as described above provided excellent group separability.

Testing the classification algorithm on another sample. The data of 21 additional ideator participants, although excluded from the main analyses because of the lower technical quality of their data, were nevertheless available to use as a test of the generalization of the classifier to another sample. The data quality was measured in terms of the low accuracy of classification of the 30 stimulus items (< .60 rank accuracy) and the generally greater head motion parameters (mean maximum = 1.81 mm) than the 17 participants in the main study (mean = 1.27 mm, t (77) = 2.73, p < 0.01). Nevertheless, the classifier developed from the first set of 17 ideators and 17 controls was used, without any modifications, to try to distinguish these 21 suicidal ideators from the 17 control participants with good data quality. As in the main classification, the classifier's features were the neural representations of the 6 most discriminating concepts. The neural representation of each concept consisted of the mean activation level of the 5 most stable voxels in each of the five most discriminating locations. The resulting classification accuracy was 0.87 (p < 0.000002; Sensitivity: 0.81; Specificity: 0.94; PPV: 0.94: NPV: 0.8), replicating the findings from the main analysis. Although high quality data from both the ideator group and the control group may be necessary for model development, once a model is developed, it can accurately classify suicidal participants with lower data quality. Thus, the findings were replicated on a second sample of ideators, supporting the generalizability of the model.

The model also did reasonably well in identifying concept alterations associated with having made an attempt within the excluded 21 suicidal ideators. Those participants who had made an attempt versus those who had not were correctly classified with an accuracy of 0.61 ($p < .04$, 13 of 21 participants correctly classified).

These results indicate that the models developed on the basis of the data of participants with less noise in their data can be successfully applied to participants with more noisy data. However, a model that is developed from the data of either ideator or control participants with noisy data does not discriminate groups well. We attribute such noise to inability to rigorously sustain attention to the task and to maintain head position in the scanner. The implication is that high quality data from both the ideator group and the control group are necessary for model development, but once a model is developed, it can achieve accurate identification of suicidal ideator participants with lower data quality.

By testing the performance of the neurosemantic classifier on the additional larger sample of independent ideator participants beyond those who provided the data for the classification algorithm, we provide a replication within this study, thus strengthening support for the generalizability of the model, which applies to *all* of the recruited participants.

Discussion

The findings from this study provide a biological foundation for altered concept representations in those with suicidal thoughts and recent suicidal behavior. The differences in the neural representations of concepts enable accurate classification of suicidal ideator versus control group membership, as well as suicidal ideator versus suicide attempter – the latter distinction being one that few risk factors are able to make [17]. These two findings show that suicidal ideation and attempt are associated with measurable alterations in the way a person thinks about death, suicide, and other positive and negative concepts. The recently-developed fMRI methods for measuring the neural representation of a concept makes it possible to compare neurotypical to clinical representations of concepts, and within a clinical population, to compare suicidal ideation with and without suicidal behavior.

The specific concepts that were altered in people with suicidal ideation, *death, cruelty, trouble, carefree, good,* and *praise,* include items from all three stimulus categories, one that is suicide-related, two that are negative, and three positive concepts. The valuation of what is important and good in life and what is not appears to be altered in ideators. Our results provide a neurally-based, quantitative measure of this alteration.

Most of the ideators showed high levels of self-reported depression that is characterized by the "cognitive triad," which includes a negative view of self, the world, and the future [25]. Pessimism about the future, or hopelessness, has been shown to be correlated with and predictive of future suicidal behavior above and beyond depression [26,27]. The observed alterations of specific concepts may be reflecting more general cognitive changes of this type.

The differences in the emotion signature components of the altered concepts provide additional information about the nature of the perspective change. As described above, the concept of *death* evoked more shame while the concept of *trouble* evoked more sadness in the suicidal group. *Trouble* also evoked *less* anger in the suicidal ideation group. The positive concept *carefree* evoked less pride in the suicidal ideation group. This pattern of differences in emotional response suggests that the altered perspective in suicidal ideation may reflect a resigned acceptance of a current or future negative state of affairs, manifested by listlessness, defeat, as well as a degree of anhedonia (less pride in *carefree*). This type of neurally-acquired information helps characterize the disorder as well as providing specific targets for intervention.

The altered perspective seems even more clear in the contrast between suicidal ideators who had made an attempt versus those who had not, where the most altered concepts were *death, lifeless,* and *carefree,* which includes two suicide related concepts and one positive concept. The finding of a meaningful difference between ideators with and without a history of a suicide attempt is consistent with previous findings showing differential reaction times in response to suicide-related words relative to neutral words [11], and in response to the paired concepts of *death* and *self* versus *life* and *self* [3]. Furthermore, the emotion signature differences show an interpretable pattern. For example, the suicide-related concept *death* evoked less sadness in the ideators who had made an attempt than in those who had not. The two subgroups of ideators differ in their emotional response to particular concepts.

Those ideators who had made an attempt may have thought of death with less sadness than those ideators who had not, whereas the overall group of ideators experienced more shame than controls when thinking about death. It has been shown that many suicidal ideators vacillate between an attraction to life and attraction to death [28], and that having moral objections to suicide is protective against engaging in a suicidal act even with suicidal ideation [29].

We speculate that for those who are conflicted about engaging in a suicidal act, the thought of facilitating death is shameful, whereas those ideators who have made an attempt show greater attraction to and acceptance of death, and hence less sadness in thinking about it. This perspective is also consistent with decreased anger associated with the concept of *lifeless* in ideators with a history of an attempt.

Neuroimaging studies also provide evidence of emotion alteration associated with suicide risk. fMRI studies have found altered processing of angry faces in suicide attempters, and anger and hostility are strongly related to suicidal behavior [24,30], and hostility is also strongly predictive of suicidal behavior [31,32].

More generally, the ability of a machine learning classifier to make discriminations *within the suicidal ideator group* speaks to the specificity of the neurosemantic assessment approach. The classifier is not simply detecting an abnormality that is likely to be present in many disorders, such as depression. It makes accurate discriminations within the ideator group, distinguishing those who had a previous history of a suicide attempt, and thus are at higher risk for future suicidal behavior. While it is possible that these findings were due to the greater severity of suicidal ideation and depression in past attempters, the specificity of the discriminating concepts, *death* and *suicide* speak to a possible application of the approach in the assessment of imminent suicidal risk. Moreover, we have identified differences in the emotions experienced by those ideators with and without a history of suicide attempt , such as differences in anger in thinking about death that are not likely to be explained merely by differences in depressive symptoms.

There are several types of evidence indicating that the activation pattern (neural signature) for a given emotion truly indexes that emotion. First, the emotion signatures are sufficiently specific to accurately identify which emotion was being experienced in the Kassam et al.[8] study. Second, in a validation check of the emotion manipulation (an instruction to drama student participants to evoke a particular verbally named emotion such as *shame*), a separate condition presented IAPS pictures (International Affective Picture System) depicting *disgust*. The classifier trained on the instruction-evoked activation patterns of the emotions correctly identified the emotion evoked by the *disgust* pictures with .91 rank accuracy, indicating the strong similarity of the *disgust* activation patterns evoked in two very different ways, which speaks to the construct validity of the measure. Third, the neural signatures of the emotions in the Kassam et al.[8] study were similar across participants, such that a classifier trained on the emotion signatures of all but one participant could identify the emotions of the left-out participant with .71 rank accuracy. This finding of the commonality of emotion signatures across participants indicates the convergent validity of these neural signatures. The current study provides additional evidence for reliability and usefulness of the approach by finding that the emotions signature weights in a concept representation are features that can identify membership in the ideator group. Given the limited previous use of this potentially powerful approach to analyzing emotional content from neural signatures, there should be caution concerning the inferences that can be made.

Thus the findings also enable progress beyond saying that one group is measurably different from another. They enable at least part of the difference to be attributed to the emotional component of a concept representation. Unlike a dictionary definition of a concept, a neural representation includes the emotional response to the concept. Some concepts, such as *snake*, have long been known to entail an emotional response. The findings here show that certain concepts evoke different emotions in people with suicidal ideation compared to controls, and also evoke different emotions in suicidal ideators dependent on whether they have ever made a suicide attempt. When used as the features of a classifier, these differences in the emotion component in the neural signature of a concept can be used to provide accurate classification of group membership (in both the ideator-control classification and the attempter-non attempter classification).

fMRI capabilities have made it possible to characterize the altered brain activity of a clinical population as having a higher or lower level of activation in some brain region (say anterior cingulate) than a control group during the performance of some task. By contrast, our approach attempts to characterize a network of altered neural activity that constitutes the representation of a concept and the emotion it evokes. At a given brain location, for some concepts the activation level is higher in the ideator group and for other concepts it is lower. The current study makes an early attempt at relating a pattern of activation values across multiple brain locations to neurotypical and altered representations of particular concepts and their emotional component in a manner that seeks consilience between brain activity and psychological states. At the same time, it remains possible to determine which brain structures are the sites of a clinical alteration.

This study is distinctive in neuroimaging research on suicidal ideation and behavior because it directly focuses on how suicidal individuals think about various concepts, rather than on responses to tasks, that however salient, do not mirror the experience of the suicidal person as directly. This neurosemantic assessment has face validity because those suicide attempters at highest risk and with the highest suicide intent engaged in suicidal ideation because they wanted to die (and thus thought about suicide as more attractive) or wanted to escape an impossible situation or feeling state, which might lead to altered responses to various death and life related concepts.

There are several potential benefits of this neurosemantic approach. The identification of differential patterns of regional activation could suggest brain regions to target using brain stimulation techniques such as transcranial magnetic stimulation (TMS) or transcranial direct current stimulation (tdcs) [33]. The identification of altered emotional responses to suicide related concepts could prove very useful to a psychotherapist in trying to heighten the patient's attraction to life, and decrease the attraction to suicide and death. If these new findings have predictive value, then they would also be useful in guiding a clinician's decisions about psychotherapeutic targets and in monitoring overall suicidal risk. The neurosemantic approach can also guide the development of less costly and more easily disseminable methods that can potentially yield similar information, such as EEG assessment of neural concept representations, as demonstrated for neurotypical participants [34]. And despite its greater cost, this approach might be effective in highly suicidal individuals who are repeatedly hospitalized for suicidal crises or those who require a higher level of care, such as an intensive outpatient program.

An unexplored prospective benefit of the approach is its potential to predict imminent suicidal risk. A longitudinal investigation of a larger cohort of individuals with suicidal ideation could repeatedly assess the altered neural representations to determine whether there is a neural signature of an imminent attempt. Such information would be invaluable in the case of the small percentage (e.g., 5%) of patients in psychiatric inpatient care who make up as much as half of suicides subsequent to discharge from a hospital [35]. In future prospective studies, it would be of great interest to learn if our neurosemantic assessments are useful in monitoring for current suicidal risk and in predicting future suicide attempts. If so, this approach could be useful for monitoring ongoing suicidal risk and response to treatment.

Study limitations. Performance of the task requires highly cooperative and focused participants (not everyone can keep their attention intensely focused for 30 minutes). However, we also showed that the models developed on the less noisy participants' data can be successfully applied to more noisy data from other participants, which substantially improves the chances for potential clinical applications. Moreover, it may be possible in the future to develop shorter batteries that focus on concepts most likely to identify altered responses associated with suicidal risk and which would require sustained attention over a shorter period.

Another limitation is that the current study does not provide a contrast between suicidal ideation and psychiatric control participants who are affected by psychopathology in general. However, the ability to distinguish within the suicidal ideation group between attempters and non-attempters suggests that our classification is more specific and not just related to psychopathology in general. Within its limitations, the current study provides a promising first step in assessing a psychiatric disorder of brain and mind that takes both of these facets into account.

Methods

Participants. Participants were 79 young adults, either affected with current suicidal ideation (*n* = 38) or healthy controls with no personal or family history of psychiatric disorder or suicide attempt (*n* = 41). Exclusion criteria included neurological disorders, anoxia history, head injuries, Wechsler verbal score < 80 [36], current use of sedative medication, pregnancy, ineligibility for magnetic resonance imaging (MRI), psychosis, substance misuse or positive urine drug/saliva alcohol screen.

Assessment. History of suicide attempt (defined as potentially self-injurious behavior with some non-zero intention of dying) was assessed with the Suicide History Form and Suicide Intent Scale [37,38]. The severity of suicidal ideation was assessed using the interview-rated Columbia-Suicide Severity Rating Scale (C-SSRS) [39], and the self-reported Adult Suicide Ideation Questionnaire(A-SIQ)[40]. General psychopathology, depression, anxiety, and history of child maltreatment were assessed using the Adult Self Report (ASR)[41,42], the Patient Health Questionnaire- 9(PHQ-9)[43], the Adult Spielberger State Trait Anxiety Inventory (STAI-T)[44], and the Childhood Trauma Questionnaire (CTQ)[45], respectively.

Participants in neurosemantic analyses. The neurosemantic analyses below are based on 34 participants, 17 per group whose fMRI data quality was sufficient for accurate (normalized rank accuracy > .6) identification of the 30 individual concepts from their fMRI signatures. The selection of participants included in the primary analyses was based only on the technical quality of the fMRI data. The data quality was assessed in terms of the ability of a classifier to identify which of the 30 individual concepts they were thinking about with a rank accuracy of at least .6, based on the concepts' neural signatures. The participants who met this criterion also showed less head motion ($t(77) = 2.73$, $p < .01$). The criterion was not based on group discriminability. The 17 selected for the primary data analysis and the 21 remaining suicidal participants did not differ on demographic data, diagnoses, clinical severity of depression, anxiety, or suicidal ideation, or history of suicide attempt. The data of the participants with poor data quality were also analyzed, as reported in the Results section.

A previous study of ASD using a similar approach[10] used 17 participants with good data quality per group, hence the target of a similar sample size. Three additional control participants who had also satisfied this criterion were selected at random and excluded to equate the group sizes. The final groups were balanced on gender ratio, age, and WASI IQ. Participants in the suicidal ideator group were significantly more symptomatic than the control group on almost all other measures, as shown in Table 1. There were no systematic differences between the 17 ideators whose data were used in the neurosemantic analysis and the 21 whose data were excluded, other than the poor classification accuracy on the 30 concepts. We attribute the sub-optimal fMRI data quality (inaccurate concept identification from its neural signature) of the excluded participants to some combination of excessive head motion and inability to sustain attention to the task of repeatedly thinking about each stimulus concept for 3 sec over a 30 min testing period. Despite their exclusion from the main neurosemantic analysis, we show below that there remains valuable information in the fMRI data of the excluded suicidal ideator participants. The comparison of self-report data between the 34 participants included in the neurosemantic analyses and the remaining (excluded) participants is reported in Supplementary Information.

The study protocol was approved by the University of Pittsburgh and Carnegie Mellon University Institutional Review Boards. All participants gave their informed written consent.

Stimuli. The stimuli were three groups of 10 words each, half of them nouns and half adjectives related to: (1) suicide (e.g., *death*, *overdose*); (2) negative affect (e.g., *sad*, *gloom*); and (3) positive affect (e.g., *happy*, *carefree*) as shown in Table 2. The set of 30 stimulus items was presented 6 times, in different random orders. Each item was displayed for 3 sec followed by a 4 sec blank interval to allow for the hemodynamic response to take its course. Seventeen sec long fixation intervals were included periodically to provide an activation baseline. The stimuli were displayed in white font and centered on a black background.

Task instructions. Participants were asked to actively think about the concepts to which the stimulus words refer while they were displayed, thinking about their main properties (and filling in details that come to mind) and attempting consistency across presentations.

Image acquisition and preprocessing. The fMRI data were acquired on a Siemens Verio 3.0 Tesla scanner (20 slices, voxel size 3.125 x 3.125 x 5 mm, repetition time 1s). The data were pre-processed and converted to a standard MNI space using SPM8 (Wellcome Dept. of Cog. Neurology), and a single mean value was computed for each voxel and stimulus item (see Supplementary Information for details).

fMRI data analytic approach. Three analyses are described here: (1) selecting voxels with stable semantic tuning curves; (2) spatial clustering of the stable voxels at the group level to determine the brain locations that contain the neural representations of the concepts; and (3) developing a resulting machine learning classification model from the reduced data, and attempting to classify participants' group membership using the model.

1. Selecting voxels with stable semantic tuning curves. These analyses focus on a subset of all the voxels (each ~50 mm^3) whose semantic tuning curve of activation over the set of stimulus items is *stable* across the multiple presentations of the set of items (see Supplementary Information for details).

2. Obtaining group-level clusters of stable voxels. A fixed number of the most stable voxels are selected in each participant (excluding bilateral occipital lobes), and a group hit map is computed and thresholded by the number of contributing participants and spatial proximity (see Supplementary Information for details). The clusters of stable voxels in the group hit maps indicate where the set of neural representations (including all of the concepts) are located for the two groups, as shown in Figure 1. Preliminary testing identified which of the clusters best discriminated between the two groups (see Supplementary Information). The classifier's features included voxels in clusters that are common between the two groups as well as voxels from unshared clusters.

3. Machine learning methods. Machine learning entails training a classifier on a subset of the data and testing the classifier on an independent subset. The cross-validation procedure iterates through all possible partitionings (folds) of the data always keeping the training and test sets separate from each other. The main machine learning here uses a Gaussian Naïve Bayes (GNB) classifier (using pooled variance). The main type of classifications performed in this study was a group membership classification that assigned each participant to one of the two groups; the accuracy was the proportion of correctly classified participants, and significance levels were obtained using a binomial distribution, and b) identification of which of the 30 concepts a participant was thinking about; in this case, rank accuracy was computed (see Supplementary Information for details) and compared to a chance level of accuracy obtained by random permutation testing.

The main reason that classification was used rather than General Linear Modeling (GLM) is that classification is multivariate whereas GLM uses univariate analysis of fMRI data (assessing each voxel independently). The phenomena of interest here (and in many fMRI studies of cognition) are inherently multivariate, in the sense that such cognitivelyrelated phenomena typically occur in a number of different voxels or voxel clusters that need not be proximal to each other. In particular, the neural representations of individual concepts such as *apple* or *death* correspond to activation in a set of spatially distributed voxel clusters, and the groups here differ in the *collective* pattern of activation levels in these spatially distributed voxels. GLM, because of its univariate nature, fails to assess both the collective pattern and the group differences in the collective pattern. By contrast, the classifier's features are the set of activation levels of a set of spatially distributed voxels. Very many other studies show greater sensitivity of classification over GLM where the phenomena of interest consist of a spatially distributed pattern of activation.

Group membership classification. Two 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.

The features used by the classifier to characterize a participant consisted of a vector of activation levels for a number of (discriminating) concepts in a set of (discriminating) brain locations. To determine how many and which concepts were most discriminating between ideators and controls, a reiterative procedure analogous to stepwise regression was used, first finding the single most discriminating concept, and then the second most discriminating concept, reiterating until the next step reduced the accuracy. A similar procedure was used to determine the most discriminating locations (clusters). The procedure is further described in Supplementary Information. The activation level in each brain location was computed as a mean activation of the five most stable voxels in that location. The classifier was trained on the data of all but one participant, and the group membership of the left out participant was predicted.

In addition to group membership classification based on the neural representations of the stimulus concepts themselves, another classification was based on the emotional content of the neural representations of the discriminating concepts. The discriminating concepts' activation was represented as a weighted sum of activation vectors characterizing the involvement of four emotions: *sadness, shame, anger* and *pride.* Each participant was characterized by a vector consisting of the weights associated with these emotions for each discriminating concept, and the participants' group membership was classified using a machine learning procedure similar to the one described above (see Supplementary Information for details).

Classification of 30 concepts*.* This procedure attempted to identify which of the 30 concepts a participant was thinking about, given an independent sample of its neural signature. This measure provided an index of the inconsistency or noise level in a participant's neural signature data.

Code availability. The custom computer code that was used in the main analysis of this study is available from the corresponding author upon reasonable request.

Data availability. The de-identified data that support the main findings of this study are available from the corresponding author upon reasonable request.

References

1. CDC (2016) – WISQARS data. at <http://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html>
2. Glenn, C. R. & Nock, M. K. Improving the short-term prediction of suicidal behavior. *Am. J. Prev. Med.* **47,** S176–S180 (2014).
3. Nock, M. K., Park, J. M., Finn, C. T., Deliberto, T. L., Dour, H. J. & Banaji, M. R. Measuring the suicidal mind: Implicit cognition predicts suicidal behavior. *Psychol. Sci.* **21,** 511–517 (2010).
4. Busch, K. A., Fawcett, J., & Jacobs, D. G. Clinical correlates of inpatient suicide. *J. Clin. Psychiatry* **64,** 14–19 (2003).
5. Mann, J. J., Arango, V. A., Avenevoli, S., Brent, D. A., Champagne, F. A., Clayton, P., Currier, D., Dougherty, D. M., Haghighi, F., Hodge, S. E., Kleinman, J., Lehner, T., McMahon, F., Mościcki, E. K., Oquendo, M. A., Pandey, G. N., Pearson, J., Stanley, B., Terwilliger, J. & Wenzel, A. Candidate endophenotypes for genetic studies of suicidal behavior. *Biol. Psychiatry* **65,** 556–563 (2009).
6. Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P. & Nock, M. K. Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychol. Med.* **46,** 225–236 (2016).
7. 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).
8. 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).
9. Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A. & Just, M. A. Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science.* **320,** 1191–1195 (2008).
10. 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).
11. Cha, C. B., Najmi, S., Park, J. M., Finn, C. T. & Nock, M. K. Attentional bias toward suicide-related stimuli predicts suicidal behavior. *J. Abnorm. Psychol.* **119,** 616–622 (2010).
12. Armey, M. F., Crowther, J. H. & Miller, I. W. Changes in Ecological Momentary Assessment Reported Affect Associated With Episodes of Nonsuicidal Self-Injury. *Behav. Ther.* **42,** 579–588 (2011).
13. Bresin, K., Carter, D. L. & Gordon, K. H. The relationship between trait impulsivity, negative affective states, and urge for nonsuicidal self-injury: A daily diary study. *Psychiatry Res.* **205,** 227–231 (2013).
14. Bryan, C. J., Morrow, C. E., Etienne, N. & Ray-Sannerud, B. Guilt, shame, and suicidal ideation in a military outpatient clinical sample. *Depress. Anxiety* **30,** 55–60 (2013).
15. Bryan, C. J., Ray-Sannerud, B., Morrow, C. E. & Etienne, N. Shame, pride, and suicidal ideation in a military clinical sample. *J. Affect. Disord.* **147,** 212–216 (2013).
16. Humber, N., Emsley, R., Pratt, D. & Tarrier, N. Anger as a predictor of psychological distress and self-harm ideation in inmates: A structured self-assessment diary study. *Psychiatry Res.* **210,** 166–173 (2013).
17. Nock, M. K., Prinstein, M. J. & Sterba, S. K. Revealing the form and function of self-injurious thoughts and behaviors: A real-time ecological assessment study among adolescents and young adults. *J. Abnorm. Psychol.* **118,** 816–827 (2009).
18. Olié, E., Ding, Y., Le Bars, E., de Champfleur, N. M., Mura, T., Bonafé, A., Courtet, P. & Jollant, F. Processing of decision-making and social threat in patients with history of suicidal attempt: A neuroimaging replication study. *Psychiatry Res. Neuroimaging* **234,** 369–377 (2015).
19. Dukart, J., Schroeter, M. L. & Mueller, K. Age correction in dementia – matching to a healthy brain. *PLoS One* **6,** e22193 (2011).
20. 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).
21. Jacobson, C., Batejan, K., Kleinman, M. & Gould, M. Reasons for attempting suicide among a community sample of adolescents. *Suicide Life-Threatening Behav.* **43,** 646–662 (2013).

22. Rogers, M. L., Kelliher-Rabon, J., Hagan, C. R., Hirsch, J. K. & Joiner, T. E. Negative emotions in veterans relate to suicide risk through feelings of perceived burdensomeness and thwarted belongingness. *J. Affect. Disord.* **208,** 15–21 (2017).
23. Pestian, J., Matykiewicz, P. & Linn-Gust, M. What's in a note: construction of a suicide note corpus. *Biomed. Inform. Insights* **5,** 1–6 (2012).
24. Pan, L. A., Hassel, S., Segreti, A. M., Nau, S. A., Brent, D. A. & Phillips, M. L. Differential patterns of activity and functional connectivity in emotion processing neural circuitry to angry and happy faces in adolescents with and without suicide attempt. *Psychol. Med.* **43,** 2129–2142 (2013).
25. Beck, A. T. & Haigh, E. A. P. Advances in cognitive theory and therapy: the generic cognitive model. *Annu. Rev. Clin. Psychol.* **10,** 1–24 (2014).
26. Adler, A., Bush, A., Barg, F. K., Weissinger, G., Beck, A. T. & Brown, G. K. A mixed methods approach to identify cognitive warning signs for suicide attempts. *Arch. Suicide Res.* **20,** 528–538 (2016).
27. Jager-Hyman, S., Cunningham, A., Wenzel, A., Mattei, S., Brown, G. K. & Beck, A. T. Cognitive distortions and suicide attempts. *Cognit. Ther. Res.* **38,** 369–374 (2014).
28. Brown, G. K., Steer, R. A., Henriques, G. R. & Beck, A. T. The internal struggle between the wish to die and the wish to live: a risk factor for suicide. *Am. J. Psychiatry* **162,** 1977–1979 (2005).
29. Bakhiyi, C. L., Calati, R., Guillaume, S. & Courtet, P. Do reasons for living protect against suicidal thoughts and behaviors? A systematic review of the literature. *J. Psychiatr. Res.* **77,** 92–108 (2016).
30. Jollant, F., Lawrence, N. S., Giampietro, V., Brammer, M. J., Fullana, M. A., Drapier, D., Courtet, P. & Phillips, M. L. Orbitofrontal cortex response to angry faces in men with histories of suicide attempts. *Am. J. Psychiatry* **165,** 740–748 (2008).
31. Mann J J, Waternaux C, Haas G L, M. K. M. Toward a clinical model of suicidal behavior in psychiatric patients. *Am J Psychiatry* **156,** 181–189 (1999).
32. Brent, D. A., Melhem, N. M., Oquendo, M., Burke, A., Birmaher, B., Stanley, B., Biernesser, C., Keilp, J., Kolko, D., Ellis, S., Porta, G., Zelazny, J., Iyengar, S. & Mann, J. J. Familial pathways to early-onset suicide attempt: a 5.6 year prospective study. *JAMA Psychiatry* **72,** 160–168 (2015).
33. Minzenberg, M. J. & Carter, C. S. Developing treatments for impaired cognition in schizophrenia. *Trends Cogn. Sci.* **16,** 35–42 (2012).
34. Suppes, P., Han, B., Epelboim, J. & Lu, Z.-L. Invariance of brain-wave representations of simple visual images and their names. *Proc. Natl. Acad. Sci.* **96,** 14658–14663 (1999).
35. Kessler, R. C., Warner, C. H., Ivany, C., Petukhova, M. V., Rose, S., Bromet, E. J., Brown, M., Cai, T., Colpe, L. J., Cox, K. L., Fullerton, C. S., Gilman, S. E., Gruber, M. J., Heeringa, S. G., Lewandowski-Romps, L., Li, J., Millikan-Bell, A. M., Naifeh, J. A., Nock, M. K., Rosellini, A. J., Sampson, N. A., Schoenbaum, M., Stein, M. B., Wessely, S., Zaslavsky, A. M. & Ursano, R. J. Predicting Suicides After Psychiatric Hospitalization in US Army Soldiers. *JAMA Psychiatry* **72,** 49–57 (2015).
36. Wechsler, D. *Wechsler Abbreviated Scale of Intelligence*. (Harcourt Assessment, 1999).
37. Beck, A. T., Schuyler, D., Herman, I. in *Predict. Suicide* (ed. Beck, A. T., Resnick, H .L. P., Lettieri, D. J.) 45–56 (Charles Press, 1974).
38. Oquendo, M. A., Halberstam, B. & Mann, J. J. in *Stand. Eval. cinical Pract.* (ed. First, M. B.) 103–129 (American Psychiatric Press, 2003).
39. Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., & Mann, J. *Columbia-suicide severity rating scale (C-SSRS)*. (Columbia University Medical Center, 2008).
40. Reynolds, W. M. *Professional manual for the suicidal ideation questionnaire*. (Psychological Assessment Resources, 1987).
41. Achenbach, T. M., Howell, C. T., McConaughy, S. H., & Stanger, C. Six-year predictors of problems in a national sample: IV. young adult signs of disturbance. *J. Am. Acad. Child Adolesc. Psychiatry* **37,** 718–727 (1998).
42. Achenbach, T. *Adult self report measure for ages 18-59*. (University of Vermont, 2003).
43. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9. *J. Gen. Intern. Med.* **16,** 606–613 (2001).
44. Spielberger, C. D. *State Trait Anxiety Inventory for Adults: Sampler Set: Manual, Test, Scoring Key;[form Y]*. (STAIS-AD. Mind Garden, 1983).
45. Bernstein, D P, Fink, L, Handelsman, L, Foote, J, Lovejoy, M, Wenzel, K, Sapareto, E, Ruggiero, J. Initial reliability and validity of a new retrospective measure of child abuse and neglect (CTQ). *Am J Psychiatry* **151,** 1132–1136 (1994).

Acknowledgements

This research was partially supported by the National Institute of Mental Health Grant MH029617 and an Endowed Chair in Suicide Studies at the University of Pittsburgh School of Medicine. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

The experiments were conceived and designed by M.A.J., L.P., V.L.C., D.M., and D.B. The experiments were performed by M.A.J., L.P., and D.B. The materials and analysis tools were contributed by M.A.J., V.L.C., C.C., and M.K.N. The data were analyzed by V.L.C. The paper was written by M.A.J., V.L.C., and D.B.

Competing Interests Statement

The authors declare no competing interests

Address Correspondence to:

Marcel Adam Just, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213 or Email: just@cmu.edu

SUPPLEMENTARY INFORMATION

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 McMakin[3], Christine Cha[4], Matthew K. Nock[5] and David Brent[2]

[1]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA
[2]Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA
[3]Department of Psychology, Florida International University, Miami, FL
[4]Clinical Psychology Department, Columbia University, New York, NY
[5]Department of Psychology, Harvard University, Cambridge, MA

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: cluster activation levels → Y_i, i=1,…,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.

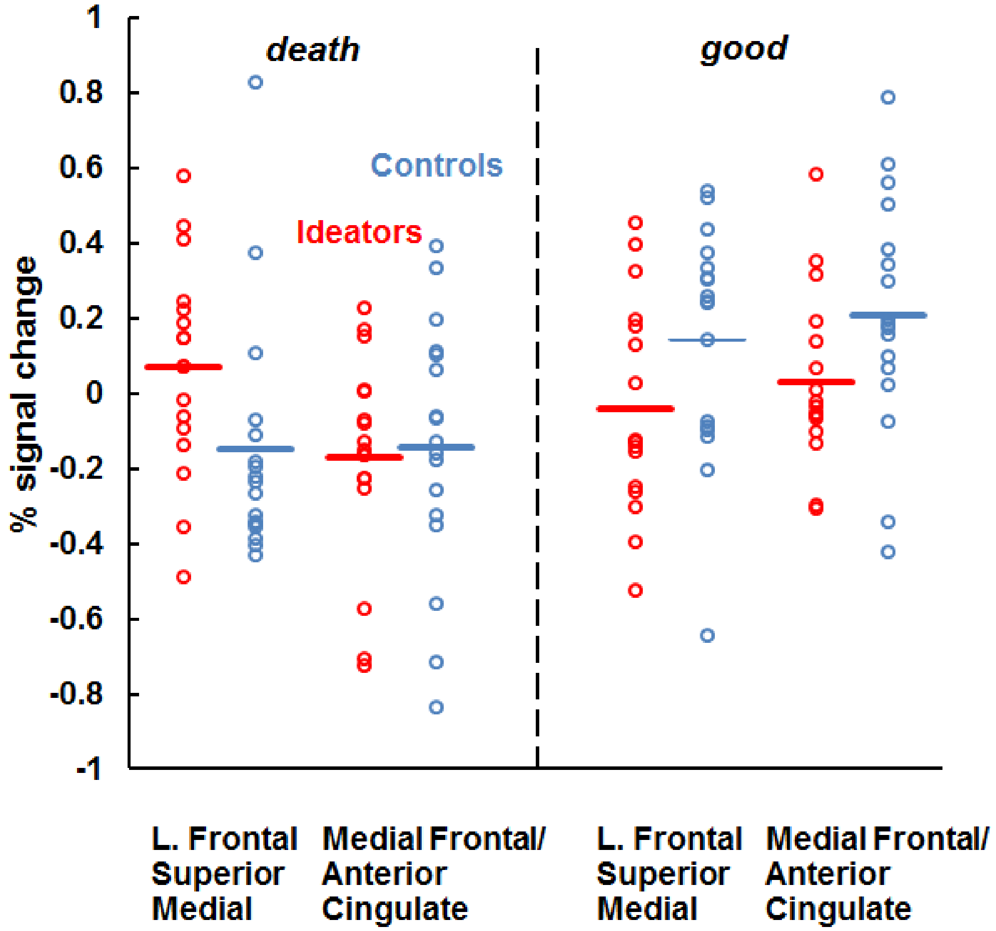
Supplementary Table

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)			

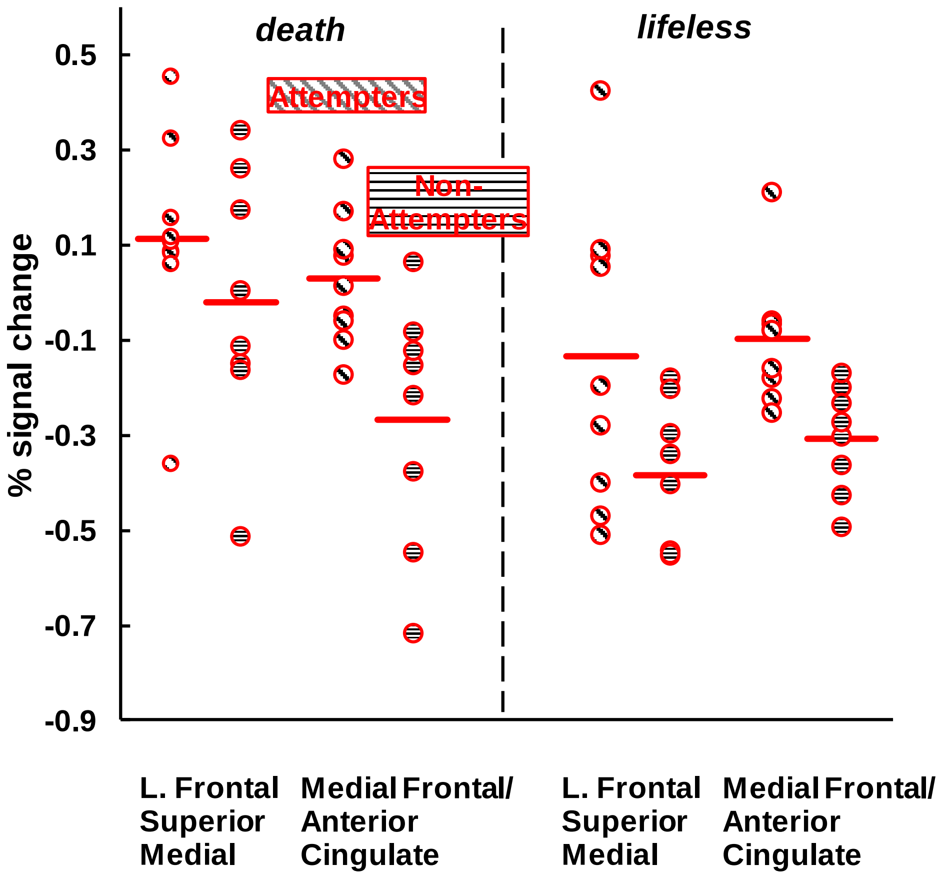
Abbreviations:
WASI IQ: Wechsler Abbreviated Scale of Intelligence;
ASIQ: Adult Suicide Ideation Questionnaire;
PHQ: Patient Health Questionnaire;
CTQ: Child Trauma Questionnaire;
ASR: Adult Self Report;
SIS: Suicidal Ideation Scale
Standard deviations are shown in parenthesis.

Supplementary Figures



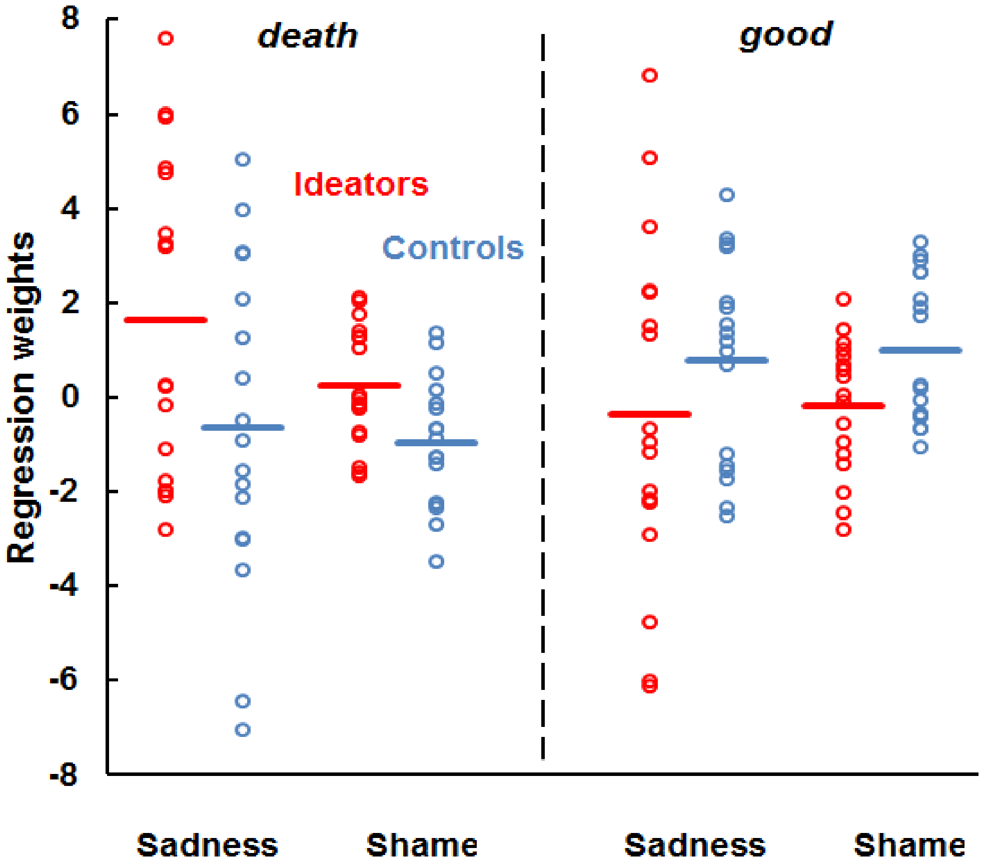
Supplementary Figure 1. Distributions of activation levels for 17 ideators and 17 controls for two concepts in two locations.

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.

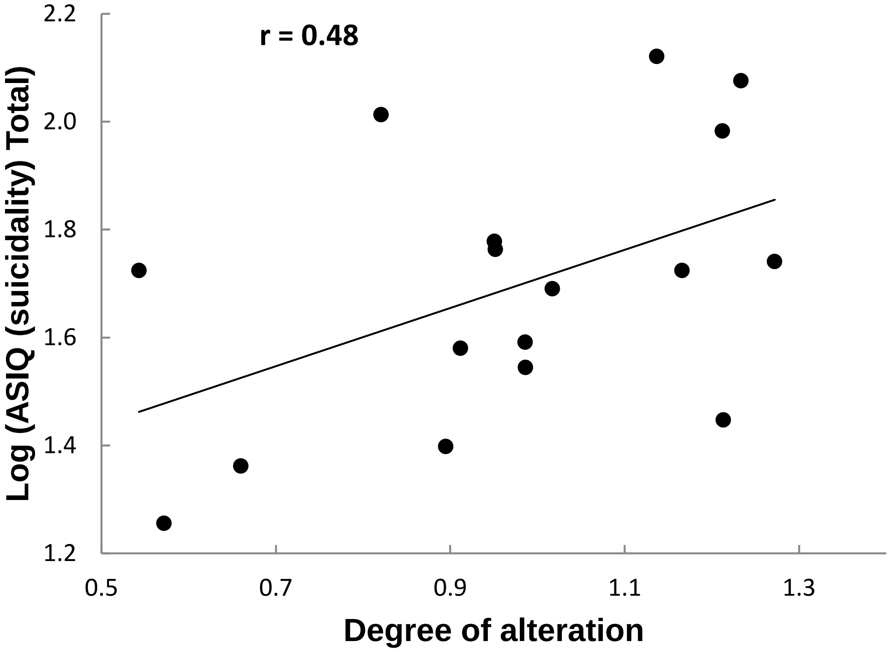


Supplementary Figure 3. Distributions of emotion regression weights for 17 ideators and 17 controls for two emotions, *sadness* and *shame*, for the two discriminating concepts (*death* and *good*).

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).